# Does Conditional Memory Help an LLM Inventory Agent under Regime Shifts?
# A Systems Evaluation Frozen before Test-Outcome Access

Fengkai Gao
Artificial Intelligence Programme, Faculty of Science and Technology
Beijing Normal-Hong Kong Baptist University
t330034007@mail.bnbu.edu.cn

## Abstract

Long-running language-model agents may reuse experience after operating conditions change. Change detection alone, however, cannot determine whether a stored strategy revision remains useful when its consequences are observed only after execution. We introduce ShiftMem, a conditional memory lifecycle that uses public change signals to trigger review, an exact quoted-lead-time condition to gate retrieval, and delayed operational outcomes to update confidence and lifecycle state. The LLM is restricted to bounded strategy revisions, while a shared deterministic controller retains responsibility for daily actions. We evaluated ShiftMem against an implemented lexical retrieval baseline in a paired study of a synthetic single-item inventory system frozen before access to Test outcomes. The evaluation spanned two provider-hosted model deployments, eight held-out scenarios, and 160 completed deployment–method–scenario–seed runs. Across 70 model-specific pairs nested in 35 shared environment clusters, the prespecified but dependence-limited decision rule did not support a ShiftMem advantage. The mean cost-gap difference (ShiftMem minus baseline) was +45.44 (negative values favor ShiftMem; nominal 95% mean CI $[-2.72, 93.60]$); the selected two-sided Wilcoxon test, which does not test the mean directly, gave $p = 0.203$. A post-hoc clustered sensitivity retained the same unfavorable mean estimate, with a 95% cluster-bootstrap CI of $[11.26, 79.09]$. Variation across deployments and test regimes remained descriptive. These findings establish neither equivalence nor the intrinsic ineffectiveness of the conditional memory lifecycle. Instead, within this testbed, they underscore that conditional memory benefits should be demonstrated rather than inferred from architecture alone. The study provides an auditable implementation and evaluation record for testing memory reuse under delayed feedback while retaining provider and structured-output failures as part of the evaluated system.

## 1 Introduction

Long-running language-model agents may rely on memories formed under conditions that no longer hold. Stored observations and reflections support persistent simulated agents [10], verbal feedback can guide later trials [13], and trajectory-level lessons can transfer across tasks [18]. Yet an experience that was useful when recorded can become stale as operating conditions change. This risk is especially consequential when the memory recommends an operational strategy and its outcome is observable only after delayed execution. The resulting question is not simply what an agent should remember, but when it should continue to trust its past.

Existing research addresses parts of this problem without resolving the operational decision studied here. Long-term memory benchmarks test temporal reasoning and knowledge updates [7, 17]. Concept-drift research distinguishes change detection from the adaptation that follows [5], while nonstationary inventory research develops control under changing or partially observed demand [2, 15]. Together, these literatures address persistence, change detection, and adaptive control, but do not establish whether an explicit memory lifecycle reduces downstream cost under delayed feedback. A detector can indicate that public observations changed. It cannot by itself establish that a retrieved experience is invalid or that suspending it will reduce downstream cost.

We study this question in a synthetic single-item lost-sales inventory system, chosen as a controlled testbed rather than a model of enterprise deployment. ShiftMem treats a change in public context as a prompt for review rather than proof of invalidity, and uses delayed outcomes to update memory confidence and state. Page–Hinkley signals provide the change trigger [9]. The LLM is confined to bounded strategy revisions, while a deterministic controller issues daily orders. We compare ShiftMem with a lexical retrieval baseline under paired demand and supply streams. This design compares the end-to-end systems. It does not isolate the effect of any lifecycle component.

The prespecified evaluation covered two provider-hosted model deployments, eight held-out scenarios, and 160

completed deployment–method–scenario–seed runs. Its dependence-limited primary decision rule did not support a ShiftMem advantage, and the mean point estimate was unfavorable. Because the 70 model-specific pairs shared 35 environment clusters, we also report a post-hoc clustered sensitivity, which retained the same unfavorable direction. Point estimates varied descriptively across test regimes and deployments without establishing generalizable effect modification. The lexical retrieval baseline always ran first, leaving method order and provider drift as potential confounders.

**Contributions.** This study makes three bounded contributions:

- It develops a controlled architecture that separates bounded LLM strategy review from deterministic daily execution, allowing memory-conditioned revision to be evaluated without changing the inventory controller.
- It provides a content-addressed evidence package with network-free reproducible integrity checks and deterministic post-outcome analyses. The package records provider and structured-output failures, fallback behavior, reservation journals, and source hashes.
- It reports the negative primary result with its uncertainty and dependence-aware sensitivity analyses, together with descriptive deployment heterogeneity and post-hoc reliability analyses.

