ISCON: A THEORETICAL FRAMEWORK FOR TESTING DESIGN RATIONALES OF PERSONALIZATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalization models are designed around different commitments to time-order invariance, recent evidence, older evidence, and distinct long- and short-term effects. Utility metrics do not establish whether behavior follows those commitments. We propose the *Information-Survival Contract* (ISCON), a theoretical framework that takes each model's rationale as given and turns it into a testable behavioral contract. ISCON specifies which controlled histories validly test the rationale (*evidence conditions*), how outputs are scored (*measurement*), how scores are combined across evaluation cases (*aggregation rule*), and how the same model's aggregated scores should compare across histories (*acceptable difference*). It distinguishes support, contradiction, and unresolved cases. Its formal properties explain why the same unchanged response can agree with a model designed to aggregate profile evidence with little order sensitivity and contradict one designed to give recent evidence greater influence. We instantiate ISCON with PREFTURB and evaluate models motivated by profile aggregation, recent-state updates, long-range retention, or separate temporal preference components on PENS news recommendation and headline generation, MovieLens-1M movie recommendation, and OpenAI-Reddit personalized summarization. PREFTURB uses two complementary tests under the evidence conditions required by each rationale. One compares the same model's responses to histories containing only long-term (LT), short-term (ST), or episodic (EP) evidence. The other moves, removes, or repeats selected evidence, or introduces distracting or competing evidence. For MAMBA4REC, whose long-range rationale motivates retaining useful older evidence, LT MRR is 0.028, below ST at 0.101 and EP at 0.062, contradicting the tested commitment. A follow-up control finds that late LT placement is 0.099 MRR worse than deletion. A separate control finds that increasing LT evidence with competing ST evidence fixed raises MRR by 0.022. PREFTURB therefore reveals a sharper diagnosis in which older evidence can help this model even though its placement can make that evidence harmful. *The resulting tests identify where a model's own commitments are supported, contradicted, or unresolved without judging which commitment is preferable.*

1 INTRODUCTION

Personalization models use past interactions to recommend items or generate text tailored to a person. A reading history can inform news ranking and personalized headline generation (Ao et al., 2021). Designs make different assumptions about how history should influence outputs. Attentive models aggregate browsed items (Wu et al., 2019b), sequential models update a state as interactions accumulate (Liu et al., 2024a), and memory-based systems can coordinate short-term and long-term information (Zhang et al., 2024). Their intended behavior may involve limited sensitivity to order, greater influence from recent evidence, continued influence from older evidence, or distinguishable long-term and short-term effects. These are behavioral expectations, not guarantees established by a model's architectural description alone. Successful recommendations and fluent generated text also cannot establish adherence to these commitments. A model may rank a target highly without using the history evidence its design intends to preserve. Likewise, reference similarity can conceal whether generated text responds to differences among users (Vansh et al., 2023; Dasgupta et al., 2024). The

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.

central question is *whether history dependence follows the model's own rationale, not merely whether its outputs score well.*

Metamorphic and behavioral testing compare responses across controlled inputs against specified expectations (Chen et al., 2018; Ribeiro et al., 2020). History shuffling has also been used to examine whether LLM recommenders use interaction order (Kim et al., 2025). Order sensitivity alone cannot establish adherence to a particular personalization rationale. Consider a reader who follows technology over months, football news recently, and travel articles during one weekend. These interactions provide long-term (LT), short-term (ST), and episodic (EP) evidence, respectively. A test of the commitment to recent evidence moves the same football interactions earlier while holding the football recommendation task fixed. An unchanged ranking score can agree with an order-invariant design while contradicting a design that requires a higher score from recent relevant evidence. The edit cannot decide which interpretation applies. Holding the target and candidate set fixed keeps the measured comparison focused on the history condition. Moving irrelevant interactions would not test the commitment. Even with suitable evidence, the test must specify what is scored, how common evaluation cases are combined, and what score difference the rationale requires.

We propose the ***I**nformation-**S**urvival **Con**tract* (ISCon), a framework for testing whether a personalization model follows its stated design rationale. ISCon specifies the valid evidence conditions, output measurement, aggregation rule, and expected relation among the same model's aggregated responses. These choices determine support, contradiction, or an unresolved outcome. Finite-cohort quantities are directly computable; population claims additionally require justified uncertainty. We establish when such comparisons support a verdict and why different aggregation rules can change the tested relation on identical outputs. ISCon *takes the model's rationale as given rather than judging whether another rationale is preferable.* We instantiate ISCon for temporal preference evidence with PrefTurb using two complementary tests. *Role isolation* measures the same model under LT-, ST-, or EP-only evidence. Controlled edits then change the position, availability, repetition, or competing context of selected evidence. Evidence and its LT/ST/EP role are frozen beforehand, so moving LT evidence toward prediction does not relabel it as ST. Across conditions, the model, task, target or reference, scoring procedure, evaluation cases, and aggregation weights remain fixed. We prove that the construction preserves retained event content and recorded roles and that its measurement and aggregation realize the specified ISCon comparison under stated assumptions.

We evaluate profile aggregation, recent-state, long-range, and temporally factorized designs on PENS news recommendation and headline generation, MovieLens-1M movie recommendation, and OpenAI-Reddit personalized summarization. Recommendation scores distinguish exact-target ranking from credit for related items. Generation scores distinguish reference similarity from correspondence between user-specific output differences and preference differences. These comparisons test observable contracts motivated by each model's stated rationale. For Mamba4Rec, the long-range commitment to useful older evidence requires LT to score at least as well as one of the two transient alternatives on the declared evidence domain. On MovieLens-1M, mean reciprocal rank (MRR) with LT evidence is 0.028, below ST at 0.101 and EP at 0.062, contradicting that tested commitment under exact-target ranking. Moving the same LT evidence late produces an MRR 0.099 below deletion. In a separate control, increasing LT evidence with competing ST evidence fixed raises MRR by 0.022. The controls vary placement and evidence quantity separately, and their effects need not agree. These comparisons show that *this model can benefit from additional older evidence even though late placement makes that evidence harmful.* They narrow the question for further investigation beyond the initial adherence failure, without identifying an internal cause.

Contributions. First, ISCon specifies adherence and establishes when score comparisons justify a verdict. Second, PrefTurb constructs controlled preference histories and realizes that specification under stated assumptions. Third, our recommendation and generation analyses connect the basic adherence test to follow-up diagnoses of history dependence and interference.

2 Related Work

Personalization Design Tests. Behavioral tests examine how recommenders respond to controlled input changes. Kim et al. (2025) test interaction-order use through history shuffling, while other studies derive transformations from algorithm principles (Mao et al., 2024; 2020) or modify ratings

and prompts (Khirbat et al., 2024). These tests do not by themselves establish whether a response satisfies a model's particular history-use rationale on appropriate preference evidence.

Formal Specification and Decisions. Formal testing specifies expected responses and supports decisions under uncertainty. Tests that require a controlled input transformation to induce a specified relationship between outputs are commonly framed as *metamorphic testing* (Chen et al., 2018). CheckList organizes capability-based behavioral tests (Ribeiro et al., 2020), while error-controlled procedures with an explicit undecided outcome have been developed as *agnostic tests* (Coscrato et al., 2020). These tools do not specify the evidence, measurement, aggregation, and score relation needed to test a personalization rationale.

Evidence Construction and Preservation. Position and memory benchmarks examine retrieval, temporal reasoning, and updates (Liu et al., 2024b; Maharana et al., 2024; Wu et al., 2025). They do not specify how to construct and preserve LT, ST, and EP preference evidence from original histories while ensuring that controlled edits isolate the intended change rather than alter the selected evidence.

Measurement and Follow-Up. Personalization evaluations measure output quality and user-specific responsiveness. EGISES and PerSEval compare user-specific output and reference differences (Vansh et al., 2023; Dasgupta et al., 2024), ExPerT assesses content and style (Salemi et al., 2025), PersonaLens tests personalized conversation (Zhao et al., 2025), and PERG examines preference alignment and correctness (Okite et al., 2025). These evaluations do not by themselves establish when a score relation supports adherence or when a follow-up establishes its retention. Detailed comparisons appear in Tables 27 through 29 in Appendix M.6.

3 ISCon: Theoretical Framework

Personalization models are designed to utilize user interaction (e.g. *click*, *skip*, *generate*, etc. actions on specific items such as *news articles*, *products*, *movies*, etc.) histories in different ways. A recency-based model is expected to give greater influence to recent relevant *evidence* (i.e., predictive cues), whereas a profile-based model may be relatively insensitive to where that evidence appears, and a long-range model is expected to preserve the influence of useful older evidence. Consider a news recommender designed around recency. Suppose a user has interacted with several football articles and the target is another *football* article. If the model behaves according to its design, these football interactions should have a stronger influence when they are recent than when they occur much earlier. An order-insensitive model need not show this difference. Thus, different design rationales imply different expectations about how the model should respond to the same relevant evidence.

Overview. We propose ISCon (**I**nformation-**S**urvival **Con**tract) as a framework for testing these expectations. For a fixed model M and one stated design rationale F, ISCon first formalizes when the evidence is relevant to F, i.e., the *validity conditions* ($\mathcal{A}_F$) of the evidence. As an example, if the target is *football* article, earlier unrelated *technology* interactions are not a useful test of whether a long-term model preserves relevant long-term evidence. ISCon then requires labeling evidence by their temporal roles (i.e., long-term, short-term, and episodic). It next enforces an acceptable *scoring function* S, for measuring model performance. It then specifies the conditions that *controlled trajectories* need to satisfy for testing a specific F. For example, valid test trajectories can be controlled positional shifts, both favorable and adversarial, in football-specific evidence when the fixed target is a football article. Given a specific F, it uses S to record the actual model performance under a controlled test and compares that with the corresponding expected performance. This expected score relationship between F and the controlled test is termed as *agreement condition* ($\mathfrak{C}_F$). For example, for a fixed F that claims recency, we expect the model's performance to be better when a favorable controlled test places evidence in recent positions than when an adversarial controlled test places the same evidence in earlier positions. It is to be noted that what is a favorable test for one F may not be so for another. For example, an F that claims order-insensitivity should not lead to performance variation. ISCon then aggregates model performance score w.r.t S using a fixed aggregation rule. We term $(\mathcal{A}_F, S, \text{aggregation}, \mathfrak{C}_F)$ as ISCon's specified *contract*. Under this, ISCon concludes whether the model performance supports, contradicts, or leaves the F unresolved.

Scoring a Controlled Test. Consider one evaluation case i and a controlled condition ω, such as placing the same football evidence in an early or recent position. Let H_i^{ω} denote the corresponding interaction history. As stated, ISCon is applicable only when the interactions are related to the

evidence and have a temporally aligned label that is consistent with the rationale F. We term this the *validity conditions* $\mathcal{A}_F$. Across H_i^ω, only the evidence position of the history changes. For a given test suite, we fix the model M, the query z_i given to the model, and the prediction/generation target T_i. S measures the model's response under condition ω as $R_i(\omega) = S(M(H_i^\omega, z_i), T_i)$. Thus, in the football example, $R_i(\text{recent})$ and $R_i(\text{early})$ are the *scores* obtained for the same football target when the same relevant football evidence is placed recently or much earlier in the history. ISCon is a comparative analysis across multiple controlled tests. For example, to test recency-based F, ISCon records performance under both *late* and *early* controlled tests and then compares the results. More generally, for i-th evaluation of F, let X_i denote the summary vector of all d pair-wise comparisons across $\omega \in \Omega$, using the comparison operator h. Then $X_i = h((R_i(\omega))_{\omega\in\Omega}) \in \mathbb{R}^d$.

Aggregation Function. The summary vector X_i is then combined using the aggregation rule fixed by the contract: $\widehat{\mu} = \sum_{i=1}^N w_i X_i,\ w_i \geq 0,\ \sum_{i=1}^N w_i = 1$. Thus, $\widehat{\mu}$ scalarizes X_i across all i evaluations of F. To generalize beyond the limited evaluation cases N, we specify the larger population these cases are intended to represent. Let $\mathbb{P}$ denote this sampling distribution. The corresponding population response is $\mu = \sum_i w_i\, \mathbb{E}_{\mathbb{P}}[X_i]$.

Agreement Condition $\mathfrak{C}_F$. This specifies what is the expected performance pattern under (Ω, F). In the football example, an order-insensitive rationale may require the early- and late-placement responses to remain invariant as $|\mu_{\text{late}} - \mu_{\text{early}}| \leq \varepsilon_{\text{inv}}$, whereas a recency-based rationale may require the recent placement to produce at least a specified advantage η, such that $\mu_{\text{late}} - \mu_{\text{early}} \geq \eta$.

Testability. For the evaluation case set (termed *cohort*), we can directly check whether $\widehat{\mu}$ satisfies the agreement condition. Note that a claim about the larger population must additionally account for uncertainty in μ. A specific j-th agreement condition satisfiability, $g_j(\mu)$, is translated into a non-negative value. For the recency example, $g_j(\mu) = \mu_{\text{late}} - \mu_{\text{early}} - \eta$, so $g_j(\mu) \geq 0$ means that late placement improves the target score by at least the required margin η. A stricter design requirement is handled analogously using $g_j(\mu) > 0$.

Let Q denote the uncertainty region for μ due to population estimation. The observed evidence *supports* this recency requirement only if $g_j(\mu) \geq 0$ for every $\mu \in Q$, and *contradicts* it only if the requirement fails for every $\mu \in Q$; otherwise, the result is *unresolved*. To make this decision executable, we need only the smallest and largest values that each agreement requirement can take in Q. For requirement j, let $L_j = \inf_{v\in Q} g_j(v), U_j = \sup_{v\in Q} g_j(v)$. If $L_j \geq 0$, then every population value allowed by Q satisfies the requirement; if $U_j < 0$, then every allowed value violates it. These two cases correspond to support and contradiction, respectively, and all remaining cases are unresolved. The following result gives the statistical guarantee behind this decision rule.

Theorem 3.1 (Sound contract decisions)**.** *Suppose the uncertainty region Q contains the true population response μ with probability at least $1 - \alpha$. The probability of incorrectly declaring support or contradiction is at most α. (Proof: Appendix E.3)*

The guarantee entails that when we say the evidence supports or contradicts the expected $\mathfrak{C}_F$, the conclusion is unlikely to be attributable solely to sampling uncertainty. Note that every ISCon verdict applies only to the evidence and evaluation domain specified by the contract, and not to all possible model behavior (Proposition A.3).

Aggregation Properties. ISCon prescribes that aggregation function has to determined by the exact question posed by the contract. Consider a long-range rationale whose corresponding performance under LT evidence test should remain competitive with at least one of the two transient roles, ST or EP. Let L, S, and E denote their scores for one evaluation case. One possible aggregation method is $C_{\text{pop}} = \mathbb{E}[L] - \min(\mathbb{E}[S], \mathbb{E}[E])$ which compares the average LT response against the weaker of the average ST and EP responses. However, this is not the same as another variant $C_{\text{inst}} = \mathbb{E}[L - \min(S, E)]$ which chooses the weaker transient role separately for each case before averaging. To generalize, for paired changes Δ_i, $|\sum_i w_i \Delta_i| \leq \sum_i w_i |\Delta_i|$. ISCon therefore distinguishes stability of the aggregate response from stability of individual responses (see Proposition D.1).

4 PrefTurb: Instantiation of ISCon

PrefTurb implements the ISCon tests for recommendation and generation. Figure 1 shows how this controlled diagnostic process works. It selects relevant interactions, places the same interactions

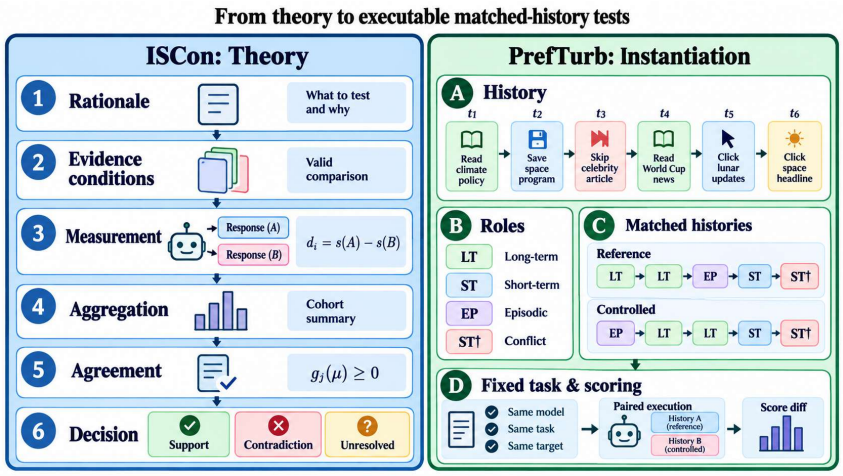

Figure 1: **From ISCon theory to PrefTurb instantiation.** ISCon specifies the rationale-conditioned comparison through evidence conditions, measurement, aggregation, agreement, and uncertainty-aware decision. PrefTurb instantiates this specification using controlled preference histories while keeping the model, task, target or reference, and scoring context fixed.

earlier or later, and scores the model's response. As a first step, PrefTurb instantiates the method of selecting valid evidence w.r.t a chosen target, as per ISCon. To this end, PrefTurb denotes an *event* as at position t in user u's source history as $e_{u,t} : (a_t, x_t)$ where a_t is a specific action (eg. *click*, *skip*) and x_t is the item on which the action is taken (eg. *article, movies*). It also denotes the target item as $y_u^\star$. It then represents $e_{u,t}$ and $y_u^\star$ using fixed encoding functions ψ_{evt} and ψ_{tar} respectively, which map the event and target into the same representation space, with the event representation also incorporating the recorded action. It then uses an *affinity function*, $s_u(t) = \text{sim}(\psi_{\text{evt}}(e_{u,t}), \psi_{\text{tar}}(y_u^\star))$, which helps identify events *relevant* to the target. Once the relevance is fixed, PrefTurb then assigns temporal labels, as required by ISCon, using predefined frequency patterns across five temporal bins (i.e., segments that divide the user trajectory). For the LT label, a relevant event must recur at least three times across at least 2/5 temporal bins, and must involve at least two distinct semantically coherent items of the same category, as determined by the sim operator. For the ST label, the relevant events must form a coherent temporal interval toward the recent end of the original history. The core interval is defined by the top-k percentile of the selected event positions, which limits the influence of isolated outliers. The resulting interval must span between 5% and 50% of the full history. It must also contain repeated, semantically coherent activity rather than a single isolated event. In contrast, EP requires a brief *local spike* containing at most four selected items spanning at most four original history positions. These events should be concentrated within that span, form a strong consecutive run, and show little recurrence of the same preference outside it. Appendix G.1 gives the full role-specific criteria.

F-Compatible Trace Construction. Once valid LT/ST evidence has been selected, PrefTurb arranges the scattered interactions into one coherent ordered *trace* that can be moved, removed, or repeated as a single unit. Moving each event separately would change many positions and local neighbors, while arbitrary concatenation could introduce unnatural semantic jumps that disadvantage the model for reasons unrelated to its temporal rationale. To give the model a reasonable opportunity to use the selected evidence, PrefTurb builds a semantically smooth trace. First, the selected events and a temporary query anchor form an affinity graph. Fiedler-vector spectral seriation (Atkins et al., 1998) provides the initial order and pairwise swaps with 2-opt reversals (Croes, 1958) refine it. The query anchor is used only to guide construction and is removed before evaluation. The resulting LT/ST trace is then frozen and moved as a unit across controlled history positions. Note that EP evidence keeps its original source-local order.Appendix G.2 gives the objective and executed settings, and Algorithm 1 summarizes the complete trace-construction procedure. **Controlled Tests.** The frozen trace is used to create the history conditions required by the ISCon contract. For *role isolation test*, the same model is evaluated separately with LT-, ST-, or EP-only trace. For *controlled tests*, the same frozen trace is moved to different positions, removed, repeated, or combined with separately

selected distracting or competing evidence. In our example, moving an LT football trace changes only where that same evidence appears, while deleting it changes only its availability, and repeating it changes only its multiplicity. The model is run under each resulting condition, and the corresponding responses are compared according to the rationale being tested. See Table 8 for the controlled-test operators.

Scoring Functions. We use scoring functions such as reciprocal rank and nDCG to measure recommendation model performance under each condition. We additionally use SR-ERR as a semantic-substitute-aware diagnostic. For generation, ROUGE and BERTScore measure similarity to a reference, while PerSEval measures whether differences among users' outputs correspond to differences among their references (Lin, 2004; Zhang et al., 2020; Dasgupta et al., 2024). Appendix C gives the scoring functions and their required inputs.

Aggregation. PREFTURB next combines the condition-wise responses in the way specified by the ISCON contract. Each evaluation case contributes to the paired difference. As an example, one paired difference can be $R_i(\text{late}) - R_i(\text{early})$. For role isolation, the corresponding LT, ST, and EP responses are retained for the required cross-role relation. These case-level quantities are then aggregated over the same applicable cases and with the same weights across conditions. Recommendation gives equal weight to each retained user-target case, while generation follows the declared case-, document-, or user-level aggregation. The resulting quantity is the empirical response $\widehat{\mu}$ defined in Section 3.

Applying the Agreement Condition. Finally, PREFTURB evaluates the aggregated response against the agreement condition $\mathfrak{C}_F$ specified for the model's rationale. For a population-level claim, PREFTURB additionally accounts for the uncertainty region Q from Section 3: the evidence *supports* the requirement only when every value allowed by Q satisfies it, *contradicts* it when every allowed value violates it, and is otherwise *unresolved*.

5 EXPERIMENTAL SETUP

Tasks, Datasets, and Models We study news recommendation and personalized headline generation on PENS (Ao et al., 2021), movie recommendation on MovieLens-1M (Harper & Konstan, 2015), and personalized summarization in the OpenAI-Reddit (OAI) setting. Recommendation pairs a user's history with a fixed candidate set and positive target. Generation pairs the history and source document with a user-specific reference. These tasks test whether responses to controlled preference evidence match the rationale under test. Models span attentive aggregation (NRMS/NAML) (Wu et al., 2019b;a), recurrent aggregation (EBNR) (Ao et al., 2021), selective state-space updates (MAMBA4REC) (Liu et al., 2024a), structured preference or memory representations (IMPerSumm, WALK2PERS, SCAPE) (Chatterjee et al., 2026b;a; Lian et al., 2025), interaction reconstruction (DiffRec) (Wang et al., 2023), and prompted histories (Qwen/Gemma). The tested commitments are weak order dependence from profile aggregation, greater influence from recent evidence through recurrent state updates, continued influence from useful older evidence, and distinguishable temporal preference components. Each model is tested against its own observable commitment, not a shared behavioral criterion. Appendix F records the assignments.

Evaluation Protocol Each comparison fixes checkpoint, inference settings, user, task input, target/reference, and scoring context. Appendix F.1 and Table 7 record evidence budgets, candidates, and consumed-history settings. Reanalysis uses saved outputs, with no training or new model inference. Recommendation reports MRR and related rank readouts. For compact tables, generation abbreviates ROUGE as RG and PerSEval as PSE. PSE-RG-L, PSE-RG-SU4, and PSE-BERT use their named base measurements. Original generation uses equal-document aggregation. Other configurations retain their declared weights. Prose uses raw scores. Table 21 rescales scores by 100 only for display.

Research Questions.

RQ1: Does the model behave as its design rationale suggests? We isolate LT, ST, and EP evidence under matched task conditions. Let $(\widehat{\mu}_L, \widehat{\mu}_S, \widehat{\mu}_E)$ denote their cohort-level responses on the same cases and weights. For compact reporting, we use

$$\begin{aligned} &\text{LNC} = \widehat{\mu}_L - \min(\widehat{\mu}_S, \widehat{\mu}_E), \qquad \text{STA} = \widehat{\mu}_S - \widehat{\mu}_L, \\ &\text{TRD} = \max(\widehat{\mu}_L, \widehat{\mu}_S, \widehat{\mu}_E) - \min(\widehat{\mu}_L, \widehat{\mu}_S, \widehat{\mu}_E), \qquad \text{MRS} = \min_{r \neq r'} |\widehat{\mu}_r - \widehat{\mu}_{r'}|. \end{aligned} \tag{1}$$

Table 1: **RQ1 differences** $\times 100$**.** Diagnostics: LNC = LT versus the weaker transient role. STA = ST minus LT. TRD/MRS = largest/smallest role gap. Readouts are MRR/nDCG/SR-ERR (Rec.) and PSE-RG-L/PSE-RG-SU4/PSE-BERT (Gen.). T2 is the second PENS interest-injection variant. Only Mamba4Rec's exact-target row supports an interval-based adherence verdict. The other rows are descriptive point estimates. Full role means: Tables 12 and 13, Appendix I.1.

