SIMULATING INDIVIDUAL DELIVERY RIDERS FROM THEIR OWN HISTORIES WITH MEASUREMENT-GROUNDED MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Delivery platforms want to simulate how an individual rider will behave by conditioning a language model on that rider's own history. When such a simulator must show LLM-written memory to its forecaster, the memory's numbers have to stay tied to the rider's measurements. We build the simulator around a measurement contract: source statistics keep their numerator, denominator, unit and window; LLM-written memory is stored as typed, unverified claims with declared dependencies; a deterministic check flags a claim only against a measurement of the same definition; and dependency closure retracts every statement built on a flagged claim. We prove that the retained memory is the unique largest conflict-free, dependency-closed sub-memory, that budgeted selection is feasible and maximal, and that a premise-preserving selector keeps both guarantees after retraction. Across 316 riders, conditioning on the rider's own history lowers offer-response log loss by 0.0892 relative to population context and raises accuracy by 15.30 and 8.98 points over wrong-identity and matched-neighbor histories, so the simulator represents the intended rider. An outcome-free audit of 590 memories for 295 riders finds 41.6% of the explicit numeric claims outside tolerance even when the memory is written from the very table the measurements come from; an independent reader of the same memories reproduces every parsed claim, source measurement and retraction set. In a preregistered two-week forecasting study with a Qwen3.5-9B writer and forecaster, typed memory with the consistency check is non-inferior and superior to the same memory shown as plain text on the 22 evaluable riders of a holdout group outside every earlier cohort, with differences of -0.2196 in offer log loss and -0.0506 in refusal-fraction error, and across all 295 riders it lowers the share of conflicting metrics on which the forecast lies closer to the stale claim than to the fresh measurement from 22.3% to 4.7%. A registered validation at five new forecast cutoffs finds checked memory non-inferior to plain memory on offer log loss, and on 789 never-exposed riders of a third sample checked memory lowers offer log loss relative to plain memory by 0.1239 with interval $[-0.1578, -0.0909]$ in a registered descriptive contrast. In a registered large-sample confirmation on a 9,350-rider cohort, combining an LLM forecast that reads the rider's memory documents and measurements with a smoothed personal statistical predictor lowers refusal-fraction error by 9% relative to that predictor alone on 1,998 confirmatory riders, -0.0055 $[-0.0080, -0.0029]$, under both registered tests. The memory layer thus provides measurement-grounded maintenance with structural guarantees for simulators that must use external memory.

1 INTRODUCTION

Delivery platforms constantly need to anticipate how a *particular* rider will behave: whether she will accept the next dispatched offer, how many days she will work in the coming two weeks, or how much of her work will fall in the evening. Online experiments answer such questions slowly and at real cost, and population-level models average away exactly the heterogeneity that matters for dispatch, scheduling and incentive design (Chen et al., 2019; Wu et al., 2022; Robbennolt et al., 2026). Large language model (LLM) agents suggest a cheaper route. Generative agents and social-simulation

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.

sandboxes populate virtual environments with agents that remember, reflect and act (Park et al., 2023; Piao et al., 2026; Zhang et al., 2025; Yang et al., 2024), and a growing line of work conditions an agent on one person's own records to simulate that person (Park et al., 2026; Wang et al., 2026b; Guo et al., 2026). This paper studies the corresponding individual-level prediction setting: *given one rider's history and an LLM, simulate the rider's real future behavior, establish that the simulator represents the intended rider, and determine how LLM-written memory must be represented so that it does not mislead the simulation with wrong numbers.*

Rider histories make the task harder than it first appears. The evidence is numeric and windowed: a refusal fraction has a numerator, a denominator and an observation window. It is partially observed: a day without records is not a day without work. And it drifts: a memory written two weeks ago may no longer describe the rider. An LLM-written persona or memory can be fluent and internally coherent while misreporting a measured quantity, and later statements can silently build on the wrong one. Existing sandboxes are mostly validated on aggregate patterns or on believability rather than on the future behavior of the individual being simulated (Park et al., 2023; Piao et al., 2026; Anthis et al., 2025; Li & Ji, 2026). General-purpose agent memories reconcile conflicting free-text facts by recency, validity time or model judgment (Rasmussen et al., 2025; Banerjee et al., 2026; Latimer et al., 2026; Xu et al., 2025), but have no notion of when two numbers measure the same thing. A rider simulator therefore has to keep its numbers tied to their definitions and has to be judged against the rider's real future, not against plausibility.

We build a single-rider simulator around a *measurement contract*, shown in Figure 1. Every source statistic carries its numerator, denominator, unit and window. LLM-written memory, organized into long-term patterns, recent context, a personal profile and decision hypotheses, is stored as typed and explicitly *unverified* claims with declared dependencies. A deterministic checker flags a memory claim only when it disagrees with a measurement *of the same definition*. Statements that depend on a flagged claim are retracted by dependency closure, and the retraction is shown to the forecaster together with its reason. A question-conditioned selector fits whole statements into a character budget. We prove that the retained memory is the unique largest conflict-free, dependency-closed sub-memory and that a premise-preserving budgeted selector keeps this guarantee.

The evaluation asks four questions a practitioner must answer before relying on such a simulator. *Does the simulator represent the intended rider?* Identity substitution with wrong-rider and matched-neighbor histories answers this on 316 riders. *Can LLM-written memory be trusted for numbers?* An outcome-free audit of 295 riders checks every numeric claim that the memory writer makes against the source measurement of the same definition and splits each disagreement into writer error and drift. *Does the consistency check improve forecasts that use memory?* A preregistered study on the same riders shows the same lagged memory as plain text, as typed claims and as typed claims after the check, with identical source measurements, and fixes non-inferiority margins before scoring; the 27 riders outside every earlier cohort form the confirmatory group. Further evaluations test transfer to five new cutoffs, to 789 never-exposed riders and to a later cutoff with a simple numeric repair as comparator. *Does an LLM forecast that reads memory and measurements add value to a simple personal statistical predictor?* A registered forecast on a fourth cohort of 9,350 riders, whose memory documents are shown verbatim, answers this at scale. The rider is the unit of analysis throughout.

Our contributions are as follows.

- **A measurement-grounded memory interface for simulators that must use external memory.** Numeric claims are bound to their definitions, conflicts are identified only between claims of the same definition, and dependency closure retracts statements built on a contradicted claim. Propositions 1, 2 and 5 state the guarantees of retraction, budgeted selection and their composition.
- **Identity fidelity.** Across 316 riders, personal context outperforms population, wrong-identity and matched-neighbor controls, so the simulator uses information that belongs to the represented rider.
- **Numeric errors of LLM-written memory arise at writing time.** For the audited writer, 41.6% of the numeric claims in memories written at the forecast cutoff, from the very table the measurements are computed on, fall outside tolerance, and memories written two weeks earlier reach 42.6%; closure retracts 527 further statements that build on a flagged claim.

- **A preregistered test of the consistency check.** On holdout riders outside every earlier cohort, checked memory is non-inferior and superior to plain memory on both primary endpoints, also with corrected rendering; across all riders it lowers the share of conflicting metrics on which the forecast lies closer to the stale claim from 22.3% to 4.7%.
- **Transfer to new cutoffs and new riders.** At five new windows, checked memory is non-inferior to plain memory on offer log loss under the registered test; on 789 never-exposed riders, every rendering of checked memory descriptively lowers both primary errors relative to plain memory; at a later cutoff, it is non-inferior to a simple numeric repair on refusal-fraction error.
- **Large-sample confirmation against a statistical predictor.** On 1,998 evaluable riders of a 9,350-rider cohort, an LLM forecast that reads memory documents and measurements, combined with a smoothed personal statistical predictor, lowers refusal-fraction error by 9% relative to that predictor under both registered tests.

2 Related Work

Generative agents and social-simulation sandboxes. Generative Agents store an agent's experiences in natural language, synthesize them into reflections and retrieve them to plan behavior in a sandbox town, and are evaluated for believability (Park et al., 2023). "Silicon sample" studies showed that conditioned LLMs reproduce aggregate survey and experimental patterns (Argyle et al., 2023; Aher et al., 2023; Filippas et al., 2024), and LLM predictions correlate with treatment effects across 70 preregistered survey experiments (Ashokkumar et al., 2026). Sandboxes have scaled to social networks, recommendation platforms, million-agent social media and population-aligned societies (Gao et al., 2023; Wang et al., 2025b; Zhang et al., 2024; Yang et al., 2024; Piao et al., 2026; Zhang et al., 2025). They are validated mainly on population-level phenomena. Critical work warns that LLM stand-ins can flatten groups (Wang et al., 2025a), that realistic-looking responses do not certify treatment-effect accuracy (Li & Ji, 2026), and that confirmatory use requires calibration against real data (Hullman et al., 2026; Anthis et al., 2025). Our simulator targets the individual: one agent per real rider, scored on that rider's own future behavior.

Simulating specific individuals from their own data. Agents built from interviews and surveys of 1,052 people reproduce those people's held-out survey answers at 82–86% of their own two-week test–retest consistency (Park et al., 2026). Centaur predicts held-out participants in laboratory tasks (Binz et al., 2025), and shopping agents predict the next action of real sessions with personas, rationales or trained policies (Lu et al., 2026; Wang et al., 2026b; 2025c). Persona benchmarks find that strong models struggle to continue behavior chains (Li et al., 2025) and fall short on facets of persona fidelity (Du et al., 2026); with ten years of one volunteer's messages, fine-tuning, retrieval and memory each fail an individual Turing test (Guo et al., 2026). LaMP retrieves from user profiles for personalized outputs (Salemi et al., 2024), and behavior-grounded personas pre-screen A/B tests (Benomar et al., 2026). These works evaluate same-period held-out answers, the next click in a session or human plausibility judgments. Rider histories are numeric, windowed and partially observed, and the target is behavior over the following weeks. Traveler agents align with individuals' route choices (Liu et al., 2026) and rider-persona agents study system-level delivery routing (Zhang & Xiao, 2026); Appendix Q covers mobility and platform work.

Memory, conflicts and revision. Agent memories page, forget, link and update free-text facts (Packer et al., 2023; Zhong et al., 2024; Xu et al., 2025; Chhikara et al., 2025). Zep invalidates contradicted facts in a bi-temporal graph (Rasmussen et al., 2025), Hindsight separates world facts from evolving beliefs (Latimer et al., 2026), APEX-MEM resolves conflicts at query time (Banerjee et al., 2026), and LongMemEval tests extraction, temporal reasoning and knowledge updates (Wu et al., 2025). Our retraction applies truth-maintenance withdrawal (Doyle, 1979; de Kleer, 1986) to declared dependencies and, unlike belief-base contraction (Alchourrón et al., 1985), yields a unique largest admissible memory. Its core is a *numerical* contract: a claim's identity is its metric, denominator and window, a conflict exists only against a measurement of the same definition, and the memory is judged by forecasts of real behavior. LLMs can forecast numeric series zero-shot (Gruver et al., 2023); our forecaster instead receives the rider's source measurements explicitly, and memory enters only as typed, checked context. Appendix Q covers further related work.

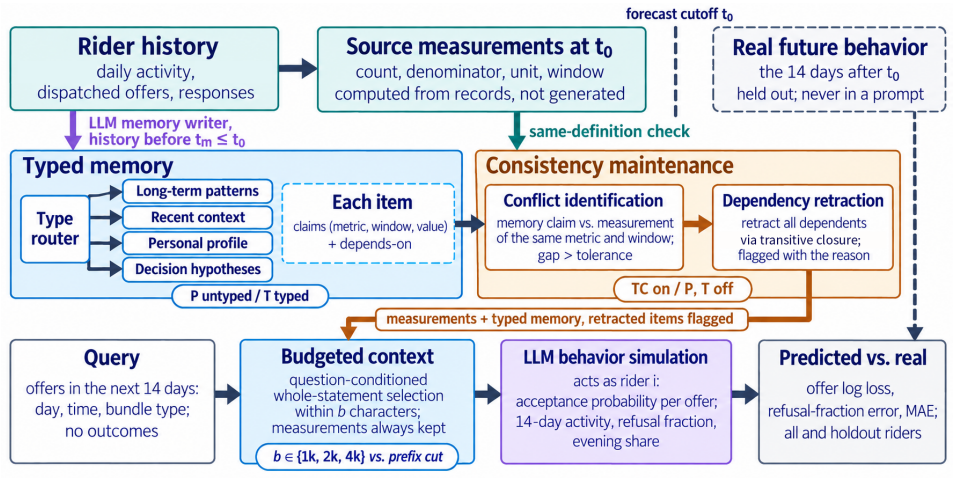


Figure 1: Rider simulation framework. Source measurements with count, denominator, unit and window are computed from the rider's history up to the forecast cutoff t_0. The LLM writes a typed four-role memory from the history before a memory cutoff $t_m \le t_0$, with numeric claims and declared dependencies. Consistency maintenance flags a claim that departs from the measurement with the same metric and window by more than a fixed tolerance and retracts every statement that depends on it; retracted items stay visible with their reason. A question-conditioned selector keeps whole statements within b characters. The same LLM simulates the rider's responses to future offers and 14-day activity, which are scored against the rider's real behavior after t_0; that behavior never enters a prompt. Tags mark the switches of the compiled conditions of Appendix O; in the forecasting study, retraction and selection are separate conditions, and Appendix T analyzes their composition.

3 Measurement-Grounded Rider Simulation

Algorithm 1 in Appendix O compiles the context of every condition, and proofs are in Appendix S.

3.1 Simulation setting

For rider i, let $H_i^{<t}$ be the records before local day t. A simulation query fixes a forecast cutoff t_0, a horizon of h days, and a *memory cutoff* $t_m = t_0 - \Delta$; we use $h = \Delta = 14$. As in generative-agent simulations (Park et al., 2023), the model has two roles. The *memory writer* Φ reads the daily activity table for $[t_m - 60, t_m)$ and writes a memory $M_i = \Phi(H_i^{<t_m})$. The *forecaster* Ψ reads a compiled context and simulates the rider over $[t_0, t_0 + h)$. The aggregate task returns the numbers of recorded days, completions per recorded day, the refusal fraction and the evening share. The offer task returns an acceptance probability for each future dispatched offer, given only the offer's time and type. Both roles use the same model with greedy decoding. Unparsable forecasts are replaced by a fallback computed from the 28-day measurements, and we count these replacements. Conditions differ only in how memory and measurements are compiled into the context, as Appendix O details.

Source measurements. A measurement definition is a key $k = (m, \ell)$: a metric $m \in \mathcal{M}$ and a trailing window of $\ell \in \{7, 28, 60\}$ calendar days. $\mathcal{M}$ contains completions per recorded day; the refusal fraction, that is, manual refusals divided by accepted plus manually refused dispatched offers; the evening share of completions, from 18:00 to 24:00; and recorded days per week. Over the recorded days in $[t - \ell, t)$, the source gives a numerator $n_k(t)$ and a denominator $d_k(t)$, and

$$\mu_k(t) = \begin{cases} \alpha_m \, n_k(t)/d_k(t), & d_k(t) > 0, \\ \perp, & d_k(t) = 0, \end{cases} \qquad (1)$$

$$f_k(t) = \big(\mu_k(t),\, n_k(t),\, d_k(t),\, u_m,\, [t - \ell, t),\, \pi,\, \textsc{measurement}\big).$$

Here $\alpha_m = 7$ for days per week, whose denominator is the number of calendar days in the window, and $\alpha_m = 1$ otherwise; u_m is the unit and π the source binding. The measurement set $F(t)$ also

contains 28-day acceptance rates in four hour-of-day bins. An undefined value $\perp$ is never replaced by zero, and a day without records does not count as a day of zero activity. Table 24 registers the event filter, statistic, claim key and scoring target of every quantity in every study.

Typed memory. The memory is a finite set $\mathcal{S}$ of statements. Each statement s has an identifier, a role $\rho(s) \in \{\text{long}, \text{recent}, \text{profile}, \text{decision}\}$, a text $x(s)$, numeric claims $\Gamma(s) \subseteq \{(k, c) : k \in \mathcal{M} \times \{7, 28, 60\},\ c \in \mathbb{R}\}$ and declared premises $\delta(s)$. A claim (k, c) says that $\mu_k(t_m) = c$: its window ends at the memory cutoff. The declared dependency graph is $G = (\mathcal{S}, E)$, where $(u, v) \in E$ iff v names u as a premise. Premises that name no statement are ignored. Cycles are allowed. Each memory item has evidence type MEMORY, meaning it was written at t_m and is unverified. Measurements computed at t_0 have type MEASUREMENT. As in agent memories that separate facts from opinions (Latimer et al., 2026), the type records how a statement was produced, not whether it is true. Assumption A1 of Appendix S.5 is that identifiers are unique within $\mathcal{S}$.

3.2 Conflict identification

A memory claim and a measurement are compared only when they share the definition k: metric, numerator population, denominator population, unit and window length. For metric m, let $\tau_m(v) = \max(a_m,\ r_m|v|)$, with $(a_m, r_m) = (2.0, 0.20)$ for completions per day, $(0.05, 0)$ for the two fractions and $(1.0, 0)$ for days per week. The *direct conflict set* is

$$D = \big\{ s \in \mathcal{S} :\ \exists (k, c) \in \Gamma(s) \text{ with } \mu_k(t_0) \neq \perp \text{ and } |c - \mu_k(t_0)| > \tau_m(\mu_k(t_0)) \big\}. \tag{2}$$

The claim describes a window ending at t_m; the measurement describes a window of the same length ending at t_0. A conflict therefore means that, under the stated definition, the claim does not describe the rider at the forecast cutoff. The discrepancy decomposes exactly into the writer's error at t_m and the drift between t_m and t_0, as Proposition 3 shows. Undefined measurements produce no conflict. Quantities with different definitions are never compared, so a refusal share of recorded responses cannot overwrite a refusal share of deliveries.

3.3 Dependency retraction

A statement should not remain available as if its premise still held once that premise has been withdrawn. The consistency operator applies dependency-directed retraction, as in truth maintenance (Doyle, 1979), to the declared graph:

$$R_0 = D, \qquad R_{j+1} = R_j \cup \{ v \in \mathcal{S} : \exists u \in R_j,\ (u, v) \in E \}, \qquad R = \textstyle\bigcup_{j \geq 0} R_j. \tag{3}$$

A retracted statement keeps its identifier and is shown as RETRACTED with a reason, without its text. The reason is the claimed and measured values for $s \in D$, and "depends on a retracted statement" for $s \in R \setminus D$. The *local* comparator retracts only D. Call $K \subseteq \mathcal{S}$ *admissible* if $K \cap D = \emptyset$ and K is dependency-closed, that is, $v \in K$ and $(u, v) \in E$ imply $u \in K$.

Proposition 1 (Dependency retraction)**.** *Under A1: (i) $R_j = R$ for all $j \geq |\mathcal{S} \setminus D|$, and R is the set of statements reachable in G from D. (ii) $\mathcal{S} \setminus R$ is admissible, and every admissible K satisfies $K \subseteq \mathcal{S} \setminus R$. The retained memory is therefore the unique largest admissible sub-memory. (iii) Every $v \in R$ has a declared path from a directly conflicting statement, so no statement outside such paths is retracted. (iv) Applying the operator to the retained memory $\mathcal{S} \setminus R$ retracts nothing. (v) R grows monotonically: enlarging D or E, or lowering any a_m or r_m, can only enlarge it. (vi) The local retained set $\mathcal{S} \setminus D$ is admissible iff $R = D$.*

Corollary 1. *Every claim (k, c) of a retained statement with $\mu_k(t_0) \neq \perp$ satisfies $|c - \mu_k(t_0)| \leq \tau_m(\mu_k(t_0))$.*

Part (ii) is where this differs from general belief-base contraction, in which several maximal subsets can survive (Alchourrón et al., 1985). Admissible sets are closed under union, so here the maximal one is unique and the result does not depend on processing order. By (vi), the local comparator leaves out no conflicting statement but, whenever $R \neq D$, keeps a conclusion whose premise was withdrawn. Comparing the two operators on the same memory therefore isolates dependency propagation from conflict detection. These guarantees cover declared premises and numeric claims with matched definitions. An undeclared dependency cannot be detected, and the text of a statement is not verified.

3.4 QUESTION-CONDITIONED SELECTION UNDER A BUDGET

For a character budget b, each statement costs its rendered length plus one. A relevance score counts its numeric claims, its matches with a fixed term list for the task and, for offer questions, whether it is a decision hypothesis, as defined in Eq. (7) of Appendix R. The selector first keeps the top-scoring statement of each role if it fits and then fills the budget greedily in score order; a prefix comparator keeps the longest document-order prefix that fits. Proposition 2 in Appendix R shows that the selection never exceeds the budget, leaves less room than any unselected statement would need, represents every role whenever the role representatives fit, and equals the full typed context once the budget covers the memory. Selection is not nested in b, so a quality–cost curve over budgets compares distinct contexts. The selector does not keep premises and, in Algorithm 1, selects from unretracted memory. Proposition 5 in Appendix T gives a premise-preserving selector applied after retraction and shows that the composed context is admissible, budget-feasible and closure-maximal.

4 STUDIES AND DATA

The longitudinal data comprise platform activity records, derived personal memories and evaluation outcomes. Table 3 maps each study to its representation, target and status. Collections may share riders, so their counts are neither added nor used as sample sizes. Quantities with different filters carry different names in the metric registry of Appendix N and are never pooled.

4.1 IDENTITY-CONTROLLED RESPONSE DATA

The identity study evaluates 316 stratified riders with at least 90 voluntary responses to system-dispatched offers, timeouts excluded, in a relative 28-day window. A 55-response prefix, four endpoints, thresholds and corrections were fixed in writing before any outcome was inspected. The personal context receives profile bins and rolling counts and rates; the population context receives fieldwise pre-window medians; wrong-identity and matched-neighbor controls cover a fixed 158-rider subset, with 52,140 calls to DeepSeek-V4-Flash at temperature zero. Contrasts average within rider and use seven-day block bootstraps with predeclared corrections, as Appendix B details.

4.2 AUDIT AND HOLDOUT COHORT

Two six-month samples of rider records supply riders at a common forecast cutoff t_0, with memory cutoff $t_m = t_0 - 14$ days. After excluding the 112-person cohort as a block, every rider with at least 10 recorded days in the 60 days before t_0, 3 in the last 14, 10 in the 60 days before t_m and 20 voluntary responses is eligible, using pre-cutoff information only. This yields 295 riders; 27 of them are outside every earlier cohort: they are not in the 316-rider identity roster or the excluded 112-person cohort, and an identity check finds no overlap with the 20-rider development cohort. For each rider, Qwen3.5-9B with greedy decoding and thinking disabled writes two four-role memories from the 60-day daily table before the respective cutoff: a lagged memory at t_m and a fresh memory at t_0. Every statement carries typed claims (m, ℓ, c) and declared premises, parsed deterministically as described in Appendix S.5. The audit compares claims with $F(t_0)$ and lagged claims also with $F(t_m)$, which splits each discrepancy into writer error and drift by Proposition 3; it uses no outcome.