Rationale	Task / Model	Diagnostic	Three readouts	Observed pattern
Long-range	PENS Rec. / NRMS	LNC	$+.316/+.322/+.095$	Positive estimate
Long-range	Movie / Mamba4Rec	LNC	$-3.371/-4.465/-.177$	Exact: contradicted
Recency	PENS Rec. / EBNR	STA	$+.026/-.010/+.012$	Mixed signs
Recency	OAI Gen. / EBNR-T2	STA	$+1.580/+1.548/+19.894$	Positive estimate
Profile	PENS Gen. / NAML-T2	TRD	$.050/.006/.083$	Small role gaps
Profile	OAI Gen. / NAML-T2	TRD	$.395/.449/7.918$	Larger role gaps
Factorized	PENS Rec. / Walk2Pers	MRS	$.232/.199/.014$	Positive estimate
Factorized	PENS Gen. / Walk2Pers	MRS	$.004/.038/.012$	Positive estimate

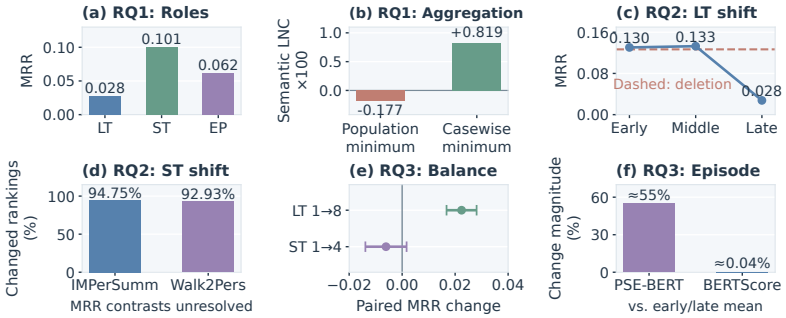

Figure 2: **Controlled tests of RQ1 (a,b), RQ2 (c,d), and RQ3 (e,f).** Mamba4Rec: $6{,}040$ MovieLens cases in (a,b,e), $3{,}436$ in (c). Panel (b)'s semantic LNC is LT relative to the weaker transient role under SR-ERR. ST suite (d): $15{,}552$ cases per model. In (e), the other evidence stays at one copy. Whiskers are exploratory simultaneous 95% intervals. In (f), middle placement is relative to each metric's early/late mean; PSE-BScore decreases. Scales differ. Full results: Tables 11- 24, Appendices I- I.4.

These summarize long-term non-collapse (LNC), short-term advantage (STA), overall role dispersion (TRD), and minimum role separation (MRS). They are descriptive until the model's rationale specifies which relation is expected, with what margin and uncertainty.

RQ2: When RQ1 is supported, does that behavior persist under changes to the same evidence? We move, delete, repeat, or age the frozen evidence while keeping the remaining test conditions fixed. If RQ1 is not supported, these tests are interpreted only as diagnostics of how the response changes.

RQ3: How does model respond to distracting or competing evidence? We vary irrelevant distractors, temporary preferences, contradictory evidence, persistence, placement, and LT/ST balance.

Statistical Analysis. Paired analyses use 10K bootstrap resamples while keeping all conditions of each resampled unit together: users for recommendation, complete documents for original generation, and the recorded grouping for other suites. We report pointwise intervals for single quantities and simultaneous intervals for declared families (Appendix J). These are exploratory summaries, not the finite-sample guarantees of Appendix B.5; a directional interval crossing zero is therefore unresolved, and not evidence of invariance.

6 Results & Analysis

RQ1: Design Adherence. The rationale-specific diagnostics reveal different adherence patterns across model families. For Mamba4Rec, the long-range commitment is tested through long-term non-collapse (LNC), which requires LT not to fall below both transient roles. Under MRR, LT is 0.028 versus 0.101 for ST and 0.062 for EP on $6{,}040$ movie recommendation cases, giving a negative LNC and adjusted LT-minus-transient intervals below zero. This contradicts the exact-target long-range relation. The paired appendix decomposition shows the same direction under nDCG, whereas under SR-ERR, which credits semantic substitutes, the LT–EP comparison remains unresolved (Table 14,

Table 2: **Cross-RQ summary for specialized temporal models on PENS.** RQ1 reports minimum role separation (MRS) using MRR for recommendation and PSE-BERT for generation. RQ2 reports early/middle/late (E/M/L) movement or its observed range. RQ3 reports the indicated controlled response. These are descriptive results, not additional adherence verdicts. All reported scores are on their raw scale. Full results are in Appendices I.1-I.4.

Model / Task	RQ1: Role separation	RQ2: Movement	RQ3: Interference
IMPerSumm / Rec.	MRS 6.60×10^{-4}	.112/.114/.114	Distractor .103→.083
IMPerSumm / Gen.	MRS 1.00×10^{-4}	.020–.023	Spike .023/.000/.021
Walk2Pers / Rec.	MRS 2.32×10^{-3}	.104/.106/.105	Spike .091/.073/.093
Walk2Pers / Gen.	MRS 1.20×10^{-4}	.031–.034	Spike .033/.015/.033

Appendix I.1). Thus, the same controlled histories contradict the required LNC under exact-target retrieval without establishing the same conclusion under the semantic readout. The exploratory interval between LT with the weaker transient mean is $[-0.041, -0.026]$ MRR, showing persistence for every nonnegative required LT advantage in the recorded threshold analysis (Appendix K).

On the other hand, NRMS has a positive LNC estimate on PENS recommendation, in contrast to the negative Mamba4Rec LNC. EBNR's short-term advantage (STA) has mixed signs across the three recommendation readouts but positive estimates in OAI generation. NAML-T2 has a small role dispersion (TRD) on PENS generation and a larger TRD on OAI, while IMPerSumm and Walk2Pers exhibit nonzero minimum role separation (MRS) in PENS recommendation and generation (Table 2). These remain descriptive role-response estimates unless the corresponding rationale supplies the required relation, margin, and uncertainty. Mamba4Rec further shows that even a fixed diagnostic depends on the aggregation rule specified by the contract. Its semantic LNC is -1.77×10^{-3} when LT is compared with the weaker transient role after averaging, but $+8.19 \times 10^{-3}$ when that role is selected separately for each case before averaging (Figure 2(b)). Case-specific comparison roles produce the reversal in Proposition D.3. Hence, *changing the aggregation rule can reverse the tested LNC conclusion without changing any prediction.*

RQ2: Retention and History Changes. The Mamba4Rec LT suite localizes the earlier contradiction to how the same frozen evidence is presented. In Figure 2(c), early and middle placement obtain MRR near 0.13, whereas late placement falls to 0.028. The late-minus-early difference is -0.103 MRR and the late-minus-deletion difference is -0.099, with the latter's adjusted interval excluding zero. Of $1{,}000$ cases recovering the target early or in the middle, 881 lose it late, corresponding to 25.64% of the full $3{,}436$-case cohort. The response therefore cannot be reduced to the selected LT evidence being uniformly helpful or unhelpful: the same frozen evidence behaves differently when its position changes. Movement changes placement, whereas deletion removes evidence Proposition B.1). Mean SR-ERR changes by only -0.013 overall but by -0.094 among the 860 cases losing an early-placement hit late, so a small average semantic change masks much larger losses for those users (Appendix J). Repetition still leaves late below early at four copies.

The broader RQ2 suites expose a second distinction: retention of an aggregate score is not retention of individual responses. Late ST placement changes more than 92% of rankings for both IMPerSumm and Walk2Pers, while their adjusted MRR contrasts remain unresolved (Figure 2(d), Appendix I.2). Their movement profiles are summarized in Table 2. Different lists can retain the target at the same rank or miss it under both conditions, so a stable aggregate rank metric need not imply unchanged outputs. In the Walk2Pers PENS ST suite, mean PSE-BScore changes by only 5.25×10^{-4} across $15{,}237$ cases, despite a mean absolute casewise change of 0.055, illustrating the cancellation permitted by Proposition D.1. The threshold profiles reinforce the same distinction: the IMPerSumm late-minus-early MRR interval is $[-3.38, 6.27] \times 10^{-4}$, while the Walk2Pers PSE-BERT interval is $[-5.255, 6.305] \times 10^{-3}$. Both remain unresolved under sufficiently small invariance tolerances even when individual outputs change substantially (Table 3). This reveals *an unresolved aggregate contrast can therefore coexist with widespread case-level changes, while the same frozen evidence can change its effect through placement alone.*

RQ3: Resistance to Interference. The interference controls separate recorded strength from persistence and placement. For Mamba4Rec, whose RQ1 relation is contradicted, these remain diagnostic rather than adherence verdicts. Increasing nominal conflict strength from 0.60 to 0.95

Table 3: **Threshold sensitivity on fixed cohorts.** Intervals are exploratory simultaneous nominal 95% bounds within the corresponding saved families. For invariance, ϵ denotes the allowed score change. Full threshold profiles are in Appendix K.

Comparison	n	Estimate [interval]	Threshold implication
Mamba4Rec: LT − weaker transient	6,040	$-.034$ $[-.041, -.026]$	Contradicted for every required LT advantage $\gamma \geq 0$.
Mamba4Rec: LT late − early MRR	3,436	$-.103$ $[-.119, -.087]$	Contradicted if $\epsilon < .087$; unresolved for $.087 \leq \epsilon < .119$; supported if $\epsilon \geq .119$.
IMPerSumm: ST late − early MRR	15,552	1.45×10^{-4} $[-3.38, 6.27]\times10^{-4}$	Unresolved for $\epsilon < 6.27\times10^{-4}$; supported thereafter.
Walk2Pers: ST late − early PSE-BERT	15,237	5.25×10^{-4} $[-5.255, 6.305]\times10^{-3}$	Unresolved for $\epsilon < 6.305\times10^{-3}$; supported thereafter.

changes MRR only from 6.38 to 6.40. By contrast, increasing conflict persistence from one to five turns reduces MRR from 5.73 to 4.26, while moving the same type of conflict from early to late reduces it from 14.71 to 5.73 (Table 4). Both adverse differences have adjusted intervals below zero (Appendix J). Thus, the response is not determined by recorded strength alone: how long competing evidence persists and where it appears produce substantially different effects.

Table 4: **RQ3 conflict controls.** Mamba4Rec, MovieLens. MRR $\times100$. Descriptive endpoints. Details in Table 21, Appendix I.4 (RQ3 results).

Controlled change	MRR $\times100$
Strength $0.60 \rightarrow 0.95$	$6.38 \rightarrow 6.40$
Conflicting turns $1 \rightarrow 5$	$5.73 \rightarrow 4.26$
Placement early $\rightarrow$ late	$14.71 \rightarrow 5.73$

Evidence balance reveals a complementary distinction between evidence amount and position. Adding LT while holding ST fixed improves MRR by 0.022 $[0.017, 0.028]$ (Figure 2(e)), whereas increasing ST with LT fixed remains unresolved. *Older evidence can therefore help this model even though placing it late is adverse.* The generation controls expose a related distinction at the measurement level. For Walk2Pers, a brief preference episode gives PSE-BERT values $0.033/0.015/0.033$ at early/middle/late placement; for IMPerSumm, the corresponding profile is $0.023/0.000/0.021$ (Table 2). Consistently, Figure 2(f) shows that Walk2Pers's middle placement reduces PSE-BScore by about 55%, while BERTScore changes by only about 0.04% relative to each metric's early/late mean. Because PerSEval compares users (Appendix C.3), this identifies a personalization-sensitive response that broad reference similarity largely misses. It is not a rationale contradiction unless the valid temporary interest has a declared expected relation (Appendix I.4). Thus, *interference is not a single scalar property: persistence, placement, evidence balance, and the measurement used to read the response expose different forms of behavioral sensitivity.*

7 Limitations

ISCon requires evidence and measurements that validly represent the tested rationale. LT/ST/EP roles are operational constructs validated against source history and semantic consistency, not direct observations of latent preference. Output comparisons also do not identify internal mechanisms. Construction may fail to produce a valid test for some histories. Empirical intervals are exploratory and conditional on saved configurations. The formal error guarantee requires its stated sampling and coverage assumptions.

8 Conclusion

We introduced ISCon to turn a model's rationale into an explicit behavioral test and PrefTurb to instantiate it with matched preference histories. Across recommendation and generation, the tests separate basic adherence from follow-up diagnosis, including harmful LT placement despite benefit from added LT and a personalization-sensitive change nearly invisible to BERTScore. They report support, contradiction, or unresolved evidence without judging which rationale is preferable.

AI USE STATEMENT

Generative AI tools assisted manuscript revision and figure polishing, LaTeX preparation, and citation checks. All scientific decisions remained with the authors, who take full responsibility for the veracity, originality, and integrity of the work, including its methods, proofs, reported results, references, and final wording.

ETHICS STATEMENT

Interaction histories, if not audited and anonymized, can encode sensitive preferences, and exploitative personalization can amplify exposure bias or support unwanted profiling. The reported recommendation and generation metrics do not establish privacy, fairness, or suitability for deployment. Applications should therefore assess consent, access to user histories, disparate effects, and user control over personalization. Appendix L.1 identifies the external datasets and models and their applicable terms.

REPRODUCIBILITY STATEMENT

Appendices A through E provide the assumptions, statements, and proofs for ISCON and PREFTURB's conformance. Appendix F identifies the models, comparison cohorts, training settings, recorded execution settings, and supplied prompt interfaces. Appendices G.2 and H specify evidence selection, trajectory construction, and history edits, including the recorded operator levels. Appendix H.5 defines measurements, aggregation, and scorer configurations. Appendix H.6 reports candidate-evidence validation and local sensitivity. Appendices I through I.4 collect the model results, and Appendices J and K describe the paired statistical analyses and threshold sensitivity. Appendix L gives the reconstruction records, claim boundaries, and asset usage conditions.

REFERENCES

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 82–92, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.7. URL https://aclanthology.org/2021.acl-long.7/.

Jonathan E. Atkins, Erik G. Boman, and Bruce Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998. doi: 10.1137/S0097539795285771.

Parthiv Chatterjee, Asish Joel Batha, Tashvi Patel, Sourish Dasgupta, and Tanmoy Chakraborty. Beyond markovian drifts: Action-biased geometric walks with memory for personalized summarization. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=HvOKarTubb.

Parthiv Chatterjee, Dhara Jhaveri, Sourish Dasgupta, and Tanmoy Chakraborty. IMPerSumm: Information-modulated user preference modeling for personalized text summarization. In *Proceedings of the ACM Web Conference 2026*, pp. 6774–6785, Dubai, United Arab Emirates, 2026b. Association for Computing Machinery. doi: 10.1145/3774904.3792642. URL https://doi.org/10.1145/3774904.3792642.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys*, 51(1):1–27, 2018. doi: 10.1145/3143561. URL https://doi.org/10.1145/3143561.

Victor Coscrato, Rafael Izbicki, and Rafael B. Stern. Agnostic tests can control the type I and type II errors simultaneously. *Brazilian Journal of Probability and Statistics*, 34(2), 2020. doi: 10.1214/19-BJPS431. URL https://doi.org/10.1214/19-BJPS431.

G. A. Croes. A method for solving traveling-salesman problems. *Operations Research*, 6(6):791–812, 1958. doi: 10.1287/opre.6.6.791.

Sourish Dasgupta, Ankush Chander, Parth Borad, Isha Motiyani, and Tanmoy Chakraborty. PerSEval: Assessing personalization in text summarizers. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=yqT7eBz1VJ.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL https://arxiv.org/abs/2408.00118.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015. doi: 10.1145/2827872.

Madhurima Khirbat, Yongli Ren, Pablo Castells, and Mark Sanderson. Metamorphic evaluation of ChatGPT as a recommender system. *arXiv preprint arXiv:2411.12121*, 2024. doi: 10.48550/arXiv.2411.12121. URL https://arxiv.org/abs/2411.12121.

Sein Kim, Hongseok Kang, Kibum Kim, Jiwan Kim, Donghyun Kim, Minchul Yang, Kwangjin Oh, Julian McAuley, and Chanyoung Park. Lost in sequence: Do large language models understand sequential recommendation? In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 1160–1171. Association for Computing Machinery, 2025. doi: 10.1145/3711896.3737035. URL https://doi.org/10.1145/3711896.3737035.

Junhong Lian, Xiang Ao, Xinyu Liu, Yang Liu, and Qing He. Panoramic interests: Stylistic-content aware personalized headline generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1109–1112, 2025. doi: 10.1145/3701716.3715539. URL https://doi.org/10.1145/3701716.3715539.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. Mamba4Rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*, 2024a. doi: 10.48550/arXiv.2403.03900. URL https://arxiv.org/abs/2403.03900.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Pingchuan Ma, Shuai Wang, and Jin Liu. Metamorphic testing and certified mitigation of fairness violations in NLP models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 458–465, 2020. doi: 10.24963/ijcai.2020/64. URL https://www.ijcai.org/proceedings/2020/64.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Chengying Mao, Xiaorong Yi, and Tsong Yueh Chen. Metamorphic robustness testing for recommender systems: A case study. In *2020 7th International Conference on Dependable Systems and Their Applications (DSA)*, 2020. doi: 10.1109/DSA51864.2020.00060. URL https://doi.org/10.1109/DSA51864.2020.00060.

Chengying Mao, Jifu Chen, Xiaorong Yi, and Linlin Wen. An empirical study on metamorphic testing for recommender systems. *Information and Software Technology*, 169:107410, 2024. doi: 10.1016/j.infsof.2024.107410. URL https://doi.org/10.1016/j.infsof.2024.107410.

Chimaobi Okite, Naihao Deng, Kiran Bodipati, Huaidian Hou, Joyce Chai, and Rada Mihalcea. Benchmarking and improving LLM robustness for personalized generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16040–16072, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.870. URL https://aclanthology.org/2025.findings-emnlp.870/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Alireza Salemi, Julian Killingback, and Hamed Zamani. ExPerT: Effective and explainable evaluation of personalized long-form text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17516–17532, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.900. URL https://aclanthology.org/2025.findings-acl.900/.

Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. General then personal: Decoupling and pre-training for personalized headline generation. *Transactions of the Association for Computational Linguistics*, 11:1588–1607, 2023. doi: 10.1162/tacl_a_00621. URL https://aclanthology.org/2023.tacl-1.90/.

Rahul Vansh, Darsh Rank, Sourish Dasgupta, and Tanmoy Chakraborty. Accuracy is not enough: Evaluating personalization in summarizers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2582–2595, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.169. URL https://aclanthology.org/2023.findings-emnlp.169/.

Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 832–841. Association for Computing Machinery, 2023. doi: 10.1145/3539618.3591663. URL https://doi.org/10.1145/3539618.3591663.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3863–3869, 2019a. doi: 10.24963/ijcai.2019/536. URL https://www.ijcai.org/proceedings/2019/536.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6389–6394, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1671. URL https://aclanthology.org/D19-1671/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Xiaoyuan Xie, Zhiyi Zhang, Tsong Yueh Chen, Yang Liu, Pak-Lok Poon, and Baowen Xu. METTLE: A METamorphic testing approach to assessing and validating unsupervised machine learning systems. *IEEE Transactions on Reliability*, 69(4):1293–1322, 2020. doi: 10.1109/TR.2020.2972266. URL https://doi.org/10.1109/TR.2020.2972266.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. LLM-based medical assistant personalization with short- and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2386–2398, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.132. URL https://aclanthology.org/2024.naacl-long.132/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.927. URL https://aclanthology.org/2025.findings-acl.927/.

APPENDIX CONTENTS

Notation Reference **17**

A Contract Properties And Proofs **17**

A.1 From Design Rationales To Score Relations 17

A.2 Construction Properties Of Frozen-Evidence Interventions 18

A.3 Finite Tests Of A Declared Behavioral Relation 18

A.4 Rationale-Dependent Movement Invariance 18

A.5 Monotone Responses And Sampled Survival Sets 19

A.6 Scope Of Finite Behavioral Tests 19

A.7 Exact And Semantic Readouts 19

A.8 Aggregation Order And Non-Collapse 20

B Formal Conformance Of PREFTURB To ISCON **20**

B.1 Specification And Assumptions 21

B.2 Formal Conformance 21

B.3 Frozen-Evidence Construction And Composition 22

B.4 Equality Of The Measured Contract Quantities 23

B.5 An Explicit Simultaneous Uncertainty Procedure 23

B.6 Sound Decisions And Resolution Away From Boundaries 24

B.7 Proof Of The Conformance And Decision-Transfer Theorem 25

B.8 Sound Follow-Up Claims For Retention And Interference 26

B.9 Examples Explaining The Assumptions 26

C Concrete Measurement And Its Formal Properties **27**

C.1 Recommendation Measurement And Its Bounds 27

C.2 Concrete Generation Configurations 28

C.3 A Specified PerSEval Measurement With Complete Peers 28

C.4 Measurement Conformance And Transfer 30

D Concrete Aggregation And Its Formal Properties **30**

D.1 Concrete Weights And Their Meaning 30

D.2 Pairing And The Order Of Linear Comparisons 31

D.3 Nonlinear Aggregation And Comparator Selection 32

D.4 Grouped Uncertainty And Aggregation 33

D.5 Aggregation Conformance And Transfer 33

E Concrete Relations And Their Decision Guarantees **34**

E.1 Explicit Relations And Their Meaning 34

E.2 Exact Bounds For Unordered Role Separation 35

E.3 Sound Decisions With Strict And Closed Boundaries 35
E.4 Threshold Profiles From One Uncertainty Set 36

F Implementation And Evaluation Specification **36**
F.1 Evaluation Units And Controls 36
F.2 Datasets And Comparison Cohorts 38
F.3 Recorded Model and Scoring Configurations 38
F.4 Prompted LLM Interfaces 38

G Target-Conditioned Trace Construction **40**
G.1 Candidate Extraction And Temporal-Role Validation 40
G.2 Event and Target Representation 40
G.3 Embedding-Space Validity 40
G.4 Temporal Role Conditions 40
G.5 Preference-Space Ordering 42
G.6 Construction Algorithm 43

H Perturbation Operator Definitions **44**
H.1 Common Operator Specification 45
H.2 Operator Realization 45
H.3 Operator-Specific Conditions 45
H.4 Applicability and Reporting 46
H.5 Readouts And Aggregation 46
H.6 Validation Of Selected Preference Evidence 47

I Cross-Setting Results Overview **48**
I.1 RQ1: Role Evidence 48
I.2 RQ2: Movement, Deletion, And Re-Exposure 48
I.3 Complete Short-Term Response Profiles 49
I.4 RQ3: Interference And Competing Evidence 49

J Supporting Statistical Analyses **57**

K Threshold Sensitivity Analysis **57**

L Reproducibility And Claim Boundaries **58**
L.1 Asset Terms And Usage 58

M Extended Related Work **58**
M.1 Recommender Reliability And Sequence Use 58
M.2 Behavioral Relations And Specification Testing 59

M.3 Context Position And Conversational Memory 59
M.4 Construction And Interpretation Of Controlled Tests 59
M.5 Personalization Measurements And Benchmarks 60
M.6 Master Comparison 60

NOTATION REFERENCE

The notation below in the Table 5 follows the contract, its history construction, and the decision made from the scored outputs. Symbols reused in different settings are qualified by their role.

Table 5: Notation for the specified adherence test. A temporal role describes selected evidence, whereas a design rationale states the model behavior being tested.

Symbol	Meaning
ISCON *specification, Section 3*	
F, $\mathcal{A}_F$, $\mathfrak{C}_F$	Model's design rationale, evidence conditions, and required relation between scored responses.
M, z_u, T_u	Fixed prediction interface, task input, and target or reference with scoring context for unit u.
Ω, H_u^ω	Tested conditions and the controlled history supplied in condition ω.
S, $R_u(\omega)$	Measurement and measured response under a controlled history.
h, X_u	Operation applied before averaging and the resulting quantity for unit u.
w_u, $\widehat{\mu}$, μ	Fixed nonnegative weights summing to one, observed aggregate, and corresponding population target.
PREFTURB *construction, Section 4 and Appendix G.2*	
$e_t = (x_t, a_t)$	An event's content and recorded action.
ψ_{evt}, ψ_{tar}, $s_u(t)$	Frozen event and target encoders and event-target affinity.
LT, ST, EP	Long-term, short-term, and episodic origin roles, fixed before movement.
$r_u(t)$, $\ell_u(t)$	Temporal role and separate target relation (supporting, irrelevant, or contradictory).
B, C, $\|$	Selected evidence, remaining history, and sequence concatenation.
$H_{p,K}$	History placing K copies of B after the first p events of C.
d, A_{ij}, λ	Event distance, semantic affinity, and weight on the path's affinity penalty.
Measurement and decisions, Appendices C–E	
m, n_d	Number of retained documents and retained users for document d, in equal-document aggregation.
I_d	Fixed set of users whose outputs and references enter a document's personalization measurement.
ϵ, η, κ	Negligible-change tolerance, minimum directional gain, and minimum role separation. These are score margins.
Q, α	Joint confidence set and its declared error level.
g_j, L_j, U_j	Signed requirement and valid lower and upper bounds over the confidence set.

The construction neighborhood radius also uses ϵ in Appendix G.2. It is an embedding-distance threshold, distinct from the score tolerance above. The cutoff K in ranking measurements is distinct from the repetition count K in $H_{p,K}$. PerSEval's parameters are likewise distinct from the decision error level α.

A CONTRACT PROPERTIES AND PROOFS

The contract properties explain what must remain explicit when a design rationale becomes a test. The expected relation determines the meaning of a response. Construction guarantees specify what the input comparison preserves. Measurement and aggregation determine which behavioral quantity is evaluated, and joint coverage determines when it can receive a verdict. The proofs below justify these links under their stated assumptions. They do not assume that trained models satisfy the resulting contracts.

A.1 FROM DESIGN RATIONALES TO SCORE RELATIONS

Keeping the same internal representation and deterministic output rule keeps the output unchanged. A nearly unchanged representation requires a sensitivity assumption before it can imply a nearly

unchanged score. Let D map an internal representation to the model output. If $\|h - h'\| \leq \epsilon$ and the composite output score $g_q(h) = S_q(D(h), y^\star)$ is L_q-Lipschitz, then $|g_q(h) - g_q(h')| \leq L_q\epsilon$. Here L_q bounds how much the score can change per unit of representation change. A small $L_q\epsilon$ would bound the observed score difference under these assumptions. Discrete ranking and decoding need not satisfy a useful global bound. Near a tie, a small logit change can swap the top item and change its exact-target score abruptly. Thus, a representation-level design claim and an output-score contract require an explicit bridge. The latter cannot be inferred merely by replacing hidden-state distance with a task metric.

For sequential updates, let F_k map a state and the next input to a new state at step k. Suppose $\|F_k(h, x) - F_k(h', x)\| \leq \lambda\|h - h'\|$ with $0 < \lambda < 1$. For two histories differing only in their state at time t and sharing the subsequent inputs, induction gives $\|\Delta h_T\| \leq \lambda^{T-t}\|\Delta h_t\|$. Here Δh_t is the state difference at step t. Each shared subsequent input shrinks its upper bound by a factor λ. More subsequent history therefore permits a smaller bound on the remaining state difference. This state-difference bound does not order task scores when an event moves. An earlier perturbation can align better with the target than a larger recent perturbation. A nonzero long-range Jacobian, the derivative linking a past input to a later state, permits dependence without guaranteeing that it improves the target score. Distinct temporal modules can also produce identical task outputs. The family relations in the main text make this bridge explicit through readout and evidence-comparability assumptions. A contradiction can then challenge the stated behavioral interpretation without purporting to disprove the architectural construction.

A.2 Construction Properties Of Frozen-Evidence Interventions

Proposition B.1 is established by a frozen-ledger argument in Appendix B.3, including finite compositions of edits. For a block of two events and a remainder of four, one copy gives six events and two copies give eight. Movement with one copy preserves all six event identities. These illustrative counts give an implementation check that is independent of model scores. Labels record the original evidence role, while rebuilt predecessor fields record the new sequence. Semantic label validity and evidence reaching the predictor remain separate requirements.

A.3 Finite Tests Of A Declared Behavioral Relation

Reliable population verdicts require bounds that cover every score comparison in the contract. Theorem 3.1 applies to a mixture of closed and strict requirements. Its full proof appears in Appendix E.3. The coverage-event argument, affine and weak non-collapse bounds, and resolution away from affine boundaries are developed in Appendix B.6. Appendix B.5 supplies an explicit sufficient confidence construction. They do not establish exact coverage of the exploratory intervals used for the saved-output analyses.

For weak non-collapse, $\mu_L - \min(\mu_S, \mu_E) = \max(\mu_L - \mu_S, \mu_L - \mu_E)$. Thus, LT must be below both transient roles to contradict the zero-margin relation. Being below just one role is insufficient. This difference matters when selecting a follow-up comparison after an initial adverse response.

A.4 Rationale-Dependent Movement Invariance

A flat positional response has no intrinsic support-or-contradiction interpretation. Its meaning depends on whether the tested design rationale calls for equal responses or a positive recency advantage. Here $\tau \in [0, 1]$ denotes placement from earlier to later, and active means that the response is above its declared readout floor.

Proposition A.1 (Contract dependence of positional invariance)**.** *Let $R(\tau) = c$ be an active position-flat response. It satisfies exact positional invariance. It cannot satisfy a strict recency condition requiring $R(\tau_j) > R(\tau_i)$ for at least one tested pair $\tau_j > \tau_i$. A weak non-decreasing condition, however, allows this same flat profile.*

Proof. Every paired difference is zero. Thus invariance and weak monotonicity hold, while a required positive difference does not. □

The strictness and practical margin belong to the declared hypothesis. Choosing them after observing a flat profile would change the tested claim to fit the result. The hypothesis must determine what counts as success before the response is interpreted. In particular, small response dispersion at a readout floor does not demonstrate active use of preference evidence.

A.5 Monotone Responses And Sampled Survival Sets

We can separately ask which tested conditions reach a required score. Let δ be that score threshold, Ω_{test} the tested conditions, and $\mu^{E^\star}_{M,q}(\omega)$ the population score for model M, readout q, and selected evidence $E^\star$. The sampled set

$$\widehat{\mathcal{Z}}^{E^\star}_{M,q}(\delta) = \{\omega \in \Omega_{\text{test}} : \mu^{E^\star}_{M,q}(\omega) \geq \delta\} \tag{2}$$

records where the population response exceeds that level. The hat marks restriction to sampled conditions, not an uncertainty estimate of μ. A utility threshold answers where a response remains high. A contract answers whether responses have the expected relationship. Their meanings remain distinct.

If moving evidence later can only increase its score, then passing a fixed score threshold at one position implies passing it at every later position. The next property states this ordering consequence.

Proposition A.2 (Terminal survival zone)**.** *Let $R : [0,1] \to \mathbb{R}$ be non-decreasing and $\mathcal{Z}_\delta = \{\tau : R(\tau) \geq \delta\}$. Then $\mathcal{Z}_\delta$ is upward closed: it is empty, all of $[0,1]$, or an upper interval with an open or closed lower endpoint. On an ordered finite placement grid, membership forms a suffix of the grid.*

Proof. If $\tau \in \mathcal{Z}_\delta$ and $\tau' \geq \tau$, then $R(\tau') \geq R(\tau) \geq \delta$. Restricting this implication to grid points gives the suffix property. □

For three ordered placements, falling below the threshold early and reaching it at middle and late placement is compatible with this property. Reaching it early and falling below it later violates the property for the response function. When scores estimate population means, an apparent reversal also requires uncertainty analysis. A compatible suffix on the tested grid establishes neither monotonicity between grid points nor the full response law.

A.6 Scope Of Finite Behavioral Tests

Two models can agree on every tested history and disagree on a history we did not test. Finite agreement therefore cannot establish unrestricted agreement.

Proposition A.3 (Finite-test non-identifiability)**.** *Suppose the history domain contains an untested history H' and the output space contains at least two outputs. In the unrestricted class of history-to-output functions, any function f has a distinct counterpart g agreeing with it on every history in a finite tested set $\mathcal{H}$ and differing at H'.*

Proof. Choose an admissible output $y' \neq f(H')$. Set $g(H) = f(H)$ for $H \neq H'$ and $g(H') = y'$. Then g agrees with f on $\mathcal{H}$ and differs outside it. This construction applies to text and ranking outputs without assuming that outputs can be added. □

The statement is about unrestricted behavioral identification. It does not exclude identification under additional restrictions on a known model class. Identifying internal storage would additionally require evidence about the model's internal computation. The useful target of a finite protocol is therefore a declared behavioral relation on its tested domain. It need not identify the unrestricted model to answer that question.

A.7 Exact And Semantic Readouts

A recommendation can miss the exact target while placing an equally relevant substitute first. The following example separates those two scoring questions.

Proposition A.4 (Layered ranking separation)**.** *There exist two rankings with different exact-target reciprocal rank and identical semantic satisfaction under a substitute-aware readout that gives full satisfaction to a semantically equivalent first-ranked item.*

Proof. Place the target first in ranking A. In ranking B, place an item assigned semantic satisfaction one first and the target tenth. Reciprocal ranks are 1 and 0.1. Under the readout, satisfaction one at the first item leaves zero probability of continuing to later items. Both semantic scores therefore equal one. first-rank case of the SR-ERR cascade defined in Equation 11 in Appendix H.5. □

This existence example does not assume that empirical substitutes attain satisfaction one. The ranking difference is real even though the substitute-aware score cannot distinguish the lists. *The example establishes why one readout cannot certify another.* Small changes in a population semantic mean likewise do not imply small changes for every individual.

A.8 Aggregation Order And Non-Collapse

The population calculation first averages each role and then selects the weaker transient role. The instance calculation first selects the weaker transient role for each case and then averages. Different users can select different comparison roles. Let (L, S, E) be the LT, ST, and EP scores of the same randomly sampled case.

Proposition A.5 (The effect of selecting a comparator)**.** *For* $C_{\mathrm{pop}} = \mathbb{E}[L] - \min(\mathbb{E}[S], \mathbb{E}[E])$ *and* $C_{\mathrm{inst}} = \mathbb{E}[L - \min(S, E)]$*,*

$$C_{\mathrm{inst}} - C_{\mathrm{pop}} = \tfrac{1}{2}\big(\mathbb{E}|S - E| - |\mathbb{E}(S - E)|\big) \geq 0. \tag{3}$$

Proof. Substitute $\min(a, b) = (a + b - |a - b|)/2$ into both definitions and cancel the linear terms. Nonnegativity follows from $|\mathbb{E}(S - E)| \leq \mathbb{E}|S - E|$. □

The gap measures cancellation induced by variation across instances in which the transient role is stronger. If the same transient role is at least as strong for every case, the gap is zero. If each transient role is stronger for some cases, the gap can be positive. With the same population and weights, the instance calculation can never be smaller than the population calculation. A nonnegative population value therefore cannot become a negative instance value. A negative population value can become positive.

For an arithmetic illustration, give each of two users LT score 0.3. Let their ST/EP scores be $(0.8, 0)$ and $(0, 0.8)$. Both transient population means are 0.4, giving $C_{\mathrm{pop}} = -0.1$. Each user has a weaker transient score of zero, giving $C_{\mathrm{inst}} = +0.3$. The sign reverses because the comparator changes across users. These invented values illustrate the identity, not an experiment. Appendix J reports the observed example while retaining the population-level RQ1 quantity.

B Formal Conformance Of PrefTurb To ISCon

The main-text conformance theorem joins four requirements that a model designer can check separately. The historians must realize the intended evidence comparison. The measurement must score the specified outputs with their complete context. Aggregation must calculate the declared quantity. A population verdict must use valid uncertainty bounds for that quantity. We establish these links below without assuming that the model follows its rationale.

The construction guarantees the preservation of the recorded evidence attributes. Their semantic interpretation remains a stated assumption. The identities concern the specified procedure, while applying them to a particular executable requires its implementation to follow that procedure. Statistical conclusions additionally require the stated sampling assumptions.

B.1 Specification And Assumptions

An evaluation unit contains everything needed to compute its scores. For recommendation, this can be a user with the target and fixed candidate set. For a personalization score that compares users, it contains the complete declared peer group, its references, and all outputs entering the score. A document and its comparison users is one possible unit. Sharing peers can require a larger unit when assessing independence. Write a unit as u, its fixed task information as z_u, and its references and scoring context as T_u. Let Ω be a finite, nonempty set of history conditions. The prediction interface M includes preprocessing, input serialization, the checkpoint, and inference settings. For stochastic inference, either the target averages over a specified randomness law or it conditions on a declared realization scheme. Saved outputs alone do not justify a claim about averaging over decoding randomness.

The following requirements specify the scope of the result.

(i) **Fixed test specification.** The model rationale, tested conditions, selection and repair rules, measurement, eligibility, aggregation rule, and practical margins are specified independently of the evaluated outcomes. A finite collection of possible follow-ups can be specified together. Retrospectively chosen specifications remain descriptive unless inference accounts for that selection. (ii) **Accepted frozen evidence.** Selection and any permitted repairs finish before intervention comparisons. The final accepted ledger records unique source-instance identifiers, content, available actions, origin roles, target relations, and ordered traces. Acceptance checks enforce the declared operational role, relation, and continuity rules. A repair can change the selected event set before it is frozen. Ambiguous or rejected cases remain inapplicable under the declared eligibility rule. No rejected case is silently scored as zero. (iii) **Declared intervention and input conditions.** Each edit specifies retained identities, copies, ordering, and permitted changes to length, adjacency, or context. Task information, references, candidates or peers, and prediction settings remain fixed within a comparison. Any required evidence-access property is checked at the appropriate input boundary. A full interface can legitimately be tested for losing evidence through truncation. A narrower claim about a predictor's retention requires that the relevant evidence reaches that predictor. (iv) **Evidence and measurement interpretation.** The selected evidence satisfies the substantive comparability and relevance conditions of the tested rationale, and the specified score relation is a justified observable commitment of that rationale in this domain. Encoder thresholds and high task scores do not prove this requirement. An operational-label-only interpretation must be described at that narrower scope. An invariance claim about useful evidence use needs an activity or evidence-sensitivity requirement where warranted. Activity alone is not proof of evidence use. (v) **One declared aggregation target.** All conditions in one contract use the same eligible population or common finite cohort, the same weights, the same score definition, and the same scoring-pool policy. Different cohorts or metric configurations define separate contracts unless their relationship is explicitly modeled. The order of per-case nonlinear operations and averaging is fixed. An evidence-based acceptance rule determines the domain before outcome comparisons. Response-dependent missingness or exclusions require an explicit treatment rather than silently changing the target population. (vi) **An identified inference basis.** Exact arithmetic on a fixed, fully observed cohort can decide its descriptive relation. Population inference instead requires a stated sampling model and simultaneous coverage. Appendix B.5 supplies one explicit procedure under independent bounded evaluation units. A different procedure can be substituted with its own coverage justification.

These requirements do not say that the model's responses satisfy the contract. In particular, the expected score relation is a proposition to test, not an assumption in the construction proof.

B.2 Formal Conformance

A conformance proof must identify the exact relation the protocol is meant to compute. For a declared rationale F, let $\mathcal{A}_F$ denote its evidence and input conditions. Let S be the specified measurement, including all peer-dependent context when needed. Collect the condition scores of unit u into R_u. A fixed map h forms the per-unit quantities that are subsequently averaged. Write the resulting mean vector as μ. For closed requirements, the contract is

$$\mathfrak{C}_F(\mu) \iff g_{F,j}(\mu) \geq 0 \quad \text{for every } j \in J_F,$$

where J_F is finite and nonempty. Margins are part of the fixed functions $g_{F,j}$. Strict requirements replace the relevant $\geq$ by $>$; Appendix E.3 proves the mixed strict/closed version used in Theorem 3.1. All random variables, bound computations, and decision events in the statistical statements are assumed measurable.

A protocol structurally instantiates this contract on an admissible domain if its accepted inputs satisfy the declared construction conditions and it computes the same measurement and averaged quantities that appear in $\mathfrak{C}_F$. An uncertainty-based verdict is statistically sound at level α if the probability of any false support or contradiction declaration in the declared family is at most α, under the stated sampling assumptions.

This definition separates computing the right test from the tested model passing it. The evidence and measurement interpretation remains an explicit scope assumption, not a conclusion obtained from event counting.

B.3 Frozen-Evidence Construction And Composition

Once evidence is selected and accepted, an edit can change its declared position or number of copies without quietly changing what that evidence is. Let $\mathcal{L}_u$ be the final frozen ledger. A source identifier i retrieves a fixed payload $v_u(i)$ consisting of the event content, available action, and frozen origin labels. Every output occurrence has a provenance identifier and, if repeated, a separate occurrence identifier. An edit is specified by an ordered list of provenance identifiers. Its history is obtained by retrieving their payloads in that order. Structural fields such as predecessor links are regenerated to encode this order and are not treated as immutable semantic payload.

Proposition B.1 (Preservation under declared edits)**.** *Every edit constructed in this way preserves each retained occurrence's payload. Its multiplicity is the number of appearances of the corresponding provenance identifier in the edit list. Its realized order is the declared list order. These properties are preserved by any finite composition of such edits, provided added evidence is entered in the ledger before the comparisons and every composition declares its resulting multiplicities and order.*

In particular, for disjoint source-instance sequences B and C of lengths m and n, respectively,

$$H_{p,K} = C_{1:p} \| B^{\| K} \| C_{p+1:n}, \qquad 0 \leq p \leq n, \quad K \in \{0, 1, \ldots\}, \tag{4}$$

has length $n + Km$. With $K = 1$, movement preserves the event multiset and the internal orders of B and C. With fixed p, changing K changes the selected multiplicities and the resulting length, positions, and adjacency. Here $\|$ denotes concatenation, and $B^{\|0}$ is empty.

Proof of Proposition B.1. At every output occurrence, construction retrieves $v_u(i)$ from the same frozen ledger. Its payload therefore cannot depend on the edit condition. Counting the appearances of i gives its output multiplicity, and traversing the list gives its output order. For a sequence of edits, the first edit has these properties. If an intermediate history has them, the next permitted edit copies, removes, or reorders existing provenance references, or inserts references already recorded in the ledger. Every resulting occurrence still retrieves the same payload. Induction proves preservation through any finite number of edits. Multiplicities and order are those of the composed list, and hence no unchanged-total-count claim is made when the composition duplicates or removes events. For $H_{p,K}$, the two pieces of C contribute $p + (n - p) = n$ events and the copies of B contribute Km. With $K = 1$, changing p changes the insertion boundary while leaving each sequence's internal order intact. The other statements follow by counting copies. □

Role isolation is a declared sublist. Deletion omits specified identifiers. Repetition duplicates them. Distractor, spike, and conflict edits insert frozen identifiers with the corresponding accepted relations. Persistence and balance use declared counts and placements. This representation covers these edit types at the sequence-construction level. It does not establish the semantic correctness of their labels or attribute a length-changing edit solely to evidence strength.

Construction guarantees apply after the final repair and freeze. They do not prove that trace search always finds an admissible sequence, that a repair loop terminates, or that approximate ordering solves an optimization problem exactly. A run that does not yield an accepted trace has no applicable test on that trace. A bounded repair budget is a possible implementation safeguard, not a property established for the existing code here.

B.4 Equality Of The Measured Contract Quantities

PrefTurb measures the exact controlled comparison stated by ISCon when it retains the declared scoring context and aggregation rule.

Lemma B.1 (Equality of specified contract quantities). *Let $Y_u^\omega = M(H_u^\omega, z_u)$ denote the output or joint output collection required for scoring. Define*

$$R_u(\omega) = S(Y_u^\omega, T_u), \qquad X_u = h\big((R_u(\omega))_{\omega\in\Omega}\big).$$

Within a fixed finite cohort of N units, let $w_u \geq 0$ and $\sum_{u=1}^N w_u = 1$. PrefTurb's computed vector is

$$\widehat{\mu} = \sum_{u=1}^{N} w_u X_u.$$

If the ISCon specification uses these histories, measurement, map h, cohort, and weights, then PrefTurb's vector equals the specified finite-cohort vector coordinate by coordinate. With fixed weights and integrable unit scores, it also satisfies

$$\mathbb{E}[\widehat{\mu}] = \sum_{u=1}^{N} w_u \mathbb{E}[X_u] = \mu$$

where independence is unnecessary for this expectation identity. Under a common unit distribution P, $\mu = \mathbb{E}_P[X]$.

Proof. For every unit and condition, substituting the frozen-history construction into the prediction interface and then into the declared score gives the ISCon response by its definition. For a peer-dependent score, the substitution includes the entire declared collection of peer outputs and references. Keeping only the focal user's output would not be the same measurement. Applying the same h therefore gives the same per-unit vector in both formulations. Multiplying identical coordinates by the same weights and summing proves finite-cohort equality. Linearity of expectation gives the population identity for fixed weights and integrable coordinates. If all units share distribution P, their expectations coincide, and the normalized weights sum to one. □

The map h explicitly records the order of operations. To compare population role means, retain $h(L, S, E) = (L, S, E)$ and evaluate $g(\mu) = \mu_L - \min(\mu_S, \mu_E)$. To average casewise non-collapse, include the coordinate $h_4(L, S, E) = L - \min(S, E)$ and evaluate $g(\mu) = \mu_4$. *These are different contract quantities, even on identical outputs.*

Shared candidate or peer membership means that pool membership and its scoring rule remain fixed. Peers' outputs can change under the intervention and enter the score accordingly. Resampling saved score contributions is a different operation from recomputing the score on newly assembled peer pools.

B.5 An Explicit Simultaneous Uncertainty Procedure

Population verdicts require an uncertainty set with a stated coverage guarantee. The following construction supplies one for independent bounded evaluation units. Its bounds can be conservative, leaving some contracts unresolved.

Proposition B.2 (Simultaneous coverage for bounded weighted units). *Let $X_1, \ldots, X_N \in \mathbb{R}^d$ be independent evaluation-unit vectors. Coordinates within the same vector can be arbitrarily dependent. Suppose each coordinate has a known deterministic bound $a_k \leq X_{u,k} \leq b_k$. Let the nonnegative weights w_u be deterministic with sum one, and let*

$$\widehat{\mu}_k = \sum_u w_u X_{u,k}, \qquad \mu_k = \sum_u w_u \mathbb{E}[X_{u,k}].$$

Choose $0 < \alpha < 1$ *and* $\delta_k > 0$ *with* $\sum_{k=1}^{d} \delta_k \le \alpha$. *Define*

$$r_k = \frac{b_k - a_k}{2}\sqrt{\frac{\sum_u w_u^2}{\delta_k}},$$

$$I_k = [\widehat{\mu}_k - r_k, \widehat{\mu}_k + r_k] \cap [a_k, b_k], \qquad Q = \prod_{k=1}^{d} I_k.$$

Then $\Pr(\mu \in Q) \ge 1 - \alpha$.

Proof. For a scalar random variable $Z \in [a, b]$, the pointwise inequality $(Z - a)(b - Z) \ge 0$ gives $\mathbb{E}[Z^2] \le (a + b)\mathbb{E}[Z] - ab$. Writing $m = \mathbb{E}[Z]$, it follows that

$$\mathrm{Var}(Z) \le (b - m)(m - a) \le \frac{(b - a)^2}{4}.$$

Independence between units removes cross-unit covariance terms, so

$$\mathrm{Var}(\widehat{\mu}_k) \le \frac{(b_k - a_k)^2}{4} \sum_u w_u^2.$$

For any centered variable V and $r > 0$, $V^2 \ge r^2 \mathbf{1}\{|V| > r\}$. Taking expectations yields $\Pr(|V| > r) \le \mathbb{E}[V^2]/r^2$. Apply this with $V = \widehat{\mu}_k - \mu_k$ and $r = r_k$. The probability of missing coordinate k is at most δ_k. If $a_k = b_k$, that coordinate is constant and is covered exactly. The true mean is in $[a_k, b_k]$, and hence clipping cannot discard it when the untrimmed interval covers it. The probability that at least one coordinate fails is bounded by the sum of its failure probabilities, at most $\sum_k \delta_k \le \alpha$. This proves simultaneous coverage. □

This elementary bounded-variance construction supplies one sufficient uncertainty procedure. With equal weights, $\sum_u w_u^2 = 1/N$, and hence more independent units narrow the bounds. Concentrating weight on a few units limits that gain. A designer therefore needs independent evidence for the declared average, not merely more correlated scores. No normality assumption or independence across history conditions is required. Its widths may be too large to resolve a practical contract. Narrower procedures need their own justification. This proof does not establish exact coverage of the exploratory bootstrap already reported in the manuscript.

The independence and deterministic-weight assumptions are substantive. Shared users or peers, crossed user/document dependence, and random weights induced by varying cluster sizes cannot be ignored. Grouping genuinely independent complete comparison units is one valid route when the target weighting is preserved. A score without a known bound is outside this proposition. The bounds must hold for the score over its declared domain. The observed sample minimum and maximum are insufficient. Predeclared clipping or normalization changes the measurement and requires corresponding contract margins. Merely conditioning on a peer pool constructed from the same sample does not prove independence.