Preregistered forecasting protocol. The same model forecasts each rider from a context compiled by Algorithm 1. Every condition receives the same source measurements $F(t_0)$ and the same lagged memory statements; the conditions of interest differ only in how that memory is shown: as plain text (P), as typed, unverified claims (T), or as typed claims after the consistency check with dependency retraction (TC). The registered design contains further memory conditions, listed in Appendix E. Each condition forecasts the four two-week aggregates and the acceptance probability of up to 24 future dispatched offers per rider, one per equal-count time stratum, given the recorded offer time and bundle type. The primary endpoints are the rider's mean clipped offer log loss and the absolute refusal-fraction error; the six tests of the primary contrasts T−P, TC−P and TC−T use a paired rider bootstrap with 10,000 resamples and Holm correction. While the forecasts were being generated and before any was scored, a written amendment fixed non-inferiority margins of 0.020 nats for log loss and 0.010 for refusal error, made the 27 holdout riders the only confirmatory group and added eight one-sided non-inferiority tests with Holm correction at one-sided 0.025. Of the 27 holdout riders, 22

Table 1: Numeric claims of LLM-written memories checked against the source measurement of the same metric and window at t_0 for 295 riders, without outcomes. Dependency-only retractions are statements removed by the closure of Eq. (3) without a conflict of their own.

Memory	Statements	Checked claims	In conflict	Riders w/ conflict	Dep.-only retracted
Fresh, at t_0	4,477	2,406	1,001 (41.6%)	274 / 295	569
Lagged, at t_m	4,363	2,308	983 (42.6%)	276 / 295	527

have future offers and refusal outcomes, so the primary endpoints use $n = 22$; results on the other 268 riders, who belong to cohorts used in the earlier studies, are descriptive. Appendix E gives the full protocol, and Table 27 the complete confirmatory family.

An outcome-free audit before scoring found that the typed rendering of recorded days per week omitted the factor $\alpha_m = 7$, a line absent from P. A second registered run on the same riders, memories and model, a sensitivity check rather than a replication, therefore repeats T and TC with the corrected line as T7 and TC7 and adds TCO, which removes the retracted statements silently; Appendix F describes its 4,130 calls.

Locked validation at new cutoffs. After these runs were scored, a fourth amendment, appended before any output of the new run existed, registered a validation at five new two-week windows with cutoffs 42, 28 and 14 days before and 14 and 28 days after t_0, disjoint from each other and from the registered window. Its 128 rider-cutoff units come from 46 second-sample riders outside every earlier cohort, namely 26 holdout riders and 20 never-exposed riders; 37 riders with 105 units have future offers. Each unit gets one lagged memory and forecasts under P, T7, TCO and TC7 with the registered code in 1,152 calls. The contrasts TC7−P and TC7−T7, the removal contrast TCO−T7 and the marker contrast TC7−TCO on both primary endpoints form Holm-corrected superiority and non-inferiority families of eight tests each, with the rider as unit; Appendix G gives the details.

5 Results

5.1 The simulator reproduces identity-specific behavior

Table 4 reports the identity contrasts with intervals under the predeclared corrections. Across 316 riders, conditioning on the rider's own history reduces log loss by 0.0892 $[-0.1132, -0.0532]$ relative to population context. On the 158-rider control subset, replacing the history with that of another rider lowers accuracy by 15.30 percentage points $[10.54, 20.83]$, and replacing it with that of the closest matched neighbor lowers it by 8.98 points $[5.09, 13.09]$. Generic rider knowledge and broad interpersonal similarity therefore do not account for the gain: the simulator uses information specific to the represented rider.

5.2 LLM-written memory misstates the rider's numbers

Table 1 applies the conflict check of Eq. (2) to every numeric claim in the 590 memories of the audit cohort. The fresh memory is written from the same 60-day table on which $F(t_0)$ is computed, so its conflicts contain no drift: 1,001 of its 2,406 checked claims, that is 41.6% with rider-cluster interval [39.1%, 44.1%], fall outside tolerance. The lagged memory, written two weeks earlier, has 983 conflicting claims out of 2,308, that is 42.6% with interval [39.7%, 45.5%]. By Proposition 3, 76.7% of the attributable lagged conflicts, with interval [73.0%, 80.3%], have a larger writer error than drift, $|e_w| > |e_d|$; this classifies magnitudes and is not a causal share of drift. As Appendix C shows, conflicts occur on every metric and at halved or doubled tolerances. These rates describe the audited writer configuration, Qwen3.5-9B with thinking disabled reading a 60-day daily table, and the claims it makes explicitly. An independent reader that uses neither the registered parser nor the stored measurements reproduces every parsed statement, all 6,974 defined source measurements and every retraction set of the 590 memories, so the check is exact on the explicit claims it covers; Appendix D shows that 27.4% of the lagged writer errors match the same metric in another window.

The writer declares premises freely: 1,150 of the 4,363 lagged statements name at least one. Closure retracts 527 statements beyond the 961 directly conflicting ones, in 247 riders, so local repair alone

Table 2: Checked memory TC versus plain memory P with identical measurements and memory statements. d is the error of TC minus that of P, so negative values favor TC; intervals are two-sided 95% paired rider bootstraps with $B = 10{,}000$ resamples, and Holm-adjusted p values cover the six primary tests of each run. $0/B$ means that no resample lies beyond the null value, so the estimate is limited by the resampling resolution; Table 42 gives the counts. Every row is non-inferior at the registered margins of 0.020 for log loss and 0.010 for refusal error, and superior. Main: registered run, confirmatory on the holdout riders. Corrected: TC7 versus P in the registered follow-up run.

		Offer log loss		Refusal-fraction error	
Riders	Run	d [95% CI]	Holm p	d [95% CI]	Holm p
Holdout, $n = 22$	main	-0.2196 $[-0.349, -0.099]$	0.0024	-0.0506 $[-0.102, -0.007]$	0.047
	corrected	-0.1767 $[-0.319, -0.052]$	0.015	-0.0488 $[-0.098, -0.009]$	0.028
All, $n = 282$	main	-0.1536 $[-0.197, -0.113]$	$0/B$	-0.0101 $[-0.017, -0.004]$	0.0018
	corrected	-0.1561 $[-0.200, -0.116]$	$0/B$	-0.0087 $[-0.016, -0.002]$	0.022

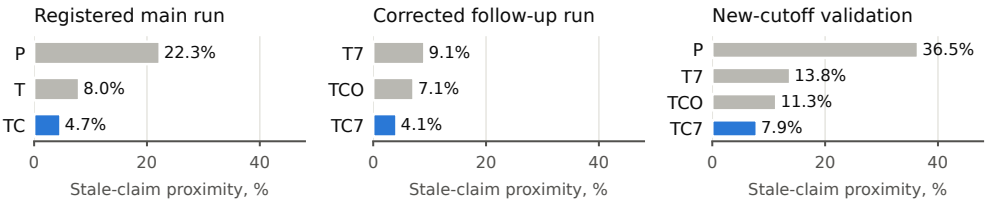


Figure 2: Stale-claim proximity: share of conflicting metrics on which the forecast lies closer to the stale claim than to the fresh measurement. Left: registered main run, 364 metrics of all riders. Middle: corrected follow-up run. Right: locked validation at five new cutoffs, 203 metrics.

would leave these conclusions in the context with a withdrawn premise, as Proposition 1(vi) implies. Answer graphs of 64 development riders in Appendix L show the same structural effect of closure.

5.3 Checked memory in the preregistered holdout forecast

Table 2 gives the confirmatory result for the full mechanism. On the 22 evaluable holdout riders, checked memory TC improves on plain memory P on both primary endpoints, and the registered non-inferiority test against P, followed in fixed sequence by superiority, holds for both. With the rider of largest gain left out, TC$-$P remains superior and non-inferior on log loss with Holm p 0.0060 and p_{NI} 0.0008, and non-inferior on refusal error with p_{NI} 0.0126, as Appendix E reports; Appendix K shows that 16 of the 22 riders gain more than the 0.020 margin on log loss. With the corrected days-per-week line, TC7 is non-inferior and superior to P on both endpoints. Table 27 in Appendix U lists all registered tests.

The check moves forecasts away from stale claims. For each rider and each metric on which the lagged memory conflicts with the fresh pre-cutoff measurement, we take the first eligible conflicting claim and record whether the forecast lies closer to that stale claim than to the measurement; we call this stale-claim proximity, registered as the adoption rate. This happens for 81/364 conflicting metrics under P, 29/364 under T and 17/364 under TC, that is for 22.3%, 8.0% and 4.7% of them, which meets the registered mechanism rule that TC lies below T; as Figure 2 shows, with corrected rendering the rate is 9.1% under T7 and 4.1% under TC7.

5.4 Transfer to new cutoffs and to new riders

New cutoffs. In the locked validation of Appendix G, TC7 is non-inferior to plain memory on offer log loss, with a difference of -0.1737 $[-0.290, -0.045]$ and Holm p_{NI} 0.017.

New riders. A fifth registered amendment adds 6,312 forecasts for 789 riders of a third sample from 124 regions whose data never entered any earlier prompt, memory, audit or score; checked memory

is shown with retraction markers as TC7, without the retracted statements as TCO, or with equal-length placeholders as TCN. In these registered descriptive contrasts, every rendering of checked memory lowers both primary errors relative to plain memory, with intervals excluding zero; for TC7 the differences are -0.1239 $[-0.1578, -0.0909]$ on log loss and -0.0072 $[-0.0125, -0.0021]$ on refusal error, and Table 13 lists all renderings. Region resampling and removal of the most favorable rider or region keep the log-loss intervals of all three renderings and the refusal-error intervals of TC7 and TCO below zero, as Table 15 shows.

Later cutoff and simple repair. A preregistered two-day forecast at a later cutoff includes 2,441 riders, 1,934 of them evaluable, as Appendix J details. Checked memory is non-inferior on refusal-fraction error to a simple numeric repair SR, which overwrites each conflicting claim value with the measurement and keeps the statement text, with $d = -0.0020$ $[-0.0041, 0.0000]$ and both Holm-adjusted $p < 0.001$; in a descriptive contrast it lowers offer log loss relative to unchecked typed memory by -0.0240 $[-0.0380, -0.0105]$. Two open-weight model judges blind to the check find more than 85% of the directly retracted statements but about 23% of the kept ones contradicted by the measurement, and the share of statements shown as memory that both judges find contradicted is 16.1% with the check against 26.2% without it.

5.5 An LLM forecast combined with a personal statistical predictor at scale

An eighth amendment, appended before any outcome was extracted, registers a forecast on 9,350 riders of four cities, each with four memory documents written from the latest 30 or 60 recorded days, as Appendix I describes. Of the 2,888 riders eligible under registered pre-cutoff rules, a fixed pseudo-random split puts 2,373 in the confirmatory set, and the 1,998 of them with a voluntary response in the seven-day window are evaluable, with 306,787 responses. The comparator S is a smoothed personal statistical predictor without an LLM, which shrinks the rider's refusal and time-slot rates from the same records toward the city rate, unlike the measurement-only LLM condition S of the holdout study. The LLM forecast M reads the memory documents verbatim, without the check of Section 3, together with a same-definition measurement pack, and STK combines M with S using weights fitted on the 416 evaluable development riders only. STK lowers the refusal-fraction error relative to S by -0.0055 $[-0.0080, -0.0029]$, about 9% of the error of S, 0.0621, and meets the registered superiority rule under both tests, with Holm-adjusted p of 6.4×10^{-5} for the bootstrap with 10^6 resamples and 8.9×10^{-5} for the paired t test. Table 17 lists all forecasters with their cost; because this cohort shows memory verbatim, it evaluates the forecast combination rather than the check.

6 Discussion and Conclusion

Scope. The simulator forecasts logged responses to dispatched offers and one- and two-week activity of regularly active riders; it does not estimate causal willingness or responses to new incentive schemes, and its outputs describe riders rather than allocate rewards or restrict any rider's opportunities. The results on checked memory refer to the tested writer and forecaster, Qwen3.5-9B with thinking disabled, and to the registered renderings of memory. The guarantees cover explicit claims bound to a registered metric definition and declared premises; dependency closure is a structural guarantee on the declared graph, and the correspondence between claims and statement text is assessed with model judges.

Conclusion. Conditioning an LLM on a rider's own history reproduces identity-specific behavior, but LLM-written memory misstates many of the rider's numbers. A simulator that must show such memory to its forecaster should therefore check it: binding memory to same-definition measurements and retracting what depends on a contradicted claim costs 0.013 ms per memory, against 81.5 s of model time to write one as Table 25 shows, and makes memory safer context. The forecast then follows the measurement instead of the stale claim; checked memory improves on plain memory in a preregistered holdout forecast and on never-exposed riders, and at a later cutoff it is non-inferior to a simple numeric repair while model judges find fewer contradicted statements in checked than in unchecked memory. On 9,350 riders, an LLM forecast combined with a smoothed personal statistical predictor improves on that predictor for the refusal fraction.

AI USE STATEMENT

In this work, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes existing activity records of delivery riders, namely dispatched offers, voluntary responses and daily activity counts. The authors use these records for this research under the authorization of the data owner, report them only in aggregate form and release no individual-level data. The study is retrospective: it collects no new data from riders and involves no intervention on them, and every forecast is only compared with recorded outcomes. The records remain under access control and are not released. The paper reports only aggregate statistics and generic categories; identifiers, contact details, precise locations and trajectories are omitted, and the supplementary material contains no rider-level data, as Appendix P describes. Simulating individual riders could be misused, for example to put pressure on riders or to make automated decisions about them; the forecasts studied here serve research on individual-level simulation, and any use of such forecasts should support planning under human oversight rather than decisions against individual riders. Because forecast quality can differ across riders, the paper reports rider-level distributions of the paired differences and results by city and by memory length in Appendices K and I.

REPRODUCIBILITY STATEMENT

Section 3, Algorithm 1 and Appendix S.5 specify the memory format, the consistency check with dependency retraction and the context compiler; Appendix S states the assumptions and gives complete proofs of the formal properties. Section 4 and Appendices B, E, G, I and J describe the samples, inclusion rules, models, decoding settings, endpoints and statistical tests, and Appendix U lists every registered test. Appendix P describes data access and the anonymized supplementary material, which contains the code, prompts, registered protocol text, scoring and statistical scripts, tests on synthetic inputs and aggregate result files; Table 25 reports the cost of each stage.

REFERENCES

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239. URL https://doi.org/10.2307/2274239.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81005–81034. PMLR, 2025. URL https://proceedings.mlr.press/v267/anthis25a.html.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2. URL https://doi.org/10.1017/pan.2023.2.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 656(8126):115–122, 2026. doi: 10.1038/s41586-026-10742-x. URL https://doi.org/10.1038/s41586-026-10742-x.

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic Semi-Structured memory with temporal reasoning for Long-Term conversational AI. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16470–16489. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.749. URL https://doi.org/10.18653/v1/2026.acl-long.749.

Ziyad Benomar, Weronika Łajewska, Leonardo Perelli, and Saab Mansour. Data-Driven Persona-Conditioned agents for A/B test simulation. arXiv preprint arXiv:2609.01038, 2026. URL https://arxiv.org/abs/2609.01038. Accepted to the EMNLP 2026 Industry Track.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644 (8078):1002–1009, 2025. doi: 10.1038/s41586-025-09215-4. URL https://doi.org/10.1038/s41586-025-09215-4.

M. Keith Chen, Judith A. Chevalier, Peter E. Rossi, and Emily Oehlsen. The value of flexible work: Evidence from Uber drivers. *Journal of Political Economy*, 127(6):2735–2794, 2019. doi: 10.1086/702171. URL https://doi.org/10.1086/702171.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. TabFact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, 2020.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building Production-Ready AI agents with scalable Long-Term memory. In *ECAI 2025 – 28th European Conference on Artificial Intelligence*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://doi.org/10.1162/tacl_a_00644.

Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986. doi: 10.1016/0004-3702(86)90080-9. URL https://doi.org/10.1016/0004-3702(86)90080-9.

Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979. doi: 10.1016/0004-3702(79)90008-0. URL https://doi.org/10.1016/0004-3702(79)90008-0.

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and Yiqun Liu. TwinVoice: A Multi-dimensional benchmark towards digital twins via LLM persona simulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19604–19628. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.981. URL https://doi.org/10.18653/v1/2026.findings-acl.981.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from Homo silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pp. 614–615, New Haven, CT, USA, 2024. Association for Computing Machinery. doi: 10.1145/3670865.3673513. URL https://doi.org/10.1145/3670865.3673513.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S^3: Social-network simulation system with large language Model-Empowered agents. arXiv preprint arXiv:2307.14984, 2023. URL https://arxiv.org/abs/2307.14984.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G. Wilson. Large language models are Zero-Shot time series forecasters. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pp. 19622–19635, 2023. doi: 10.52202/075280-0861. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3eb7ca52e8207697361b2c0fb3926511-Abstract-Conference.html.

Minghao Guo, Ziyi Ye, Wujiang Xu, Xi Zhu, Wenyue Hua, and Dimitris N. Metaxas. Individual Turing test: A case study of LLM-based simulation using longitudinal personal data. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3770–3775. ACM, 2026. doi: 10.1145/3805712.3809868. URL https://doi.org/10.1145/3805712.3809868.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. arXiv preprint arXiv:2602.15785, 2026. URL https://arxiv.org/abs/2602.15785.

Xiaochong Lan, Jie Feng, Jiahuan Lei, Xinlei Shi, and Yong Li. Benchmarking and advancing large language models for local life services. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 4566–4577. ACM, 2025. doi: 10.1145/3711896.3737196. URL https://doi.org/10.1145/3711896.3737196.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 275–285, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-demo.27. URL https://aclanthology.org/2026.acl-demo.27/.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. How far are LLMs from being our digital twins? a benchmark for Persona-Based behavior chain simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15738–15763. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.813. URL https://doi.org/10.18653/v1/2025.findings-acl.813.

Xuchuan Li, Fei Huang, Jianrong Lv, Zhixiong Xiao, Guolong Li, and Yang Yue. Be more real: Travel diary generation using LLM agents and individual profiles. arXiv preprint arXiv:2407.18932, 2024. URL https://arxiv.org/abs/2407.18932.

Zonghan Li and Feng Ji. Statistical realism is not evidence that LLMs can estimate treatment effects in social science experiments. arXiv preprint arXiv:2604.02458, 2026. URL https://arxiv.org/abs/2604.02458.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. arXiv preprint arXiv:2605.05583, 2026. URL https://arxiv.org/abs/2605.05583.

Tianming Liu, Jirong Yang, Yafeng Yin, Manzi Li, Linghao Wang, and Zheng Zhu. Aligning LLM agents with human learning and adjustment behavior: A dual agent approach. *Transportation Research Part C: Emerging Technologies*, 191:105818, 2026. doi: 10.1016/j.trc.2026.105818. URL https://doi.org/10.1016/j.trc.2026.105818.

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Yaochen Xie, Yisi Sang, Qi He, and Dakuo Wang. Can LLM agents simulate Multi-Turn human behavior? evidence from real online customer behavior data. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43961–43977. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2034. URL https://doi.org/10.18653/v1/2026.acl-long.2034.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://doi.org/10.18653/v1/2024.acl-long.747.

Baichuan Mo, Hanyong Xu, Ruoyun Ma, Jung-Hoon Cho, Dingyi Zhuang, Xiaotong Guo, and Jinhua Zhao. Large language models for travel behavior prediction. *Transportation Research Interdisciplinary Perspectives*, 38:102124, 2026. doi: 10.1016/j.trip.2026.102124. URL https://doi.org/10.1016/j.trip.2026.102124.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via Timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8631–8661. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.435. URL https://doi.org/10.18653/v1/2025.naacl-long.435.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. arXiv preprint arXiv:2310.08560, 2023. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM agents grounded in Self-Reports enable General-Purpose simulation of individuals. arXiv preprint arXiv:2411.10109, 2026. URL https://arxiv.org/abs/2411.10109. Version 3 (2026-06-28); earlier versions titled "Generative Agent Simulations of 1,000 People".

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. AgentSociety: Large-Scale simulation of LLM-Driven generative agents advances understanding of human behaviors and society. *iFuture*, 2026. doi: 10.26599/IF.2026.9710004. URL https://doi.org/10.26599/IF.2026.9710004.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. arXiv preprint arXiv:2501.13956, 2025. URL https://arxiv.org/abs/2501.13956.

Dale Robbennolt, Jake Robbennolt, Zhaomiao Guo, Stephen D. Boyles, and Chandra R. Bhat. Understanding delivery driver order acceptance: A dynamic discrete choice approach with Markov decision processes. *Transportation Research Part C: Emerging Technologies*, 190:105794, 2026. doi: 10.1016/j.trc.2026.105794. URL https://doi.org/10.1016/j.trc.2026.105794.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.399. URL https://doi.org/10.18653/v1/2024.acl-long.399.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025a. doi: 10.1038/s42256-025-00986-z. URL https://doi.org/10.1038/s42256-025-00986-z.

Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. Large language models as urban residents: An LLM agent

framework for personal mobility generation. In *Advances in Neural Information Processing Systems 37*, pp. 124547–124574. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3957. URL https://doi.org/10.52202/079017-3957.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language Model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025b. doi: 10.1145/3708985. URL https://doi.org/10.1145/3708985.

Shu Wang, Edwin Yu, Oscar Love, Tom Zhang, Tom Wong, Steve Scargall, and Charles Fan. MemMachine: A Ground-Truth-Preserving memory system for personalized AI agents. arXiv preprint arXiv:2604.04853, 2026a. URL https://arxiv.org/abs/2604.04853.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. Customer-R1: Personalized simulation of human behaviors via RL-based LLM agent in online shopping. arXiv preprint arXiv:2510.07230, 2025c. URL https://arxiv.org/abs/2510.07230.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. OPeRA: A dataset of observation, persona, rationale, and action for evaluating LLMs on human online shopping behavior simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43942–43960. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.2033. URL https://doi.org/10.18653/v1/2026.acl-long.2033.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10480–10496. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.586. URL https://doi.org/10.18653/v1/2024.emnlp-main.586.

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 (ICLR 2025)*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Zhuolin Wu, Li Wang, Fangsheng Huang, Linjun Zhou, Yu Song, Chengpeng Ye, Pengyu Nie, Hao Ren, Jinghua Hao, Renqing He, and Zhizhao Sun. A framework for Multi-stage bonus allocation in meal delivery platform. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4195–4203. ACM, 2022. doi: 10.1145/3534678.3539202. URL https://doi.org/10.1145/3534678.3539202.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://doi.org/10.18653/v1/2024.emnlp-main.486.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun

Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. OASIS: Open agent social interaction simulations with one million agents. arXiv preprint arXiv:2411.11581, 2024. URL https://arxiv.org/abs/2411.11581.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1807–1817. ACM, 2024. doi: 10.1145/3626772.3657844. URL https://doi.org/10.1145/3626772.3657844.

Chengbo Zhang and Zuopeng Xiao. Large language models as delivery rider: Generating instant food delivery riders' routing decision with LLM agent framework. arXiv preprint arXiv:2603.12559, 2026. URL https://arxiv.org/abs/2603.12559. Proceedings of the 3rd International Conference on Urban Science and Intelligence.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. SocioVerse: A world model for social simulation powered by LLM agents and a pool of 10 million Real-World users. arXiv preprint arXiv:2504.10157, 2025. URL https://arxiv.org/abs/2504.10157.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with Long-Term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via Multi-Hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://doi.org/10.18653/v1/2023.emnlp-main.971.