B.6 Sound Decisions And Resolution Away From Boundaries

An uncertainty set becomes a contract verdict by checking which required relations hold throughout it. We first consider closed requirements. The strict-boundary extension appears in Appendix E.3. Support means every value still allowed by uncertainty satisfies the declared relation. Contradiction means at least one required relation fails throughout that uncertainty set.

Let Q be a nonempty uncertainty set satisfying $\Pr(\mu \in Q) \ge 1 - \alpha$. For each fixed requirement g_j, obtain valid bounds

$$L_j \le \inf_{v \in Q} g_j(v), \qquad U_j \ge \sup_{v \in Q} g_j(v).$$

Declare support if every $L_j \geq 0$, contradiction if some $U_j < 0$, and unresolved otherwise. Then the probability of an incorrect declaration is at most α. The bound covers the union of false support and false contradiction events.

Proof of Theorem 3.1 for closed requirements. On the event $\mu \in Q$, support implies $g_j(\mu) \geq L_j \geq 0$ for every requirement, and hence the contract is true. Contradiction implies $g_j(\mu) \leq U_j < 0$ for at least one requirement, and hence the contract is false. Neither erroneous declaration is possible on this event. Their union is therefore contained in $\{\mu \notin Q\}$, whose probability is at most α. Both labels cannot be issued simultaneously because a nonempty Q implies $L_j \leq U_j$ for every j. All other cases receive the unresolved label. □

For a rectangular Q with coordinate endpoints ℓ_k, u_k, an affine requirement $g(v) = c_0 + \sum_k c_k v_k$ has exact bounds

$$L = c_0 + \sum_{c_k \geq 0} c_k \ell_k + \sum_{c_k < 0} c_k u_k, \qquad U = c_0 + \sum_{c_k \geq 0} c_k u_k + \sum_{c_k < 0} c_k \ell_k.$$

Each summand is minimized at its lower or upper endpoint according to the coefficient's sign, and maximized at the opposite endpoint. Independent coordinate choices within a rectangle attain these extrema. Equivalence $|\mu_a - \mu_b| \leq \epsilon$ is the conjunction of the two affine requirements $\epsilon - (\mu_a - \mu_b) \geq 0$ and $\epsilon + (\mu_a - \mu_b) \geq 0$. Recency advantage by η is the affine requirement $\mu_{\text{late}} - \mu_{\text{early}} - \eta \geq 0$. These examples make the rule computable for central contract types.

For the unadjusted weak non-collapse function $g(v) = v_L - \min(v_S, v_E)$, exact rectangular bounds are $L = \ell_L - \min(u_S, u_E)$ and $U = u_L - \min(\ell_S, \ell_E)$. Monotonicity in each coordinate and the corresponding corner choices establish both bounds. A margin is subtracted from both bounds when the contract requires it. One unfavorable LT comparison alone does not contradict this weak relation.

Soundness does not require every false contract to be resolved. Nonlinear requirements can remain unresolved even when different requirements fail at different points in Q. Numerically estimated extrema need certified conservative bounds for this theorem to apply.

The explicit procedure can resolve sufficiently separated linear requirements. For an affine $g_j(v) = c_{j,0} + \sum_k c_{j,k} v_k$, define $t_j = 2 \sum_k |c_{j,k}| r_k$. On the coverage event, both μ and every $v \in Q$ lie within r_k of $\widehat{\mu}_k$, and hence $|v_k - \mu_k| \leq 2r_k$. The triangle inequality gives $|g_j(v) - g_j(\mu)| \leq t_j$. Therefore exact bounds support the contract if every $g_j(\mu) \geq t_j$, and contradict it if some $g_j(\mu) < -t_j$. For fixed coordinate bounds and error allocations, $\sum_u w_u^2 \to 0$ implies $r_k \to 0$. For a fixed true mean vector and fixed finite affine contract, every sufficiently large sample design therefore resolves a strictly satisfied contract, or one with a strictly negative requirement, whenever that design's coverage event holds. No single coverage event over all sample sizes is asserted. This argument establishes possible resolution away from the boundary without asserting practical power for the present experiments.

B.7 Proof Of The Conformance And Decision-Transfer Theorem

The construction, measurement, aggregation, and decision form one justified test of a model's stated observable commitment. Consider the formal PrefTurb procedure that selects and accepts evidence, freezes its ledger, constructs declared histories as in Proposition B.1, obtains outputs through a fixed prediction interface, measures them as in Appendix B.4, and evaluates the declared requirements $g_{F,j}$. Suppose the specification and domain meet conditions (i) through (v) in Appendix B.1.

Theorem B.1 (PrefTurb conformance and decision transfer)**.** *If the accepted evidence satisfies the declared interpretation and input conditions, and* PrefTurb *constructs histories by Proposition B.1, applies the specified concrete measurement, and retains the contract's common cases, scoring context, weights, and operations, it computes exactly the finite-cohort quantities and relations specified by* ISCon*. With a jointly covering uncertainty set, its population verdicts inherit the error guarantee of Theorem 3.1.*

Proof of Theorem B.1. The accepted ledger and evidence conditions specify which controlled histories belong to the declared test domain. Proposition B.1 establishes the construction's preserved

attributes and its stated multiplicity and ordering changes. Condition (iii) supplies the task and input-boundary controls. The substantive interpretation of this domain is exactly the one stated in condition (iv). Appendix B.4 identifies the resulting responses and averaged vector with the contract's measurement and aggregation ingredients. Evaluating the fixed functions $g_{F,j}$ therefore evaluates the same acceptable-difference relation as the ISCON specification. This establishes structural instantiation and equality of the finite-cohort relation. For population inference, Theorem 3.1 transfers simultaneous coverage to the probability bound for incorrect declarations. Appendix B.5 provides one sufficient constructive route to that coverage. No step assumes $g_{F,j}(\mu) \geq 0$; its truth is the outcome being assessed. □

The theorem is a compositional correctness result for the specified protocol. The equality of its measurements with ISCON is partly definitional. Its substantive content lies in the explicit construction invariants, exact treatment of scoring and aggregation, and finite-decision guarantee. It does not independently establish semantic label validity, correctness of an unverified executable, statistical power, architecture-level causality, or behavior outside the tested domain. These limits prevent a correct mathematical instantiation from becoming an unsupported certification of every saved experiment.

B.8 Sound Follow-Up Claims For Retention And Interference

A supported baseline can be followed by tests of whether adherence continues, provided the comparisons and uncertainty cover the entire declared claim.

Corollary B.1 (Sound follow-up claims)**.** *Specify a finite family of baseline and follow-up contracts. Each follow-up has an expected relation derived from the same model's rationale under that intervention. The relation may allow adaptation to valid new preferences. Let one joint uncertainty set cover all quantities entering this family with probability at least* $1 - \alpha$.

Use Theorem 3.1 for each contract. A claim of retained satisfaction, or resistance understood as continued satisfaction, requires support for its baseline and every follow-up included in the claim on the same declared domain. A claim of lost satisfaction requires baseline support and contradiction of the relevant follow-up contract. Then the probability of any false such claim in the declared family is at most α*, even if the analyst chooses which members to report after seeing the results.*

Proof. On the single joint-coverage event, Theorem 3.1 makes every support and contradiction declaration in the finite family correct simultaneously. A retained-satisfaction claim on that event has a true baseline and true required follow-ups. A lost-satisfaction claim has a true baseline and a false specified follow-up. Reporting any selected subset cannot create an erroneous declaration on an event where all candidate declarations are correct. Consequently, any false reported claim requires failure of the joint-coverage event, with probability at most α. □

This is an unconditional familywise probability bound. It does not assert a bound of α conditional on the baseline having passed. The result permits selecting a report from the declared family, not choosing a new cohort, margin, condition, or hypothesis outside that family after inspecting responses. Such changes require an analysis that covers the new target. A baseline unresolved or contradicted on the follow-up domain permits response diagnostics without a claim of loss of previously established satisfaction. If an earlier baseline uses a different cohort, establish the baseline relation on the follow-up domain before making the retention claim. Separate baseline and follow-up error budgets can also be combined through a stated joint failure bound.

B.9 Examples Explaining The Assumptions

These constructed examples diagnose logical failure modes. They are not empirical findings.

Changing the cohort can manufacture a contrast. Suppose unit A always scores 1 and unit B always scores 0 in every history condition. Using only A for the early condition and only B for the late condition gives a mean change of -1. Each unit's actual paired change is zero. Thus, condition-specific cohorts do not generally compute the matched contract quantity.

Passing a baseline by selection can manufacture apparent decline. Let independent baseline and follow-up scores B, F each be Bernoulli with mean $1/2$. Their population mean difference is zero. Among units selected because $B = 1$, $\mathbb{E}[F - B \mid B = 1] = 1/2 - 1 = -1/2$. The selected comparison answers a different question and cannot establish a population-wide loss of adherence without accounting for selection.

A fixed metric name can conceal a changed measurement. If a score depends on peer outputs, changing the peer pool can change the focal score while leaving the focal output unchanged. A per-user output-only formulation therefore omits a real argument of that score.

Frozen labels do not establish semantic truth. A lookup table can retain a wrong label perfectly. The construction proof establishes retention of the assigned relation; its semantic adequacy comes from the declared validation and scope.

A supplied-history guarantee does not imply a predictor-input guarantee. A preprocessing map that retains only a suffix can remove a moved block. The constructed histories differ while the inputs to the downstream predictor coincide. The full interface can be diagnosed for this loss; the predictor cannot be claimed to have processed the removed evidence.

C Concrete Measurement And Its Formal Properties

The measurement ingredient determines which output changes can affect a contract. We make the recommendation and generation scorers explicit and establish the applicable bounds needed for statistical transfer.

C.1 Recommendation Measurement And Its Bounds

Consider an ordered recommendation list $v_1, \ldots, v_\ell$, a single target item t, and a declared cutoff $K \geq 1$. All identities share the same item-ID convention. Let $K' = \min(K, \ell)$. For the single-target nDCG formula below, the list has no duplicate item identities. Define the target's first position r within the evaluated prefix, or $r = \infty$ if it is absent. The selected measurements are

$$\mathrm{RR}\,@K = \begin{cases} 1/r, & r \leq K', \\ 0, & r = \infty, \end{cases} \qquad \mathrm{nDCG}\,@K = \begin{cases} 1/\log_2(1+r), & r \leq K', \\ 0, & r = \infty. \end{cases}$$

The retained-list reciprocal-rank function in the inspected source evaluates its supplied list, which may already have been truncated. The prefix convention here makes that scope explicit. MRR is the declared average of casewise reciprocal ranks, not a separate per-case response.

To credit related recommendations through semantic-substitute expected reciprocal rank, fix the distance function $\sigma \geq 0$, the embeddings it uses, and $\zeta > 0$. Set

$$s_k = \exp[-\sigma(v_k, t)/\zeta], \qquad \mathrm{SERR}\,@K = \sum_{k=1}^{K'} \frac{s_k}{k} \prod_{m<k} (1 - s_m).$$

In the inspected ranking configuration, $\sigma = 1 - \mathrm{clip}(\cos, -1, 1)$, where $\cos$ is embedding cosine similarity, and $\zeta = 0.5$. Its missing-embedding branch assigns $s_k = 0$. Embeddings, cutoff, and item-ID mapping remain part of this configuration. A changed missingness policy defines a different measurement.

Lemma C.1 (Bounds for the specified ranking measurements)**.** *All three measurements lie in* $[0, 1]$*, including the empty evaluated list, which receives zero.*

Proof. For a retrieved target, $r \geq 1$, and hence $1/r \in (0, 1]$ and $\log_2(1 + r) \geq 1$. The absent-target branch is zero. This proves the reciprocal-rank and single-target nDCG bounds.

For the semantic score, nonnegative distance and positive temperature imply $0 \leq s_k \leq 1$, including the declared zero branch. Write $q_0 = 1$ and $q_k = \prod_{m \leq k}(1 - s_m)$. Then $q_{k-1} s_k = q_{k-1} - q_k \geq 0$. Therefore

$$0 \le \sum_{k=1}^{K'} \frac{q_{k-1} s_k}{k} \le \sum_{k=1}^{K'} (q_{k-1} - q_k) = 1 - q_{K'} \le 1.$$

The empty sum is zero. Each function is thus a specified bounded scalar measurement of the ranked output. □

C.2 Concrete Generation Configurations

Generation measurements compare output texts with references. We fix the following concrete scorer configuration. ROUGE-L uses stemmed F-measure and returns distance $1 - \mathrm{round}(F, 10)$. The ROUGE-SU4 implementation combines unigram and skip-bigram overlap with a maximum skip of four and the same rounding. BERTScore uses the English `roberta-base` model through the `evaluate` implementation, returning $\mathrm{round}(1 - F_1, 10)$ with its declared empty-text replacement. The exact text normalization, component, dependency versions, and encoder weights belong to the measurement configuration. A higher-is-better similarity is obtained by subtracting the declared distance from one. The contract's direction and practical margin use the resulting scale.

This configuration converts missing text to an empty string and strips surrounding whitespace. ROUGE returns distance one if either text is empty. BERTScore substitutes the literal word `empty` for an empty string. These settings define an explicit functional specialization. They do not establish which scorer version generated every archived result. Other recorded branches, including different encoders, rounding, peer limits, or missing-input handling, retain their own configuration identities. In particular, no blanket $[0, 1]$ bound is assumed for every BERTScore variant or exported generation score. The bounded result below states the distance-domain assumptions needed by its particular PerSEval function. A separate recorded branch uses Jensen–Shannon distance in its native lower-is-better direction, while the corresponding PerSEval accuracy-similarity input uses one minus that stored distance (Appendix H.5). Because the present reconstruction does not establish a single normalized JSD implementation and domain for all archived uses of that branch, we do not assign it an additional formal bound here. Its recorded scorer configuration remains distinct from the ROUGE/BERTScore specialization above.

C.3 A Specified PerSEval Measurement With Complete Peers

PerSEval's measurement is concretely a combination of agreement between users' output differences and reference differences, adjusted for output-reference accuracy. All relevant peers enter the measurement. We specify the complete-peer configuration below. Its guarantee applies when the scorer implements these functions on the stated domain. Configurations that change the peer pool, distance, or missing-data rule require their own specialization.

For a document d, let I_d be its fixed peer set with at least two matched users. Let a_i be the user-specific references, y_i the generated outputs, and c_d the fixed source representation used by the scorer. In the inspected local class, the source representation concatenates the document summary and document text. Fix a distance function δ. On the domain of this statement, all required distances are finite and in $[0, 1]$, and the source distances below are strictly positive. These are domain assumptions, not bounds established for every imported base metric.

For each user, the scorer first measures distances to the other users and weights those distances relative to the distance from the source. It applies the same operation to references and generated outputs. For $s \in \{a, y\}$, define

$$x_{ij}^s = \delta(s_i, s_j), \qquad b_i^s = \delta(s_i, c_d) > 0,$$

$$p_{ij}^s = \frac{\exp(x_{ij}^s / b_i^s)}{\sum_{k \in I_d \setminus \{i\}} \exp(x_{ik}^s / b_i^s)}, \qquad z_{ij}^s = p_{ij}^s x_{ij}^s.$$

The simplified branch uses $z^s_{ij} = x^s_{ij}$ instead. Agreement is high when a pair of generated outputs differs by a similar amount to the corresponding references. Fix $\varepsilon_p = 10^{-5}$ and define pair agreement and responsiveness as

$$A_{ij} = \frac{\min(z^y_{ij}, z^a_{ij}) + \varepsilon_p}{\max(z^y_{ij}, z^a_{ij}) + \varepsilon_p}, \qquad D_i = \frac{1}{|I_d| - 1} \sum_{j \neq i} A_{ij}.$$

Reproducing differences between users is insufficient if every output is far from its own reference. The accuracy adjustment therefore also uses each output-reference distance, the smallest such distance, and their mean. Let

$$e_i = \delta(y_i, a_i), \qquad e_{\min} = \min_{i \in I_d} e_i, \qquad \bar{e} = \frac{1}{|I_d|} \sum_{i \in I_d} e_i.$$

With $\varepsilon = 10^{-7}$, define

$$t_{1i} = \frac{e_i - e_{\min}}{\bar{e} - e_{\min} + \varepsilon}, \qquad t_2 = \frac{e_{\min}}{1 - e_{\min} + \varepsilon}, \qquad G_{a,b}(t) = \frac{1}{1 + 10^a \exp(-10^b t)}.$$

The accuracy components and resulting personalization measurement are

$$\mathrm{ACP}_i = G_{4,1}(t_{1i}), \qquad \mathrm{ADP} = G_{4,1}(t_2),$$

$$E_i = 1 - G_{3,1}(\mathrm{ACP}_i + \mathrm{ADP}) + \varepsilon, \qquad P_i = D_i E_i.$$

Document-level aggregation in this branch is $P_d = |I_d|^{-1} \sum_i P_i$. The cross-document score is the mean of document scores. Alternative row-weighted or capped-peer configurations define different measurements or aggregation targets and retain their own identities.

Proposition C.1 (Well-defined and bounded PerSEval measurement)**.** *On the declared domain, this is a well-defined peer-dependent scalar measurement with* $0 \leq P_i \leq 1 + \varepsilon$*. Fixed normalized averaging preserves this bound.*

Proof. Positive b^s_i and finite distances make every exponent finite in exact real arithmetic. The exponential terms are positive, and hence each softmax denominator is positive and $p^s_{ij} \in (0, 1]$. Thus $z^s_{ij} \geq 0$. With positive ε_p, every agreement denominator is positive, and the numerator is positive and no larger than its denominator. Therefore $0 < A_{ij} \leq 1$, and $0 < D_i \leq 1$.

Since $\bar{e} \geq e_{\min}$ and $0 \leq e_{\min} \leq 1$, both denominators defining t_{1i} and t_2 are positive. Their values are finite. For finite real t, $0 < G_{a,b}(t) < 1$. Hence $\varepsilon < E_i < 1 + \varepsilon$. Multiplication gives $0 < P_i < 1 + \varepsilon$, which implies the stated closed bound. Convex averaging preserves that closed interval. The formula explicitly includes every peer in the pair comparisons and in the accuracy adjustment, and hence it is a joint peer-dependent measurement. □

This bound makes the specified personalization measurement eligible for the bounded-unit uncertainty construction when its sampling assumptions also hold. It does not transfer to a different scorer merely sharing the PerSEval name. The proof concerns exact real arithmetic on the stated domain. An actual implementation must handle overflow, underflow, undefined base distances, and missing peers consistently with the specified measurement. A missing-peer fallback is a different branch. The zero-source-distance behavior is excluded here because the inspected division helper and numeric scalar types require their own execution-specific verification.

C.4 Measurement Conformance And Transfer

The concrete functions above fill the measurement slot in the general contract. The following result identifies what must remain fixed for their computed scores to be the quantities specified by that contract.

Proposition C.2 (Concrete measurement conformance and transfer)**.** *For each comparison, choose one concrete measurement configuration above, including its component and full scoring context. If its domain requirements hold, replace the generic S in Appendix B.4 with its RR, nDCG, SR-ERR, configured ROUGE/BERTScore, or fully specified PerSEval function. Keep the model, metric configuration, targets/references, scoring-pool membership and declared aggregation rule fixed across the controlled histories.*

The resulting responses and averages are exactly a measurement instantiation of the corresponding ISCon *contract. The original theorem's evidence-to-rationale interpretation conditions still apply. Where the chosen measurement and all averaged coordinates have proven bounds, the bounded-unit uncertainty result is available under its additional sampling assumptions.*

Proof. Appendix C.1 explicitly defines the ranking functions. The generation configuration fixes the base scorer, and Appendix C.3 explicitly defines one PerSEval branch using its fixed base distance and full peer collection. These give concrete substitutions for the measurement argument in Appendix B.4. For each condition, evaluating the specified function on the condition's model outputs and fixed scoring context produces its defined response. Applying the contract's same per-unit transformation and weights gives the same mean vector. Appendix B.4 therefore applies to those particular functions. Substituting them does not assume any required score ordering and hence does not assume model adherence.

Appendix C.1 gives $[0, 1]$ bounds for its ranking scores. Appendix C.3 gives $[0, 1 + \varepsilon]$ for the specified personalization branch on its stated distance domain. If the three role-score coordinates $X_L, X_S, X_{\mathrm{EP}}$ lie in $[0, B]$, paired differences and $X_L - \min(X_S, X_{\mathrm{EP}})$ lie in $[-B, B]$, and absolute differences lie in $[0, B]$. These follow by subtracting two quantities in $[0, B]$. Fixed normalized averages preserve those intervals. Thus these bounds can be supplied for the actual coordinates averaged in Appendix B.5. Ratios and other transformations need their own bounds. No bound transfers to an unverified metric variant merely because its name is similar. Once all required boundedness, independence and fixed-specification assumptions hold, Appendices B.5 and B.6 give the stated decision guarantee. □

D Concrete Aggregation And Its Formal Properties

After scoring each output, aggregation determines how much each case contributes to a comparison. Within one comparison, let A be the fixed nonempty set of recommendation cases, with $n = |A|$. For generation, let D be the fixed nonempty document set, $m = |D|$, and let I_d contain the retained user-document cases for document d. Write $n_d = |I_d| > 0$ and $N = \sum_{d \in D} n_d$. Repeated users remain distinct cases across different targets when the evaluation configuration specifies this. All scores are finite and include their complete peer context. Membership, weights, and scoring context remain fixed across the conditions of one contract.

The deterministic identities below require no independence. A common observed set defines a descriptive comparison. Generalization after outcome-dependent exclusions requires a missingness argument. The uncertainty result additionally requires independence between complete groups with fixed weights, or a justified conditional sampling model with those properties.

D.1 Concrete Weights And Their Meaning

Giving every case equal influence and giving every document equal influence are explicit, different choices of aggregation rule. The recommendation rule is

$$\widehat{\mu}_{\mathrm{rec}}(\omega) = \frac{1}{n} \sum_{i \in A} R_i(\omega).$$

The generation case rule and document rule are

$$\bar{R}_d(\omega) = \frac{1}{n_d} \sum_{i \in I_d} R_{di}(\omega), \qquad \widehat{\mu}_{\mathrm{case}}(\omega) = \frac{1}{N} \sum_d \sum_i R_{di}(\omega) = \sum_d \frac{n_d}{N} \bar{R}_d(\omega),$$

$$\widehat{\mu}_{\mathrm{doc}}(\omega) = \frac{1}{m} \sum_d \bar{R}_d(\omega) = \sum_d \sum_i \frac{1}{m n_d} R_{di}(\omega).$$

Lemma D.1 (Normalized weights and the difference between generation averages)**.** *Thus recommendation uses weight $1/n$, generation case weighting uses $1/N$, and generation document weighting uses $1/(mn_d)$ per case. Normalizing raw inverse-count weights $1/n_d$ also yields $1/(mn_d)$.*

The difference between these averages depends on how document size varies with the measured quantity. For any scalar document quantity x_d, set $\bar{n} = N/m$, $\bar{x} = m^{-1} \sum_d x_d$, and define finite-set covariance by

$$\mathrm{Cov}_D(n, x) = \frac{1}{m} \sum_d (n_d - \bar{n})(x_d - \bar{x}).$$

Then

$$\sum_d \frac{n_d}{N} x_d - \frac{1}{m} \sum_d x_d = \frac{\mathrm{Cov}_D(n, x)}{\bar{n}}.$$

Proof. Each weight is nonnegative. Its sum is 1, since the recommendation weights sum to n/n, case weights to N/N, and document weights to $\sum_d n_d/(mn_d) = 1$. The regrouping identities follow by distributing each document sum. Raw inverse-count weights sum to $\sum_d n_d/n_d = m$, giving the stated normalized weights. Expanding the centered covariance gives $\mathrm{Cov}_D(n, x) = m^{-1} \sum_d n_d x_d - \bar{n}\bar{x}$. Division by $\bar{n} = N/m$ yields the final identity. □

Equal document sizes are sufficient for the two generation averages to agree. Agreement can also occur through zero covariance. In a signed contrast, a document with more cases has more total influence under the case rule. With document sizes 3 and 1, and document-average score changes $1/2$ and $-3/4$, the case-weighted change is $3/16$, while the document-weighted change is $-1/8$. This arithmetic example changes no score. *The two rules answer different questions, and hence their difference is not evidence that either rule is intrinsically incorrect.*

The inverse-count identity requires n_d to count the actual retained scoring contributions. Weights frozen using a larger set do not automatically give documents equal total influence after exclusions. Such saved weights still define their own normalized aggregation rule; they require an explicit configuration check before being described as the equal-document rule above.

D.2 Pairing And The Order Of Linear Comparisons

With the same cases and weights, a change in the average score equals the average of each case's score change.

Lemma D.2 (Paired changes commute with fixed averaging)**.** *For any of Appendix D.1's normalized weights w_i and two conditions a, b,*

$$\widehat{\mu}(b) - \widehat{\mu}(a) = \sum_i w_i \big(R_i(b) - R_i(a)\big).$$

The same identity holds when first calculating changes within documents and then applying the declared document weights. If all case scores are unchanged between conditions, the aggregate change is zero.

Proof. Distributing the common w_i over subtraction gives the first identity. Applying the identity within each document and distributing the fixed document weights proves the grouped version. When $R_i(b) = R_i(a)$ for each i, every summand is zero. □

If the condition weights instead differ, write their values on the union of eligible cases, using zero for an excluded case. Algebra gives

$$\widehat{\mu}_b(b) - \widehat{\mu}_a(a) = \sum_i w_i^a \big(R_i(b) - R_i(a)\big) + \sum_i (w_i^b - w_i^a) R_i(b).$$

This decomposition presumes scores exist on that union. It does not impute missing scores. *The second term is a composition effect, which can be nonzero even with unchanged case scores.* All conditions used in one joint role comparison therefore need common membership and weights. Separate pairwise intersections do not automatically define one joint vector for a multi-condition contract.

D.3 Nonlinear Aggregation And Comparator Selection

A nonlinear comparison can define different estimands depending on whether it is applied before or after aggregation. Weak long-term non-collapse gives the central example. Selecting the weaker transient role separately within each case and then averaging is not the same quantity as first aggregating each temporal role and then selecting the weaker transient aggregate.

Proposition D.1 (Weighted comparator selection and cancellation)**.** *For normalized nonnegative weights, let (L_i, S_i, E_i) be three finite role scores and define*

$$C_{\text{means}} = \sum_i w_i L_i - \min\left(\sum_i w_i S_i, \sum_i w_i E_i\right), \qquad C_{\text{cases}} = \sum_i w_i \big(L_i - \min(S_i, E_i)\big).$$

Then

$$C_{\text{cases}} - C_{\text{means}} = \frac{1}{2}\left(\sum_i w_i |S_i - E_i| - \left|\sum_i w_i (S_i - E_i)\right|\right) \geq 0.$$

Equality holds exactly when all nonzero differences $S_i - E_i$ with positive weight have the same sign. The all-zero case also gives equality. For signed case changes Δ_i,

$$\left|\sum_i w_i \Delta_i\right| \leq \sum_i w_i |\Delta_i|.$$

Proof. Substitute $\min(s, e) = (s + e - |s - e|)/2$ into both expressions for C. The linear terms cancel, giving the displayed gap. Write p for the weighted sum of positive differences and q for the weighted magnitudes of negative differences. The gap equals $(p + q - |p - q|)/2 = \min(p, q)$, which is nonnegative and is zero exactly when $p = 0$ or $q = 0$. This proves the equality characterization. For changes Δ_i, the triangle inequality and nonnegative weights give the last inequality. □

This is the finite weighted specialization of Proposition A.5. Equal document weighting supplies weights $1/(mn_d)$ when the nonlinear operation is case-wise. Applying a minimum after forming document means introduces another aggregation rule and must be declared separately. *Opposing user-level changes can cancel in a mean even when many individual scores change.* Signed net movement, absolute net movement, and mean absolute case movement therefore answer different stability questions.

D.4 Grouped Uncertainty And Aggregation

The aggregation unit and the independent sampling unit need not coincide. In particular, several user-specific scores associated with one document may jointly determine a document-level personalization measurement without constituting several independent observations. A sufficient construction is therefore to form the complete group-level vector required by the declared estimand and apply the uncertainty procedure only across groups for which independence is justified. Counting their scores separately does not itself justify narrower uncertainty.

Proposition D.2 (Coverage with complete independent groups)**.** *Let X_d be a complete document-level vector after the declared within-document computation, with coordinate bounds $a_k \le X_{d,k} \le b_k$. Suppose the vectors are independent across documents and the weights W_d are deterministic, nonnegative, and satisfy $\sum_d W_d = 1$. Set*

$$\widehat{\mu}_k = \sum_d W_d X_{d,k}, \qquad \mu_k = \sum_d W_d \mathbb{E}[X_{d,k}].$$

For $0 < \alpha < 1$ and allocations $\delta_k > 0$, $\sum_k \delta_k \le \alpha$, the intervals with radii

$$r_k = \frac{b_k - a_k}{2} \sqrt{\frac{\sum_d W_d^2}{\delta_k}}$$

jointly cover μ with probability at least $1 - \alpha$, after intersection with the known coordinate bounds.

For document means, take $W_d = 1/m$. For generation case means with fixed document sizes, take $W_d = n_d/N$. These choices preserve exactly the respective averages in Appendix D.1. Within a document, users and score coordinates may be dependent. A bound on every case coordinate is preserved by a within-document mean of that coordinate.

Proof. A within-document mean is a convex combination, and hence a coordinate whose case contributions all lie in $[a_k, b_k]$ remains in that interval. By Appendix D.1, applying the stated document weights gives the declared aggregate. The complete vectors X_d now satisfy the independence, deterministic-weight, and boundedness premises of Appendix B.5. Substitution of d, X_d, W_d for its evaluation units, vectors, and weights proves the coverage statement. In detail, independence bounds the aggregate variance by $(b_k - a_k)^2 \sum_d W_d^2/4$; the squared-tail inequality and a union bound give the result. Constant coordinates are covered exactly. □

The same construction permits grouping recommendation cases by user if those user groups are independent and their fixed weights preserve the target case mean. A document label is not proof that documents are independent. Repeated users across documents, shared metric context across groups, or other dependence require an appropriate larger-group or other valid sampling argument. Conditioning on group sizes is sufficient only if the assumed independence and bounds actually hold under that conditional law.

For the document rule, $\sum_d W_d^2 = 1/m$. For the case rule it is $\sum_d n_d^2/N^2 \ge 1/m$, because $m \sum_d n_d^2 \ge (\sum_d n_d)^2$. This compares the radius factors under the stated construction. The underlying estimands can differ, and hence it does not rank one aggregation rule as scientifically better. This proposition does not establish exact coverage of the reported exploratory bootstrap.

D.5 Aggregation Conformance And Transfer

The concrete aggregation rules calculate the quantities that the contract declares, without assuming that the model satisfies the required relations.

Proposition D.3 (Concrete aggregation conformance and transfer)**.** *Choose one of Appendix D.1's schemes for a specified evaluation configuration. Fix its complete common set, scoring context, and the position of any nonlinear transformation as in Appendix D.3. Combined with the accepted evidence construction and valid concrete measurements in the conformance argument, this gives the aggregation ingredient required by Appendix B.4 and Theorem B.1.*

For the exact finite set, the calculated vector equals the specified finite-cohort vector. For population analysis, the computed estimator equals the estimator specified by the contract. With fixed weights and integrable coordinates, $\mathbb{E}[\widehat{\mu}] = \sum_i w_i \mathbb{E}[X_i] = \mu$. This expectation identity does not assert that the realized estimate equals the population mean. A support or contradiction guarantee additionally requires the sampling and coverage premises of the relevant theorem. Appendix D.4 is one sufficient grouped construction when its assumptions hold.

Proof. For recommendation, substitute $w_i = 1/n$ and the actual per-case score vector in Appendix B.4. For generation, either use row weights $1/N$ or $1/(mn_d)$ with each score's complete context, or group the resulting per-case contributions within documents and use weights n_d/N or $1/m$. Appendix D.1 establishes that these representations agree. Appendix D.3 fixes whether a nonlinear transformation is applied to each case before grouping or to the resulting means afterward. Keeping that transformation fixed ensures coordinate equality with the ISCon specification. Appendix D.2 verifies that linear contrasts preserve pairing and can be calculated before or after these fixed weighted means. The finite equality in Appendix B.4 and the structural conclusion of Theorem B.1 therefore apply. For integrable X_i, linearity of expectation through the fixed weights gives the displayed expectation identity without requiring independence. Where Appendix D.4's statistical premises hold, it supplies the coverage needed by Theorem 3.1. None of these substitutions asserts that any required score relation is true. □

E Concrete Relations And Their Decision Guarantees

Once the controlled histories, measurement, cohort, and aggregation rule are fixed, the remaining contract component is the observable score relation declared for the model rationale on that domain. The architectural rationale motivates this relation but does not, by itself, mathematically imply it. The evidence and bridge assumptions stated elsewhere in the contract provide that interpretation. We represent each behavioral requirement by a signed margin function whose sign records its position relative to the declared boundary. This common representation makes invariance, directional advantage, activity, role separation, and non-collapse computable without treating them as the same scientific hypothesis. The subsections below specify these relations and the conditions under which uncertainty gives support, contradiction, or an unresolved verdict.

E.1 Explicit Relations And Their Meaning

To apply one decision procedure across rationales, write each commitment as a quantity that must be nonnegative or strictly positive. Subtracting an allowed margin expresses how far the response is from the required boundary. The following identities preserve the original scientific comparison.

Lemma E.1 (Scalar representations of the specified commitments)**.** *Let μ_ω be the declared aggregate score under condition ω, on a common domain with fixed measurement and aggregation rule. Assume score direction has been fixed and hence that larger values have the intended response interpretation. Each tested condition set and the role-pair set $\mathcal{P}$ are finite and nonempty, and each pair contains distinct roles. The following predicates have exact scalar-constraint representations.*

(i) An invariance commitment $|\mu_b - \mu_a| \le \epsilon$, with $\epsilon \ge 0$, is the conjunction of $g_1 = \epsilon - (\mu_b - \mu_a) \ge 0$ and $g_2 = \epsilon + (\mu_b - \mu_a) \ge 0$. (ii) A recent-evidence advantage of at least $\eta > 0$ is $g = \mu_{\text{late}} - \mu_{\text{early}} - \eta \ge 0$. Weak monotonicity uses $\eta = 0$ and is a different commitment. (iii) Activity of older evidence strictly above a declared floor a_k, after subsequent history condition k, is $g_k = \mu_{\text{old},k} - a_k > 0$. Requiring all specified k forms their conjunction. If a commitment also limits loss from its reference condition to $\tau_k \ge 0$, add $g'_k = \mu_{\text{old},k} - \mu_{\text{old},0} + \tau_k \ge 0$. That extra commitment is used only when it belongs to the declared test. (iv) For specified role pairs $\mathcal{P}$ with positive separation margins κ_{rs}, an unordered separation commitment is $g_{rs} = |\mu_r - \mu_s| - \kappa_{rs} \ge 0$ for every $(r, s) \in \mathcal{P}$. If the rationale specifies which role should have the larger response, use the corresponding signed difference instead. (v) Weak long-term non-collapse with margin $\gamma \ge 0$ is $g = \mu_L - \min(\mu_S, \mu_E) - \gamma \ge 0$. Being below only one transient role does not contradict its zero-margin version.

Proof. The first equivalence follows from $-\epsilon \le \mu_b - \mu_a \le \epsilon$. Subtracting the declared advantage gives the second. Subtracting the floor preserves the strict activity comparison, while rearranging $\mu_{\text{old},0} - \mu_{\text{old},k} \le \tau_k$ gives the optional loss constraint. Pairwise separation of at least κ_{rs} is exactly

the displayed absolute-difference inequality. Subtracting the chosen weaker transient response and the margin gives the non-collapse constraint. A conjunction retains every relation required by the declared commitment. □

With a common $\kappa > 0$, separation for all selected pairs is equivalent to $\min_{(r,s)\in\mathcal{P}} |\mu_r - \mu_s| \geq \kappa$. A positive estimated minimum alone does not establish that inequality for the population. Separation of aggregate scores is an observable commitment on the declared domain, not proof that internal memory components are identifiable. Activity is likewise a score condition and does not by itself prove use of the selected evidence.

Activity included in a behavioral commitment differs from an activity floor used solely to decide whether a readout is applicable. Failing an applicability floor alone does not establish contradiction of the model's rationale. The test specification must identify which role the floor serves.

E.2 Exact Bounds For Unordered Role Separation

A difference that could still be zero cannot certify separation by a positive amount.

Lemma E.2 (Exact rectangular bounds for unordered separation)**.** *Suppose the aggregate means belong to a nonempty rectangular uncertainty set with $\ell_r \leq \mu_r \leq u_r$. For distinct roles r, s, define*

$$l = \ell_r - u_s, \qquad u = u_r - \ell_s.$$

For $g = |\mu_r - \mu_s| - \kappa$, exact bounds over the rectangle are

$$L = \begin{cases} -\kappa, & l \leq 0 \leq u, \\ \min(|l|, |u|) - \kappa, & \textit{otherwise}, \end{cases} \qquad U = \max(|l|, |u|) - \kappa.$$

Consequently, $L \geq 0$ supports this closed separation requirement throughout the rectangle. If $U < 0$, every mean vector in the rectangle violates it. Other bound configurations remain unresolved by this rule.

Proof. The image of the rectangle under $(v_r, v_s) \mapsto v_r - v_s$ is the full interval $[l, u]$. Its endpoints are attained at opposite corners, and every intermediate value is attained by continuity along the line between those corners. On this interval, the smallest absolute value is zero if zero is included, and otherwise is the smaller endpoint magnitude. The largest absolute value is the larger endpoint magnitude. Subtracting κ gives the stated extrema. The decision implications follow by comparing these extrema to zero. □

For several role pairs, bound each requirement on the same joint rectangle. The means and pair differences need not be independent. Support requires every specified pair to meet its bound. Contradiction through one pair is sufficient. These conservative decisions need not resolve every false conjunction.

E.3 Sound Decisions With Strict And Closed Boundaries

Whether equality counts as satisfaction must be stated. An activity threshold that must be exceeded differs from a minimum advantage that may be met exactly. Let a nonempty random uncertainty set Q cover the declared mean vector with probability at least $1 - \alpha$, where $0 < \alpha < 1$. For each requirement, fix either $g_j(\mu) \geq 0$ or $g_j(\mu) > 0$, and obtain bounds

$$L_j \leq \inf_{v\in Q} g_j(v), \qquad U_j \geq \sup_{v\in Q} g_j(v).$$

For a closed requirement, its support certificate is $L_j \geq 0$ and its contradiction certificate is $U_j < 0$. For a strict requirement, use $L_j > 0$ and $U_j \leq 0$, respectively. Declare overall support when every requirement has its support certificate, contradiction when at least one has its contradiction certificate, and unresolved otherwise. The probability of any false support or contradiction declaration is at most α.

Proof of Theorem 3.1. On the event $\mu \in Q$, a support certificate implies $g_j(\mu) \geq L_j$, which satisfies its fixed closed or strict predicate. A contradiction certificate implies $g_j(\mu) \leq U_j$, which violates its fixed predicate. Thus every overall support is true and every overall contradiction is justified on the coverage event. A nonempty Q ensures $L_j \leq U_j$, and hence no requirement can have both certificates and the overall labels cannot conflict. Any false declaration lies outside the coverage event and therefore has probability at most α. □

The strict support rule is sufficient and may be conservative if an infimum of zero is not attained. On a compact rectangle and for the continuous functions above, extrema are attained. The rule also handles a singleton exact finite-cohort set. In that descriptive case, it decides the declared finite-cohort predicates without granting population generalization.

For a retention claim, include the baseline and follow-up requirements in the same declared, jointly covered family and use Corollary B.1 in Appendix B. The same coverage-event argument remains valid with the mixed predicates above. A baseline support remains necessary for claiming retained or lost satisfaction. Interference tests retain the relation appropriate to the model's own rationale and may allow a response to meaningful new evidence.

E.4 Threshold Profiles From One Uncertainty Set

A designer may want to know how much practical tolerance a conclusion needs. Fix one target vector μ, one nonempty uncertainty set Q with coverage at least $1 - \alpha$, and a declared family of threshold-indexed requirements. Suppose the lower and upper bounds are valid over that same Q for every threshold. On the event $\mu \in Q$, the preceding proof certifies every reported support or contradiction throughout the threshold family. Any false certificate therefore implies failure of that single coverage event. Evaluating additional threshold grid points alone needs no further multiplicity correction. For an arbitrary uncountable family, the overall false-certificate event must be measurable. The scalar relations used here have finitely many measurable interval endpoints, and hence their analytic profiles satisfy this requirement.

This consequence permits a sensitivity profile without changing histories, scores, or the aggregation rule. It does not validate an approximate bootstrap confidence set, combine separately calibrated families, or justify choosing a scientific margin after seeing favorable scores. The probability statement remains unconditional. It is not a probability of correctness conditional on having reported a decisive verdict.

F Implementation And Evaluation Specification

The formal guarantees above apply only to a specified experimental interface. This section records the implementation-level choices that determine the controlled comparison: evaluation unit, fixed task context, model interface, temporal-role extraction, evaluator-side target relation, diagnostic-trace construction, perturbation operator, measurement, and aggregation configuration. These details specify the scientific comparison rather than certify an arbitrary executable. Run-specific configurations that differ in history consumption, candidate pools, scorer variants, peer sets, or aggregation remain distinct evaluation settings.

F.1 Evaluation Units And Controls

Recommendation cases pair a user history with a positive target and a fixed candidate set. Generation cases pair a history with a source/task input and a user-specific reference used only for scoring. Within every controlled comparison, we keep the fitted checkpoint, tokenizer, or serialization, inference settings, task input, target/reference, candidate pool, cutoff, and scorer configuration fixed. Controlled histories are evaluation inputs rather than additional training data. Target-conditioned evidence selection is, therefore, a diagnostic intervention and is not interpreted as ordinary unselected next-item accuracy. Table 6 summarizes the model-specific design rationales that motivate these conditional tests and the corresponding question posed by each audit.

Table 6: Model rationales motivate conditional tests. Cited papers support the stated design, without proving the tested score inequalities or identifying the internal cause of observed responses.

System	Primary interface	Design rationale	Question in the audit
EBNR	PENS recommendation / generation	GRU aggregation of browsed-news representations (Ao et al., 2021)	Tests a recurrent browsing summary under controlled evidence edits.
NRMS	PENS recommendation / generation	Multi-head self-attention for news and user representations (Wu et al., 2019b; Ao et al., 2021)	Tests whether all-item access yields the expected movement and distractor responses.
NAML-T2	PENS recommendation / generation	Attentive multi-view news representation with the second PENS user-interest injection (Wu et al., 2019a; Ao et al., 2021)	Tests attentive feature aggregation with injected user information under history edits.
Mamba4Rec	Movie recommendation	Selective state-space sequence encoder followed by direct next-item prediction (Liu et al., 2024a)	Tests adherence to the long-range behavioral commitment motivated by the selective state-space design.
DiffRec	Movie recommendation	Denoising generative reconstruction of user/item interaction probabilities (Wang et al., 2023)	Contrasts output changes and near-zero exact-target scores under the same controls.
GTP	PENS headline generation	General headline generator followed by a user-conditioned customizer with late control fusion (Song et al., 2023)	Compares reference similarity with measured responsiveness to different users.
IMPerSumm	Recommendation / personalized generation	Action-aware trajectories in a user interaction graph with information modulation and memory-sensitive dynamics (Chatterjee et al., 2026b)	Motivates tests of how evidence role, action, and position relate to the observed output.
Walk2Pers	Recommendation / personalized generation	Action-conditioned geometric steps with magnitude/orientation decomposition and reinforcement/suppression memory lanes (Chatterjee et al., 2026a)	Tests output responses to movement, temporary preferences, and repeated conflict. Reinforcement and suppression lanes are not automatically LT and ST lanes.
SCAPE	Personalized headline generation	Content/style preference extraction and hierarchical long/short fusion (Lian et al., 2025)	Motivates tests of whether the stated LT/ST combination rule has the expected output response under conflicting evidence.
Qwen3-30B-A3B	In-context recommendation / generation	Mixture-of-experts instruction-tuned LLM (Yang et al., 2025)	Prompted-history comparator with no automatic assignment to a temporal preference contract.
Gemma-2-27B	In-context recommendation / generation	Decoder-only Transformer with interleaved local/global attention and grouped-query attention (Gemma Team et al., 2024)	Tests prompted-history position dependence alongside task-native and later Gemma models.
Gemma-3-27B	In-context recommendation / generation	Longer-context Gemma family with a higher local-to-global attention ratio (Gemma Team et al., 2025)	Tests position dependence in a related model with a different context design.

F.2 Datasets And Comparison Cohorts

The recorded MovieLens-1M evaluation uses 6,040 matched role cases for the original Mamba4Rec comparison and 3,436 matched movement cases. Original generation comparisons retain common user/document cases and aggregate after preserving each document's eligible peer set. The additional PENS role-generation export contains 10,000 rows per LT/ST/EP condition per model. The aligned PENS ST suite starts from 15,552 eligible ranking cases with the same 500 candidates across compared conditions and retains top-20 predictions, and its PSE-BERT movement comparison has 15,237 eligible cases. The EBNR ST suite contains 2,240 ranking and 1,401 generation cases for each T1/T2 comparison. Counts are comparison units, not automatically independent users.

F.3 Recorded Model and Scoring Configurations

Table 7: Task-model settings that materially determine which supplied history can reach the fitted predictor. Controlled comparisons keep each listed model and its inference interface fixed.

System/suite	History/input constraint	Recorded setting relevant to the diagnostic
Mamba4Rec	up to 200 most recent interactions	one Mamba layer, hidden dimension 64, dropout 0.2; same next-item head within comparisons.
DiffRec	released MovieLens noisy-training condition	hidden dimension 1,000; five-step diffusion; x_0 prediction and linear-variance schedule.
SCAPE	ten most recent supplied interactions; 512-token cap	Flan-T5-base; PENS then SMP training, followed by frozen-backbone personalization training; beam decoding with at most 32 new tokens.
PENS encoder path	last 50 mapped history items in the inspected implementation	mapped IDs are retained from the end and shorter histories are left-padded; this limit is implementation-specific and is not transferred to other models.
Aligned PENS ST suite	fixed 500-candidate pool; top 20 retained	frozen IMPerSumm/Walk2Pers components; float32 ranking and semantic ERR arithmetic, with model forward precision unchanged.
Additional role generation	fixed role-conditioned histories	fixed Walk2Pers/pointer-generator or IMPerSumm components; BERTScore with RoBERTa-base; PerSEval parameters $(3, 1, 4)$ with peer cap 4.

F.4 Prompted LLM Interfaces

The supplied news-recommendation prompt explicitly gives later clicks greater importance, so its recency commitment belongs to the prompted interface rather than to an unprompted Qwen or Gemma architecture. The inspected recommendation and headline-generation interfaces serialize history with the task input under a fixed context budget and may shorten history or candidate text when the budget is exceeded (Appendix F.4 and Appendix F.4). The prompt, preprocessing, model checkpoint, and inference settings are held constant in a controlled comparison. Thus, the tested object is the complete prompted prediction interface, including preprocessing.

Headline Prompt. The system message is "You are a faithful and preference-aware personalized news headline generation assistant." The user template is reproduced below.

```
You are a personalized news headline generation model.

Your task is to generate one personalized headline for the target
news article.

The user's click history represents the user's reading preference.
Use the click history to understand what type of news, topics,
wording, and angle the user usually prefers.

However, the generated headline must be faithful to the target news
article.
Do not add any fact that is not supported by the target article.

User click history:
{history_text}
```

```
Target news article metadata:
Category: {target_category}
Topic: {target_topic}

Target news article:
{target_body}

Instructions:
1. Generate exactly one headline.
2. The headline must be concise, natural, and news-style.
3. The headline must be personalized according to the user's click
history.
4. Prefer the angle that is most relevant to the user's past
interests.
5. Do not copy any headline from the click history.
6. Do not add unsupported facts.
7. Do not generate explanation.
8. Do not use bullet points.
9. Do not write "Personalized headline:" in the answer.
10. Output only the final headline.

Now generate the personalized headline.
```

Recommendation Prompt. The system message is "You are a strict personalized news recommender. You rank only the given candidate NewsIDs and output valid JSON only." The user template is reproduced below. The numbered example IDs illustrate the required output structure and do not identify a target label.

```
You are a personalized news recommendation model.

Task:
Given one user's previously clicked news history, rank the candidate
 news documents by how likely the user is to click them next.

User clicked history:
The clicked items are in chronological order. Later clicks are more
recent and should receive more importance.

{history_text}

Candidate bucket:
The following candidate news documents are possible next-click
options.
You must recommend only from this candidate bucket.

{candidate_text}

Ranking rules:
1. Infer the user's interests from repeated categories, topics,
entities, tone, and writing style in clicked history.
2. Give more importance to recent clicked news.
3. Use candidate headline, category, topic, and body snippet if
available.
4. Recommend only NewsIDs present in the candidate bucket.
5. Do not invent NewsIDs.
6. Do not explain.
7. Return exactly {topk} NewsIDs if at least {topk} candidates are
available.
8. If fewer than {topk} candidates are available, return all
available NewsIDs.

Output format:
Return valid JSON only:
{"recommended_news_ids": ["N123", "N456", "N789"]}
```

G TARGET-CONDITIONED TRACE CONSTRUCTION

G.1 CANDIDATE EXTRACTION AND TEMPORAL-ROLE VALIDATION

Extraction And Construction The history-only extractor uses DeepSeek-R1-Distill-Qwen-14B on each user's interactions in chronological order, with available item identifiers, categories, topics, and text. It does not receive the downstream query document, candidate set, reference, or task target. LT extraction seeks recurrent evidence distributed across the history and requests 40–45 interactions when available. ST extraction seeks 12–16 interactions forming a coherent temporary interval in recent history rather than simply selecting the newest events. EP extraction seeks a brief coherent burst, permits an empty selection, and imposes no fixed profile length. Candidate profiles subsequently undergo preference-space refinement and are frozen before controlled history edits. Target-conditioned relation assignment is performed separately; thus history-only extraction does not make the complete diagnostic construction target-blind.