A STUDIES AT A GLANCE

Table 3 lists every study with its riders, input representation, target, number of model calls and status.

Table 3: Studies, representations and status. The audit and the holdout forecast share riders and memories; the development studies share 20 riders and one future window.

Study	Riders	Input representation	Target, calls and status
Identity fidelity	316	Profile bins and rolling counts	Logged offer responses; 52,140 calls; fixed prefix
Memory-claim audit	295	Typed memories written at t_m and t_0	Numeric claims vs. same-definition measurements; no outcomes
Holdout forecast	295	Measurements + plain, typed or checked memory	Offers and two-week aggregates; 5,900 calls; preregistered, confirmatory on 27 riders; follow-up run of 4,130 calls
Locked validation	46	Lagged memory as P, T7, TCO or TC7	Same targets at five new cutoffs; 1,152 calls; preregistered, confirmatory
Never-exposed riders	789	Lagged memory as P, TC7, TCO or TCN	Same primary endpoints; 6,312 calls; preregistered; checked vs. plain memory descriptive
Large-cohort forecast	2,888	Unchecked memory documents shown verbatim + measurements; statistical predictor	Seven-day offers and refusal fraction; 5,776 calls; preregistered, confirmatory on 1,998 riders
Later cutoff	2,441	Lagged memory as TC7, SR, TL, T7 or none; two readout orders	Two-day offers and refusal fraction; 19,528 calls; preregistered, four-test family on 1,934 riders; model-judge check of 200 memories
Development forecast	20	Free-text memories $\pm$ statistics	Event-based aggregates; descriptive; Appendix M

What each study supports. The identity study supports only the claim that the simulator uses information that belongs to the represented rider. The audit supports the claim that the tested writer misstates numbers and that memory therefore has to be checked; it uses no outcome. The holdout forecast, which compares checked with plain memory under identical measurements and statements, is the confirmatory test of the complete mechanism. The locked validation and the never-exposed riders test transfer of the same mechanism to new cutoffs and new riders. The large-cohort forecast shows memory documents verbatim, without the check, and compares an LLM forecast combined with a statistical predictor against that predictor; it is not a test of the check. The later-cutoff study compares checked memory with a simple numeric repair and with unchecked memory, and its model-judge check reads the memories without seeing the check. Propositions 1, 2 and 5 are structural guarantees on the declared graph and the budget, not forecasting results.

B IDENTITY-CONTROLLED RESULTS

Sampling and admissible information. The 316 riders come from an eligible 490-person frame after a separate 112-person cohort was excluded as a whole, stratified by completions in the preceding 28 days into strata of 96, 110 and 110 riders. Personal and population contexts each have 316×55 planned calls and the two identity controls 158×55 each, giving $2(316 \times 55) + 2(158 \times 55) = 52{,}140$ call records, of which 17 returned no response. Contrasts pair the two arms on the planned evaluation point, identified by rider and point index, and use 10,000 bootstrap replicates. The sampling frame is partitioned by preceding-window completion tertiles. A fixed random ordering targets 110 riders per stratum, skips people with fewer than 90 eligible responses, and retains all attainable eligible selections; the lowest-activity stratum has a 14-person shortfall. The 90-point construction samples one response from each equal-count block of the time-ordered eligible list, so the analyzed 55-point prefix need not span the complete 28-day window. The prefix, its four endpoints, thresholds and corrections were fixed in a written interim protocol after the 90-point run had started and before any outcome was inspected; the run itself continued unchanged. The binary label denotes voluntary acceptance versus refusal of a system-dispatched offer; delivery completion is not its definition.

The personal model receives pre-window bins for activity, history span and daily completions; rolling 7- and 28-day counts, response rates, active days and elapsed time since the last event; and the current hour, weekday, weekend flag and dispatch-source indicator. It does not receive destination, price or an unobserved opportunity model. The population control uses fieldwise medians of profiles and pre-window summaries, fixed thereafter, so that contrast changes freshness as well as identity. The wrong-person control is a fixed derangement. The neighbor control minimizes squared distance over standardized pre-window acceptance rate, log response frequency, history span and completions per day, excluding self. Both preserve the target rider's evaluation point and context and truncate donor history at the earlier of that point and the donor's last event. A fixed shuffled-code half selects the 158-person control subset.

Replay and decoding. Probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ for log loss. Hours use local time. Replay makes each record usable half a second after its event time. Decoding uses temperature zero, top-p one and at most 512 output tokens. Descriptive denominators differ across arms because of calls without a response and the control subset; paired inference uses the contrast-specific intersections.

The fourth predeclared endpoint of the fixed-prefix protocol compares personal context with a non-LLM shrinkage predictor under a non-inferiority margin; it is listed in Table 32 of Appendix U.

Table 4: Identity study, fixed-prefix contrasts reported in Section 5. Negative log-loss and positive accuracy differences favor personal context. Intervals use the predeclared two- or three-comparison correction.

Personal minus comparator	Riders	Difference	Corrected interval
Population, log loss	316	-0.0892	$[-0.1132, -0.0532]$
Wrong identity, accuracy, pp	158	$+15.30$	$[10.54, 20.83]$
Matched neighbor, accuracy, pp	158	$+8.98$	$[5.09, 13.09]$

Table 5 gives the descriptive metrics of the four model arms.

Table 5: Descriptive metrics of the four model arms of the identity study.

Arm	Points	Riders	Log loss	Accuracy
Personal context	17,373	316	0.5373	0.7273
Population context	17,378	316	0.6270	0.7010
Wrong identity	8,685	158	0.8531	0.5774
Matched-neighbor identity	8,687	158	0.7195	0.6400

Table 6: Calls with and without a response per arm; a missing response is not a rejection prediction.

Model arm	Response	No response	Planned
Personal context	17,373	7	17,380
Population context	17,378	2	17,380
Matched-neighbor identity	8,687	3	8,690
Wrong identity	8,685	5	8,690

Contrasts pair the two arms on the planned evaluation point, that is, on rider and point index. Of the 17,380 planned points of each full arm, 57 in 54 groups share their response time with another planned point of the same rider, and 26 of the 8,690 points of each control arm do, in 25 groups; labels agree within every group. On the complete records, index pairing keeps 17,372 points for personal versus population, since one point has no response in both arms, and 8,679 and 8,681 for the wrong-identity and neighbor controls, whereas pairing on rider and response time would merge the tied points and leave 17,315, 8,653 and 8,655. The contrasts of Table 4 use index pairing on the response records available for re-analysis, which cover 17,172 planned points of each full arm and

8,641 and 8,640 of the two control arms; every pair has equal response times and labels. As Table 7 shows, fourteen-day blocks and rider-only resampling retain the three identity effects.

Table 7: Counting flow of the identity contrasts under point-index pairing on the records available for re-analysis, with the 14-day-block and rider-only intervals. One arm only counts the points with a response in only the first or only the second arm.

Personal minus	Planned	One arm only	Pairs	14-day blocks	Rider only
Population, log loss	17,380	1 / 6	17,164	$[-0.1125, -0.0561]$	$[-0.1111, -0.0688]$
Wrong identity, acc. in pp	8,690	5 / 7	8,629	$[10.64, 20.57]$	$[10.73, 20.07]$
Matched neighbor, acc. in pp	8,690	3 / 6	8,631	$[5.17, 12.93]$	$[5.64, 12.76]$

C MEMORY-CLAIM AUDIT DETAILS

The audit parses every memory with the deterministic parser of Appendix S.5. All 295 lagged and 295 fresh memories parse into typed statements, and none needs identifier repair. A claim is checked when its metric and window match a defined measurement; 19 lagged and 9 fresh claims had no defined measurement and are not counted. For each lagged conflict the measurement of the same definition at t_m is also computed, and the conflict is attributed, following Proposition 3, to writer error if $|e_w| > |e_d|$ and to drift otherwise; 10 conflicts have an undefined measurement at t_m and are not attributed. In 329 attributed conflicts the drift alone exceeds the tolerance. No conflict violates Proposition 3(i). Intervals below resample riders as clusters with 10,000 draws. The writer-dominated share is 76.7% $[73.0\%, 80.3\%]$ on all riders and 65.5% on holdout riders. Table 9 gives conflict rates and the writer-dominated share for each metric, Table 8 the conflict shares at halved and doubled tolerances, and Table 10 the audit by rider group.

Table 8: Tolerance sensitivity: share of checked claims in conflict when every tolerance τ_m is halved, kept or doubled.

Memory	$\tau_m \times 0.5$	τ_m	$\tau_m \times 2$
Fresh, all riders	57.5%	41.6%	22.7%
Lagged, all riders	60.8%	42.6%	26.2%
Fresh, holdout riders	71.8%	52.1%	28.6%
Lagged, holdout riders	71.7%	55.7%	37.7%

Table 9: Claims in conflict by metric on all riders and share of lagged conflicts dominated by writer error.

Metric	Fresh memory	Lagged memory	Writer-dominated, lagged
Completions per recorded day	67/642 (10.4%)	113/617 (18.3%)	38/108 (35.2%)
Refusal fraction	353/631 (55.9%)	336/598 (56.2%)	271/334 (81.1%)
Evening share of completions	357/581 (61.4%)	335/564 (59.4%)	286/332 (86.1%)
Recorded days per week	224/552 (40.6%)	199/529 (37.6%)	151/199 (75.9%)

Table 10: Audit by rider group. Holdout riders are outside every earlier cohort.

Group	Riders	Fresh in conflict	Lagged in conflict	Writer-dominated
All riders	295	1,001/2,406 (41.6%)	983/2,308 (42.6%)	746/973 (76.7%)
Riders used in earlier studies	268	890/2,193 (40.6%)	865/2,096 (41.3%)	670/857 (78.2%)
Holdout riders	27	111/213 (52.1%)	118/212 (55.7%)	76/116 (65.5%)

D INDEPENDENT IMPLEMENTATION OF THE MEMORY CHECK

A tenth amendment, appended before the analysis ran, registers a second implementation of the parser, the source measurements, the check and the closure, run on the 590 audit memories; the registered implementation is imported only as the reference. The second implementation reads the same structured claims and declared premises as the registered one, so it verifies the implementation, not the semantics of the claims or of the premises. It uses the raw memory outputs, the daily tables that the writer saw and the registered measurements, and no outcome. The reader parses each raw output with its own parser (A), recomputes every source measurement from the daily table (B) and repeats the check and the closure with its own parse and measurements (F). For each structured claim it then reads the statement text with fixed lexicons (C): whether the claimed value appears in the text within $\max(0.005,\ 0.01|v|)$, with percentages divided by 100, whether the text names the claimed metric or only another metric, whether a named window is the claimed one, and which numeric mentions of the text, excluding clock times and window lengths, are carried by a claim of the same statement. Claims outside tolerance of the measurement at the writer's own cutoff are compared with alternative definitions from the same table (D): the same metric in another window, another denominator or unit, namely completions per calendar day, the count or share of recorded days, refusals per acceptance, evening completions per recorded day or the percentage scale, and the complement. Declared premises are examined for resolvable targets, cycles and shared metrics (E). Table 11 gives every quantity for the lagged and the fresh memories.

Table 11: Independent implementation of the memory check on the 590 audit memories of 295 riders, without outcomes. Letters refer to the steps described above. Counts are statements, claims, measurements or declared premises as named in each row.

	Quantity	Lagged memory	Fresh memory
A	Memories parsed identically by the independent reader	295/295	295/295
B	Defined source measurements recomputed from the daily table and equal to the registered value	3,478/3,478	3,496/3,496
F	Memories with identical direct and retracted sets under the independent reader	295/295	295/295
C	Structured claims: value stated in the statement text	1,200/2,327 (51.6%)	1,180/2,415 (48.9%)
C	Structured claims: claimed metric named in the text	2,080/2,327 (89.4%)	2,139/2,415 (88.6%)
C	Structured claims: only another metric named	50/2,327 (2.1%)	51/2,415 (2.1%)
C	Claims whose text names a window: the claimed window	645/812 (79.4%)	603/808 (74.6%)
C	Numeric mentions in statement text carried by a claim	1,216/3,239 (37.5%)	1,201/3,199 (37.5%)
D	Claims outside tolerance at the writer's own cutoff	891/2,313 (38.5%)	1,001/2,406 (41.6%)
D	matching the same metric in another window	244/891 (27.4%)	266/1,001 (26.6%)
D	matching another denominator or unit	71/891 (8.0%)	41/1,001 (4.1%)
D	matching the complement	50/891 (5.6%)	36/1,001 (3.6%)
E	Declared premises that name an existing statement	2,216/2,216	2,340/2,340
E	Decision statements with at least one declared premise	1,138/1,138	1,157/1,157
E	Declared premises linking statements that share a metric	1,368/2,216 (61.7%)	1,483/2,340 (63.4%)
E	Decision statements naming a metric claimed elsewhere: at least one declared carrier for every named metric	238/993 (24.0%)	274/1,004 (27.3%)
E	some declared	392/993 (39.5%)	420/1,004 (41.8%)
F	Retracted statements when only claims with value and metric in the text are checked	628 of 1,488 (42.2%)	661 of 1,553 (42.6%)

The independent reading reproduces the registered parse, all defined source measurements and every direct and retracted set exactly, so the checker's decisions follow from the structured claims and the source table alone. The check covers the explicit structured claims that the writer attaches to its statements: 89.4% of the lagged claims name their metric in the text, and 37.5% of the numeric mentions in statement text are carried by a claim. Among the lagged claims that disagree with the measurement at the writer's own cutoff, 27.4% match the same metric in another window and 8.0% another denominator or unit, so part of what the check flags is a binding of a correct number to the wrong definition. Checking only claims whose value and metric are both stated in the text retracts 42.2% of the statements that the registered check retracts, all of them within the registered retraction

set. The lexical steps C–E are descriptive: a lexical mismatch between claim and text is not a semantic error, and a shared metric does not show that a declared premise is needed.

E PREREGISTERED HOLDOUT FORECAST: PROTOCOL

Registered protocol. The design, arms, endpoints and decision rules were written down before any forecast was generated. Three amendments followed. The first, a parser fix that assigns unique statement identifiers, came before any forecast job was built. The second, which fixed the non-inferiority margins, the confirmatory group, the eight-test family, the joint reading of TC$-$T and the metric registry, and the third, which recorded the measured extent of two implementation limits and added the follow-up run of Appendix F, came while the forecasts were being generated and before any forecast was scored or any outcome viewed. The log-loss margin lies below one quarter of the 0.0892 identity gain; as an illustration only, for a calibrated Bernoulli forecast with acceptance probability 0.60–0.85, lowering the predicted probability by 0.08–0.10 raises the expected log loss by 0.013–0.030 nats, and the excess depends on the direction of the shift: from 0.8 it is 0.026 when the probability is lowered to 0.7 and 0.044 when it is raised to 0.9; the refusal margin is one fifth of the checker's refusal tolerance.

Conditions. Offer forecasts receive the offer time and bundle type recorded on the rider's response event. All ten registered conditions receive the measurements $F(t_0)$ and differ only in memory: none (S); plain (P), typed (T) and typed with retraction (TC) from the lagged memory; typed fresh memory (TF); another rider's typed memory (W); question-conditioned selection from T at $b \in \{1000, 2000, 4000\}$ characters (B_b); and a 2,000-character document-order prefix of T (PRE_{2000}), giving $295 \times 10 \times 2$ calls. The budgeted conditions select from T without retraction, and their selector does not keep premises; the premise-preserving composition of Appendix T is part of the follow-up run.

Groups and execution. The 268 exposed riders are the 209 eligible riders of the first sample and the 59 eligible riders of the second sample who were in the identity roster; the 27 holdout riders are the remaining eligible riders of the second sample, with no overlap with the identity roster, the excluded 112-person cohort or the development cohort. The riders have 6,620 sampled future offers in total. Qwen3.5-9B in bf16 with greedy decoding, thinking disabled and one seed writes the memories and forecasts; all 590 memories parsed without identifier repair. The registered fallback replaced 13 of the 2,950 offer forecasts, namely 3 under S, 4 under TF, 3 under W, 1 under P, 1 under B_{1000} and 1 under PRE_{2000}, and no aggregate forecast. TC retracts 1,488 statements in 276 riders, 961 of them directly and 527 through dependency only; mean memory characters are 1,881 under P, 3,166 under T and 2,621 under TC. Every call of the conditions in the confirmatory family was generated and parsed on the holdout riders, so a complete-case analysis gives identical numbers.

Confirmatory family. The registered confirmatory family on the 22 evaluable holdout riders consists of the six primary superiority tests and eight non-inferiority tests. The registered verdict rules are: non-inferior if the Holm-adjusted one-sided $p_{\mathrm{NI}} < 0.025$; inferior if the lower 95% bound exceeds the margin; inconclusive otherwise. The registered paired-t sensitivity, with p values p_t, uses the same eight-test Holm correction. Table 27 in Appendix U lists every test of the family. The aggregate score files of the supplement contain every registered analysis on all groups, including the block-Holm non-inferiority adjustment of the follow-up mechanism block.

Descriptive readings on all riders. On all 295 riders the registered rule "typing helps", which requires T$-$P below zero with Holm $p < 0.05$, holds on both primary endpoints with T$-$P of -0.0824 $[-0.124, -0.040]$ and -0.0095 $[-0.017, -0.002]$; TC$-$P is -0.1536 $[-0.197, -0.113]$ and -0.0101 $[-0.017, -0.004]$, and "consistency helps" holds for log loss with TC$-$T of -0.0712 $[-0.117, -0.033]$ and Holm $p\,0/B$. The mechanism rule holds: the forecast lies closer to the stale claim on 17/364 conflicting metrics under TC, 29/364 under T and 81/364 under P, that is, on 4.7%, 8.0% and 22.3% of them. On the secondary completion-based evening error, TC versus P is non-inferior and superior on all riders, with a difference of -0.0096 $[-0.017, -0.003]$. On the 260 evaluable exposed riders, disjoint from the holdout group, TC is non-inferior and superior to P on both primary endpoints under the registered rule, with differences of -0.1480 $[-0.194, -0.105]$ and

-0.0066 $[-0.013, -0.001]$. On the holdout riders, typing alone lowers refusal error, with T$-$P of -0.0382 $[-0.082, -0.007]$ and Holm p 0.0088.

Rider-level distribution and influence. A ninth amendment, appended before these analyses ran, registers the rider-level distribution of TC$-$P and the influence of single riders. On the 22 evaluable holdout riders, TC$-$P on log loss is negative for 18 riders, with median -0.0858 and quartiles -0.4334 and -0.0245, and leaving out any one rider keeps the mean between -0.2492 and -0.1875. With the most favorable rider removed, leaving 21 riders, TC$-$P on log loss is -0.1875 $[-0.3026, -0.0738]$ and remains superior with Holm p 0.0060 over the six tests and non-inferior with Holm p_{NI} 0.0008 over the eight tests. On refusal-fraction error, TC$-$P is negative for 12 riders and zero for 5, with median -0.0035; leave-one-out means range from -0.0564 to -0.0347, and non-inferiority holds with the most favorable rider removed, with Holm p_{NI} 0.0126. In a descriptive analysis of all 282 evaluable riders, both intervals stay below zero with the most favorable rider removed, at -0.1445 $[-0.1834, -0.1079]$ for log loss and -0.0087 $[-0.0156, -0.0026]$ for refusal error. This cohort carries no region attribute; region clustering is reported for the never-exposed riders in Appendix H.

F Registered follow-up run

The third amendment registers a follow-up run on the same 295 riders, lagged memories, measurements, model and decoding, with the corrected days-per-week rendering $7 \times n_k/d_k = \mu_k$ in every condition. All of its readings are supplementary: they are outside the confirmatory family and cannot change a verdict of Table 27.

Conditions. The run makes $7 \times 295 \times 2 = 4{,}130$ calls. T7 and TC7 equal T and TC except for the corrected measurement line. BC_{2000} applies the premise-preserving selector of Appendix T to T7 at 2,000 characters, and TCB_{2000} applies it after TC retraction, omitting retracted statements. TL retracts only the direct conflicts D and is the local comparator of Proposition 1(vi). TCO removes the statements of R without markers. TDM removes $|R|$ statements drawn at random from the unflagged set $\mathcal{S} \setminus R$ only, matched in characters over 1,000 seeded draws; for 70 riders $\mathcal{S} \setminus R$ has fewer than $|R|$ statements and all of it is removed. TDM is thus a matched removal of unflagged statements, not generic random deletion. Outcome-free build checks confirmed that T7 and TC7 prompts differ from T and TC only in the days-per-week lines, that TCB_{2000} and TCO never show a retracted statement as memory, and that TDM removes the registered number of statements. All 4,130 calls were generated; the registered fallback replaced 7 offer forecasts.

Rendering. Under the registered rule, a main-run conclusion is robust to the rendering defect if the recomputed contrast keeps its sign and decision. This holds for the confirmatory TC versus P result on both endpoints: in Table 2, TC7$-$P is -0.1767 $[-0.319, -0.052]$ and -0.0488 $[-0.098, -0.009]$, with Holm p 0.015 and 0.028 over the six recomputed tests, non-inferior and superior. It also holds for the refusal-error gain of typing, with T7$-$P of -0.0526 $[-0.102, -0.012]$ and Holm p 0.013. Across all riders, T7$-$P is -0.0745 $[-0.117, -0.031]$ and -0.0095 $[-0.017, -0.001]$, and TC7$-$T7 is -0.0816 $[-0.129, -0.041]$ on log loss. In a descriptive analysis of the 260 evaluable exposed riders, TC7 is non-inferior to P on both primary endpoints and lowers log loss by -0.1544 $[-0.201, -0.112]$.

Proximity and composition. Across all riders, the forecast lies closer to the stale claim on 33/364 conflicting metrics under T7, 26/364 under TCO and 15/364 under TC7, that is, on 9.1%, 7.1% and 4.1% of them. Retraction before premise-preserving selection is non-inferior to the same selection without retraction on both primary endpoints across all riders, with $\mathrm{TCB}_{2000}-\mathrm{BC}_{2000}$ of -0.0356 $[-0.082, 0.005]$ and -0.0009 $[-0.005, 0.003]$ and block-Holm p_{NI} 0.0074 and $0/B$, so adding retraction before selection as in Proposition 5 keeps the accuracy of premise-preserving selection within the registered margins. Mean memory lengths are 3,166 characters under T7, 2,621 under TC7 and 2,067 under TCO.

Retraction and removal conditions side by side. Table 12 lists the mean errors of the conditions that differ only in how the checked statements are handled: T7, typed without the check; TC7, closure

retraction with markers; TL, direct-only retraction; TCO, silent omission of the closure set; and TDM, matched removal of unflagged statements, next to plain memory. The registered comparisons among them are listed in Table 34 of Appendix U.

Table 12: Follow-up run: mean errors of the retraction and removal conditions. Lower is better; every typed condition uses the corrected rendering.