G.2 EVENT AND TARGET REPRESENTATION

Preference candidates are extracted from the original history before the evaluated model is queried. The history-only extraction stage uses chronological interactions and their available item metadata/text to propose recurrent LT evidence, bounded recent ST evidence, and compact EP evidence. Target relation is assigned separately. For event $e_t = (x_t, a_t)$ and target $y_u^\star$, frozen encoders give $\mathbf{e}_t = \psi_{\mathrm{evt}}(e_t)$ and $\mathbf{q}_u^\star = \psi_{\mathrm{tar}}(y_u^\star)$, with affinity $s_u(t) = \mathrm{sim}(\mathbf{e}_t, \mathbf{q}_u^\star)$. Fixed thresholds assign supporting, irrelevant, and contradictory evidence:

$$\ell_u(t) = \begin{cases} \mathrm{sup}, & s_u(t) \geq \theta_{\mathrm{sup}}, \\ \mathrm{irr}, & |s_u(t)| \leq \theta_{\mathrm{irr}}, \\ \mathrm{con}, & s_u(t) \leq -\theta_{\mathrm{con}}. \end{cases} \tag{5}$$

Ambiguous events outside these calibrated regions are excluded from relation-specific controls. LT requires recurrence over a broad source-history span, ST requires recency without satisfying the LT condition, and EP requires a compact semantically coherent block with bounded support outside the block. Overlapping assignments follow a fixed precedence rule.

G.3 EMBEDDING-SPACE VALIDITY

Source Positions And Semantic Consistency Let $H_u = (h_{u,1}, \ldots, h_{u,T_u})$ denote the original history. A preference group p is identified by topic, with category as fallback, and has source positions

$$P_{u,p} = \{i : g(h_{u,i}) = p\}.$$

All temporal quantities below use these original positions.

For semantic consistency, all-MiniLM-L6-v2 embeds each item's category, topic, headline, and article text. Normalized sums of item embeddings form centroids over occupied temporal bins. Let C^{raw} be their median pairwise cosine similarity and C_u^{null} the corresponding same-user background similarity across different preference groups. We use

$$C = \mathrm{clip}_{[0,1]}\left(\frac{C^{\mathrm{raw}} - C_u^{\mathrm{null}}}{1 - C_u^{\mathrm{null}}}\right),$$

with $\mathrm{clip}_{[0,1]}(C^{\mathrm{raw}})$ when the background estimate is unavailable. This encoder is used only for evidence validation. The semantic recommendation scoring uses the separate encoder specified in Appendix H.5.

G.4 TEMPORAL ROLE CONDITIONS

Long-Term Evidence LT evidence should recur across the history rather than arise from one concentrated burst. For one user and preference group, let $Q_a(P)$ denote the empirical a-quantile of occurrence positions, b the number of occupied bins among five equal temporal bins, k the number of

successive occurrences separated by at least one intervening interaction, and ℓ the longest consecutive run. We define

$$\begin{aligned} F &= \frac{|P|}{\max_q |P_{u,q}|}, \qquad S = \frac{Q_{0.9}(P) - Q_{0.1}(P)}{T_u - 1}, \\ R &= \tfrac{1}{2}\left(\frac{b}{5} + \min(k/3, 1)\right), \qquad A = 1 - \frac{\ell}{|P|}, \\ V_{\mathrm{LT}} &= F^{0.1} S^{0.3} R^{0.3} A^{0.15} C^{0.15}. \end{aligned} \tag{6}$$

A group is resolved only when it contains at least three occurrences, occupies at least two temporal bins, contains at least two distinct items, and has an estimable semantic score. The geometric combination penalizes weak persistence, recurrence, anti-burst, or semantic components rather than allowing frequency alone to determine LT validity.

Short-Term Evidence ST evidence should persist within a temporary interval while remaining concentrated within that interval. The selected candidate positions define a window W from their 10th to 90th percentiles. All original occurrences of the same preference group are then counted. Let f_{in} and f_{out} denote occurrences inside and outside W, and let $d = |W|/T_u$. Define

$$\begin{aligned} D_{\mathrm{in}} &= f_{\mathrm{in}}/|W|, \qquad D_{\mathrm{out}} = f_{\mathrm{out}}/(T_u - |W|), \\ E &= \frac{D_{\mathrm{in}}}{D_{\mathrm{in}} + D_{\mathrm{out}} + \varepsilon}, \qquad \Gamma = \sqrt{\frac{f_{\mathrm{in}}}{f_{\mathrm{in}} + f_{\mathrm{out}}}(1 - d)}, \\ V_{\mathrm{ST}} &= E^{0.15} (S^{\mathrm{ST}})^{0.25} (R^{\mathrm{ST}})^{0.2} \Gamma^{0.25} (C^{\mathrm{ST}})^{0.15}. \end{aligned} \tag{7}$$

Here $\varepsilon > 0$ is a fixed stabilizer. Local persistence S^{ST} normalizes the 10th–90th percentile span of in-window occurrences by $|W| - 1$. Local recurrence is

$$R^{\mathrm{ST}} = \tfrac{1}{2}\left(b^{\mathrm{ST}}/3 + \min(k^{\mathrm{ST}}/2, 1)\right),$$

using three equal local bins and returns within W. The semantic component is computed from occupied local bins. Resolution requires at least three in-window occurrences, two occupied local bins, two distinct items, an estimable semantic score, and a duration satisfying $d_{\min} \le d \le d_{\max}$. LT and ST therefore use different temporal reference scales: one tests persistence across the full history, while the other tests persistence within a bounded local interval.

Episodic Evidence The EP validator treats the extracted candidate sequence as containing one or more brief local spikes. Candidate IDs are aligned in extracted order to increasing positions in the source history. Within each preference group, aligned items are partitioned into maximal spikes containing at most four selected items and spanning at most four source positions. This validation reconstruction does not modify the histories supplied to the tested model.

For a spike spanning w positions, let f_{in} and f_{out} count the group's occurrences inside and outside the spike interval, and let ℓ be its longest consecutive run. With $L = 4$,

$$D = \frac{f_{\mathrm{in}}}{w}, \qquad K = \frac{L - w + 1}{L}, \qquad B = \frac{\ell}{f_{\mathrm{in}}}, \qquad P = \frac{f_{\mathrm{in}}}{f_{\mathrm{in}} + f_{\mathrm{out}}}. \tag{8}$$

These quantities measure local density, compactness, concentration in one run, and lack of recurrence outside the episode. For multiple metadata-bearing items, semantic coherence C is the background-corrected median pairwise cosine similarity defined above. With

$$z = (D, K, B, P, C), \qquad w^0 = (0.15, 0.35, 0.10, 0.25, 0.15),$$

and observable-component set $\mathcal{A}$,

$$V_{\mathrm{EP}} = \prod_{j \in \mathcal{A}} z_j^{w_j^0 / \sum_{k \in \mathcal{A}} w_k^0}.$$

For a singleton, semantic coherence is unobservable and the remaining weights are renormalized. Such an event can establish a brief observed episode, but not a persistent latent preference.

Coverage And Aggregation For a nonempty candidate row, let X be its unique selected IDs and $X^+ \subseteq X$ the IDs found in the user's source history. Source faithfulness is $|X^+|/|X|$. EP chronological alignment is the fraction of candidate IDs that can be mapped, in extracted order, to increasing source positions.

A resolved LT/ST item receives its preference-group score, while a resolved EP item receives its reconstructed spike score when it is source-supported and chronologically alignable. All other items receive zero validated mass, $v(x) = 0$. The row-level validation score is

$$\mathrm{SoftV}_r = \frac{100}{|X|} \sum_{x \in X^+} v(x), \qquad r \in \{\mathrm{LT}, \mathrm{ST}, \mathrm{EP}\}. \tag{9}$$

Keeping unresolved items in the denominator makes the score sensitive to validation coverage rather than only to the successfully resolved subset. SoftV measures consistency with the specified temporal construct; it is neither extraction accuracy nor a calibrated cross-role preference score.

Sensitivity And Uncertainty The validator weights are specified independently of downstream model scores. Local sensitivity changes each component weight by $\delta \in \{-0.05, +0.05\}$ in turn and proportionally rescales the remaining weights, producing ten alternatives per role. EP additionally considers equal weights and maximum spike spans of three or five positions. Stability is summarized by Spearman correlation with the default validator and absolute score movement on the 0–100 scale.

These checks assess the stability of the evidence validators and are separate from the acceptable-difference analysis used for model-adherence decisions in Appendix K. EP uncertainty resamples users as clusters, retaining each user's rows together, with an additional user-macro summary obtained after within-user averaging. The reported validation intervals characterize evidence-validation uncertainty and are not simultaneous uncertainty bounds for ISCon contract verdicts.

G.5 Preference-Space Ordering

Once an accepted role–relation evidence set has been obtained, PrefTurb converts it into a controlled trace that can be moved, removed, or repeated as one experimental unit. This stage does not redefine the temporal role or target relation established during evidence selection. Its purpose is only to organize the accepted evidence into a semantically coherent diagnostic object before any behavioral intervention is applied.

Preference-Space Ordering. For temporal role r and target relation ℓ, let

$$V_u^{(r,\ell)} = \{t : r_u(t) = r, \ell_u(t) = \ell\}$$

denote the accepted evidence indices, and let $\mathcal{H}(V)$ denote all visit-once orders of V. For an order π, we define the construction cost

$$\begin{aligned} J_u^{(r,\ell)}(\pi) = & \sum_{k=1}^{|\pi|-1} d(\mathbf{e}_{u,\pi_k}, \mathbf{e}_{u,\pi_{k+1}}) \\ & + \lambda_{\mathrm{ord}} \sum_{a<b} A_{ab} \left|\mathrm{pos}_\pi(a) - \mathrm{pos}_\pi(b)\right|, \end{aligned} \tag{10}$$

where $A_{ab} \geq 0$ is a fixed affinity and $\lambda_{\mathrm{ord}} \geq 0$ controls the affinity-separation penalty. The first term discourages large semantic transitions between adjacent selected events, whereas the second discourages strongly related events from being placed far apart in the constructed trace.

The realized ordering $\widehat{\pi}_u^{(r,\ell)}$ is produced by the recorded heuristic using Equation 10. We do not claim that it is the exact minimizer of the objective, nor do we claim an approximation ratio unless an executed solver record establishes one. A lower value of J therefore means only a smoother ordering under this fixed construction objective; it does not imply higher downstream task quality, stronger preference, recovery of natural chronology, or correspondence with the evaluated model's internal representation.

The task anchor is used only to orient the entry point into the accepted evidence. It is temporary construction context and is removed before the resulting preference trace is supplied to the evaluated model. Accordingly, the anchor does not become part of the model-facing user history.

Recorded PENS Ordering Configuration. For the recovered PENS construction, the ordering space uses `all-MiniLM-L6-v2`. Each temporary query anchor and selected history item is represented from the available headline, news body, topic, and category fields. The resulting representation has 384 `float32` coordinates and is not L_2 normalized. Pairwise distance is root-mean-square coordinate difference.

A Gaussian affinity is constructed using an instance-specific scale equal to the median non-diagonal pairwise distance. The affinity graph is then restricted to the symmetric union of each node's three nearest neighbors, and the recorded objective uses $\lambda_{\mathrm{ord}} = 0.3$.

The search considers three candidate successors to the temporary query anchor. For each candidate entry point, the remaining evidence is initialized by Fiedler-vector spectral seriation of the affinity graph and subsequently refined using pairwise swaps and 2-opt reversals. Refinement terminates after a complete pass with no improvement or after 30 passes. A proposed update is retained only when it decreases the objective by more than 10^{-12}. No random sampling is used. The query anchor is removed after orientation, before the constructed preference trace is passed to the evaluated model. These values identify the recorded PENS construction and are not transferred to another dataset or run family without a corresponding execution record.

G.6 Construction Algorithm

Continuity Refinement. A favorable value of the global ordering objective does not guarantee that every adjacent transition is locally smooth. We therefore inspect the adjacent distances of the ordered LT/ST trace before the trace is frozen. Let

$$d_j = d_{\mathrm{RMSD}}\left(\mathbf{e}_{u,\widehat{\pi}_j}, \mathbf{e}_{u,\widehat{\pi}_{j+1}}\right), \qquad j = 1, \ldots, m-1,$$

and let μ_u and σ_u denote their mean and standard deviation. For the recovered PENS construction, the fixed cutoffs are

$$\tau_1 = \mu_u + 0.5\sigma_u, \qquad \tau_2 = \mu_u + 0.7\sigma_u, \qquad \tau_3 = \mu_u + 1.0\sigma_u.$$

They define four construction regimes: retain, local smoothing, bridge insertion, and replacement, respectively. The distances and cutoffs are computed once from the ordered pre-refinement trace and are not recomputed after an individual refinement.

EP follows a different construction path. Because source-local concentration and discontinuity are part of the episodic construct, accepted EP evidence preserves its relative source-history order and is not subjected to the LT/ST preference-space ordering or continuity-refinement procedure.

Recorded PENS Repair Provenance. The recovered PENS implementation requires a qualification on bridge insertion and replacement. Repair candidates are drawn from the construction embedding index after excluding identifiers already present in the trace and the local items involved in the refinement decision. The recovered implementation does not additionally require a repair candidate to occur in the same user's source history or to receive a new LT/ST role assignment. Consequently, an inserted or replacement item has continuity-repair provenance and should not automatically be interpreted as having the same source-role provenance as the focal LT/ST evidence.

This distinction limits the construction claim. The remaining execution record must establish how often bridge insertion or replacement was realized in the evaluated traces, identify the corresponding repair identifiers and their origins, and distinguish auxiliary continuity-repair items from the focal evidence used by the behavioral contract. Until that linkage is established, repair membership itself is not treated as evidence that the inserted item independently satisfies the LT/ST construct.

After construction, the resulting trace is frozen. It is not independently reordered, smoothed, bridged, or replaced after movement, deletion, repetition, or another PREFTURB intervention. Otherwise the evidence itself would change together with the experimental manipulation. The construction stage therefore establishes a fixed diagnostic input; whether the evaluated model satisfies the corresponding behavioral relation is assessed only from its responses to the subsequent controlled histories.

Algorithm 1 F-compatible construction of an operator-ready preference trace

Require: Accepted role–relation evidence $V_u^{(r,\ell)}$; temporary task anchor a_u; frozen construction encoder; ordering objective $J_u^{(r,\ell)}$; recorded ordering and continuity-refinement settings
Ensure: Frozen operator-ready trace $\widetilde{E}_u^{r,\ell}$ and construction metadata

if $V_u^{(r,\ell)} = \varnothing$ **then**
  Mark the construction as *inapplicable*
  return
end if
if $r = \mathrm{EP}$ **then**
  $\widetilde{E}_u^{\mathrm{EP},\ell} \leftarrow V_u^{(\mathrm{EP},\ell)}$ in relative source-history order ▷ no LT/ST ordering or continuity refinement
else

F-compatible ordering
  Encode the temporary task anchor and accepted LT/ST evidence
  Construct the recorded distance and affinity graph
  Select the recorded candidate anchor successors
  for each candidate successor **do**
    Initialize the remaining evidence by Fiedler-vector spectral seriation
    Refine the visit-once order using pairwise swaps and 2-opt reversals
    Retain an update only when it decreases $J_u^{(r,\ell)}$
  end for
  Select the lowest-cost realized ordering $\widehat{\pi}_u^{(r,\ell)}$
  Remove the temporary task anchor

Continuity refinement
  Compute adjacent RMSD distances $d_1, \ldots, d_{m-1}$ on the ordered pre-refinement trace
  Compute μ_u, σ_u, and fixed cutoffs τ_1, τ_2, τ_3
  for each declared refinement location **do**
    if $d_j < \tau_1$ **then**
      Retain the transition
    else if $\tau_1 \le d_j < \tau_2$ **then**
      Apply the recorded local-smoothing rule
    else if $\tau_2 \le d_j < \tau_3$ **then**
      Apply the recorded bridge-insertion rule when an eligible candidate exists
    else
      Apply the recorded replacement rule when an eligible candidate exists
    end if
    If no eligible repair candidate exists, retain the original trace item
  end for
  $\widetilde{E}_u^{r,\ell} \leftarrow$ the resulting refined LT/ST trace
end if
Freeze trace membership, payloads, origin labels, internal order, and construction metadata before any PREFTURB operator is applied
return $\widetilde{E}_u^{r,\ell}$ and metadata

H PERTURBATION OPERATOR DEFINITIONS

Appendix G constructs and freezes the role–relation evidence trace before any controlled intervention is applied. This section specifies how PREFTURB transforms that frozen trace into matched history conditions for behavioral testing. The operators vary a declared property of the evidence such as position, availability, multiplicity, distracting context, or competing preference evidence while recording the objects that must remain fixed for the comparison to remain interpretable. The operator itself defines only the controlled input change; the expected response to that change is supplied separately by the corresponding ISCON contract. We therefore specify each operator through its edited evidence, intervention parameters, preserved invariants, applicability conditions, and realized history construction.

Table 8: PREFTURB operator catalogue. Each operator specifies an input comparison. The model-specific contract supplies its expected score relation.

Family	Operation	Controlled Evidence	Primary Parameters	Behavioral Quantity
Access	Movement	selected role, supporting	depth, percentile	response under movement
Dependence	Deletion	selected role, supporting	removed span	response to removal
Frequency	Duplication	selected role, supporting	repeat count, placement	exposure response
Isolation	LT-only, ST-only, EP-only	supporting	retained role	role-conditioned response
Distractor	Insertion	irrelevant	placement, strength, duration	response to irrelevant events
Spike	Insertion/scaling	ST or EP, supporting	placement, scale, repetition	response to a temporary interest
Spike	Deletion	ST or EP, supporting	removed burst	agreement with the deletion reference
Conflict	Insertion	ST or EP, contradictory	strength, persistence, placement	response to contradiction
Balance	Ratio sweep	LT support and ST conflict	$n_{\mathrm{LT}}/n_{\mathrm{ST}}$	response to evidence counts

H.1 COMMON OPERATOR SPECIFICATION

An operator defines an input intervention, not its expected behavioral verdict. Operator j is represented by

$$C_j = (r_j, \ell_j, o_j, \vartheta_j, \mathcal{I}_j),$$

where $r_j \in \{\mathrm{LT}, \mathrm{ST}, \mathrm{EP}\}$ is the origin role of the controlled evidence, $\ell_j \in \{\mathrm{sup}, \mathrm{irr}, \mathrm{con}\}$ is its frozen target relation, o_j identifies the edit, ϑ_j contains the operator-specific settings, and $\mathcal{I}_j$ lists objects that must remain fixed. We use ϑ_j rather than α to avoid confusing operator parameters with the statistical error level in Section 3. Typical invariants include the user, task input, target/reference, model checkpoint, inference settings, candidate or peer pool, and scorer configuration.

The tuple answers three questions before the model is scored: *which evidence is edited, exactly how is it edited, and what is held fixed?* The ISCON contract then answers a fourth, separate question: *what score relation should this model's rationale imply under that edit?* Two operations with the same informal name can therefore be different tests when their placement, multiplicity, or invariants differ.

H.2 OPERATOR REALIZATION

Algorithm 2 applies one declared edit to an accepted frozen trace. The output pair is usable only when its record confirms every invariant required by $\mathcal{I}_j$. An intended change in length, multiplicity, or cross-block adjacency is recorded as part of the intervention rather than treated as an invariant failure.

H.3 OPERATOR-SPECIFIC CONDITIONS

Movement. Movement changes the insertion boundary of the same frozen segment. It preserves that segment's event identities, internal order, and multiplicity, and preserves the untouched events' relative order. Cross-block neighbors necessarily change. They are part of the intervention, not a preservation claim. Placement levels must be defined by exact indices or a deterministic percentile-to-index rule.

Deletion. Deletion removes only the declared frozen evidence. A response difference can establish behavioral dependence on that supplied evidence under the fixed interface. It does not identify which internal component used the evidence or whether removal improved the model for another reason.

Algorithm 2 Matched preference-evidence perturbation

Require: Original trajectory τ_u, task input z_u, target/reference T_u, frozen traces, operator specification C_j

Ensure: Controlled trajectory $\tau_u^{(j)}$ and invariant record

Verify that the requested frozen role-relation trace is available
Copy τ_u to $\tau_u^{(j)}$
Apply edit o_j to the declared trace using settings ϑ_j
Preserve provenance and relative order of every event outside the edited span unless the operator explicitly changes them
Preserve the task input, target/reference, model checkpoint, inference settings, and scorer configuration
Preserve candidate, source, query, and peer objects required by the declared measurement
Record every intended change in length, multiplicity, position, or adjacency
Run the operator-specific validity and invariant checks
if any required invariant fails **then**
    Mark the pair inapplicable for this operator. Do not assign it a zero response
end if
return $\tau_u^{(j)}$ and the invariant record

Duplication. Duplication adds copies of the same frozen evidence at declared placements and counts. Content and recorded action remain unchanged. Sequence length and exposure count change together. Both original and resulting lengths are therefore part of the operator record.

Distractor Insertion. A distractor must satisfy the operator's accepted irrelevant-evidence condition before insertion and must remain a task-native event. Placement, amount, duration, and the operational irrelevance threshold are fixed configuration fields. Sensitivity to such an event is not automatically a contract failure. The rationale determines the acceptable relation.

Spike Controls. A spike is a compact ST/EP segment with a declared target relation. Repetition changes the number of frozen occurrences and is covered by Proposition B.1. A variant that changes an event weight, salience field, or payload is a different operator and requires its mutable field to be declared explicitly. Spike deletion supports a reversion claim only when the remaining serialized history is the designated reference construction.

Conflict and Balance Controls. Conflict evidence must satisfy the accepted contradictory relation under the same target-specific construction. Strength, persistence, and placement are distinct intervention coordinates. Likewise, increasing LT support while ST is fixed and increasing ST conflict while LT is fixed are separate balance interventions. A ratio value alone does not identify which side changed. See Table 9 for the recorded perturbation levels and matched comparisons.

H.4 Applicability and Reporting

For every operator level, the record should contain eligible and matched counts, exclusions, realized history lengths, parameter values, and invariant failures. Appendix J reports the counts and serialized-history checks recoverable from the saved outputs. An operator with missing invariants is inapplicable for the corresponding claim rather than negative evidence about the model.

Correlations over a small set of aggregate conflict doses or evidence ratios are descriptive unless an instance-level inferential design is specified. Re-exposure also needs construction detail: placing all copies together and placing one copy at the original position plus additional terminal copies are different histories even when both are called "repetition." `ISCon` therefore interprets only the relation attached to the exact recorded operator, not to its informal name.

H.5 Readouts And Aggregation

For case u, operator j, and readout q, the paired response is $\Delta_{u,j,q} = S_{u,j,q}^{(1)} - S_{u,q}^{(0)}$. MRR and nDCG measure exact-target ranking. SR-ERR additionally credits semantically related ranked items. With

Table 9: Recorded perturbation levels retained in the compact implementation specification. Construction thresholds define the input intervention; decision margins are treated separately.