	Holdout, $n = 22$		All, $n = 282$	
Condition	Log loss	Refusal error	Log loss	Refusal error
P, plain	0.8922	0.1454	0.7530	0.0819
T7, typed	0.8254	0.0928	0.6784	0.0723
TC7, closure with markers	0.7155	0.0966	0.5968	0.0731
TL, direct only with markers	0.7519	0.0848	0.6163	0.0711
TCO, closure omitted	0.7256	0.1087	0.6329	0.0723
TDM, matched removal	0.7002	0.0953	0.6446	0.0743

The aggregate follow-up score files in the supplement contain every registered reading of this run, with the registered block-Holm non-inferiority adjustment of the mechanism block computed from the stored probabilities.

G LOCKED VALIDATION AT NEW CUTOFFS: PROTOCOL

Locking. The fourth amendment was appended after the main and follow-up runs had been scored and before any memory or forecast of the validation run was generated. At that time only a count summary of the new inputs, with units, riders and offers per cutoff and no outcome values, had been read, and no outcome of the new run had been viewed. The amendment fixes riders, cutoffs, conditions, endpoints, both test families, the decision rules and the sensitivity analyses. It uses the registered code throughout: the inputs come from the registered input construction, imported unchanged; on the inputs of the registered runs, the job builder reproduces all 2,360 prompts of P, T7, TCO and TC7 exactly; the model is the registered snapshot; decoding is greedy with thinking disabled.

Riders, cutoffs and units. The candidates are the 78 riders of the second sample who are outside every earlier cohort: they are not in the identity roster or the excluded 112-person cohort and do not overlap with the development cohort; all 27 holdout riders are among them. The cutoffs lie 42, 28 and 14 days before and 14 and 28 days after the registered t_0, each with horizon 14 days, so the five windows are disjoint from each other and from the registered window. Excluded before any output were the registered cutoff itself, the cutoff 56 days before it, whose window contains a major public holiday and whose 60-day memory table would extend before the start of the records, and a cutoff 35 days after it, whose window overlaps the last one; no later cutoff fits in the observation period. A rider-cutoff unit is eligible by the registered pre-cutoff rules of Section 4.2. This gives 128 units, namely 22, 28, 28, 26 and 24 per cutoff, from 46 riders: 26 holdout riders with 91 units and 20 riders with 37 units who were not eligible at the registered cutoff and whose data never entered any prompt, memory, audit or score of the earlier studies. 105 units of 37 riders, 11 of them new riders, have at least one voluntary response in their window; the units have 2,240 sampled future offers. For the holdout riders, days of the three earlier windows were part of the 60-day history shown to the model in the registered run, as inputs and never as scored outcomes.

Conditions and execution. Qwen3.5-9B writes one lagged memory per unit at $t_m = t_0 - 14$ with the registered prompt in 128 calls, of which 126 parsed and 2 used the registered text fallback. P, T7, TCO and TC7 are defined as in Appendix F, all with corrected rendering, and each forecasts the two-week aggregates and up to 24 offers per unit; all 1,024 forecasts were generated, and the registered fallback replaced 1 offer forecast of P. TC7 retracts 859 statements, 547 direct conflicts and 312 dependency-only retractions, in 123 of the 128 units; mean memory lengths are 1,781 characters under P, 3,021 under T7, 1,558 under TCO and 2,299 under TC7. 127 units have every call generated and parsed in all four conditions.

Analysis and decision rules. Endpoints, clipping and fallback are those of the registered scorer. The rider is the unit of analysis: for a contrast, a rider's value is the mean, over the rider's units where both conditions have the endpoint, of the unit difference, and intervals come from a paired rider bootstrap with 10,000 resamples and a fixed seed. A condition is superior on an endpoint if the Holm-adjusted two-sided p over the eight superiority tests is below 0.05 and $d < 0$, and non-inferior if the Holm-adjusted one-sided p_{NI} over the eight non-inferiority tests is below 0.025; inferior requires the lower 95% bound to exceed the margin; otherwise the verdict is inconclusive. An effect of TC7 over T7 may be attributed to a component only through the removal contrast TCO−T7 or the marker contrast TC7−TCO. Table 28 in Appendix U lists every test of both families, together with the registered paired-t sensitivity analysis; the complete-case analysis, with 104 units for TC7−P, changes no verdict.

Mechanism reading. With the registered definition applied to the 203 conflicting metrics of the TC7 context, the forecast lies closer to the stale claim on 74/203 metrics under P, 28/203 under T7, 23/203 under TCO and 16/203 under TC7, that is, on 36.5%, 13.8%, 11.3% and 7.9% of them; TC7 retracts statements in 123 of the 128 units. The aggregate score file of the validation run in the supplement contains every registered reading, including the secondary endpoints, the period, cutoff and rider-group breakdowns of all eight contrasts, and the token and time use per condition.

H FORECAST FOR NEVER-EXPOSED RIDERS

Riders and units. The third sample contains 1,973 riders with records over 73 consecutive days. The 11 riders who appear in any earlier sample are excluded, which leaves 1,962 riders whose data never entered any prompt, memory, audit or score of the earlier studies. The registered pre-cutoff eligibility rules, applied to 45-day histories, admit 861 riders; 72 of them have no voluntary response in the forecast window, so neither primary endpoint is defined, and 789 riders from 124 regions remain, one unit each. The cutoff is common to all riders; the memory is written 14 days earlier from a 45-day table by the registered prompt, whose window names are the only textual change.

Table 13: Descriptive results on the 789 new riders: mean error per condition and d, the error minus that of plain memory P, with 95% paired rider-bootstrap intervals; proximity over 841 conflicting metrics.

	Offer log loss		Refusal-fraction error		
Condition	Mean	d [95% CI]	Mean	d [95% CI]	Proximity
P	0.6921	–	0.1002	–	28.7%
TC7	0.5682	-0.1239 $[-0.1578, -0.0909]$	0.0930	-0.0072 $[-0.0125, -0.0021]$	8.0%
TCO	0.5709	-0.1212 $[-0.1537, -0.0902]$	0.0943	-0.0059 $[-0.0107, -0.0011]$	11.5%
TCN	0.5604	-0.1318 $[-0.1641, -0.1005]$	0.0947	-0.0055 $[-0.0105, -0.0006]$	12.5%

Counting flow. Table 14 traces every count from planned calls to the scored contrasts. Each contrast pairs the unit-level errors of two conditions by the rider code; with one unit per rider, the rider is both the pairing key and the resampling unit. Offer log loss averages the clipped loss over the unit's sampled offers, and the refusal-fraction error is the absolute error of the predicted two-week refusal fraction. Intervals come from 10,000 paired rider-bootstrap resamples with a fixed seed; the paired-t interval, a registered sensitivity analysis, gives the same signs and excludes zero for all six contrasts with P. For TC7 versus P, log loss is lower for 456 riders and higher for 271.

Table 14: Counting flow of the never-exposed-rider forecast.

Stage	Count
Riders in the third sample	1,973
Excluded as present in an earlier sample	11
Eligible before the cutoff	861
Eligible without a response in the window	72
Units built, one per rider	789
Memories written	789
Forecast calls planned / generated	6,312 / 6,312
Offer forecasts replaced by the registered fallback under P, TC7, TCO, TCN	2, 5, 4, 5
Sampled future offers	18,300
Riders in every contrast with P	789
Memories with a flagged conflict	723
Retracted statements, direct / dependency-only	2,565 / 1,445
Conflicting metrics scored for stale-claim proximity	841

Secondary endpoints. The Brier score is 0.2124 under P and 0.1787, 0.1799 and 0.1771 under TC7, TCO and TCN; the error of the simulated acceptance rate is 0.1520 under P and 0.1257, 0.1271 and 0.1238. Mean prompt lengths for the offer forecasts are 4,573 characters under P, 7,700 under TC7 and TCN and 7,142 under TCO. Table 29 in Appendix U lists the registered confirmatory family of this study.

Distribution, influence and region clustering. Under the ninth amendment, the rider differences of the three renderings against P were also examined for skew, single-rider influence and regional clustering. The differences are skewed: for TC7−P on log loss the median is -0.0070, and 76 riders improve by more than 0.5 nats. The 789 riders come from 124 regions, with a median of 2 riders per region; the largest region holds 6.1%. Table 15 gives region-cluster bootstrap intervals from 10,000 region resamples, cluster-robust t tests with 123 degrees of freedom, and intervals after removing the most favorable rider or the most favorable region. For all three renderings on log loss and for TC7 and TCO on refusal-fraction error, every interval lies below zero.

Table 15: Never-exposed riders, 789 riders in 124 regions: robustness of the contrasts with P that meet the registered robustness criterion. Region bootstrap: 95% interval from resampling regions; CR1: two-sided p of the cluster-robust t test; removal: d and 95% interval without the most favorable rider or region.

Contrast	Endpoint	Region bootstrap	CR1 p	Rider removed	Region removed
TC7 − P	log loss	$[-0.1551, -0.0930]$	1.4×10^{-12}	-0.1194 $[-0.1522, -0.0873]$	-0.1190 $[-0.1497, -0.0891]$
TCO − P	log loss	$[-0.1509, -0.0927]$	1.2×10^{-13}	-0.1167 $[-0.1481, -0.0864]$	-0.1164 $[-0.1439, -0.0892]$
TCN − P	log loss	$[-0.1636, -0.1016]$	4.3×10^{-14}	-0.1273 $[-0.1585, -0.0976]$	-0.1256 $[-0.1548, -0.0975]$
TC7 − P	refusal error	$[-0.0131, -0.0020]$	0.011	-0.0066 $[-0.0117, -0.0015]$	-0.0062 $[-0.0118, -0.0012]$
TCO − P	refusal error	$[-0.0112, -0.0014]$	0.018	-0.0053 $[-0.0101, -0.0006]$	-0.0049 $[-0.0097, -0.0006]$

I LARGE-COHORT FORECAST AGAINST A PERSONAL STATISTICAL PREDICTOR

Cohort, cutoff and inclusion. The fourth cohort consists of the 9,350 riders of four cities (C1–C4) for whom an independent distillation run outside this study had written four memory documents, on long-term patterns, decision model, short-term context and persona, from the rider's latest 60 recorded days for 7,126 riders or latest 30 recorded days for 2,224 riders; a different LLM, DeepSeek-V4-Flash, wrote them, and all four are non-empty for every rider. The inputs of the study are exactly the records that the distillation read. Every input row and every memory input date precedes the first day of the forecast window, the seven calendar days after the cutoff; this leakage rule excludes no rider. A rider is eligible, using pre-cutoff information only, with at least 10 recorded days in the memory history, at least 3 recorded days in the 14 days before the cutoff and at least 20 voluntary responses in the history; 2,888 riders qualify, namely 1,307 in C1, 777 in C2, 231 in C3 and 573 in C4. The first rule that the others fail is fewer than 10 recorded days for 3,497 riders, fewer than 3 recent days for 2,959 and fewer than 20 responses for 6. Eligible riders have 10–60 recorded days in the memory history, with

median 60, and their four documents have a median of 12,039 characters. A fixed pseudo-random function of the rider identifier sets, before any outcome, a development split of 515 riders, used only to fit the combination below, and a confirmatory split of 2,373 riders. Decision points are voluntary responses to system-dispatched offers in the seven-day window, timeouts excluded. A rider with no such response is not evaluable, because neither primary endpoint is defined; 2,414 of the 2,888 riders are evaluable, 1,998 of them with 306,787 responses in the confirmatory split and 416 with 64,095 responses in the development split. Table 16 gives the counting flow.

Table 16: Counting flow of the large-cohort forecast.

Stage	Count
Riders with four memory documents	9,350
Excluded by the leakage rule	0
Eligible before the cutoff, one unit each	2,888
Development / confirmatory split	515 / 2,373
Forecast calls planned / generated / fully parsed	5,776 / 5,776 / 5,776
Forecasts replaced by the registered fallback	0
Evaluable riders with a voluntary response in the window	2,414
Confirmatory analysis set, riders / responses	1,998 / 306,787
Development fitting set, riders / responses	416 / 64,095

Forecasters. Both LLM conditions use the registered Qwen3.5-9B snapshot with greedy decoding, thinking disabled, at most 160 new tokens and one call per rider and condition, and ask for the refusal probability in each of five time slots, 00–10, 10–14, 14–17, 17–21 and 21–24 h, and the refusal fraction of the window. Both receive the same measurement pack, computed from the input records only: recorded days, days in the last 14, refusal fractions with numerator and denominator over the last 7 and 28 days and all recorded days, slot-level refusal fractions and response shares, completions per recorded day and timeouts. M adds the four memory documents verbatim; N is the identical prompt without them. The statistical predictor S uses the same records and no LLM. With R_r refusals among n_r responses of rider r, $R_{r,s}$ and $n_{r,s}$ their counts in slot s, $\pi_{r,s}$ the rider's pre-cutoff share of responses in slot s and p_c the pooled pre-cutoff refusal fraction of the eligible riders of the same city,

$$p_r = \frac{R_r + 20\,p_c}{n_r + 20}, \qquad p_{r,s} = \frac{R_{r,s} + 10\,p_r}{n_{r,s} + 10}, \\ q_r^{\mathrm{S}} = \sum_s \pi_{r,s}\,p_{r,s}. \tag{4}$$

The combination STK is fitted on the development split only and applied unchanged to the confirmatory split. For offers it is a rider-weighted logistic regression of acceptance on the logits of the two acceptance probabilities,

$$\operatorname{logit} P^{\mathrm{STK}}(\text{accept}) = b_0 + b_1 \operatorname{logit} P^{\mathrm{S}}(\text{accept}) \\ + b_2 \operatorname{logit} P^{\mathrm{M}}(\text{accept}), \tag{5}$$

and for the refusal fraction $q^{\mathrm{STK}} = w\,q^{\mathrm{S}} + (1-w)\,q^{\mathrm{M}}$ with w on a grid of step 0.01 minimizing the development error. The fit gives $b = (-0.482,\ 0.811,\ 0.350)$ after five Newton iterations and $w = 0.02$. $\mathrm{STK_N}$ is the same combination fitted with N instead of M, with $b = (-0.443,\ 0.717,\ 0.435)$ and $w = 0.02$, and S28 applies the rule of Eq. (4) to slot counts of the last 28 days, shrunk to the rider rate p_r.

Endpoints and tests. Offer log loss averages $-\log P(\text{observed response})$ over the rider's window responses, with the slot probability of the offer's hour clipped to $[0.01, 0.99]$; the refusal-fraction error is the absolute error of the predicted window refusal fraction against manual refusals over accepted plus manually refused offers. The Brier score is descriptive. The registered confirmatory family has four tests, M versus N and STK versus S on both endpoints, each with the paired rider bootstrap, using 10,000 resamples and a fixed seed, and the paired t test, both Holm-adjusted over the four tests; A is superior to B on an endpoint only if $d < 0$ and both adjusted p values lie below 0.05. Table 30 in Appendix U lists the family.

Results. Table 17 lists every forecaster of the cohort side by side, with mean errors on the confirmatory split and paired differences to S. M reads the four memory documents verbatim, without the check of Section 3. STK lowers the refusal-fraction error relative to S by -0.0055 $[-0.0080, -0.0029]$. In the bootstrap, 0 of the $B = 10{,}000$ resamples lie at or above zero and the Holm-adjusted p is 6.4×10^{-5} with 10^6 resamples; the paired t test gives 8.9×10^{-5}. Under STK, 1,316 riders have a lower error, against 682 under S; 336 riders gain more than the registered margin of 0.010, as Appendix K shows. The same contrast on the 416 riders of the development split is -0.0057 $[-0.0113, -0.0010]$. By stratum, the point estimate is negative in every city, with -0.0124 $[-0.0191, -0.0063]$ for the 402 riders of C4, -0.0058 for the 511 of C2, -0.0032 for the 911 of C1 and -0.0005 for the 174 of C3; in every quartile of memory length, with -0.0101, -0.0046, -0.0038 and -0.0043 from the shortest to the longest; and for both history lengths, with -0.0052 $[-0.0079, -0.0025]$ for the 1,672 riders with 60-day documents and -0.0068 for the 326 with 30-day documents. STK also lowers the Brier score relative to S, by -0.0014 $[-0.0027, -0.0002]$. The LLM forecasts carry the refusal-fraction gain: M$-$S is -0.0055 $[-0.0081, -0.0029]$ and N$-$S -0.0050 $[-0.0074, -0.0028]$, and STK$_{\mathrm{N}}-$S is -0.0050 $[-0.0073, -0.0028]$ on refusal-fraction error and -0.0045 $[-0.0082, -0.0009]$ on log loss, so the same-definition measurement pack read by the LLM already moves the forecast beyond the smoothed rates. On offer log loss, STK improves on M by -0.0067 $[-0.0137, -0.0011]$, and the 28-day statistical rule improves on S, with S28$-$S of -0.0082 $[-0.0119, -0.0048]$ on log loss and -0.0045 $[-0.0061, -0.0030]$ on refusal-fraction error. Because every call was fully parsed, the complete-case analysis is identical to the registered one.

Reading of the registered result. The 9% reduction is a mean over riders. The median rider difference of STK$-$S is -6.8×10^{-6}; the rider-level differences are heterogeneous, with 336 riders or 16.8% below -0.010 and 233 or 11.7% above $+0.010$, as Table 20 shows. The measurement-only LLM forecast N lowers the refusal-fraction error relative to S by a similar amount, -0.0050, with 813 instead of 8,111 input tokens per rider, and the registered comparison of M with N is listed in Table 30. The result therefore supports an LLM forecast that reads same-definition measurements, combined with S, as an improvement on S for the refusal fraction; it is not evidence about the consistency check, which this study does not apply.

Table 17: Large-cohort forecast: every forecaster side by side on the confirmatory split of $n = 1{,}998$ riders with 306,787 responses; lower errors are better. Mean errors, the paired difference to S with its two-sided 95% paired rider-bootstrap interval, and the mean LLM input tokens and seconds per rider, with one call per LLM forecaster and rider; the statistical predictors and the combination use no model call. S: smoothed personal refusal and slot rates; S28: 28-day slot rates; N: LLM reading the measurement pack; M: LLM reading the memory documents verbatim, without the consistency check, and the measurement pack; STK$_{\mathrm{N}}$ and STK: N and M combined with S; NL: non-LLM predictor fitted on the 416 development riders only, registered in the eleventh amendment, with an interval from 2,000 resamples as in Table 40.

	Mean error			Difference to S [95% CI]		LLM cost
Forecaster	Offer log loss	Refusal-fraction error	Brier	Offer log loss	Refusal-fraction error	tokens / s
S	0.2700	0.0621	0.0834	—	—	—
S28	0.2618	0.0576	0.0808	-0.0082 [-0.0119, -0.0048]	-0.0045 [-0.0061, -0.0030]	—
NL	0.2576	0.0565	0.0794	-0.0124 [-0.0175, -0.0076]	-0.0056 [-0.0075, -0.0038]	—
N	0.2680	0.0571	0.0812	-0.0021 [-0.0077, 0.0038]	-0.0050 [-0.0074, -0.0028]	813 / 6.8
M	0.2736	0.0566	0.0824	0.0036 [-0.0037, 0.0120]	-0.0055 [-0.0081, -0.0029]	8,111 / 12.2
STK$_{\mathrm{N}}$	0.2655	0.0571	0.0814	-0.0045 [-0.0082, -0.0009]	-0.0050 [-0.0073, -0.0028]	813 / 6.8
STK	0.2669	0.0566	0.0819	-0.0031 [-0.0064, 0.0001]	-0.0055 [-0.0080, -0.0029]	8,111 / 12.2

Cost. M averages 8,111 input tokens, 101 output tokens and 12.2 s per call, and N 813 input tokens, 89 output tokens and 6.8 s. The combination adds no model call.

J LATER CUTOFF: SIMPLE REPAIR, READOUT ORDER AND MODEL JUDGES

Design. An eleventh amendment, appended before any outcome of its window was extracted and before any forecast beyond a 20-unit format check existed, registers a forecast at the cutoff t_0 on the day after the large-cohort outcome window ends, with memory cutoff $t_m = t_0 - 14$ days and a two-day forecast window. Riders come from the four-city cohort of Appendix I; riders of every other

earlier cohort of this study are excluded. Of the 9,350 riders, 2,441 meet the registered pre-cutoff rules of at least 10 recorded days in the 60 days before t_0, 3 in the last 14, 10 in the 60 days before t_m and 20 voluntary responses; 1,934 have a voluntary response in the window and form the analysis set with 95,583 window responses; it contains 1,889 riders of the large-cohort study, whose earlier outcome window now lies in the history, and 45 riders outside every earlier analysis. Qwen3.5-9B with greedy decoding and thinking disabled writes one lagged memory per rider with the registered prompt and produces every forecast. All conditions receive the same system prompt, the same typed measurement block at t_0 and the same task, which asks for per-slot refusal probabilities and the window refusal fraction; they differ only in the memory block: TC7, checked memory with retraction markers as in Appendix F; SR, simple numeric repair, in which every structured claim that fails the same-definition check carries the measured value instead of the claimed one and every statement, its text and its premises stay in the memory; TL, direct-only marking, in which the conflicting statements are marked and the statements that depend on them stay; T7, typed memory as written; S7, no memory. TC7b and T7b repeat TC7 and T7 with the memory block placed before the measurement block. Across riders, the check flags 2.69 statements directly and 1.56 through dependency on average, and SR replaces 2.74 claim values. Mean input tokens per call are 1,967 under TC7, 2,075 under SR and T7, 2,027 under TL and 1,298 under S7. All 17,087 forecast calls returned an answer; in 15 of them it was incomplete, and the registered scorer fills a missing slot from the reported refusal fraction or a missing fraction from the slot probabilities, so no call needed the fallback. Endpoints, fallback and scorer are those of the large-cohort study; the rider is the unit of analysis.

Registered family and claims. Four tests form the registered family of Table 35: superiority of TC7 over SR on offer log loss, non-inferiority of TC7 to SR at 0.010 on refusal-fraction error, STK versus a non-LLM predictor NL fitted only on the development riders of the large cohort, and TC7b versus S7 on log loss. Each test uses the paired rider bootstrap and the paired t test, with a separate Holm adjustment for each method, and a verdict requires both. Checked memory is non-inferior to the simple repair on refusal-fraction error, with $d = -0.0020$ $[-0.0041, 0.0000]$ and zero of 10,000 bootstrap resamples at or beyond the margin, also in the registered sensitivity families of Table 37 with the 1, 3 and 5 most favorable riders removed and on complete cases. Table 38 gives the mean errors of every condition. The descriptive contrasts of the amendment in Table 39 include checked versus unchecked typed memory: TC7 lowers offer log loss by -0.0240 $[-0.0380, -0.0105]$ and refusal-fraction error by -0.0030 $[-0.0057, -0.0007]$ relative to T7.

Model-judge check of the memory interface. Two open-weight judges that differ from the writer and forecaster, Qwen3-32B with thinking disabled and Llama-3.3-70B-Instruct with greedy decoding, read the lagged memories of 200 riders together with the measurements at t_0; the 2,910 statements comprise 541 flagged directly, 321 retracted through dependency only and 2,048 kept. The sample is fixed by code order before any judgment, and the judges see neither which statements the check flagged or retracted nor any condition label. For each claim they judge whether it encodes a level that the text asserts with the same metric, denominator, unit and window; for each statement, the quantitative assertions that no claim encodes and whether it is materially contradicted by the same-definition measurement; and for each declared premise, whether the statement relies on it. Table 18 reports every registered rate for each judge and, for binding, dependency and contradiction, for the items on which both judges answer yes, with rider-bootstrap intervals from 2,000 resamples where registered. For unnecessary retraction, the fourteenth amendment forms each judge's event first: both judges classify 4.6% [3.0, 6.4] of the 862 retracted statements as unnecessary, namely 3.0% of the directly and 7.5% of the dependency-only retracted ones, and at least one judge does so for 34.2%, with $\kappa = 0.08$. Each judge finds more than 85% of the directly retracted statements contradicted by the measurement, against about 23% of the kept statements, and the share of statements shown as memory that both judges find contradicted is 16.1% [14.2, 18.0] with the check under TC7 against 26.2% [24.3, 28.0] without it under T7. The judges are language models, not a human reference standard; their agreement is given in Table 18.

Fixed statement sets and uncertain labels. Table 19 keeps the original three disjoint statement sets fixed. Both judges label 432 directly conflicting and 329 kept statements contradicted, and no dependency-only statement receives two positive labels; the second judge answers unknown for 318 of the 321 dependency-only statements, and an unknown label is not counted as a negative. Direct-only checking TL and full closure TC7 both retain the 329 jointly positive kept statements.

Table 18: Model-judge check of the memory interface on 200 riders with 2,910 statements, in percent, with rider-bootstrap 95% intervals from 2,000 resamples for the rates for which they were registered. Per-judge columns use each judge's own answers. Both: for binding, dependency and contradiction, the claims, premises or statements on which both judges answer yes. †: fourteenth amendment; each judge's composite event is formed first, and Both is the intersection, that is, the retracted statements that each judge separately classifies as unnecessary, shown with the per-judge shares and the union on the 862 retracted statements judged by both. Coverage is reported per judge only. The registered aggregate formed these two composite entries after combining the judges' atomic answers: it counted a direct retraction unless both judges find it contradicted and a dependency-only retraction unless both judges find it contradicted or both identify a retracted premise that it relies on, giving 34.3%, and it took the smaller of the two unencoded-assertion counts, giving 78.0%; those two entries are not rates of agreement and are not shown in the Both column. The lower part gives the agreement of the two judges. Binding precision: claims judged to encode the asserted level; coverage: claims over claims plus unencoded quantitative assertions; dependency precision: declared premises judged to be relied on. Judged contradicted: by the same-definition measurement at t_0. Shown as memory: statements the forecaster sees as ordinary memory in each condition. Retractions judged unnecessary: directly retracted statements judged not contradicted and dependency-only retractions whose retracted premises are all judged not relied on.

Rate	Qwen3-32B	Llama-3.3-70B-Instruct	Both
Claim binding precision	47.6 [43.1, 52.2]	47.2 [42.7, 51.8]	29.5 [25.8, 33.2]
Coverage of quantitative assertions	65.3 [63.1, 67.6]	55.6 [52.7, 58.6]	–
Dependency precision	52.1 [47.1, 57.1]	91.8 [88.9, 94.5]	49.9 [45.0, 54.7]
Judged contradicted: directly retracted	91.1	85.8	79.9
Judged contradicted: dependency-only retracted	2.5	0.6	0.0
Judged contradicted: kept	22.9	23.2	16.1
Shown as memory and judged contradicted: T7	33.3 [31.4, 35.2]	32.4 [30.4, 34.3]	26.2 [24.3, 28.0]
Shown as memory and judged contradicted: TL	20.1 [18.2, 22.1]	20.2 [18.2, 22.0]	13.9 [12.2, 15.6]
Shown as memory and judged contradicted: TC7	22.9 [20.8, 25.0]	23.2 [21.0, 25.3]	16.1 [14.2, 18.0]
Retractions judged unnecessary (share of R)	26.2 [23.1, 29.4]	12.6 [10.0, 15.4]	4.6 [3.0, 6.4]†
directly retracted	8.9	14.2	3.0 [1.6, 4.4]†
dependency-only retracted	55.5	10.0	7.5 [4.1, 11.3]†
Retractions judged unnecessary by at least one judge	–	–	34.2 [30.8, 37.6]†
Recall of the check (judged contradicted)	50.9	49.3	56.8
Recall of the closure (judged contradicted)	51.7	49.5	56.8

Judgment	Raw agreement	Cohen's κ	n
binding	64.1%	0.28	1,543
dependency	55.8%	0.08	1,498
contradiction	41.9%	0.23	2,910
unnecessary retraction†	70.4%	0.08	862

Relative to the retained sets these are $329/2369$ under TL and $329/2048$ under TC7, and on the common original denominator both are $329/2910$.

Table 19: Contradiction labels on the fixed original statement sets. Entries are the counts of yes, no and unknown for each judge; the final column counts the statements with yes from both judges. An unknown label is not counted as a negative.

Original set	Statements	Qwen3-32B	Llama-3.3-70B	Both yes
Direct conflict D	541	493/48/0	464/67/10	432
Dependency only $R \setminus D$	321	8/307/6	2/1/318	0
Kept $S \setminus R$	2048	468/1556/24	476/522/1050	329

K Rider-level distributions and resampling resolution

Practical threshold. The tenth amendment fixes, before the analysis ran, a practical threshold equal to the registered non-inferiority margins: 0.020 nats for offer log loss and 0.010 for the refusal-fraction error. With the registered scorers executed unchanged, Table 20 gives the median and quartiles of the rider-level differences of the main contrasts and the numbers of riders whose difference lies below $-$threshold, within $\pm$threshold or above $+$threshold; the large-cohort scorer reproduces the four registered confirmatory means exactly. The supplement contains the same summaries, with fixed-bin histogram counts and further quantiles, for every contrast of these studies.

Table 20: Rider-level differences d of the first minus the second forecaster, so negative values favor the first. Median with quartiles, and riders with d below $-$threshold, within the threshold and above $+$threshold, with thresholds 0.020 for log loss and 0.010 for refusal error.

Riders	Contrast	Endpoint	n	Median [quartiles]	Below $-$threshold	Within	Above $+$threshold
Holdout	TC$-$P	log loss	22	-0.0858 [-0.4334, -0.0245]	16 (72.7%)	3 (13.6%)	3 (13.6%)
Holdout	TC$-$P	refusal error	22	-0.0035 [-0.0740, 0.0000]	9 (40.9%)	9 (40.9%)	4 (18.2%)
All 282	TC$-$P	log loss	282	-0.0126 [-0.2083, 0.0023]	130 (46.1%)	115 (40.8%)	37 (13.1%)
All 282	TC$-$P	refusal error	282	0.0000 [-0.0057, 0.0070]	62 (22.0%)	159 (56.4%)	61 (21.6%)
Never-exposed	TC7$-$P	log loss	789	-0.0070 [-0.1415, 0.0070]	340 (43.1%)	305 (38.7%)	144 (18.3%)
Never-exposed	TC7$-$P	refusal error	789	0.0000 [-0.0100, 0.0060]	199 (25.2%)	427 (54.1%)	163 (20.7%)
Large cohort	STK$-$S	log loss	1,998	0.0000 [-0.0008, 0.0016]	204 (10.2%)	1,577 (78.9%)	217 (10.9%)
Large cohort	STK$-$S	refusal error	1,998	-6.8×10^{-6} [-0.0023, 0.0005]	336 (16.8%)	1,429 (71.5%)	233 (11.7%)

Resolution of bootstrap p values. Every registered bootstrap p value is computed from $B = 10{,}000$ paired rider resamples of the mean difference. Let k be the number of exceedances. For a two-sided test, k is the smaller of the numbers of resampled means at or below zero and at or above zero, and the estimate is $\min(1, 2k/B)$. For a one-sided non-inferiority test, k is the number of resampled means at or above the margin, and the estimate is k/B. Holm's step-down is applied to these estimates within each registered family. When $k = 0$, the estimate and its Holm adjustment are both 0, and the tables write $0/B$ for them instead of a p value. This entry marks the resampling resolution of the run; it is neither a deterministic upper bound on the tail probability of the ideal bootstrap nor a bound on a Holm-adjusted value. Table 42 lists k, B, the side, the estimate, the family size and the Holm-adjusted value of every bootstrap test in the tables. For the large-cohort contrast STK$-$S on refusal-fraction error, where $k = 0$, 1,000,000 resamples of the same rider differences place 7 resampled means at or above zero; the add-one two-sided estimate $2(k+1)/(B+1)$ is 1.6×10^{-5}, and its Holm adjustment over the four registered tests is 6.4×10^{-5}, next to 8.9×10^{-5} from the paired t test.

L Dependency retraction on model-written answer graphs

This diagnostic applies the closure operator to structured answers rather than to memories. For 64 development riders and four questions each, Qwen3.5-9B in bf16 with greedy decoding writes an answer graph whose nodes are source facts, memory quotes and provisional judgments with declared premises, once as plain structured context and once with typed evidence, giving 256 answers per

condition. The executor withholds nodes that fail its source or format checks and then either retracts only those nodes, the local rule, or also every declared descendant, the closure rule, on the same answers. Of the 256 answers per condition, 175 plain and 167 typed answers are executable; all 256 enter the per-question differences. Table 21 gives the paired differences per original question with rider-level bootstrap intervals.

Table 21: Closure minus local retraction on the same answers of 64 riders, per original question, with unadjusted 95% intervals.

Quantity	Plain answers	Typed answers
Retracted declared descendants	$+0.0703$ $[0.0352, 0.1094]$	$+0.0547$ $[0.0273, 0.0859]$
Provisional judgments kept	-0.0703 $[-0.1094, -0.0352]$	-0.0547 $[-0.0859, -0.0273]$
Retained memory quotes	0 $[0, 0]$	0 $[0, 0]$
Source facts replaced	0 $[0, 0]$	0 $[0, 0]$
Executable answers	0 $[0, 0]$	0 $[0, 0]$

Closure removes exactly the judgments built on a withdrawn node, 18 in plain and 14 in typed answers in total, and nothing else, as Proposition 1(iii) requires. The diagnostic establishes how the operator acts on real model output; Section 5.3 evaluates the forecasts of the complete checked-memory bundle.

M DEVELOPMENT COHORT

Twenty riders have four memory documents and 60 observed historical days each; one of 21 candidate bundles was incomplete and was excluded before any comparison. The 1,200 historical person-days include 96,290 events, 21,856 completed units and 8,298 refusals. The following two-week window contains 275 observed person-days, 13–14 per rider, as Figure 3 shows. With D_i the recorded future days, c_{id} completions, r_{id} recorded refusals of any type, g_{id} grabs, e_{id} events and e^{eve}_{id} events in the last six hours of the day, the development targets are

$$
\begin{aligned}
y_i^{\mathrm{count}} &= \frac{\sum c_{id}}{|D_i|}, \\
y_i^{\mathrm{refusal}} &= \frac{\sum r_{id}}{\sum (g_{id} + r_{id})}, \\
y_i^{\mathrm{eve_e}} &= \frac{\sum e^{\mathrm{eve}}_{id}}{\sum e_{id}},
\end{aligned}
\tag{6}
$$

summed over $d \in D_i$; missing days are not zero activity. The refusal and evening targets differ from the registered quantities of Table 24. The model, DeepSeek-V4-Flash at temperature zero, receives the complete memory, or the complete memory together with source statistics, namely 60-, 28- and 7-day histories and a weighted estimate; both arms share model, decoding and targets, and all calls returned valid outputs. Table 22 reports per-rider MAE. Adding the statistics lowers the refusal error from 0.1299 to 0.0370 and the evening error from 0.0481 to 0.0442.

The 20-rider studies reuse retrospective 60-day memory versions whose event cutoffs precede the outcome window; the memories of the audit and holdout cohort, by contrast, are written inside the study from tables that end at t_m or t_0, so they cannot contain information from any later window.

Table 22: Development cohort: MAE of the memory contexts on the same 20 riders and two-week future window; lower is better. Refusal is the all-refusal fraction and evening the event-based share $\mathrm{eve_e}$.

Memory context	Refusal	Evening
Complete memory	0.1299	0.0481
Complete memory + source statistics	0.0370	0.0442

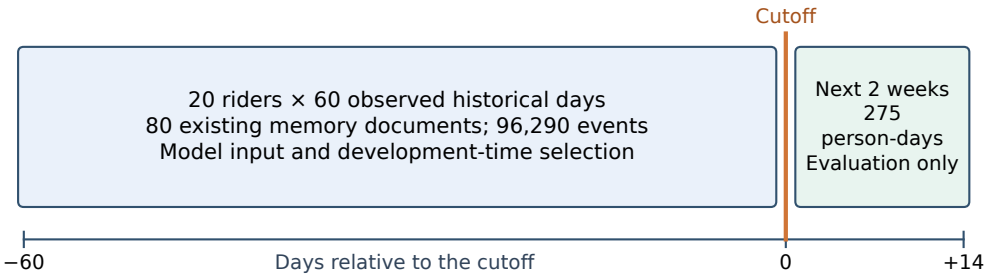


Figure 3: Temporal protocol of the 20-rider studies; days are counted from the historical cutoff, which is day 0. Of 280 possible future person-days, 275 have records; missing days are not imputed as zero.

Table 23: Complete four-document bundles and longitudinal coverage. Collection membership may overlap; a complete bundle does not certify every assertion. Full coverage means the target historical duration is observed.

Collection	Complete bundles	Full coverage	Observed duration
Reference collection	5,975	not inferred	not inferred
Up-to-60-day collection	7,156	2,548	1–60 days
Up-to-30-day collection	2,203	420	1–30 days
Temporal development cohort	20	20	60 days

N METRIC REGISTRY

All quantities of the audit and the holdout forecast are computed by one implementation, included in the supplement; Table 24 lists them together with the targets of the earlier studies. Days and hours are local; a recorded day is a day with at least one event of any type, and a day without records is never counted as zero activity. Windows are $[c - w, c)$ with $w \in \{7, 28, 60\}$ days, where the cutoff c is the forecast cutoff t_0 for measurements and the writer's cutoff for memory claims, so a flagged claim from a two-week-old memory mixes writer error and drift. Refusal fractions and offer labels of the audit and holdout forecast use only voluntary responses to system-dispatched offers, with acceptance coded 1 and manual refusal coded 0; timeout rejections and self-selected orders are excluded. The holdout evening target is the completion-based share $\mathrm{eve_c}$; the development cohort's event-based share $\mathrm{eve_e}$ and its all-refusal fraction are different estimands and are reported under their own names. Recorded days per week are $7\, n_{\mathrm{rec}}/n_{\mathrm{cal}}$.

Displayed measurements. The corrected renderer shows every measurement as $\alpha_m \times n_k/d_k = \mu_k$ whenever $\alpha_m \neq 1$, for example "$7 \times 14/28 = 3.50$" recorded days per week, so that the displayed numerator, denominator, factor and value satisfy Eq. (1); round-trip tests on synthetic tables check this relation for every rendered line, including undefined denominators. The holdout forecast of Section 5.3 used the earlier renderer, which printed the equation $n_k/d_k = \mu_k$, for example "$14/28 = 3.50$", false for this metric. The value itself was computed correctly, and the plain context of P, which states no equation, is unchanged. Every condition received the same measurement lines, in the plain format for P and the typed format otherwise, so the conditions remained information-matched; the follow-up run of Appendix F uses the corrected renderer.

Table 24: Metric registry: event filter, history statistic, memory claim key with conflict tolerance, and scoring target. The last four rows are the targets of the earlier studies.

Name	Event filter	History statistic	Claim key, tolerance	Target over the next 14 days
completions/rec. day	delivery completions	completions / recorded days	(m, w), $\max(2, 0.2\|x\|)$	same ratio; MAE
refusal fraction	system-dispatched accept / manual refusal, no timeouts	refusals / responses	(m, w), 0.05	same ratio; MAE, primary
$\mathrm{eve_c}$	completions, local hour ≥ 18	evening / all completions	(m, w), 0.05	same ratio; MAE
rec. days/week	any event	$7\, n_{\mathrm{rec}}/n_{\mathrm{cal}}$	(m, w), 1.0	recorded-day count, 0–14; MAE
hour-bin acceptance	as refusal, 4 local-hour bins	accepted / responses, 28 d	none, never checked	–, baseline input
offer acceptance	future system-dispatched responses	–	none	≤ 24 stratified offers; mean clipped log loss, primary
dev. count	completed deliveries	completions / recorded days	none	same ratio; MAE
dev. all-refusal fraction	every recorded grab and refusal, incl. timeouts and self-selected orders	refusals / (grabs + refusals)	none	same ratio; MAE
$\mathrm{eve_e}$	all events, local hour ≥ 18	evening / all events	none	same ratio; MAE
identity response	55 sampled dispatched offers, no timeouts	–	none	rider-mean log loss, accuracy

O CONTEXT COMPILATION AND MATCHED CONDITIONS

Every condition receives the same task and the same measurements $F(t_0)$. Algorithm 1 compiles the context deterministically. The conditions are: no memory (S); plain memory (P), which shows the statements and claims of the lagged memory as untyped bullets and the measurements without type labels; typed memory (T), which adds identifiers, roles, the memory cutoff, declared premises, units and evidence types; typed memory with consistency (TC), which applies Eqs. (2)–(3) to T; selected typed memory at $b \in \{1000, 2000, 4000\}$; and a prefix comparator at $b = 2000$. Two reference conditions are not information-matched: a memory written at t_0 (TF) and another rider's memory (W); neither is interpreted. The S, P, T, TC and budgeted contexts are deterministic functions of the same input $X = (M_i, F(t_0), q)$ by Proposition 4. The contrasts T$-$P and TC$-$T therefore change presentation and computation, not the information available about the rider. Retraction also removes text, so TC shows less than T.

Algorithm 1 Context compilation for rider i, condition a and task q

Require: lagged memory $M = \Phi(H_i^{<t_m})$; fresh memory $M^{\mathrm{TF}} = \Phi(H_i^{<t_0})$; donor memory M^{W}; measurements $F(t_0)$; task q

$C \leftarrow \textsc{Render}(F(t_0))$ ▷ untyped if $a = \mathrm{P}$
if $a = \mathrm{S}$ **then**
  return $C \,\|\, q$
end if
if $a = \mathrm{TF}$ **then**
  $M \leftarrow M^{\mathrm{TF}}$
else if $a = \mathrm{W}$ **then**
  $M \leftarrow M^{\mathrm{W}}$
end if
$\mathcal{S} \leftarrow$ statements of M; $K \leftarrow \mathcal{S}$; $R \leftarrow \emptyset$
if $a = \mathrm{TC}$ **then**
  $R \leftarrow D$ ▷ Eq. (2)
  while some $v \notin R$ has a premise $u \in R$ **do**
    $R \leftarrow R \cup \{v\}$
  end while ▷ Eq. (3)
else if $a = \mathrm{B}_b$ **then**
  $K \leftarrow \emptyset$
  for each role r with statements **do** ▷ phase 1
    $s_r \leftarrow \prec$-first statement of role r
    if $c(K) + c(s_r) \le b$ **then**
      $K \leftarrow K \cup \{s_r\}$
    end if
  end for
  for $s \in \mathcal{S}$ in $\prec$ order **do** ▷ phase 2
    if $s \notin K$ and $c(K) + c(s) \le b$ **then**
      $K \leftarrow K \cup \{s\}$
    end if
  end for
else if $a = \mathrm{PRE}_b$ **then**
  $K \leftarrow$ longest document-order prefix of $\mathcal{S}$ with $c(K) \le b$
end if
$C \leftarrow C \,\|\, \textsc{Render}(K)$: $s \in R$ as RETRACTED, other s as MEMORY, untyped if $a = \mathrm{P}$
return $C \,\|\, q$ ▷ forecast $\hat{y}_a = \Psi(C \| q)$

P DATA ACCESS AND REPRODUCTION

Models. Qwen3.5-9B, an open-weight model, writes the memories and produces the forecasts of the memory-claim audit, the holdout forecast, the follow-up run, the five-cutoff validation and the never-exposed-rider forecast with greedy decoding and thinking disabled; it produces the LLM forecasts M and N of the large cohort and the answer graphs of Appendix L. DeepSeek-V4-Flash produces the forecasts of the identity study and the readouts of the development cohort at temperature zero, and it wrote the four memory documents of every large-cohort rider in a distillation run that preceded the study; the study uses these documents verbatim. The statistical predictors and the consistency check use no model.

Models and material of the later-cutoff study. In the later-cutoff study of Appendix J, Qwen3.5-9B writes the memories and produces the forecasts, and Qwen3-32B with thinking disabled and Llama-3.3-70B-Instruct are the two model judges, both with greedy decoding. The supplement contains the prompt construction, job builders, scorers and synthetic end-to-end tests of this study and of the registered second-model replication listed in Table 26, the aggregation code of the model judges, and the aggregate result files of the later-cutoff study, together with the run status of the five-cutoff validation, that is, memory parse status and construction checks.

Privacy. The paper reports aggregate statistics and generic categories. Identifiers, contact details, precise locations and trajectories are omitted, and relative windows preserve temporal meaning. Raw event records and rider-level data are not released.

Reproduction. The anonymized supplement provides memory and forecast prompt construction, numerical definitions, statistical analysis code, registered protocols, aggregate results and synthetic examples. It includes the memory parser, conflict and dependency retraction, budget selection and the premise-preserving compiler. Synthetic checks verify the measurement definitions, undefined-denominator handling and formal properties; they test implementation behavior rather than forecast accuracy.

Published aggregates recover the displayed tables, including identity contrasts, model conditions, complete confirmatory families, sensitivity families, robustness analyses, model-judge summaries and the fixed-denominator analysis of Table 19. Paired rider-level differences are summarized by quantiles, fixed-bin histogram counts and threshold counts, without any per-rider value. Regenerating predictions requires the restricted histories, memories and labels. The identity study's verbatim system prompts and the data access layer are also withheld; Appendix B specifies its inputs and controls, and data holders supply the access interface.

Table 25 reports stage-wise token and time costs. One memory serves subsequent forecasts for its rider until refresh; the amortized per-forecast cost is the writing cost divided by forecasts per refresh, plus checking, selection and readout. Code and aggregate results remain separate from the restricted individual records.

Table 25: End-to-end cost records by pipeline stage. Tokens and generation time are summed over calls from the recorded usage of each run; generation time is the sum of per-call wall times over concurrently running calls, so it is compute time, not elapsed time. Deterministic operators were timed single-threaded over all 590 memories of the audit cohort. API readouts report the recorded request latency. The development studies reuse retrospective memories, so only their readout cost is listed. Q9B: Qwen3.5-9B in bf16 with greedy decoding; DSV4F: DeepSeek-V4-Flash; det.: deterministic operator without a model call. Answer graphs cover 64 riders $\times$ 4 questions $\times$ 2 arms of Appendix L; the development readouts cover all calls of the two development runs on the 20 riders of Appendix M.