Family	Recorded levels	Matched comparison
Role isolation	LT, ST, EP	role-specific selected evidence under the same declared case and target.
Movement	early/middle/late; 20/50/80 percentiles where used	same frozen block moved through the same remaining history.
Deletion/repetition	deletion; two/four copies at recorded positions	removal or multiplicity change relative to the matched retained/one-copy condition.
Distractor	thresholds $0.50, 0.70, 0.90$; recorded position/persistence	task-native evidence classified as irrelevant under the frozen construction.
Conflict	thresholds $0.60, 0.80, 0.95$; persistence 1/2/3/5; early/middle/late	same selected contradictory evidence with one declared strength, duration, or placement change.
LT/ST balance	$(1,1), (1,2), (1,4), (2,1), (4,1), (8,1)$	ordered pair gives LT/ST evidence multiplicities.

$K' = \min(K, \ell_u)$, semantic distance σ, fixed temperature ζ_s, and $r_{u,k} = \exp[-\sigma(v_{u,k}, v_u^+)/\zeta_s]$,

$$\text{SR-ERR}\,@K(u) = \sum_{k=1}^{K'} \frac{1}{k} \left[\prod_{m=1}^{k-1} (1 - r_{u,m}) \right] r_{u,k}. \tag{11}$$

The supplied news/MovieLens implementation uses normalized BGE-M3 item embeddings, cosine distance, and $\zeta_s = 0.5$. Original MovieLens exports contain ten predictions, so a field named SR-ERR@15 there is evaluated over the available ten items.

Generation uses reference-similarity readouts (ROUGE-L, ROUGE-SU4, BERTScore, and the recorded JSD branch) together with PerSEval variants that compare user-specific output differences with corresponding reference differences. Original generation results preserve complete document peer pools and equal document weighting after within-document aggregation. Additional role-generation and aligned-ST exports retain their recorded capped-peer/row-weighted configurations. Readout configurations are therefore not pooled across suites merely because a metric family shares a name. Table 10 summarizes the scoring and aggregation configurations retained across the reported suites. Contract quantities use the common eligible set for the compared conditions; signed mean change measures net population movement, whereas mean absolute change measures casewise response magnitude.

Table 10: Scoring and aggregation configurations retained across the reported suites. The configuration accompanies the metric name because peer sets, encoders, and weighting can change the estimand.

Configuration	Measurement	Aggregation/context
Original generation	reference similarity and complete-peer PerSEval	fixed document peer pools; equal document weight after averaging eligible users.
Additional role generation	ROUGE-L, ROUGE-SU4, RoBERTa-base BERTScore and saved PerSEval variants	PerSEval $(3, 1, 4)$; peer cap 4; row-weighted estimand.
Additional JSD branch	stored Jensen–Shannon distance (lower is better)	PerSEval accuracy-similarity input uses one minus the stored distance.
Aligned ST generation	saved reference-similarity and capped-peer personalization readouts	common eligible cases; equal row weights.
News/MovieLens ranking	MRR/nDCG and BGE-M3 SR-ERR	explicit retained-list cutoff; missing semantic embeddings contribute zero satisfaction.

H.6 Validation Of Selected Preference Evidence

Preserving an assigned role during a history edit is distinct from establishing that the selected source evidence exhibits that role. We therefore validate extracted LT, ST, and EP candidates against their original histories, independently of the tested model's recommendation or generation scores. The validation measures recurrence, temporal extent or concentration, and semantic consistency. Results appear in Appendix H.6.

The source-grounded audit covers 2,862 matched LT/ST user–target cases and a separate raw EP audit of 2,866 rows from 103 users. Selected LT and ST identifiers have 100% source faithfulness;

resolved-item coverage is 96.98% for LT and 89.39% for ST, with mean SoftV scores 65.26 and 60.32 on the stated 0–100 construct-consistency scale. Raw EP source faithfulness is 97.92%; chronological alignment and resolved-item coverage are both 84.13%, with mean SoftV 56.83. These values assess consistency with the specified temporal constructs, not extraction accuracy or downstream model performance.

Local ± 0.05 one-at-a-time validator-weight changes preserve near-identical rankings on these cohorts: the lowest reported Spearman correlations are approximately 0.995 (LT), 0.991 (ST), and 0.997 (EP). The evidence audit therefore supplies a construction check for the frozen traces; model adherence is still determined only from the declared behavioral relation and its measured uncertainty.

Together, these checks validate the operational LT, ST, and EP assignments along the dimensions used by the construction: source faithfulness, temporal structure, semantic consistency, and local sensitivity to the validator parameters. These validated operational roles are then frozen before any model-response comparison and form the evidence domain of the corresponding ISCon contract.

I Cross-Setting Results Overview

The supplementary tables below collect the reported model comparisons without repeating the main-text interpretation. Each row retains its own task, cohort, measurement, and weighting. Signed differences are descriptive unless an interval-based decision is stated explicitly. Table 11 provides a compact cross-RQ overview before the RQ-specific results below.

Table 11: Reported specialized-model results across RQ1, RQ2, and RQ3 on PENS. MRS is the smallest role-pair gap. E/M/L denotes early/middle/late placement. Raw MRR is used for recommendation and PSE-BERT for generation. Movement ranges are not confidence intervals. Comparisons retain their own original cohorts and do not establish a sequential retention verdict. Full traces appear in Appendices I.1, I.2, and I.4.

Model / task	**Test**	**Recorded result**
Walk2Pers / Rec.	RQ1 role separation	MRS 2.32×10^{-3}
Walk2Pers / Gen.	RQ1 role separation	MRS 1.20×10^{-4}
IMPerSumm / Rec.	RQ2 movement E/M/L	$0.112/0.114/0.114$
Walk2Pers / Rec.	RQ2 movement E/M/L	$0.104/0.106/0.105$
IMPerSumm / Gen.	RQ2 movement range	0.020 to 0.023
Walk2Pers / Gen.	RQ2 movement range	0.031 to 0.034
IMPerSumm / Rec.	RQ3 strong middle distractor	$0.103 \rightarrow 0.083$
Walk2Pers / Rec.	RQ3 spike E/M/L	$0.091/0.073/0.093$
IMPerSumm / Gen.	RQ3 spike E/M/L	$0.023/0.000/0.021$
Walk2Pers / Gen.	RQ3 spike E/M/L	$0.033/0.015/0.033$

I.1 RQ1: Role Evidence

RQ1 compares isolated LT, ST, and EP evidence under each model's declared rationale. Recommendation and generation results use their recorded task-specific readouts and matched supports. Tables 12 and 13 report the corresponding RQ1 recommendation and generation results, respectively. while Table 14 gives the paired Mamba4Rec decomposition.

I.2 RQ2: Movement, Deletion, And Re-Exposure

RQ2 moves the same selected evidence, removes it, or changes its multiplicity. These controls localize a response; when RQ1 is not initially satisfied they are not interpreted as loss of prior adherence. Tables 15 and 16 summarize the RQ2 recommendation and generation comparisons, respectively. Table 17 records the additional PENS short-term configurations. Table 18 further records cases where movement changes the output without resolving the mean-score difference.

Table 12: RQ1 recommendation overview using MRR. Values are raw means on the same admitted cases within each row. n counts cases. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. Recommendation gives cases equal weight. OAI ranking is an auxiliary saved evaluation setting. The comparison definitions accompany these summaries in this appendix.

Dataset / model	n	**LT**	**ST**	**EP**
MovieLens-1M				
Mamba4Rec	6040	0.028	0.101	0.062
DiffRec	6040	0.003	0.003	0.003
PENS				
NRMS	2866	0.006	0.007	0.003
NAML	2866	0.005	0.001	0.005
EBNR	2866	0.002	0.002	3.78×10^{-4}
GTP	2866	0.006	0.008	0.014
Qwen3-30B-A3B	1408	0.086	0.161	0.125
Gemma-2-27B	979	0.083	0.111	0.108
Gemma-3-27B	8	0.208	0.028	0.086
OpenAI-Reddit				
NRMS	2116	0.338	0.338	0.333
NAML	2116	0.367	0.363	0.363
EBNR	2116	0.334	0.335	0.336
GTP	2116	0.002	0.007	0.016

I.3 Complete Short-Term Response Profiles

An endpoint difference can conceal where a response changes. The full early, middle, and late ST profiles in Table 19 in Appendix I.3 distinguish a gradual change from a change confined to the latest position. They also compare a task score with its companion measurement. For example, EBNR-T1 generation PSE-BERT rises between early and middle placement and then remains unchanged at the displayed precision. Its BERTScore remains nearly constant. IMPerSumm and Walk2Pers recommendation means vary little across positions, although their ranked lists often change, as discussed in Appendix I.2. *The profile distinguishes changes in the output, the average task score, and the personalization measurement instead of treating them as one response.*

Deletion and repeated exposure further separate evidence availability from placement (Table 20 in Appendix I.3). EBNR's mean MRR after deletion is 7.82×10^{-2}, compared with 7.37×10^{-2} for one late ST exposure. At four exposures the late mean is 7.25×10^{-2}. These descriptive values do not by themselves contradict a recency rationale. The paired late-minus-early MRR estimate is approximately -0.004, and its adjusted interval $[-0.021, 0.012]$ includes zero (Table 17 in Appendix I.2). Initial satisfaction has not been established for this comparison, and hence the profile remains a history-change diagnosis rather than a claim of lost satisfaction.

The EBNR exposure layout differs from the aligned IMPerSumm/Walk2Pers suite, which places copies contiguously at the named position. Its raw ranking exports and scored outputs also retain the correspondence question documented in Appendix J. These distinctions prevent the new profile from being treated as a pooled repetition effect or a newly certified adherence failure. They do not alter the reported means.

I.4 RQ3: Interference And Competing Evidence

RQ3 keeps irrelevant distractors, temporary valid episodes, and contradictory evidence conceptually separate. A valid temporary preference can warrant adaptation, whereas an irrelevant distractor can warrant stability; the same signed score change therefore has no universal interpretation outside its declared relation. Table 21 summarizes the MovieLens conflict-control comparisons. Table 22 reports the corresponding RQ3 recommendation comparisons across the evaluated settings. Tables 23 and 24 report the RQ3 generation comparisons on PENS and OpenAI-Reddit, respectively.

Table 13: RQ1 generation overview using PSE-BERT. Values are raw means on the same admitted cases within each row. n counts cases and d counts documents. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. PENS denotes headline generation and OAI denotes summarization. Each retained document receives equal weight after averaging its users. T1/T2 retain the reported interest-injection variants. The comparison definitions accompany these summaries in this appendix.

Dataset / model	n / d	**LT**	**ST**	**EP**
PENS				
NRMS-T1	1543 / 665	0.016	0.009	0.016
NRMS-T2	1543 / 665	0.003	0.003	0.003
NAML-T1	1539 / 664	0.032	0.041	0.044
NAML-T2	1539 / 664	0.008	0.007	0.006
GTP	1538 / 663	0.001	0.001	0.001
SCAPE	1541 / 665	0.002	0.002	0.002
Qwen3-30B-A3B	1541 / 665	0.252	0.265	0.257
Gemma-2-27B	1541 / 665	0.128	0.133	0.129
Gemma-3-27B	1541 / 665	0.241	0.235	0.234
OpenAI-Reddit				
NRMS-T1	419 / 177	0.011	0.010	0.007
NRMS-T2	419 / 177	4.30×10^{-6}	4.30×10^{-6}	4.30×10^{-6}
NAML-T1	22 / 11	3.56×10^{-6}	0.214	0.089
NAML-T2	22 / 11	4.17×10^{-6}	0.290	0.218
EBNR-T1	22 / 11	0.006	0.293	0.307
EBNR-T2	22 / 11	3.62×10^{-6}	0.222	0.207
GTP	419 / 177	1.91×10^{-6}	1.91×10^{-6}	1.91×10^{-6}
SCAPE	350 / 156	8.72×10^{-7}	8.72×10^{-7}	8.72×10^{-7}
Qwen3-30B-A3B	625 / 236	0.178	0.184	0.257
Gemma-2-27B	625 / 236	0.211	0.192	0.195
Gemma-3-27B	625 / 236	0.283	0.286	0.274

Table 14: Paired Mamba4Rec RQ1 decomposition on MovieLens ($N = 6{,}040$). Scores and interval endpoints are multiplied by 100. Pointwise 95% intervals use 10,000 paired bootstrap resamples.

Readout	**LT**	**ST**	**EP**	**LT–ST Δ [CI]**	**LT–EP Δ [CI]**
MRR	2.80	10.07	6.17	$-7.27\,[-7.95, -6.58]$	$-3.37\,[-3.92, -2.82]$
nDCG@10	4.06	13.08	8.52	$-9.02\,[-9.78, -8.23]$	$-4.47\,[-5.11, -3.84]$
SR-ERR@10	69.40	70.61	69.58	$-1.21\,[-1.44, -0.99]$	$-0.18\,[-0.37, 0.01]$

Table 15: RQ2 recommendation overview using MRR. Each row gives a signed difference of means and its matched case count n. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. Recommendation gives cases equal weight. OAI ranking is an auxiliary saved evaluation setting. The comparison definitions accompany these summaries in this appendix.

Dataset / model	**Comparison**	**Difference**	n
MovieLens-1M			
Mamba4Rec	Late − early	-0.103	3,436
Mamba4Rec	Late − deletion	-0.099	3,436
DiffRec	Late − early	0	3,434
DiffRec	Late − deletion	-5.36×10^{-5}	3,434
PENS			
NRMS	Late − early	-0.004	2,844
NRMS	Late − deletion	-0.004	2,844
NAML	Late − early	-0.002	2,844
NAML	Late − deletion	-0.005	2,844
EBNR	Late − early	-0.005	2,844
EBNR	Late − deletion	-0.002	2,844
GTP	Late − early	-0.003	2,857
GTP	Late − deletion	-0.002	2,857
Qwen3-30B-A3B	Late − early	0.031	1,408
Qwen3-30B-A3B	Late − deletion	0.038	1,408
Gemma-2-27B	Late − early	0.036	979
Gemma-2-27B	Late − deletion	-0.010	979
Gemma-3-27B	Late − early	0	8
Gemma-3-27B	Late − deletion	-0.031	8

Table 16: RQ2 generation overview using PSE-BERT. Each row gives a signed difference of means and its matched case count n. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. PENS denotes headline generation and OAI denotes summarization. Each retained document receives equal weight after averaging its users. T1/T2 retain the reported interest-injection variants. The comparison definitions accompany these summaries in this appendix.

Dataset / model	**Comparison**	**Difference**	n
PENS			
NRMS-T1	Late − early	≈ 0	1,524
NRMS-T2	Late − early	≈ 0	1,524
NAML-T1	Late − early	≈ 0	1,530
NAML-T2	Late − early	≈ 0	1,530
EBNR-T1	Late − early	-1.29×10^{-4}	1,538
EBNR-T2	Late − early	≈ 0	1,538
GTP	Late − early	≈ 0	1,538
GTP	Late − deletion	≈ 0	1,538
SCAPE	Late − early	≈ 0	1,530
SCAPE	Late − deletion	≈ 0	1,530
Qwen3-30B-A3B	Late − early	-0.009	1,536
Qwen3-30B-A3B	Late − deletion	0.056	1,536
Gemma-2-27B	Late − early	0.020	1,536
Gemma-2-27B	Late − deletion	0.024	1,536
Gemma-3-27B	Late − early	0.030	1,536
Gemma-3-27B	Late − deletion	0.030	1,536
OpenAI-Reddit			
NAML-T1	Late − early	≈ 0	109
NAML-T2	Late − early	0	109
EBNR-T1	Late − early	≈ 0	109
EBNR-T2	Late − early	≈ 0	109
SCAPE	Late − early	≈ 0	109

Table 17: RQ2 additional PENS ST configurations. Each difference compares the same selected ST evidence placed late with its early or absent condition. Cells give the raw mean difference and matched case count. The aligned suite uses case means, MRR for recommendation, and capped-peer PSE-BERT for generation. EBNR generation uses equal-document PSE-BERT. Candidate protocols and peer definitions remain specific to each configuration. These estimates supplement the LT tests without establishing retention of previously satisfied adherence.

Setting / model	**Task / score**	**Late − early**	**Late − absent**
Aligned / IMPerSumm	Rec. / MRR	1.45×10^{-4} (15,552)	1.45×10^{-4} (15,552)
Aligned / IMPerSumm	Gen. / PSE-BERT	2.05×10^{-4} (15,237)	2.05×10^{-4} (15,237)
Aligned / Walk2Pers	Rec. / MRR	5.78×10^{-5} (15,552)	5.78×10^{-5} (15,552)
Aligned / Walk2Pers	Gen. / PSE-BERT	5.25×10^{-4} (15,237)	5.25×10^{-4} (15,237)
EBNR suite / EBNR	Rec. / MRR	-0.004 (2,240)	-0.004 (2,240)
EBNR suite / EBNR-T1	Gen. / PSE-BERT	0.003 (1,401)	2.14×10^{-10} (1,401)
EBNR suite / EBNR-T2	Gen. / PSE-BERT	7.42×10^{-4} (1,401)	≈ 0 (1,401)

Table 18: Movement can change rankings without resolving mean-score differences. Mamba denotes Mamba4Rec. Raw scores are ordered early/middle/late. A ranking change means a different saved ordered list. Original MovieLens and additional ST rows retain separate candidate protocols. Adjusted ST mean-MRR contrasts remain unresolved.

Probe	Observed response	Matched cases	Diagnostic meaning
Mamba LT movement	MRR .130/.133/.028	3,436	Late contraction
Mamba LT movement	SR-ERR .707/.708/.694	3,436	Smaller semantic mean change
Mamba late versus deletion	-0.099 MRR	3,436	Below the deletion reference
IMPerSumm ST early-to-late	94.75% ranking changes	15,552	Output response despite small mean change
Walk2Pers ST early-to-late	92.93% ranking changes	15,552	Output response despite small mean change

Table 19: Complete ST movement profiles on PENS. All entries are scores multiplied by 100, not percentage gains. E/M/L denote early/middle/late placement. Recommendation uses MRR and SR-ERR@10; generation uses PSE-BERT and BERTScore. Rows retain their own task, cohort, and aggregation from Table 17 in Appendix I.2. These are descriptive means, not additional interval-based verdicts.

		MRR or PSE-BERT			SR-ERR@10 or BERTScore		
Model	Task	E	M	L	E	M	L
EBNR	Recommendation	7.810	7.630	7.370	50.510	50.470	50.310
IMPerSumm	Recommendation	1.470	1.470	1.490	92.670	92.670	92.680
Walk2Pers	Recommendation	1.630	1.630	1.640	93.520	93.520	93.520
EBNR-T1	Generation	1.720	2.040	2.040	83.430	83.420	83.420
EBNR-T2	Generation	0.230	0.300	0.300	83.360	83.360	83.360
IMPerSumm	Generation	8.020	8.020	8.040	88.200	88.200	88.200
Walk2Pers	Generation	9.160	9.160	9.210	82.740	82.740	82.750

Table 20: EBNR ST deletion and exposure profiles on 2,240 PENS ranking cases. Scores are multiplied by 100. Deletion ($K = 0$) has MRR 7.820 and SR-ERR@10 50.550 and has no placement. For $K > 1$ at early/middle placement, the additional copies are appended at the end; all late copies are at the end. Thus K varies both exposure and the documented layout.

	MRR			SR-ERR@10		
Copies K	Early	Middle	Late	Early	Middle	Late
1	7.810	7.630	7.370	50.510	50.470	50.310
2	7.740	7.500	7.530	50.520	50.440	50.320
4	7.600	7.500	7.250	50.360	50.340	50.280

Table 21: Mamba4Rec responds differently to distinct conflict controls on MovieLens. All scores and arrow endpoints are multiplied by 100. Persistence compares one and five conflicting turns. LT/ST subscripts count copies. The mixed balance sweep requires fixed-one-side follow-ups. SR-ERR@15 uses ten available predictions.

Mamba4Rec conflict axis	Control	MRR $\times 100$	SR-ERR@15 $\times 100$
Strength	.60/.80/.95	6.38/6.39/6.4	approximately flat
Persistence	1/2/3/5 turns	$5.73 \rightarrow 4.26$	$69.74 \rightarrow 69.45$
Placement	Early/Mid/Late	14.71/14.83/5.73	72.2/72.26/69.74
LT/ST balance	LT_1:ST_4 $\rightarrow$$LT_8$:$ST_1$	$4.15 \rightarrow 7.01$	$69.36 \rightarrow 70.19$

Table 22: RQ3 recommendation overview using MRR. Each row gives a signed difference of means and its matched case count n. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. Recommendation gives cases equal weight. OAI ranking is an auxiliary saved evaluation setting. Distractor means strong minus moderate at middle placement, episode means present minus removed, and conflict means late minus early.

Dataset / model	**Comparison**	**Difference**	n
MovieLens-1M			
Mamba4Rec	Distractor	0	6,040
Mamba4Rec	Episode	0.009	1,123
Mamba4Rec	Conflict	−0.090	6,040
DiffRec	Distractor	0	6,040
DiffRec	Episode	-5.69×10^{-4}	1,123
DiffRec	Conflict	0	6,040
PENS			
NRMS	Distractor	0	2,853
NRMS	Episode	0	1,858
NAML	Distractor	0	2,853
NAML	Episode	0	1,858
EBNR	Distractor	0	2,853
EBNR	Episode	0	1,858
GTP	Distractor	0	2,865
GTP	Episode	0	1,867
Qwen3-30B-A3B	Distractor	−0.005	1,200
Qwen3-30B-A3B	Episode	0.008	1,496
Qwen3-30B-A3B	Conflict	0.046	2,866
Gemma-2-27B	Distractor	−0.012	1,000
Gemma-2-27B	Episode	0.005	999
Gemma-2-27B	Conflict	−0.015	1,000
Gemma-3-27B	Distractor	0.006	114
Gemma-3-27B	Episode	-9.76×10^{-4}	185
Gemma-3-27B	Conflict	−0.002	270
OpenAI-Reddit			
Qwen3-30B-A3B	Distractor	6.86×10^{-4}	1,486
Qwen3-30B-A3B	Episode	0.012	1,464
Qwen3-30B-A3B	Conflict	0.003	1,472
Gemma-2-27B	Distractor	−0.012	924
Gemma-2-27B	Episode	0.005	937
Gemma-2-27B	Conflict	−0.005	932
Gemma-3-27B	Distractor	3.70×10^{-5}	986
Gemma-3-27B	Episode	0.014	949
Gemma-3-27B	Conflict	0.013	980

Table 23: RQ3 generation overview on PENS using PSE-BERT. Each row gives a signed difference of means, and its matched case count n. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. PENS denotes headline generation, and OAI denotes summarization. Each retained document receives equal weight after averaging its users. T1/T2 retain the reported interest-injection variants. The comparison definitions accompany these summaries in this appendix.

Dataset / model	**Comparison**	**Difference**	n
PENS			
NRMS-T1	Distractor strong − moderate	-5.15×10^{-4}	1,537
NRMS-T1	Conflict late − early	6.14×10^{-4}	1,540
NRMS-T2	Distractor strong − moderate	≈ 0	1,537
NRMS-T2	Conflict late − early	≈ 0	1,540
NAML-T1	Distractor strong − moderate	0.010	1,541
NAML-T1	Episode present − removed	−0.029	561
NAML-T1	Conflict late − early	≈ 0	1,541
NAML-T2	Distractor strong − moderate	0.014	1,541
NAML-T2	Episode present − removed	−0.034	561
NAML-T2	Conflict late − early	-1.51×10^{-4}	1,541
EBNR-T1	Distractor strong − moderate	-7.08×10^{-4}	1,549
EBNR-T1	Episode present − removed	−0.013	745
EBNR-T1	Conflict late − early	0.002	1,540
EBNR-T2	Distractor strong − moderate	≈ 0	1,549
EBNR-T2	Episode present − removed	0	745
EBNR-T2	Conflict late − early	≈ 0	1,540
GTP	Distractor strong − moderate	≈ 0	1,530
GTP	Episode present − removed	0	559
GTP	Conflict late − early	≈ 0	1,540
SCAPE	Distractor strong − moderate	≈ 0	1,541
SCAPE	Episode present − removed	0	561
SCAPE	Conflict late − early	≈ 0	1,540
Qwen3-30B-A3B	Distractor strong − moderate	−0.003	1,549
Qwen3-30B-A3B	Episode present − removed	0.007	745
Qwen3-30B-A3B	Conflict late − early	−0.010	1,541
Gemma-2-27B	Distractor strong − moderate	9.24×10^{-4}	1,549
Gemma-2-27B	Episode present − removed	0.005	745
Gemma-2-27B	Conflict late − early	0.008	1,541
Gemma-3-27B	Distractor strong − moderate	≈ 0	1,549
Gemma-3-27B	Episode present − removed	0.005	745
Gemma-3-27B	Conflict late − early	0.004	1,541

Table 24: RQ3 generation overview on OpenAI-Reddit using PSE-BERT. Each row gives a signed difference of means and its matched case count n. These descriptive estimates do not assign adherence verdicts. ≈ 0 denotes a magnitude below 10^{-12}. PENS denotes headline generation and OAI denotes summarization. Each retained document receives equal weight after averaging its users. T1/T2 retain the reported interest-injection variants. The comparison definitions accompany these summaries in this appendix.