Stage	Model	Calls	Input tokens	Output tokens	Time
Memory writing, 295 riders $\times$ 2	Q9B	590	1,013,484	697,361	13.2 h
mean per memory, median time	Q9B	1	1,718	1,182	81.5 s
Conflict check + closure, per memory	det.	0	0	0	0.013 ms
Selection, 8 settings, per memory	det.	0	0	0	0.258 ms
Holdout forecasts, 10 conditions $\times$ 2 tasks	Q9B	5,900	11,519,145	1,072,637	17.5 h
mean per forecast, mean time	Q9B	1	1,952	182	10.7 s
Follow-up run, 7 conditions $\times$ 2 tasks	Q9B	4,130	8,476,705	751,976	15.9 h
Answer graphs	Q9B	512	8,222,688	503,213	8.1 h
Development readouts, memory contexts	DSV4F	100	471,579	4,514	131 s
Development readouts, with source statistics	DSV4F	60	406,113	2,820	86 s

Q EXTENDED RELATED WORK

Mobility, travel and platform work. LLM agents generate activity trajectories from real mobility data (Wang et al., 2024) and travel diaries from individual profiles (Li et al., 2024); zero-shot prompting reaches accuracy comparable to discrete-choice models in travel-behavior prediction (Mo et al., 2026). Traveler agents with memory and a learnable persona align with individuals' day-to-day route choices (Liu et al., 2026). For delivery work, LLM agents instantiated from four clustered rider personas study system-level routing outcomes (Zhang & Xiao, 2026), dynamic discrete-choice models explain order acceptance (Robbennolt et al., 2026), and ride-hailing drivers' reservation wages vary over time (Chen et al., 2019). Platform-side work optimizes rider bonus budgets (Wu et al., 2022) and benchmarks LLMs on local-life services (Lan et al., 2025).

Claims against tables. Table-based fact verification links natural-language statements to structured evidence, including by parsing statements into executable programs (Chen et al., 2020). Our check is deterministic and definition-matched: claims are already typed by the writer, and the evidence is a time-varying personal measurement whose metric, denominator and window must match.

Memory, conflicts and revision. MemGPT pages information between memory tiers (Packer et al., 2023), MemoryBank applies a forgetting curve (Zhong et al., 2024), A-Mem links notes that revise one another (Xu et al., 2025), and Mem0 extracts and updates salient facts (Chhikara et al., 2025). Zep keeps a bi-temporal knowledge graph and invalidates contradicted facts (Rasmussen et al., 2025); Hindsight separates world facts, experiences, summaries and beliefs (Latimer et al., 2026); APEX-MEM stores append-only events and resolves conflicts at query time (Banerjee et al., 2026); THEANINE links memories into timelines (Ong et al., 2025); MemMachine keeps raw episodes beside profile memory (Wang et al., 2026a); BeliefMem retains competing conclusions (Liao et al., 2026). LoCoMo and LongMemEval test extraction, temporal reasoning, updates and abstention (Maharana et al., 2024; Wu et al., 2025). Truth-maintenance systems withdraw conclusions whose support is removed (Doyle, 1979; de Kleer, 1986), the AGM postulates characterize belief revision (Alchourrón et al., 1985), and LLMs show knowledge conflicts (Xu et al., 2024; Xie et al., 2024), unpropagated edits (Cohen et al., 2024; Zhong et al., 2023) and difficulty revising conclusions (Wilie et al., 2024). Our retraction brings dependency-directed withdrawal to declared dependencies between LLM-written memory statements and combines it with a *numerical* contract: a claim's identity is its metric, denominator and window, a conflict exists only against a measurement of the same definition, and the memory is judged by forecasts of real behavior. LLMs can forecast numeric series zero-shot (Gruver et al., 2023); here the source measurements are given to the forecaster explicitly and memory is added as typed, checked context.

R QUESTION-CONDITIONED SELECTION UNDER A BUDGET

For a character budget b, the cost $c(s) \geq 1$ of a statement is the length of its typed rendering plus one. Role headers are not charged. Let $c(A) = \sum_{s \in A} c(s)$. For task q with a fixed term list $\mathcal{W}_q$, the relevance score is

$$\begin{aligned} \sigma_q(s) = 2|\Gamma(s)| + \big|\{w \in \mathcal{W}_q : \ w \text{ occurs in } x(s)\}\big| \\ + 2 \cdot \mathbf{1}[q = \text{offer} \wedge \rho(s) = \text{decision}]. \end{aligned} \tag{7}$$

Let $\prec$ order statements by decreasing σ_q, breaking ties by document position. Phase 1 visits the roles in turn and keeps the $\prec$-first statement of each role if it still fits. Phase 2 scans all statements in $\prec$ order and keeps each one that still fits. The prefix comparator keeps the longest prefix, in document order, whose cost is at most b. Let P_1 be the phase-1 set and $\mathcal{P}_j$ the first j statements under $\prec$.

Proposition 2 (Budget guarantees)**.** *Under A1, the selected set $\widehat{K}_b$ satisfies:*

(i) $c(\widehat{K}_b) \leq b$*;*

(ii) $b - c(\widehat{K}_b) < c(s)$ *for every unselected s; hence, unless $\widehat{K}_b = \mathcal{S}$,*

$$c(\widehat{K}_b) > b - \max_s c(s);$$

(iii) if the $\prec$-first statements of all nonempty roles have total cost at most b, every nonempty role is represented;

(iv) $\widehat{K}_b \supseteq \mathcal{P}_{j_0}$*, where*

$$j_0 = \max\{j : \ c(P_1 \cup \mathcal{P}_j) \leq b\};$$

(v) if $b \geq c(\mathcal{S})$, then $\widehat{K}_b = \mathcal{S}$ and the compiled context equals the full typed context.

Selection is not nested in b. A larger budget can admit a role representative that displaces statements kept at a smaller budget, and it can even lower the total relevance score, as Appendix S shows. A quality–cost curve over budgets therefore compares distinct contexts and need not be monotone. The prefix comparator is nested and satisfies (i) but not (ii), because it stops at the first statement that does not fit. Budgets count characters of rendered statements, not model tokens.

S FORMAL PROPERTIES OF THE SIMULATION METHOD

Throughout, $G = (\mathcal{S}, E)$ is the declared dependency graph of a memory with unique identifiers, which is assumption A1. For $A \subseteq \mathcal{S}$,

$$\mathrm{Reach}(A) = \big\{v : \ G \text{ has a directed path } u_0 \to u_1 \to \cdots \to u_\ell = v \text{ with } u_0 \in A,\ \ell \geq 0\big\}.$$

A set A' is *successor-closed* if $u \in A'$ and $(u, v) \in E$ imply $v \in A'$. Then $\mathrm{Reach}(A)$ is the smallest successor-closed superset of A: it contains A, it is successor-closed, and induction on path length places every $\mathrm{Reach}(A)$ element in any successor-closed superset of A.

S.1 DEPENDENCY RETRACTION

Proof of Proposition 1. Part (i): the iteration computes the reachable set. By induction on j, R_j is the set of statements reachable from D by a path of length at most j. This holds for $j = 0$. For the step,

$$\begin{aligned} v \in R_{j+1} &\iff v \in R_j \text{ or } v \text{ has a premise in } R_j \\ &\iff v \text{ has a path from } D \text{ of length at most } j+1. \end{aligned}$$

Hence $R = \mathrm{Reach}(D)$. The sequence is nondecreasing, and once $R_{j+1} = R_j$ it is constant. Every strict increase adds at least one element of $\mathcal{S} \setminus D$, so there are at most $|\mathcal{S} \setminus D|$ strict increases and

$$R_j = R \qquad \text{for all } j \geq |\mathcal{S} \setminus D|.$$

In-place computation. The implementation updates R in place: it sweeps the statements and adds any v with a premise already in R. Every added statement has a premise in the current set, so by induction the set stays inside $\mathrm{Reach}(D)$. Each sweep before the last one adds a statement, so at most $|\mathcal{S} \setminus D| + 1$ sweeps run. After the last sweep, no statement outside the set has a premise inside it. The set is then a successor-closed superset of D and therefore contains $\mathrm{Reach}(D)$. The in-place computation thus returns the same R, independent of sweep order.

Part (ii), step 1: the retained memory is admissible. $D \subseteq R$, so $(\mathcal{S} \setminus R) \cap D = \emptyset$. Suppose $v \in \mathcal{S} \setminus R$ and $(u, v) \in E$. If $u \in R$, then v would be reachable from D and hence in R. So $u \in \mathcal{S} \setminus R$, and $\mathcal{S} \setminus R$ is dependency-closed.

Part (ii), step 2: every admissible set lies inside it. Let K be admissible and suppose $v \in K \cap R$. Take a path $u_0 \to \cdots \to u_\ell = v$ with $u_0 \in D$. Because K is dependency-closed, backward induction along the path gives

$$u_\ell,\ u_{\ell-1},\ \ldots,\ u_0 \in K.$$

Then $u_0 \in K \cap D$, contradicting admissibility. Hence $K \subseteq \mathcal{S} \setminus R$.

Part (ii), step 3: uniqueness. Uniqueness of the largest admissible set also follows because admissible sets are closed under union: a union of dependency-closed sets is dependency-closed, and a union of sets disjoint from D is disjoint from D.

Part (iii). This is immediate from $R = \mathrm{Reach}(D)$.

Part (iv). Let $\mathcal{S}' = \mathcal{S} \setminus R$ with unchanged claims and edges $E' = E \cap (\mathcal{S}' \times \mathcal{S}')$; premises in R become dangling and are ignored. Whether a statement conflicts depends only on its own claims and on $F(t_0)$, so

$$\begin{aligned} D' &= D \cap \mathcal{S}' = \emptyset, \\ \text{hence} \quad R' &= \mathrm{Reach}_{G'}(\emptyset) = \emptyset. \end{aligned}$$

Part (v). If $a'_m \leq a_m$ and $r'_m \leq r_m$, then for every claim

$$|c - \mu| > \max(a_m, r_m|\mu|) \quad \Longrightarrow \quad |c - \mu| > \max(a'_m, r'_m|\mu|).$$

Lowering the tolerances can therefore only enlarge D. $\mathrm{Reach}(A)$ is monotone in A, and adding edges adds paths, so it is also monotone in E.

Part (vi). If $R = D$, then $\mathcal{S} \setminus D = \mathcal{S} \setminus R$ is admissible by (ii). If $R \neq D$, pick $v \in R \setminus D$ and a path from D to v. Let w be the first vertex on the path outside D, and let $u \in D$ be its predecessor. Then

$$w \in \mathcal{S} \setminus D,$$
$$(u, w) \in E,$$
$$u \notin \mathcal{S} \setminus D.$$

So $\mathcal{S} \setminus D$ is not dependency-closed. □

Proof of Corollary 1. A retained statement lies outside $R \supseteq D$. By Eq. (2), none of its claims with a defined measurement exceeds the tolerance. □

S.2 Attribution of a conflict

The measurement set at the memory cutoff, $F(t_m)$, is computed by the same Eq. (1). Each claim can therefore be compared both with the quantity it describes, $\mu_k(t_m)$, and with the quantity the forecast uses, $\mu_k(t_0)$. This comparison needs no model calls and no outcomes.

Proposition 3 (Writer error and drift)**.** *Let $(k, c) \in \Gamma(s)$ with $\mu_k(t_m) \neq \perp$ and $\mu_k(t_0) \neq \perp$, and let $\tau = \tau_m(\mu_k(t_0))$. Define the writer error and the drift*

$$e_w = c - \mu_k(t_m),$$
$$e_d = \mu_k(t_m) - \mu_k(t_0).$$

Then $c - \mu_k(t_0) = e_w + e_d$ and:

(i) if the claim conflicts, then $\max(|e_w|, |e_d|) > \tau/2$;

(ii) if $|e_w| + |e_d| \leq \tau$, the claim does not conflict;

(iii) if $\big||e_w| - |e_d|\big| > \tau$, the claim conflicts;

(iv) for a faithful writer with $e_w = 0$, the claim conflicts iff $|e_d| > \tau$.

Proof. The identity is telescoping. The triangle inequality gives

$$\big||e_w| - |e_d|\big| \;\leq\; |e_w + e_d| \;\leq\; |e_w| + |e_d|.$$

The upper bound gives (ii). Statement (i) follows from (ii), since

$$\max(|e_w|, |e_d|) \leq \tau/2 \quad \Longrightarrow \quad |e_w| + |e_d| \leq \tau.$$

The lower bound gives (iii), and (iv) is the case $e_w = 0$. □

Conflicts can therefore be split into writer-dominated cases, where $|e_w| > |e_d|$, and drift-dominated cases. When the writer is faithful, the operator retracts exactly those claims whose quantity moved by more than the tolerance between the two cutoffs.

S.3 Selection under a budget

Proof of Proposition 2. Part (i). Under A1, the running total u equals $c(K)$, because each statement is added at most once and charged once. Every addition is guarded by $u + c(s) \leq b$, and u starts at 0. This proves (i).

Part (ii). Phase 2 examines every statement. If $s \notin \widehat{K}_b$, then when s was examined the total u_s satisfied $u_s + c(s) > b$. Totals never decrease, so $c(\widehat{K}_b) \geq u_s$ and

$$c(\widehat{K}_b) + c(s) \;\geq\; u_s + c(s) \;>\; b.$$

If $\widehat{K}_b \neq \mathcal{S}$, choosing any unselected s gives

$$c(\widehat{K}_b) \;>\; b - c(s) \;\geq\; b - \max_{s'} c(s').$$

Part (iii). Let T be the set of role representatives s_r, and suppose $c(T) \le b$. When phase 1 reaches role r, the total is at most the cost of the representatives of the roles before r, which is at most $c(T) - c(s_r)$. Hence

$$u + c(s_r) \;\le\; c(T) \;\le\; b,$$

and s_r is kept.

Part (iv). The index j_0 is well defined, because $j = 0$ is feasible: $c(P_1) \le b$ by (i) at the end of phase 1. Write $s_1 \prec s_2 \prec \cdots$. We show by induction that, for $j \le j_0$, the kept set equals $P_1 \cup \mathcal{P}_{j-1}$ when phase 2 reaches s_j. This holds for $j = 1$. If $s_j \in P_1$, it is already kept. Otherwise the guard reads

$$\begin{aligned} c(P_1 \cup \mathcal{P}_{j-1}) + c(s_j) &= c(P_1 \cup \mathcal{P}_j) \\ &\le c(P_1 \cup \mathcal{P}_{j_0}) \qquad \text{(the sets are nested and costs are positive)} \\ &\le b. \end{aligned}$$

Either way the kept set becomes $P_1 \cup \mathcal{P}_j$. Later steps only add statements.

Part (v). If $b \ge c(\mathcal{S})$, every guard satisfies

$$u + c(s) \;\le\; c(\mathcal{S}) \;\le\; b,$$

so every statement is kept. The renderer filters statements only by membership and keeps role and document order, so the context equals the full typed context. □

Selection is not nested in the budget. Take a long-term role with one statement a, and a recent role with statements β, γ, where

long-term role:	$\sigma(a) = 0,$	$c(a) = 300;$
recent role:	$\sigma(\beta) = 2,\ \sigma(\gamma) = 1,$	$c(\beta) = c(\gamma) = 150.$

Budget $b = 250$. Phase 1 skips a because it does not fit and keeps β. Phase 2 cannot add γ, since $300 > 250$, or a, since $450 > 250$, so $\widehat{K}_{250} = \{\beta\}$. *Budget* $b = 300$. Phase 1 keeps a and then cannot add β because $450 > 300$, so $\widehat{K}_{300} = \{a\}$. The two selections are disjoint, and the total relevance falls from 2 to 0 as the budget grows. Role coverage is thus bought at the expense of relevance, and the context at a larger budget is not a superset of the context at a smaller one. The prefix comparator is nested: a prefix that fits b also fits any $b' \ge b$. It is not maximal, however, because a statement that does not fit ends the scan even if a later, shorter statement would fit.

S.4 Information matching of the conditions

Proposition 4 (Information matching)**.** *Fix a rider and a task q. Treat the model weights, prompts, tolerances, term lists and renderer as constants, and let $X = (M, F(t_0), q)$ with $M = \Phi(H^{<t_m})$. For every condition*

$$a \in \mathcal{A}_0 = \{\mathrm{S}, \mathrm{P}, \mathrm{T}, \mathrm{TC}, \mathrm{B}_{1000}, \mathrm{B}_{2000}, \mathrm{B}_{4000}, \mathrm{PRE}_{2000}\},$$

the context is $C_a = g_a(X)$ for a deterministic map g_a, and C_{S} depends on X only through $(F(t_0), q)$. Consequently, for any joint law of X and a future outcome Y and any loss ℓ,

$$\inf_{\psi}\ \mathbb{E}\,\ell\big(Y, \psi(C_a)\big) \;\ge\; \inf_{\psi}\ \mathbb{E}\,\ell\big(Y, \psi(X)\big), \qquad a \in \mathcal{A}_0.$$

With greedy decoding, the forecast $\hat{y}_a = \Psi(C_a)$ is itself a function of X.

Proof. Step 1: each context reads only X. Every branch of Algorithm 1 for $a \in \mathcal{A}_0$ reads only M, $F(t_0)$, q and constants. The conflict set of Eq. (2) uses the claims in M and the values in $F(t_0)$. The closure of Eq. (3) uses the premises in M. Selection uses the texts, claims, roles and rendered lengths of the statements in M.

Step 2: the inequality. For any ψ, the map $\psi \circ g_a$ is a function of X, so the left infimum is taken over a subset of the functions available on the right. □

The proposition makes explicit which contrasts are information-matched. TF uses $\Phi(H^{<t_0})$, which reads the daily table for $[t_0 - 60, t_0)$, and W uses another rider's memory. Neither is a function of X, so both serve as references rather than matched contrasts. Within $\mathcal{A}_0$, a difference in forecast quality reflects how the forecaster uses the context: typing, retraction or truncation. It does not reflect access to additional data about the rider. Information matching bounds what each context can contain; it does not make the contexts equally informative or equally costly. A deterministic map can drop information, as retraction and truncation do, the contexts differ in length and hence in readout cost, as Table 25 shows, and the proposition says nothing about which context a given forecaster uses best.

S.5 Implementation correspondence

Assumption A1. The parser enforces unique identifiers in two passes. The first occurrence of each identifier that carries its role prefix is kept. Every other statement receives the next unused identifier of its role, and premises that name a renamed identifier are remapped. Memories that needed any renaming are counted. Premises that name no statement are ignored, and self-premises are dropped. Cycles are allowed and are handled by reachability. The conflict check uses unrounded values, whereas the context shows fractions to three decimals and counts or days to two. Claims are accepted only for the four metrics in $\mathcal{M}$ and the three windows; the hour-of-day acceptance rates are shown but never compared with memory. The same closure operator is used by the declared-graph executor of the development diagnostic. Its seed set B_0 contains replaced source facts and withheld claims, including cyclic and malformed ones. Statements (i)–(iii) and (vi) of Proposition 1 hold for it with D replaced by B_0.

T Retraction followed by budgeted selection

Propositions 1 and 2 each describe one operator. In Algorithm 1 they are separate conditions: TC retracts without a budget, and B_b selects from the unretracted memory. Neither guarantee carries over to a context that is first retracted and then cut to a budget. B_b can show a retracted statement as ordinary memory. For example, a statement claiming a 28-day refusal fraction of 0.9 against a measured 0.1 is retracted by TC but kept as MEMORY by B_{1000}. Proposition 2 also says nothing about premises. With a conflict-free memory $L_1 \to D_1$ whose statements cost 384 and 126, the selector of Eq. (7) at $b = 126$ keeps D_1 without its premise L_1. The same happens when that selector is applied to the retained memory $\mathcal{S} \setminus R$. This section defines a selector that keeps premises and proves one guarantee for the composed context.

Premise closure and eligibility. For $s \in \mathcal{S}$ let

$$\downarrow s = \{u \in \mathcal{S} :\ G \text{ has a path } u \to \cdots \to s\}, \qquad \downarrow A = \textstyle\bigcup_{s \in A} \downarrow s, \tag{8}$$

so that $\downarrow s$ contains s and all its direct and indirect declared premises. A set K is dependency-closed iff $\downarrow K = K$. For an excluded set $X \subseteq \mathcal{S}$, a statement is *eligible* if $s \notin X$ and $\downarrow s \cap X = \emptyset$; write $\mathcal{E}(X)$ for the eligible set.

Closed selection. The joint compiler of Algorithm 2 first computes D and R as in TC. In condition TCB_b, retracted statements are omitted. In condition TCBM_b, they are shown as RETRACTED lines with identifier and reason, and $c^{\times}(s)$ denotes the length of such a line plus one. Phase 0 of TCBM_b keeps these markers, direct conflicts first and then dependency-only retractions, each in document order and each only if it still fits. Call the kept set Q, with $Q = \emptyset$ for TCB_b, and let $b' = b - c^{\times}(Q)$. Phases 1 and 2 then run the selector of Section 3.4 with budget b'. The score, order $\prec$ and costs are unchanged, and there are two differences. A candidate s is added together with $\downarrow s \setminus K$, and only if this whole block fits. A candidate outside $\mathcal{E}(R)$ is skipped. Phase 1 therefore tries the $\prec$-first statement s_r of each role of $\mathcal{S}$, as the selector of Proposition 2 does. Let $\widetilde{K}_b$ be the selected memory, $\widetilde{P}_1$ the phase-1 set and $\mathcal{P}^{\circ}_j$ the first j statements of $\mathcal{E}(R)$ under $\prec$. The same selector without retraction, that is with $R = \emptyset$, is condition BC_b. The closed prefix comparator TCPRE_b scans $\mathcal{E}(R)$ in document order, adds $\downarrow s \setminus K$ for each s and stops at the first block that does not fit.

Proposition 5 (Retraction followed by selection)**.** *Under A1, for every budget $b \geq 0$:*

Algorithm 2 Joint compilation TCB_b / TCBM_b: retraction, then dependency-closed selection

Require: memory $\mathcal{S}$ with graph G; measurements $F(t_0)$; task q; budget b

$D \leftarrow$ Eq. (2); $\quad R \leftarrow$ Eq. (3); $\quad Q \leftarrow \emptyset$; $\quad u \leftarrow 0$
for $s \in D$, then $s \in R \setminus D$, each in document order **do** ▷ phase 0, TCBM_b only
  if $u + c^{\times}(s) \le b$ **then**
    $Q \leftarrow Q \cup \{s\}$; $\quad u \leftarrow u + c^{\times}(s)$
  end if
end for
$K \leftarrow \emptyset$; $\quad \mathcal{E} \leftarrow \{s \notin R : {\downarrow}s \cap R = \emptyset\}$
procedure ADD(s)
  if $s \in \mathcal{E}$ and $u + c({\downarrow}s \setminus K) \le b$ **then**
    $u \leftarrow u + c({\downarrow}s \setminus K)$; $\quad K \leftarrow K \cup {\downarrow}s$
  end if
end procedure
for each role r with statements in $\mathcal{S}$ **do** ▷ phase 1
  ADD($\prec$-first $s \in \mathcal{S}$ of role r)
end for
for $s \in \mathcal{E} \setminus K$ in $\prec$ order **do** ▷ phase 2
  ADD(s)
end for
return render of $K \cup Q$ in document order: Q as RETRACTED, K as MEMORY

(i) No re-entry. $\widetilde{K}_b \cap R = \emptyset$. *A retracted statement appears in the context at most as a* RETRACTED *line without its text, and every claim of a statement in $\widetilde{K}_b$ satisfies Corollary 1.*

(ii) Admissibility. *$\widetilde{K}_b$ is dependency-closed in G and disjoint from D. It is therefore an admissible sub-memory in the sense of Proposition 1, contained in $\mathcal{S} \setminus R$.*

(iii) Budget. $c(\widetilde{K}_b) + c^{\times}(Q) \le b$.

(iv) Closure-maximality. *For every $s \in (\mathcal{S} \setminus R) \setminus \widetilde{K}_b$,*

$$c(\widetilde{K}_b) + c^{\times}(Q) + c({\downarrow}s \setminus \widetilde{K}_b) > b.$$

(v) Role coverage. *If every s_r lies in $\mathcal{S} \setminus R$ and $c\big(\bigcup_r {\downarrow}s_r\big) \le b'$, every nonempty role is represented.*

(vi) Priority prefix. $\widetilde{K}_b \supseteq \widetilde{P}_1 \cup {\downarrow}\mathcal{P}^{\circ}_{j_0}$, *where*

$$j_0 = \max\{j : \ c(\widetilde{P}_1 \cup {\downarrow}\mathcal{P}^{\circ}_j) \le b'\}.$$

(vii) Full budget. *For TCB_b, $b \ge c(\mathcal{S} \setminus R)$ gives $\widetilde{K}_b = \mathcal{S} \setminus R$, which is the TC context without its* RETRACTED *lines. For TCBM_b, $b \ge c(\mathcal{S} \setminus R) + c^{\times}(R)$ gives the TC context.*

(viii) Idempotence. *Compiling the memory $\widetilde{K}_b$ again with the same $F(t_0)$, q and b retracts nothing and returns $\widetilde{K}_b$.*

(ix) Reduction. *If $E = \emptyset$ and $R = \emptyset$, then $\widetilde{K}_b = \widehat{K}_b$ of Proposition 2.*

(x) Seed independence. $\mathcal{E}(X) = \mathcal{S} \setminus \mathrm{Reach}(X)$ *for every $X \subseteq \mathcal{S}$. In particular $\mathcal{E}(D) = \mathcal{E}(R) = \mathcal{S} \setminus R$, so excluding only the direct conflicts gives the same selected memory as excluding the whole closure.*

The closed prefix comparator satisfies (i)–(iii) and (vii)–(viii). Without markers, its selection is nested in b. If phase 1 instead takes s_r as the $\prec$-first statement of role r within $\mathcal{E}(R)$, (v) holds for every role with a statement in $\mathcal{S} \setminus R$ without the first condition, and all other parts are unchanged.

Proof. We prove (x) first, because (i) and (ii) use it.

Part (x): seed independence. A statement s lies in $\mathrm{Reach}(X)$ iff some $u \in X$ has a path to s, that is, iff $\downarrow s \cap X \neq \emptyset$. Since $s \in \downarrow s$, this gives

$$\mathcal{E}(X) = \mathcal{S} \setminus \mathrm{Reach}(X).$$

Moreover $\mathrm{Reach}(D) = R$ and $\mathrm{Reach}(R) = R$ by Proposition 1(i).

Part (ii): admissibility. Each $\downarrow s$ is dependency-closed: if $v \in \downarrow s$ and $(u, v) \in E$, then u has a path to s. A union of dependency-closed sets is dependency-closed, and K starts empty and only grows by whole sets $\downarrow s$, so $\widetilde{K}_b$ is dependency-closed. Every added s lies in $\mathcal{E}(R) = \mathcal{S} \setminus R$, and $\mathcal{S} \setminus R$ is dependency-closed by Proposition 1(ii), so $\downarrow s \subseteq \mathcal{S} \setminus R$. Hence

$$\begin{aligned} \widetilde{K}_b &\subseteq \mathcal{S} \setminus R, \\ (\mathcal{S} \setminus R) \cap D &= \emptyset \qquad\qquad (\text{since } D \subseteq R). \end{aligned}$$

Part (i): no re-entry. This follows from (ii). The renderer prints a statement of R either not at all, under TCB_b, or in its RETRACTED form, under TCBM_b, which carries the identifier and the reason but not the text. Corollary 1 applies to every statement outside R.

Part (iii): budget. By A1, each statement is charged once, so the running total is

$$u = c^{\times}(Q) + c(K).$$

Every addition in phases 0–2 is guarded by $u + (\text{added cost}) \leq b$, and u starts at 0.

Part (iv): closure-maximality. Let $s \in (\mathcal{S} \setminus R) \setminus \widetilde{K}_b$. Phase 2 examined s with some kept set $K_s \subseteq \widetilde{K}_b$ and rejected it:

$$c^{\times}(Q) + c(K_s \cup \downarrow s) > b.$$

Costs are positive and $K_s \cup \downarrow s \subseteq \widetilde{K}_b \cup \downarrow s$, so

$$\begin{aligned} c^{\times}(Q) + c(\widetilde{K}_b) + c(\downarrow s \setminus \widetilde{K}_b) &= c^{\times}(Q) + c(\widetilde{K}_b \cup \downarrow s) \\ &\geq c^{\times}(Q) + c(K_s \cup \downarrow s) \\ &> b. \end{aligned}$$

Part (v): role coverage. Every s_r is eligible. When phase 1 reaches role r, the kept set lies in $\bigcup_{r' \text{ before } r} \downarrow s_{r'}$. Therefore

$$c(K \cup \downarrow s_r) \;\leq\; c\Big(\bigcup_{r'} \downarrow s_{r'}\Big) \;\leq\; b',$$

and s_r is added. The same argument applies to the variant, whose representatives are eligible by construction.

Part (vi): priority prefix. $j = 0$ is feasible by (iii), so j_0 is well defined. Write $s_1 \prec s_2 \prec \cdots$ for the statements of $\mathcal{E}(R)$. By induction, for $j \leq j_0$ the kept set equals $\widetilde{P}_1 \cup \downarrow \mathcal{P}^{\circ}_{j-1}$ when phase 2 reaches s_j. If s_j is already kept, then $\downarrow s_j$ is kept too, because the kept set is dependency-closed. Otherwise the guard reads

$$\begin{aligned} c(\widetilde{P}_1 \cup \downarrow \mathcal{P}^{\circ}_j) &\leq c(\widetilde{P}_1 \cup \downarrow \mathcal{P}^{\circ}_{j_0}) \qquad (\text{the sets are nested and costs are positive}) \\ &\leq b'. \end{aligned}$$

Either way the kept set becomes $\widetilde{P}_1 \cup \downarrow \mathcal{P}^{\circ}_j$.

Part (vii): full budget. For TCBM_b, every running total in phase 0 is at most $c^{\times}(R) \leq b$, so

$$\begin{aligned} Q &= R, \\ b' &\geq c(\mathcal{S} \setminus R). \end{aligned}$$

For TCB_b, $Q = \emptyset$ and $b' = b \geq c(\mathcal{S} \setminus R)$. In phases 1–2 every guard is then at most $c(\mathcal{S} \setminus R) \leq b'$, and every eligible statement is examined in phase 2, so $\widetilde{K}_b = \mathcal{S} \setminus R$. The renderer filters only by membership and keeps role and document order. It therefore prints $\mathcal{S} \setminus R$ as MEMORY, and for TCBM_b also R as RETRACTED, which is the TC context.

Part (viii): idempotence. Whether a statement conflicts depends only on its own claims and on $F(t_0)$. By (ii), no statement of $\widetilde{K}_b$ conflicts, so the recompilation has

$$D' = R' = \emptyset \qquad \text{(so no marker is shown)},$$
$$b \geq c(\widetilde{K}_b) \qquad \text{(by (iii))}.$$

Premises of $\widetilde{K}_b$ lie in $\widetilde{K}_b$, so rendered costs are unchanged. The argument of (vii) returns all of $\widetilde{K}_b$.

Part (ix): reduction. With $E = \emptyset$, $\downarrow s = \{s\}$. With $R = \emptyset$, $\mathcal{E}(R) = \mathcal{S}$ and phase 0 is empty. Phases 1 and 2 then coincide step by step with those of Proposition 2.

Closed prefix comparator. For the closed prefix, (i)–(iii) and (vii)–(viii) follow by the same arguments. For nestedness, let $b \leq b''$ and omit the markers. Up to the first block that does not fit at b, the scans at b and b'' see the same kept set and therefore the same blocks. Each of these blocks fits at b'' as well, so the selection at b is contained in the selection at b''. □

Scope. Maximizing total relevance subject to dependency closure and a budget is a precedence-constrained knapsack problem. It contains the 0–1 knapsack problem as the case $E = \emptyset$, so the greedy selector guarantees closure-maximality (iv), not optimal relevance. Part (x) has a consequence for the design of experiments. Under closed selection, local retraction and closure retraction yield the same memory and differ only in their markers. The contrast between direct retraction and dependency propagation must therefore be measured on full contexts, with the local retraction TL of Proposition 1(vi) against TC. The conditions B_b and PRE_b of Algorithm 1 remain ablations of selection without retraction, and BC_b differs from B_b only in keeping premises. TCB_b, TCBM_b and TCPRE_b are the composed pipeline to which Proposition 5 applies. Like TC, the joint conditions are deterministic functions of $X = (M, F(t_0), q)$, so Proposition 4 extends to them. The guarantees cover declared premises and matched numeric claims only. An undeclared dependency is not detected, and the text of a statement is not verified.

U REGISTERED CONFIRMATORY TESTS

Tables 27, 28, 29, 30, 35 and 36 list every test of the registered confirmatory families of the holdout forecast of Appendix E, the locked validation of Appendix G, the never-exposed-rider forecast of Appendix H, the large-cohort forecast of Appendix I, the later-cutoff study of Appendix J and the replication of the twelfth amendment; Tables 37, 39, 18 and 40 list the registered sensitivity families and descriptive analyses of the later-cutoff study. Tables 31 and 33 list the registered robustness criteria of the ninth amendment, Table 32 the fourth predeclared endpoint of the identity study, Table 34 the registered comparisons of the retraction and removal conditions of the follow-up run, and Table 41 each registered result stated in the main text next to its sensitivity analyses. Table 26 summarizes the registered protocol in the order in which it was fixed, and Tables 42 and 43 list the exceedance counts, the number of resamples, the side, the estimate and the Holm-adjusted value of every bootstrap test. An entry $0/B$ marks zero exceedances among the $B = 10{,}000$ resamples: the computed estimate and its Holm adjustment are 0, and the entry records the resampling resolution, not a bound on the tail probability, as Appendix K explains.

Registered protocol: summary and order. The registered protocol is a single append-only text. Its base text was written before any memory or forecast of the holdout study was generated, and each amendment was appended at the end of the text before the outputs it governs existed or were viewed. At every append the complete text before and after was recorded together with a check that the earlier text was unchanged. Table 26 summarizes each step in this order; the supplement contains the registered text of every step.

Protocol deviations. The later-window study retained its registered model, prompt, sample, window, outcomes and decision rules. Before evaluation, some semantic judgments were collected again item by item with the same deterministic settings; the registered analysis was unchanged. Its outcomes were extracted once after the registered earliest time, and it was scored once. The replication of the twelfth amendment was scored with the unchanged registered scorers on all completed forecasts, with the sensitivity sets of the fourteenth amendment; Table 36 lists its two tests.

Table 26: Steps of the registered protocol in the order in which they were appended, with the registered content and the outputs each step preceded.

Step	Registered content	Fixed before
Base	Holdout forecast of 295 riders, 27 of them holdout riders: conditions S, P, T, TC, TF, W, B_{1000}–B_{4000}, PRE_{2000}; offer log loss and refusal-fraction error; six primary tests with paired rider bootstrap and Holm; decision rules	any memory or forecast of the holdout study
Amendment 1	Parser correction for duplicated identifiers; conditions, endpoints and rules unchanged	any forecast job of the holdout study
Amendment 2	Holdout riders confirmatory; margins 0.020 and 0.010; eight-test non-inferiority family with fixed-sequence superiority; metric registry; declared rendering defect	scoring of any forecast and viewing of any outcome
Amendment 3	Follow-up run with corrected rendering and conditions T7, TC7, BC_{2000}, TCB_{2000}, TL, TCO, TDM; rendering-sensitivity rule; retraction and removal block	viewing or scoring of any holdout result
Amendment 4	Validation at five new cutoffs with 128 units of 46 riders: P, T7, TCO, TC7; eight superiority and eight non-inferiority tests	any memory or forecast of that validation
Amendment 5	789 never-exposed riders: TC7, TCO, TCN and P; four-test family	any memory or forecast of that run
Amendment 6	Cross-fitted combination of checked-memory forecasts with statistical baselines on the never-exposed riders; its descriptive re-analysis of the registered main run was carried out, and on 282 riders the combination of the statistical baselines with TC7 minus their combination alone is -0.0013 $[-0.0033, 0.0007]$ in offer log loss; the family on the never-exposed riders is outside the scope of this paper	any output of the amendment; scoring of the never-exposed riders
Amendment 7	Direct-only retraction TL on the never-exposed riders; outside the scope of this paper	any output of the amendment; scoring of the never-exposed riders
Amendment 8	Large cohort of 9,350 riders, 2,888 of them eligible, with development and confirmatory splits: M, N, S, STK; four-test family	extraction of any large-cohort outcome
Amendment 9	Point-index pairing of the identity study; single-rider influence; region clustering of the never-exposed riders	any output of these analyses
Amendment 10	Memory-interface check; rider-level distributions and practical threshold; resampling resolution	any output of these analyses
Amendment 11	Later cutoff with 2,441 riders and a two-day window: TC7, SR, TL, T7, S7 and two readout orders; non-LLM predictor NL fitted on the large-cohort development split; four-test family; model-judge check of 200 memories	extraction of any outcome of its window; any forecast beyond a 20-unit format check; any judgment; fitting of NL
Amendment 12	Replication with a second model family, Llama-3.1-8B-Instruct as writer and forecaster, on the units and window of Amendment 11: TC7, P7, T7; two-test family, listed in Table 36	any forecast of that replication beyond a 10-unit format check; reading of any outcome of the window
Amendment 13	Correction of the planning power figures of Amendment 12; tests and rules unchanged	collection or scoring of any output of Amendment 12
Amendment 14	Sensitivity sets of Amendments 11 and 12 computed as families with the margin, the doubled one-sided p, a Holm adjustment per method and the both-methods rule; joint judge rates formed from each judge's own events	any output of these analyses; collection of any output of Amendment 12

Table 27: Complete registered confirmatory family on the 22 holdout riders. d is the error of the first condition minus the second, with two-sided 95% paired rider-bootstrap intervals. Bootstrap with $B = 10{,}000$; Table 42 gives the counts. Superiority: Holm-adjusted two-sided p over the six tests. Non-inferiority: margins 0.020 and 0.010; each verdict is followed by the bootstrap and paired-t Holm-adjusted one-sided p over the eight tests; †superiority in fixed sequence.

	Offer log loss		Refusal-fraction error	
Test	d [95% CI]	p / verdict	d [95% CI]	p / verdict
T − P	-0.0677 $[-0.193, 0.051]$	0.54	-0.0382 $[-0.082, -0.007]$	0.0088
TC − P	-0.2196 $[-0.349, -0.099]$	0.0024	-0.0506 $[-0.102, -0.007]$	0.047
TC − T	-0.1519 $[-0.280, -0.045]$	0.0080	-0.0124 $[-0.045, 0.011]$	0.54
N1: TC vs. P	as above	non-inferior† $(0/B; 0.0056)$	as above	non-inferior† (0.0035; 0.082)
N2: TC vs. S	$+0.0489$ $[-0.076, 0.208]$	inconclusive(1.00; 1.00)	$+0.0024$ $[-0.009, 0.014]$	inconclusive(0.50; 0.54)
N3: B_{2000} vs. PRE_{2000}	-0.0094 $[-0.198, 0.154]$	inconclusive(1.00; 1.00)	$+0.0106$ $[0.002, 0.022]$	inconclusive(1.00; 1.00)
N4: B_{2000} vs. T	-0.1022 $[-0.265, 0.012]$	inconclusive(0.072; 0.34)	$+0.0176$ $[-0.005, 0.055]$	inconclusive(1.00; 1.00)

Table 28: Complete registered families of the locked validation on 37 riders with 105 units. d is the error of the first condition minus the second, with two-sided 95% paired rider-bootstrap intervals. Bootstrap with $B = 10{,}000$; Table 43 gives the counts. Superiority: Holm-adjusted two-sided p over eight tests from the bootstrap, next to its paired-t counterpart p_t. Non-inferiority: margins 0.020 for log loss and 0.010 for refusal error, Holm-adjusted one-sided p_{NI} over eight tests from the bootstrap, next to its paired-t counterpart.

Contrast	Endpoint	d [95% CI]	Superiority: p (p_t), verdict	Non-inferiority: p_{NI} ($p_{\mathrm{NI},t}$), verdict
TC7 − P	log loss	-0.1737 $[-0.290, -0.045]$	0.101 (0.077), not established	0.017 (0.013), non-inferior
	refusal error	-0.0150 $[-0.042, 0.010]$	1.000 (1.000), not established	0.108 (0.184), inconclusive
TC7 − T7	log loss	-0.0499 $[-0.170, 0.140]$	1.000 (1.000), not established	0.592 (0.618), inconclusive
	refusal error	-0.0032 $[-0.019, 0.010]$	1.000 (1.000), not established	0.106 (0.184), inconclusive
TCO − T7	log loss	-0.0078 $[-0.121, 0.173]$	1.000 (1.000), not established	0.609 (0.618), inconclusive
	refusal error	$+0.0057$ $[-0.011, 0.021]$	1.000 (1.000), not established	0.609 (0.618), inconclusive
TC7 − TCO	log loss	-0.0421 $[-0.077, -0.012]$	0.016 (0.119), superior	$0/B$ (0.0029), non-inferior
	refusal error	-0.0089 $[-0.021, 0.0009]$	0.493 (0.774), not established	$0/B$ (0.0077), non-inferior

Table 29: Complete registered confirmatory family of the never-exposed-rider forecast on 789 riders. d is the error of TC7 minus that of the second condition, with two-sided 95% paired rider-bootstrap intervals. Bootstrap with $B = 10{,}000$; Table 43 gives the counts. Holm-adjusted two-sided p over the four tests from the bootstrap, next to its paired-t counterpart p_t; verdict by the registered rule.

Contrast	Endpoint	d [95% CI]	p (p_t)	Verdict
TC7 − TCO	log loss	-0.0027 $[-0.0146, 0.0099]$	0.764 (0.764)	not established
	refusal error	-0.0013 $[-0.0044, 0.0016]$	0.764 (0.764)	not established
TC7 − TCN	log loss	0.0079 $[-0.0031, 0.0207]$	0.612 (0.643)	not established
	refusal error	-0.0017 $[-0.0042, 0.0006]$	0.612 (0.643)	not established

Table 30: Complete registered confirmatory family of the large-cohort forecast on the 1,998 riders of the confirmatory split. d is the error of the first forecaster minus that of the second, with two-sided 95% paired rider-bootstrap intervals. Bootstrap with $B = 10{,}000$; Table 43 gives the counts. Holm-adjusted two-sided p over the four tests from the bootstrap, next to its paired-t counterpart p_t; verdict by the registered rule.

Contrast	Endpoint	d [95% CI]	p (p_t)	Verdict
M − N	log loss	0.0056 $[-0.0014, 0.0144]$	0.272 (0.318)	not established
	refusal error	-0.0005 $[-0.0026, 0.0019]$	0.702 (0.688)	not established
STK − S	log loss	-0.0031 $[-0.0064, 0.0001]$	0.181 (0.187)	not established
	refusal error	-0.0055 $[-0.0080, -0.0029]$	$0/B$ (8.9×10^{-5})	superior

Table 31: Registered robustness criteria of the ninth amendment, identity and holdout studies. Identity: registered estimator, threshold and decision rule under point-index pairing, described in Appendix B. Holdout: TC−P with the most favorable rider removed, leaving 21 riders; Holm-adjusted superiority p over six tests and non-inferiority p_{NI} over eight tests.

Contrast	Endpoint	Estimate [95% CI]	Tests	Criterion
Personal − population	log loss	-0.0892 $[-0.1132, -0.0532]$	threshold -0.020	met
Personal − wrong identity	accuracy, pp	$+15.30$ $[10.54, 20.83]$	threshold $+5$	met
Personal − matched neighbor	accuracy, pp	$+8.98$ $[5.09, 13.09]$	threshold $+3$	met
TC − P, rider removed	log loss	-0.1875 $[-0.3026, -0.0738]$	p 0.0060; p_{NI} 0.0008	met
TC − P, rider removed	refusal error	-0.0347 $[-0.0765, 0.0001]$	p 0.152; p_{NI} 0.0126	not met

Table 32: Fourth predeclared endpoint of the identity study under the fixed-prefix protocol, with timestamp pairing of the interim analysis on 316 riders. The comparator is the non-LLM shrinkage predictor of the protocol, which shrinks the rider's decayed acceptance rate toward the population rate. Interval: seven-day block bootstrap with the predeclared three-comparison correction; criterion: upper limit below the margin.

Endpoint	Contrast	Estimate [corrected CI]	Margin	Criterion
S3, log loss	Personal context − shrinkage predictor	$+0.0227$ $[0.0073, 0.0519]$	0.020	not met

Table 33: Registered robustness criteria of the ninth amendment, never-exposed riders, 789 riders in 124 regions. d is the error of the rendering minus that of P. Region: 95% region-bootstrap interval; CR1: two-sided p of the cluster-robust t test; rider and region removed: 95% intervals without the most favorable rider or region; equal weight: estimate and interval with equal region weights, reported but not part of the criterion.

Contrast	Endpoint	d	Region	CR1 p	Rider removed	Region removed	Equal weight
TC7 − P	log loss	-0.1239	$[-0.1551, -0.0930]$	1.4×10^{-12}	$[-0.1522, -0.0873]$	$[-0.1497, -0.0891]$	-0.1204 $[-0.1712, -0.0759]$
TCO − P	log loss	-0.1212	$[-0.1509, -0.0927]$	1.2×10^{-13}	$[-0.1481, -0.0864]$	$[-0.1439, -0.0892]$	-0.1194 $[-0.1693, -0.0760]$
TCN − P	log loss	-0.1318	$[-0.1636, -0.1016]$	4.3×10^{-14}	$[-0.1585, -0.0976]$	$[-0.1548, -0.0975]$	-0.1224 $[-0.1746, -0.0759]$
TC7 − P	refusal error	-0.0072	$[-0.0131, -0.0020]$	0.011	$[-0.0117, -0.0015]$	$[-0.0118, -0.0012]$	-0.0065 $[-0.0184, 0.0050]$
TCO − P	refusal error	-0.0059	$[-0.0112, -0.0014]$	0.018	$[-0.0101, -0.0006]$	$[-0.0097, -0.0006]$	-0.0053 $[-0.0178, 0.0063]$
TCN − P	refusal error	-0.0055	$[-0.0114, -0.0004]$	0.049	$[-0.0098, 0.0000]$	$[-0.0100, 0.0004]$	-0.0047 $[-0.0169, 0.0067]$

Criterion: met for the first five rows; not met for TCN − P on refusal error.