Dataset / model	**Comparison**	**Difference**	n
OpenAI-Reddit			
NRMS-T1	Distractor strong − moderate	-0.008	16
NRMS-T1	Episode present − removed	0.026	2
NRMS-T1	Conflict late − early	-1.16×10^{-4}	1,092
NRMS-T2	Distractor strong − moderate	0	16
NRMS-T2	Episode present − removed	0	2
NRMS-T2	Conflict late − early	≈ 0	1,092
NAML-T1	Distractor strong − moderate	0.002	318
NAML-T1	Episode present − removed	0	300
NAML-T1	Conflict late − early	≈ 0	318
NAML-T2	Distractor strong − moderate	≈ 0	318
NAML-T2	Episode present − removed	0	300
NAML-T2	Conflict late − early	0.001	318
EBNR-T1	Distractor strong − moderate	8.39×10^{-4}	318
EBNR-T1	Episode present − removed	≈ 0	300
EBNR-T1	Conflict late − early	0.008	1,092
EBNR-T2	Distractor strong − moderate	≈ 0	318
EBNR-T2	Episode present − removed	≈ 0	300
EBNR-T2	Conflict late − early	≈ 0	1,092
GTP	Conflict late − early	≈ 0	1,092
SCAPE	Distractor strong − moderate	≈ 0	318
SCAPE	Episode present − removed	0	300
SCAPE	Conflict late − early	≈ 0	1,092
Qwen3-30B-A3B	Episode present − removed	-0.026	378
Gemma-2-27B	Distractor strong − moderate	0.005	753
Gemma-2-27B	Episode present − removed	0.008	652
Gemma-2-27B	Conflict late − early	5.56×10^{-4}	872
Gemma-3-27B	Distractor strong − moderate	-8.45×10^{-4}	753
Gemma-3-27B	Episode present − removed	≈ 0	652
Gemma-3-27B	Conflict late − early	9.80×10^{-4}	872

J SUPPORTING STATISTICAL ANALYSES

The paired analyses condition on the saved checkpoints, decoded outputs, and recorded metric contributions. The main reanalysis uses 10,000 paired bootstrap resamples: recommendation resamples matched users, original generation resamples complete documents with their comparison users, and additional suites preserve their user/document or user/target indexing. Shared resampling keeps compared conditions paired. Simultaneous intervals apply within each declared exploratory family rather than across the paper, and they do not include training-seed variation.

For the 6,040-case Mamba4Rec RQ1 comparison, exact-target LT–ST and LT–EP differences remain below zero after exploratory simultaneous adjustment. The semantic LT–EP interval crosses zero, so the weak semantic non-collapse relation remains unresolved even though the exact-target relation is contradicted. The population semantic LNC and mean instance-level LNC can have opposite signs because the minimum operator does not commute with averaging, as formalized in Appendix A.

For the 3,436-case Mamba4Rec movement cohort, late placement has lower MRR than early placement and deletion. Among cases recovering the target early or in the middle, 881 lose it under late placement; the semantic mean changes less than exact-target recovery, while the subgroup losing a recovered target exhibits a larger semantic decrease. These paired contrasts distinguish a population mean response from heterogeneous casewise effects.

The aligned PENS ST suite additionally separates output equality from score equality. Early and middle conditions can produce identical saved outputs in a suite even when late placement changes most rankings; an unresolved average target-score contrast therefore does not imply an unchanged output. Conversely, near-floor or flat task scores cannot by themselves establish useful invariance. The paper reports these diagnostics only to interpret the declared score relations, not as independent adherence criteria.

K THRESHOLD SENSITIVITY ANALYSIS

Threshold sensitivity holds the saved model outputs, comparison cohort, weighting, and readout fixed while varying the acceptable margin in a declared relation. It is a sensitivity analysis, not a procedure for choosing a margin after observing which value makes a model pass. Table 25 reports the resulting threshold regions on the fixed admitted cohorts.

Table 25: Threshold sensitivity on fixed admitted cohorts. Intervals are exploratory simultaneous nominal 95% bounds within each saved family, not paper-wide guarantees. For invariance, a is the interval's nearest distance to zero and b its farthest distance. The displayed endpoints are rounded, while the range rules apply to the full-precision endpoints. For a fixed valid uncertainty set, inspecting more thresholds does not require additional error allocation. Model adherence still requires the specified relation and evidence conditions. The full analysis is in Appendix K.

Comparison / cases	**Estimate / interval**	**Threshold regions**
Mamba4Rec LT − weaker transient / 6,040	-0.034 $[-0.041, -0.026]$	For required advantage $\gamma \geq 0$, contradiction throughout. No supported or unresolved nonnegative margin.
Mamba4Rec LT late − early MRR / 3,436	-0.103 $[-0.119, -0.087]$	Invariance $\lvert\Delta\rvert \leq \epsilon$. Contradiction for $\epsilon < a$, unresolved for $a \leq \epsilon < b$, support for $\epsilon \geq b$. $a \approx 0.087$, $b \approx 0.119$.
IMPerSumm ST late − early MRR / 15,552	1.45×10^{-4} $[-3.38, 6.27] \times 10^{-4}$	Invariance unresolved for $0 \leq \epsilon < b$ and supported for $\epsilon \geq b$, with $b \approx 6.27 \times 10^{-4}$. No contradiction region.
Walk2Pers ST late − early PSE-BERT / 15,237	5.25×10^{-4} $[-5.255, 6.305] \times 10^{-3}$	Invariance unresolved for $0 \leq \epsilon < b$ and supported for $\epsilon \geq b$, with $b \approx 6.305 \times 10^{-3}$. Mean absolute case change is 0.055.
IMPerSumm ST early repetition − clean MRR / 15,552	0 Degenerate paired uncertainty	Saved-cohort equality holds at every $\epsilon \geq 0$. Population invariance is withheld.

For Mamba4Rec's MovieLens RQ1 MRR comparison, the weak LT non-collapse quantity has exploratory simultaneous bounds approximately $[-0.041, -0.026]$. Because the upper endpoint is below zero, every nonnegative required LT advantage is contradicted under that uncertainty family. For the original LT-movement comparison, late-minus-early MRR is -0.103 with interval $[-0.119, -0.087]$. An illustrative invariance relation $|\Delta| \leq \epsilon$ is contradicted for $\epsilon < 0.087$, unresolved for approximately $0.087 \leq \epsilon < 0.119$, and supported at larger tolerances under the same interval.

Aggregate stability and casewise stability are different contracts. In the aligned Walk2Pers ST placement suite, the saved capped-peer PSE-BERT late-minus-early mean is 5.251×10^{-4} over 15,237 cases, while 6,421 paired scores differ by more than 10^{-12} and the mean absolute change is 0.055. A small net mean can therefore coexist with substantial individual movement. Threshold profiles are interpreted only for the population relation they actually bound.

L Reproducibility And Claim Boundaries

A reproducible comparison identifies the dataset/split, evaluation unit, fitted checkpoint, history preprocessing, target/reference, candidate pool or source input, controlled history, operator parameters, scorer configuration, aggregation weights, and matched support. The scientific claim is limited to that specified interface. Architecture descriptions motivate the tested relation; they do not identify the internal component that caused an observed behavioral response. Paired saved-output analyses likewise condition on the available checkpoints and decoded outputs and do not quantify training-seed variability. Table 26 separates conclusions supported by the present protocol from claims that require additional evidence.

Table 26: What the measured comparisons establish and what would require additional evidence. Each behavioral conclusion is conditional on valid constructed inputs, matched evaluation units, and the declared readout. It does not identify a hidden model component.

Supported by the present protocol	**Not supported without additional evidence**
Observable sensitivity to a declared evidence intervention under a fixed final task	Causal attribution to an internal state, gate, attention head, memory lane, or decoder block
Different exact-target and semantic score responses on the same cases	Proof that the model internally represents the same semantic neighborhood measured by the external readout
Personalization-score movement with little reference-score movement	Claim that all aspects of personalization changed while content remained completely invariant
Response curves for the tested model, task, and evidence counts	A universal LT overwrite threshold or population-level change point
Flat scores at a metric floor	Robustness or successful evidence rejection

L.1 Asset Terms And Usage

Reproduction requires respecting the separate terms attached to datasets, model weights, and software. Raw interaction histories, references, or model weights should not be redistributed when their upstream terms do not permit it. The reproduction package should instead provide acquisition instructions, construction settings, and analysis code sufficient to recreate the controlled comparisons from legitimately acquired assets. The upstream notices cited by the project source remain applicable to PENS, MovieLens-1M, Qwen, Gemma, BGE-M3, DeepSeek-R1-Distill-Qwen-14B, all-MiniLM-L6-v2, and Flan-T5-base; a code license does not supersede the terms of separately acquired data or weights.

M Extended Related Work

M.1 Recommender Reliability And Sequence Use

Sequence and recommender tests. History shuffling tests whether recommendation models use interaction order (Kim et al., 2025). This directly probes sequence use, but sequence sensitivity alone

does not specify how distinct forms of preference evidence should influence a target under a model's design rationale.

Recommender metamorphic testing instead specifies expected relations between source and transformed executions. Mao et al. (2024) develop a recommender-specific metamorphic-testing framework with relations derived from prediction-algorithm properties and evaluate its concrete realizations across multiple recommender libraries. Their earlier study provides a narrower recommender robustness case study (Mao et al., 2020), while Khirbat et al. (2024) test rating and prompt transformations. These works establish useful transformation–response relations, but their scientific object is recommender reliability or invariance rather than rationale-conditioned preference adherence.

PREFTURB addresses the latter problem by identifying LT, ST, and EP evidence before intervention and preserving the selected evidence while changing its placement, availability, repetition, or competition. The resulting responses are interpreted through an ISCON contract rather than through perturbation sensitivity alone.

M.2 BEHAVIORAL RELATIONS AND SPECIFICATION TESTING

General testing abstractions. Metamorphic testing provides a general abstraction for testing systems through relations between executions when a direct output oracle is unavailable (Chen et al., 2018). Domain-specific methods instantiate that abstraction by choosing appropriate relations and transformations. CheckList similarly operationalizes behavioral testing through capability-specific minimum-functionality, invariance, and directional tests (Ribeiro et al., 2020), while METTLE applies user-selected metamorphic expectations to clustering systems (Xie et al., 2020).

ISCON specifies a different abstraction. Its object is whether observed behavior is consistent with a model's declared design rationale. An adherence contract therefore jointly specifies admissible evidence, measurement context, aggregation, and an acceptable response relation. This distinction matters because the same observed response can be compatible with one rationale and incompatible with another.

Decision uncertainty. Agnostic tests provide a complementary statistical abstraction with support, rejection, and an unresolved outcome under explicit error control (Coscrato et al., 2020). They do not determine which evidence, measurements, or response relations constitute a scientifically valid design-rationale test. ISCON supplies that interpretation for its declared contract. Theorem 3.1 formalizes the resulting decision conditions, while Theorem B.1 transfers them to the PREFTURB instantiation under its assumptions.

Certified fairness concerns a different guaranteed object: properties of a modified predictor under specified mutations rather than correctness of an adherence conclusion about a fixed tested model (Ma et al., 2020).

M.3 CONTEXT POSITION AND CONVERSATIONAL MEMORY

Context and memory benchmarks. Lost in the Middle studies whether relevant information remains usable at different context positions (Liu et al., 2024b). LoCoMo and LongMemEval construct extended conversational histories to evaluate retrieval, reasoning, updating, and long-term memory (Maharana et al., 2024; Wu et al., 2025).

These benchmarks provide substantial controlled history construction, but their shared objective is memory or context use. They do not specify the relative influence that a particular personalization rationale assigns to matched LT, ST, and EP preference evidence. ISCON places that expected relation inside the contract, while PREFTURB preserves selected evidence as its placement or competing context changes.

M.4 CONSTRUCTION AND INTERPRETATION OF CONTROLLED TESTS

Evidence construction. Controlled testing also differs in how the compared evidence is obtained. Contrast sets use expert edits of source examples to probe local task competence (Gardner et al., 2020). Recommender metamorphic tests construct follow-up inputs from source inputs according to specified transformations (Mao et al., 2024; 2020).

PREFTURB instead constructs target-conditioned preference traces from observed histories, separates temporal role from target relation, and freezes accepted events and origin labels before intervention (Algorithm 1 in Appendix G.2). Proposition B.1 establishes preservation of retained payloads and recorded roles under the declared edits. Thus, controlled transformation alone and preservation of the particular evidence whose influence is being tested are distinct requirements.

Instantiation of the abstraction. The distinction between ISCON and PREFTURB is therefore structural. ISCON defines the obligations and interpretation of a design-rationale adherence test; it does not require one particular trace constructor, encoder, metric, or task. A concrete instantiation must supply those ingredients and show that they realize the specified contract.

PREFTURB is the demonstrated personalization instantiation in this paper. It supplies the evidence construction, interventions, measurements, aggregation, and rationale-specific relations required by its ISCON contract. Theorem B.1 establishes this connection under the stated assumptions.

Other frameworks in the comparison can likewise admit multiple concrete realizations. Metamorphic testing, for example, is instantiated through domain-specific metamorphic relations and source/follow-up inputs; ISCON is instantiated through a rationale-specific evidence, measurement, aggregation, and relation specification. The difference is therefore not whether a framework has instantiations, but what scientific object those instantiations are designed to test.

M.5 PERSONALIZATION MEASUREMENTS AND BENCHMARKS

Preference evaluation. PersonaLens evaluates personalized dialogue using profiles and interaction histories, while PERG studies factual correctness and preference following under relevant and irrelevant preference conditions (Zhao et al., 2025; Okite et al., 2025). These benchmarks provide personalization tasks and controlled preference conditions, but they do not by themselves define adherence to a model-specific temporal rationale.

Personalization measurements. EGISES and PerSEval quantify whether differences among user-specific outputs correspond to differences among references, while ExPerT evaluates personalized content and style (Vansh et al., 2023; Dasgupta et al., 2024; Salemi et al., 2025). Such measurements can serve as components of an adherence test, but a metric alone does not determine which evidence should be compared, how cases should be aggregated, or which score relation constitutes adherence. ISCON specifies those responsibilities, and PREFTURB supplies one concrete realization.

M.6 MASTER COMPARISON

Tables 27, 28, and 29 separate three responsibilities: specifying the adherence problem, constructing admissible evidence for it, and justifying the resulting diagnosis. The first two rows should therefore be read hierarchically: ISCON is the general adherence abstraction and PREFTURB is its demonstrated personalization instantiation.

The remaining rows need not occupy the same level of abstraction. They include general testing frameworks, domain- or task-specific realizations, benchmarks, and measurement procedures. The comparison asks which responsibilities each cited formulation supplies for the adherence problem; it does not rank their overall quality or usefulness. In the "Any model" column, generality refers to whether the adherence question is formulated beyond personalization, not to a claim that every possible model or rationale is admissible.

A tick denotes a supplied responsibility, a triangle a related or narrower capability, and a cross a responsibility not supplied by the cited formulation. The symbol **I** denotes an ingredient intentionally supplied by a concrete instantiation or external specification rather than by the abstraction itself.

Formal Specification. An ISCON contract jointly specifies valid evidence, measurement context, aggregation, and the acceptable response relation induced by the tested rationale. Behavioral and metamorphic frameworks can supply explicit expected relations—including the recommender-specific relations of Mao et al. (2024)—without thereby supplying the complete rationale-conditioned adherence object. Agnostic testing supplies a general decision principle but likewise leaves the scientific object of the test to the application.

Evidence Construction. PREFTURB supplies the concrete preference-trace construction and frozen-evidence interventions required by its ISCON instantiation. Recommender metamorphic tests construct legitimate source/follow-up input pairs, and profile or preference controls in other benchmarks provide related interventions. These receive partial credit where appropriate, but they do not supply the same temporal preference-trace definition and preservation guarantee.

Justified Diagnosis. A justified instantiation must connect its concrete procedure to the separately specified contract. For PREFTURB, the conformance result establishes that connection. Subsequent tests distinguish continued adherence from diagnostics after an unsupported baseline, while the aggregation results characterize how declared comparator, grouping, or weighting choices can alter the adherence verdict on the same outputs. Other methods may aggregate violations or task scores for their own objectives without providing these adherence-specific guarantees.

Table 27: Master comparison (a). Specifying the adherence problem. The first two columns identify whether the cited formulation addresses adherence to a model's own design rationale. The remaining columns concern a formal rationale-specific test and its interpretation. Generic scoring or significance testing alone does not satisfy these criteria. Error-control ticks refer to formal guarantees under their stated assumptions; the empirical bootstrap intervals in this paper remain exploratory.

Work, source, and task	Tests adherence to own design rationale		Formal specification		
	Personalization models	Any model	Formal adherence contract	Formal verdict interpretation	Adherence verdict error control
ISCon (this work) (general adherence abstraction)	✓	✓	✓	✓	✓
PrefTurb (this work) (personalization instantiation)	✓	✓	✓	✓	✓
Lost in Sequence (sequence use in recommendation) Kim et al., KDD 2025	✓	✓	△	×	×
Recommender reliability (metamorphic-testing framework) Mao et al., Inf. Softw. Technol. 2024	△	△	△	×	×
Recommender robustness (earlier recommender MT case study) Mao et al., DSA 2020	△	△	△	×	×
Metamorphic recommendation (rating/prompt invariance) Khirbat et al., arXiv preprint 2024	△	△	△	×	×
CheckList (NLP capability tests) Ribeiro et al., ACL 2020	×	△	△	×	×
METTLE (clustering behavior) Xie et al., IEEE Trans. Reliability 2020	×	△	△	×	×
Agnostic tests (hypothesis decisions) Coscrato et al., Braz. J. Probab. Stat. 2020	I	I	I	×	△
PersonaLens (personalized dialogue) Zhao et al., Findings of ACL 2025	×	×	×	×	×
PERG (correctness and preference following) Okite et al., Findings of EMNLP 2025	×	×	×	×	×
Certified fairness (text-classifier fairness) Ma et al., IJCAI 2020	×	×	×	×	×
Contrast sets (local task competence) Gardner et al., Findings of EMNLP 2020	×	×	×	×	×
Lost in the Middle (positional information use) Liu et al., TACL 2024b	×	×	×	×	×
LoCoMo (conversational memory) Maharana et al., ACL 2024	×	×	×	×	×
LongMemEval (memory capabilities) Wu et al., ICLR 2025	×	×	×	×	×
PerSEval (personalization quality) Dasgupta et al., TMLR 2024	×	×	×	×	×
ExPerT (personalized content and style) Salemi et al., Findings of ACL 2025	×	×	×	×	×

✓ supplied △ partial × not supplied **I** instantiation/specification input.

ISCon is the adherence abstraction and PrefTurb its demonstrated personalization instantiation. Other rows may likewise denote a general testing framework, a task-specific realization, or a supporting benchmark or measurement. Symbols therefore indicate whether the cited formulation supplies the stated scientific responsibility, not an overall ranking. A partial mark denotes a related or narrower capability rather than full coverage of the adherence-specific criterion.

Table 28: Master comparison (b). Constructing evidence for the adherence problem. The repeated leading columns retain the scientific objective while the remaining columns compare preference-trace construction, controlled interventions, and their preservation guarantee.

Work, source, and task	Tests adherence to own design rationale		Evidence construction		
	Personalization models	Any model	Preference trace design	Frozen evidence edits	Evidence preservation proof
ISCon (this work) (general adherence abstraction)	✓	✓	I	I	I
PrefTurb (this work) (personalization instantiation)	✓	✓	✓	✓	✓
Lost in Sequence (sequence use in recommendation) Kim et al., KDD 2025	✓	✓	×	△	×
Recommender reliability (metamorphic-testing framework) Mao et al., Inf. Softw. Technol. 2024	△	△	×	△	×
Recommender robustness (earlier recommender MT case study) Mao et al., DSA 2020	△	△	×	△	×
Metamorphic recommendation (rating/prompt invariance) Khirbat et al., arXiv preprint 2024	△	△	×	△	×
CheckList (NLP capability tests) Ribeiro et al., ACL 2020	×	△	×	×	×
METTLE (clustering behavior) Xie et al., IEEE Trans. Reliability 2020	×	△	×	×	×
Agnostic tests (hypothesis decisions) Coscrato et al., Braz. J. Probab. Stat. 2020	I	I	×	×	×
PersonaLens (personalized dialogue) Zhao et al., Findings of ACL 2025	×	×	×	△	×
PERG (correctness and preference following) Okite et al., Findings of EMNLP 2025	×	×	×	△	×
Certified fairness (text-classifier fairness) Ma et al., IJCAI 2020	×	×	×	×	×
Contrast sets (local task competence) Gardner et al., Findings of EMNLP 2020	×	×	×	×	×
Lost in the Middle (positional information use) Liu et al., TACL 2024b	×	×	×	×	×
LoCoMo (conversational memory) Maharana et al., ACL 2024	×	×	×	×	×
LongMemEval (memory capabilities) Wu et al., ICLR 2025	×	×	×	×	×
PerSEval (personalization quality) Dasgupta et al., TMLR 2024	×	×	×	×	×
ExPerT (personalized content and style) Salemi et al., Findings of ACL 2025	×	×	×	×	×

✓ supplied △ partial × not supplied I instantiation/specification input.

ISCon specifies the responsibilities that a valid instantiation must discharge; it does not prescribe one concrete preference-trace constructor. PrefTurb supplies and proves one such construction. Other transformation-based methods can provide related controlled edits without supplying the same preference-evidence definition or preservation guarantee. Symbols compare these responsibilities rather than overall method quality.

Table 29: Master comparison (c). Justifying and extending adherence conclusions. The repeated leading columns identify the tested objective. The remaining columns concern formal conformance, follow-up after the initial adherence test, and the effect of aggregation choices on its verdict.

Work, source, and task	Tests adherence to own design rationale		Justified diagnosis		
	Personalization models	Any model	Contract realization proof	Adherence follow-up	Aggregation effects
ISCon (this work) (general adherence abstraction)	✓	✓	I	✓	✓
PrefTurb (this work) (personalization instantiation)	✓	✓	✓	✓	✓
Lost in Sequence (sequence use in recommendation) Kim et al., KDD 2025	✓	✓	×	△	×
Recommender reliability (metamorphic-testing framework) Mao et al., Inf. Softw. Technol. 2024	△	△	×	×	×
Recommender robustness (earlier recommender MT case study) Mao et al., DSA 2020	△	△	×	×	×
Metamorphic recommendation (rating/prompt invariance) Khirbat et al., arXiv preprint 2024	△	△	×	×	×
CheckList (NLP capability tests) Ribeiro et al., ACL 2020	×	△	×	×	×
METTLE (clustering behavior) Xie et al., IEEE Trans. Reliability 2020	×	△	×	×	×
Agnostic tests (hypothesis decisions) Coscrato et al., Braz. J. Probab. Stat. 2020	I	I	×	×	×
PersonaLens (personalized dialogue) Zhao et al., Findings of ACL 2025	×	×	×	×	×
PERG (correctness and preference following) Okite et al., Findings of EMNLP 2025	×	×	×	×	×
Certified fairness (text-classifier fairness) Ma et al., IJCAI 2020	×	×	×	×	×
Contrast sets (local task competence) Gardner et al., Findings of EMNLP 2020	×	×	×	×	×
Lost in the Middle (positional information use) Liu et al., TACL 2024b	×	×	×	×	×
LoCoMo (conversational memory) Maharana et al., ACL 2024	×	×	×	×	×
LongMemEval (memory capabilities) Wu et al., ICLR 2025	×	×	×	×	×
PerSEval (personalization quality) Dasgupta et al., TMLR 2024	×	×	×	×	×
ExPerT (personalized content and style) Salemi et al., Findings of ACL 2025	×	×	×	×	×

✓ supplied △ partial × not supplied I instantiation/specification input.

Here **I** separates an obligation of the abstraction from a property that must be established by its concrete realization. For this work, PrefTurb provides the demonstrated realization of the ISCon contract. Other frameworks may also admit multiple concrete realizations, but their guaranteed or diagnosed object can differ from design-rationale adherence. A partial mark denotes a related or narrower capability; the symbols are not intended as a performance ranking.