Table 34: Registered comparisons of the retraction and removal conditions of the follow-up run, registered as a supplementary block of the third amendment; Table 12 gives the mean errors. d is the error of the first condition minus the second, with two-sided 95% paired rider-bootstrap intervals; p_{H}: Holm-adjusted two-sided p within the block and group. Registered reading: the first condition is better on an endpoint when the interval lies below zero.

		Holdout, $n = 22$			All, $n = 282$		
Contrast	Endpoint	d [95% CI]	p_{H}	Interval	d [95% CI]	p_{H}	Interval
TC7 − TL	log loss	-0.0364 $[-0.1468, 0.0654]$	1.000	contains 0	-0.0195 $[-0.0465, 0.0051]$	0.648	contains 0
	refusal error	0.0118 $[0.0005, 0.0322]$	0.086	above 0	0.0021 $[-0.0004, 0.0046]$	0.620	contains 0
TC7 − TCO	log loss	-0.0101 $[-0.0342, 0.0073]$	1.000	contains 0	-0.0361 $[-0.0626, -0.0113]$	0.032	below 0
	refusal error	-0.0121 $[-0.0686, 0.0290]$	1.000	contains 0	0.0008 $[-0.0044, 0.0050]$	1.000	contains 0
TC7 − TDM	log loss	0.0153 $[-0.0700, 0.1083]$	1.000	contains 0	-0.0477 $[-0.0991, -0.0050]$	0.167	below 0
	refusal error	0.0013 $[-0.0075, 0.0101]$	1.000	contains 0	-0.0012 $[-0.0078, 0.0043]$	1.000	contains 0
TCO − TDM	log loss	0.0253 $[-0.0425, 0.1097]$	1.000	contains 0	-0.0116 $[-0.0637, 0.0340]$	1.000	contains 0
	refusal error	0.0134 $[-0.0282, 0.0699]$	1.000	contains 0	-0.0020 $[-0.0094, 0.0047]$	1.000	contains 0

Table 35: Complete registered family of the later-cutoff study of Appendix J, registered in the eleventh amendment. d is the error of the first condition minus the second, with two-sided 95% paired rider-bootstrap intervals and $B = 10{,}000$. k: bootstrap exceedances, which for superiority are the smaller count of resampled means at or below and at or above zero and for non-inferiority the resampled means at or above the margin. Raw p: two-sided bootstrap estimate $\min(1, 2k/B)$, the non-inferiority one-sided estimate entering the family doubled as registered, with the paired-t counterpart in parentheses. Holm: adjustment over the four tests, separately for the bootstrap and the t test. Verdict by the registered rule, which requires both methods. T3 uses the confirmatory split of the large cohort.

Test	Contrast	Endpoint	Rule	n	d [95% CI]	k/B	Raw p (p_t)	Holm p (p_t)	Verdict
T1	TC7 − SR	log loss	sup.	1,934	-0.0139 $[-0.0260, -0.0026]$	83/10,000	0.0166 (0.0193)	0.0498 (0.0579)	not established
T2	TC7 − SR	refusal error	NI, 0.010	1,934	-0.0020 $[-0.0041, 0.0000]$	0/10,000	$0/B(< 10^{-6})$	$0/B(< 10^{-6})$	non-inferior
T3	STK − NL	refusal error	sup.	1,998	0.0001 $[-0.0019, 0.0022]$	4452/10,000	0.890 (0.898)	0.890 (0.898)	not established
T4	TC7b − S7	log loss	sup.	1,934	0.0042 $[-0.0043, 0.0130]$	1643/10,000	0.329 (0.331)	0.657 (0.661)	not established

Table 36: Complete registered family of the replication of the twelfth amendment, with Llama-3.1-8B-Instruct as memory writer and forecaster on the 1,934 evaluable riders and the two-day window of the later-cutoff study. Columns as in Table 35; Holm adjustment over the two tests, separately for the bootstrap and the t test; verdict by the registered rule, which requires both methods.

Test	Contrast	Endpoint	Rule	n	d [95% CI]	k/B	Raw p (p_t)	Holm p (p_t)	Verdict
U1	TC7 − P7	log loss	sup.	1,934	$0.1542\ [0.1205, 0.1891]$	0/10,000	$0/B(9.3 \times 10^{-18})$	$0/B(1.9 \times 10^{-17})$	not established
U2	TC7 − P7	refusal error	NI, 0.010	1,934	$0.0015\ [-0.0023, 0.0053]$	0/10,000	$0/B(8.9 \times 10^{-6})$	$0/B(8.9 \times 10^{-6})$	non-inferior

Table 37: Registered sensitivity families of the later-cutoff study, registered as item A11.9 of the eleventh amendment and computed as specified by the fourteenth amendment, which was appended before this computation. Each set recomputes the four-test family with the registered tests: paired rider bootstrap with $B = 10{,}000$ and paired t test, non-inferiority margin 0.010 with the one-sided p doubled, a separate Holm adjustment for each method, and a verdict only when both methods agree. For each test, a removal set drops the k riders with the most negative difference on that test's contrast; the complete case keeps the riders whose forecasts in both conditions parsed completely. T3, on the confirmatory split of the large cohort, enters every set with its registered values. The scorer of the eleventh amendment reported these sets as separate unadjusted contrasts, listed in Table 41. In all four sensitivity sets, the verdicts on T1–T4 agree with those of the primary family.

Set	Test	Contrast	Endpoint, rule	n	d [95% CI]	k/B	Holm p (p_t)	Verdict
1 most favorable removed	T1	TC7 − SR	log loss, sup.	1,933	$-0.0118\ [-0.0229, -0.0009]$	157/10,000	0.0942 (0.102)	not established
	T2	TC7 − SR	refusal error, NI 0.010	1,933	$-0.0017\ [-0.0036, 0.0003]$	0/10,000	$0/B(< 10^{-6})$	non-inferior
	T3	STK − NL	refusal error, sup.	1,998	$0.0001\ [-0.0019, 0.0022]$	–	0.890 (0.898)	not established
	T4	TC7b − S7	log loss, sup.	1,933	$0.0065\ [-0.0001, 0.0145]$	277/10,000	0.111 (0.157)	not established
3 most favorable removed	T1	TC7 − SR	log loss, sup.	1,931	$-0.0080\ [-0.0175, 0.0018]$	511/10,000	0.204 (0.202)	not established
	T2	TC7 − SR	refusal error, NI 0.010	1,931	$-0.0011\ [-0.0029, 0.0008]$	0/10,000	$0/B(< 10^{-6})$	non-inferior
	T3	STK − NL	refusal error, sup.	1,998	$0.0001\ [-0.0019, 0.0022]$	–	0.890 (0.898)	not established
	T4	TC7b − S7	log loss, sup.	1,931	$0.0079\ [0.0017, 0.0157]$	41/10,000	0.0246 (0.0822)	not established
5 most favorable removed	T1	TC7 − SR	log loss, sup.	1,929	$-0.0049\ [-0.0133, 0.0038]$	1337/10,000	0.535 (0.523)	not established
	T2	TC7 − SR	refusal error, NI 0.010	1,929	$-0.0006\ [-0.0023, 0.0011]$	0/10,000	$0/B(< 10^{-6})$	non-inferior
	T3	STK − NL	refusal error, sup.	1,998	$0.0001\ [-0.0019, 0.0022]$	–	0.890 (0.898)	not established
	T4	TC7b − S7	log loss, sup.	1,929	$0.0089\ [0.0027, 0.0165]$	5/10,000	0.0030 (0.0334)	not established
complete case	T1	TC7 − SR	log loss, sup.	1,928	$-0.0140\ [-0.0258, -0.0026]$	69/10,000	0.0414 (0.0579)	not established
	T2	TC7 − SR	refusal error, NI 0.010	1,928	$-0.0020\ [-0.0040, 0.0001]$	0/10,000	$0/B(< 10^{-6})$	non-inferior
	T3	STK − NL	refusal error, sup.	1,998	$0.0001\ [-0.0019, 0.0022]$	–	0.890 (0.898)	not established
	T4	TC7b − S7	log loss, sup.	1,933	$0.0042\ [-0.0043, 0.0130]$	1634/10,000	0.654 (0.662)	not established

Table 38: Later-cutoff study: mean errors of every condition on the 1,934 riders of the analysis set. S and S28 are the non-LLM predictors of the large-cohort study.

Condition	Offer log loss	Refusal-fraction error	Brier
S	0.2322	0.0631	0.0713
S28	0.2281	0.0607	0.0696
S7	0.2446	0.0585	0.0719
SR	0.2607	0.0622	0.0745
T7	0.2708	0.0632	0.0764
TL	0.2505	0.0603	0.0730
TC7	0.2468	0.0602	0.0722
T7b	0.2559	0.0603	0.0738
TC7b	0.2488	0.0605	0.0724

Table 39: Registered descriptive contrasts of the later-cutoff study on 1,934 riders. d is the error of the first condition minus the second, with 95% paired rider-bootstrap intervals from 2,000 resamples.

Contrast	Offer log loss	Refusal-fraction error	Brier
TC7 − TL	-0.0037 $[-0.0146, 0.0065]$	-0.0001 $[-0.0016, 0.0012]$	-0.0008 $[-0.0030, 0.0014]$
TC7 − T7	-0.0240 $[-0.0380, -0.0105]$	-0.0030 $[-0.0057, -0.0007]$	-0.0042 $[-0.0072, -0.0014]$
TL − T7	-0.0203 $[-0.0319, -0.0086]$	-0.0029 $[-0.0050, -0.0009]$	-0.0034 $[-0.0059, -0.0010]$
SR − T7	-0.0101 $[-0.0192, -0.0023]$	-0.0011 $[-0.0027, 0.0004]$	-0.0019 $[-0.0038, -0.0004]$
TL − SR	-0.0102 $[-0.0201, -0.0014]$	-0.0018 $[-0.0037, 0.0000]$	-0.0015 $[-0.0034, 0.0004]$
TC7 − SR	-0.0139 $[-0.0259, -0.0026]$	-0.0020 $[-0.0041, 0.0000]$	-0.0022 $[-0.0047, 0.0001]$
TC7 − S7	0.0023 $[-0.0071, 0.0118]$	0.0017 $[0.0002, 0.0034]$	0.0003 $[-0.0018, 0.0024]$
SR − S7	0.0162 $[0.0022, 0.0307]$	0.0037 $[0.0015, 0.0062]$	0.0025 $[-0.0004, 0.0055]$
T7 − S7	0.0263 $[0.0111, 0.0417]$	0.0048 $[0.0022, 0.0076]$	0.0045 $[0.0014, 0.0077]$
TC7b − TC7	0.0020 $[-0.0054, 0.0102]$	0.0003 $[-0.0016, 0.0020]$	0.0001 $[-0.0015, 0.0019]$
T7b − T7	-0.0149 $[-0.0258, -0.0053]$	-0.0030 $[-0.0052, -0.0009]$	-0.0026 $[-0.0048, -0.0007]$
TC7b − T7b	-0.0071 $[-0.0180, 0.0042]$	0.0002 $[-0.0018, 0.0021]$	-0.0014 $[-0.0037, 0.0008]$
T7b − S7	0.0113 $[-0.0015, 0.0247]$	0.0018 $[-0.0001, 0.0038]$	0.0019 $[-0.0008, 0.0046]$
TC7b − S7	0.0042 $[-0.0043, 0.0132]$	0.0020 $[0.0005, 0.0036]$	0.0004 $[-0.0013, 0.0023]$
TC7 − S	0.0146 $[0.0071, 0.0232]$	-0.0029 $[-0.0051, -0.0008]$	0.0009 $[-0.0010, 0.0029]$
TC7 − S28	0.0187 $[0.0105, 0.0280]$	-0.0005 $[-0.0023, 0.0014]$	0.0026 $[0.0008, 0.0045]$
S7 − S	0.0123 $[0.0025, 0.0233]$	-0.0046 $[-0.0068, -0.0025]$	0.0006 $[-0.0018, 0.0031]$
S28 − S	-0.0041 $[-0.0079, -0.0004]$	-0.0024 $[-0.0036, -0.0012]$	-0.0017 $[-0.0027, -0.0006]$

Table 40: Registered descriptive comparisons with the non-LLM predictor NL on the 1,998 riders of the confirmatory split of the large cohort; NL is fitted on the 416 development riders only. d is the error of the first forecaster minus the second, with 95% paired rider-bootstrap intervals from 2,000 resamples; the last rows give the mean errors.

Contrast	Offer log loss	Refusal-fraction error	Brier
STK − NL	0.0093 $[0.0054, 0.0131]$	0.0001 $[-0.0018, 0.0022]$	0.0026 $[0.0013, 0.0038]$
M − NL	0.0160 $[0.0086, 0.0247]$	0.0001 $[-0.0018, 0.0022]$	0.0030 $[0.0011, 0.0051]$
N − NL	0.0104 $[0.0050, 0.0162]$	0.0006 $[-0.0010, 0.0022]$	0.0018 $[0.0003, 0.0033]$
$\mathrm{STK_N}$ − NL	0.0079 $[0.0040, 0.0118]$	0.0006 $[-0.0009, 0.0022]$	0.0021 $[0.0009, 0.0033]$
NL − S	-0.0124 $[-0.0175, -0.0076]$	-0.0056 $[-0.0075, -0.0038]$	-0.0040 $[-0.0059, -0.0023]$
NL − S28	-0.0042 $[-0.0075, -0.0009]$	-0.0012 $[-0.0022, -0.0001]$	-0.0014 $[-0.0027, -0.0002]$
STK − S28	0.0051 $[0.0014, 0.0086]$	-0.0010 $[-0.0030, 0.0010]$	0.0012 $[-0.0001, 0.0024]$
STK − $\mathrm{STK_N}$	0.0014 $[-0.0010, 0.0038]$	-0.0005 $[-0.0026, 0.0017]$	0.0005 $[-0.0005, 0.0014]$
STK − S	-0.0031 $[-0.0064, 0.0000]$	-0.0055 $[-0.0080, -0.0030]$	-0.0014 $[-0.0028, -0.0002]$
S, mean	0.2700	0.0621	0.0834
S28, mean	0.2618	0.0576	0.0808
NL, mean	0.2576	0.0565	0.0794
M, mean	0.2736	0.0566	0.0824
N, mean	0.2680	0.0571	0.0812
STK, mean	0.2669	0.0566	0.0819
$\mathrm{STK_N}$, mean	0.2655	0.0571	0.0814

Table 41: Registered results stated in the main text, next to their sensitivity analyses. Registered test: the Holm-adjusted decision statistic of the registered family; p_t: the paired-t counterpart with the same correction. Sensitivity analyses are those registered in the same or a later amendment.

Result	Registered test	p_t	Sensitivity analyses
Holdout, TC − P, log loss, $n = 22$	superiority p 0.0024; non-inferiority p_{NI} $0/B$	$p_{\mathrm{NI},t}$ 0.0056	corrected rendering: TC7 − P Holm p 0.015; most favorable rider removed: p 0.0060, p_{NI} 0.0008; 16 of 22 riders below -0.020
Holdout, TC − P, refusal error, $n = 22$	superiority p 0.047; non-inferiority p_{NI} 0.0035	$p_{\mathrm{NI},t}$ 0.082	corrected rendering: TC7 − P Holm p 0.028; most favorable rider removed: p 0.152, p_{NI} 0.0126; 9 of 22 riders below -0.010
New cutoffs, TC7 − P, log loss, 37 riders	non-inferiority p_{NI} 0.017	$p_{\mathrm{NI},t}$ 0.013	superiority in the same family: p 0.101, p_t 0.077
Large cohort, STK − S, refusal error, $n = 1{,}998$	superiority: 0 of $B = 10{,}000$ resamples at or above 0	8.9×10^{-5}	10^6 resamples: 7 at or above 0, p 6.4×10^{-5}; complete-case analysis identical; development split -0.0057 $[-0.0113, -0.0010]$; negative point estimate in 4 of 4 cities and 4 of 4 memory-length quartiles; 336 of 1,998 riders below -0.010
Later cutoff, TC7 − SR, refusal error, $n = 1{,}934$	non-inferiority at 0.010: 0 of $B = 10{,}000$ resamples at or above the margin, Holm $0/B$	$< 10^{-6}$	complete case with $n = 1{,}928$: -0.0020 $[-0.0040, 0.0001]$; 1, 3 and 5 most favorable riders removed: -0.0017 $[-0.0035, 0.0004]$, -0.0011 $[-0.0028, 0.0008]$, -0.0006 $[-0.0023, 0.0011]$; every upper limit below the margin; registered families with these removals and on complete cases in Table 37: non-inferior in each

Table 42: Bootstrap test details, part 1: every bootstrap test whose p value appears in Tables 2, 27 and 31; with part 2 it covers Table 41. k: exceedances among the B paired rider resamples of the mean difference. Two-sided: k is the smaller of the numbers of resampled means at or below and at or above zero, estimate $\min(1, 2k/B)$. One-sided non-inferiority (NI): k resampled means at or above the margin, estimate k/B. One-sided superiority in fixed sequence: k resampled means at or above zero, estimate k/B, tested at 0.025 without Holm adjustment. m: tests in the Holm family; Holm: step-down adjustment of the estimates. $0/B$: $k = 0$, where the estimate and its Holm adjustment are both 0; the entry records the resampling resolution and is not a deterministic upper bound on the tail probability. Rider removed: the Holm values are those of the robustness analysis, which reran the six- and eight-test families on the 21 remaining riders; the other members of these two families are not among the released aggregates.

Test	Endpoint	Side	k	B	Estimate	m	Holm
Holdout: T − P	log loss	two	1,350	10,000	0.2700	6	0.5400
	refusal error	two	11	10,000	0.0022	6	0.0088
Holdout: TC − P	log loss	two	2	10,000	0.0004	6	0.0024
	refusal error	two	78	10,000	0.0156	6	0.0468
Holdout: TC − T	log loss	two	8	10,000	0.0016	6	0.0080
	refusal error	two	2,052	10,000	0.4104	6	0.5400
Holdout N1: TC vs. P	log loss	one, NI	0	10,000	$0/B$	8	$0/B$
	refusal error	one, NI	5	10,000	0.0005	8	0.0035
Holdout N2: TC vs. S	log loss	one, NI	6,236	10,000	0.6236	8	1.0000
	refusal error	one, NI	1,007	10,000	0.1007	8	0.5035
Holdout N3: B_{2000} vs. PRE_{2000}	log loss	one, NI	3,879	10,000	0.3879	8	1.0000
	refusal error	one, NI	5,067	10,000	0.5067	8	1.0000
Holdout N4: B_{2000} vs. T	log loss	one, NI	120	10,000	0.0120	8	0.0720
	refusal error	one, NI	6,492	10,000	0.6492	8	1.0000
Holdout N1, fixed sequence	log loss	one, sup.	2	10,000	0.0002	—	—
	refusal error	one, sup.	78	10,000	0.0078	—	—
Holdout TC − P, rider removed	log loss	two	5	10,000	0.0010	6	0.0060
	refusal error	two	254	10,000	0.0508	6	0.1524
	log loss	one, NI	1	10,000	0.0001	8	0.0008
	refusal error	one, NI	18	10,000	0.0018	8	0.0126
All riders: TC − P	log loss	two	0	10,000	$0/B$	6	$0/B$
	refusal error	two	3	10,000	0.0006	6	0.0018
Follow-up, holdout: TC7 − P	log loss	two	15	10,000	0.0030	6	0.0150
	refusal error	two	35	10,000	0.0070	6	0.0280
Follow-up, all: TC7 − P	log loss	two	0	10,000	$0/B$	6	$0/B$
	refusal error	two	36	10,000	0.0072	6	0.0216

Table 43: Bootstrap test details, part 2: the tests of Tables 28, 29, 30 and 34, with the definitions of Table 42.

Test	Endpoint	Side	k	B	Estimate	m	Holm
Follow-up, holdout: TC7 − TL	log loss	two	2,489	10,000	0.4978	8	1.0000
	refusal error	two	54	10,000	0.0108	8	0.0864
Follow-up, holdout: TC7 − TCO	log loss	two	1,775	10,000	0.3550	8	1.0000
	refusal error	two	3,612	10,000	0.7224	8	1.0000
Follow-up, holdout: TC7 − TDM	log loss	two	3,715	10,000	0.7430	8	1.0000
	refusal error	two	3,733	10,000	0.7466	8	1.0000
Follow-up, holdout: TCO − TDM	log loss	two	2,748	10,000	0.5496	8	1.0000
	refusal error	two	3,261	10,000	0.6522	8	1.0000
Follow-up, all: TC7 − TL	log loss	two	648	10,000	0.1296	8	0.6480
	refusal error	two	517	10,000	0.1034	8	0.6204
Follow-up, all: TC7 − TCO	log loss	two	20	10,000	0.0040	8	0.0320
	refusal error	two	3,448	10,000	0.6896	8	1.0000
Follow-up, all: TC7 − TDM	log loss	two	119	10,000	0.0238	8	0.1666
	refusal error	two	3,811	10,000	0.7622	8	1.0000
Follow-up, all: TCO − TDM	log loss	two	3,250	10,000	0.6500	8	1.0000
	refusal error	two	3,000	10,000	0.6000	8	1.0000
New cutoffs: TC7 − P	log loss	two	72	10,000	0.0144	8	0.1008
	refusal error	two	1,339	10,000	0.2678	8	1.0000
New cutoffs: TC7 − T7	log loss	two	2,484	10,000	0.4968	8	1.0000
	refusal error	two	3,461	10,000	0.6922	8	1.0000
New cutoffs: TCO − T7	log loss	two	4,103	10,000	0.8206	8	1.0000
	refusal error	two	2,256	10,000	0.4512	8	1.0000
New cutoffs: TC7 − TCO	log loss	two	10	10,000	0.0020	8	0.0160
	refusal error	two	411	10,000	0.0822	8	0.4932
New cutoffs: TC7 vs. P	log loss	one, NI	28	10,000	0.0028	8	0.0168
	refusal error	one, NI	270	10,000	0.0270	8	0.1080
New cutoffs: TC7 vs. T7	log loss	one, NI	1,972	10,000	0.1972	8	0.5916
	refusal error	one, NI	212	10,000	0.0212	8	0.1060
New cutoffs: TCO vs. T7	log loss	one, NI	3,044	10,000	0.3044	8	0.6088
	refusal error	one, NI	3,054	10,000	0.3054	8	0.6088
New cutoffs: TC7 vs. TCO	log loss	one, NI	0	10,000	$0/B$	8	$0/B$
	refusal error	one, NI	0	10,000	$0/B$	8	$0/B$
New riders: TC7 − TCO	log loss	two	3,299	10,000	0.6598	4	0.7640
	refusal error	two	1,910	10,000	0.3820	4	0.7640
New riders: TC7 − TCN	log loss	two	887	10,000	0.1774	4	0.6120
	refusal error	two	765	10,000	0.1530	4	0.6120
Large cohort: M − N	log loss	two	679	10,000	0.1358	4	0.2716
	refusal error	two	3,512	10,000	0.7024	4	0.7024
Large cohort: STK − S	log loss	two	301	10,000	0.0602	4	0.1806
	refusal error	two	0	10,000	$0/B$	4	$0/B$