# 2 Related Work

**External agent memory and long-term memory evaluation.** Language agents externalize experience in several forms: Generative Agents stores observations and higher-level reflections [10], Reflexion reuses verbal feedback from prior trials [13], and ExpeL extracts transferable insights from successful and failed trajectories [18]. Long-term memory benchmarks complement these systems by testing multi-session recall, event summarization, temporal reasoning, knowledge updates, and abstention [7, 17]. Together, this literature establishes that persistent experience can guide later behavior and that retrieval and updating require explicit evaluation. Our setting examines a complementary operational question: whether a previously validated strategy revision should still be reused after public demand or supply conditions have changed and outcome feedback arrives only after a delayed window. ShiftMem therefore evaluates memory through downstream inventory cost and recovery while making suspension, revalidation, and reactivation auditable, rather than treating recall accuracy alone as evidence of operational applicability.

**Concept drift and nonstationary inventory control.** Concept-drift detection and adaptive inventory control address different parts of decision making under change. ADWIN statistically compares subwindows [3], cumulative-sum procedures detect sustained process-level departures [9], and broader drift research separates detection from the adaptation strategy that follows [5]. Inventory research instead develops domain-specific control under changing or partially observed demand [2, 15], including lost-sales settings with censored demand and replenishment delay [20] and online control under unknown changing demand and lead times [1]. ShiftMem neither treats a detector as a validity oracle nor claims a new optimal inventory policy. A Page–Hinkley signal computed from public observations triggers an additional review and, when detector and experience variable names match, can return an active memory to probation. Stored lead-time conditions gate retrieval, while delayed outcomes update lifecycle confidence and state. Holding the bounded deterministic controller fixed allows an end-to-end comparison of the two implemented memory systems; it does not identify the effect of an individual lifecycle component.

**Operational LLM reliability and reproducible evaluation.** LLM inventory agents can coordinate or propose supply-chain decisions [12], but their behavior varies across models and uncertain inventory settings [19]. ShiftMem consequently separates language-model reasoning from routine numerical execution: the LLM reviews only three bounded strategy parameters, cannot issue daily orders or observe hidden regime and Oracle state, and falls back to the previous valid strategy after one failed schema-repair attempt. This reliability path remains part of the measured system because format restrictions can affect model performance [14], while grammar-constrained decoding treats output validity as a distinct systems problem [6]. The evaluation also follows work on pre-specification, reproducibility, and transparent reporting [4, 8, 11]. We use the narrower term *pre-Test-frozen*, not formal preregistration, and report the unsupported primary hypothesis with its effect estimate, uncertainty, heterogeneity, and post-hoc sensitivity analyses rather than reducing the result to a binary $p$-value conclusion [16]. The resulting contribution is an empirical test of a conditional memory lifecycle within a bounded operational review system, not a claim of universal memory superiority.

# 3 Method

## 3.1 Inventory Task

We use a finite-horizon, single-item lost-sales system with delayed and potentially partial replenishment. Let $I_t$ denote on-hand inventory before day-$t$ arrivals, $D_t$ realized demand, and $q_t \in \mathbb{Z}_{\geq 0}$ the order selected from the public state. An earlier order $i$ of size $q_i$, due on day $t$, arrives

as $A_{i,t} \sim \text{Binomial}(q_i, \rho_i)$ for its quoted fill probability $\rho_i$ (and equals $q_i$ when $\rho_i = 1$). The daily dynamics are

$$\begin{aligned} A_t &= \textstyle\sum_{i:\tau_i=t} A_{i,t}, \quad I_t^+ = I_t + A_t, \\ Y_t &= \min\{I_t^+, D_t\}, \quad L_t = D_t - Y_t, \\ I_{t+1} &= I_t^+ - Y_t, \\ C_t &= c_p q_t + c_h I_{t+1} + c_s L_t + c_f \mathbb{1}[q_t > 0]. \end{aligned} \tag{1}$$

Here $Y_t$, $L_t$, and $C_t$ are sales, lost sales, and operating cost. Demand follows a scenario-specific Poisson or negative-binomial distribution. Under the implementation's negative-binomial parameterization, mean $\mu_t$ and dispersion $r_t$ imply success probability $p_t = r_t/(r_t + \mu_t)$. Regime identity, future demand, supplier fill, and Oracle state are hidden. The public observation contains the day, current inventory, nominal pipeline inventory and due-day schedule, quoted lead time, last demand and sales, costs, and the latest 14 days of realized history.

## 3.2 Bounded Strategy Review and Daily Control

Given this task, ShiftMem treats memory reuse as bounded revision of a three-parameter ordering strategy. The LLM acts only at scheduled or change-triggered reviews, whereas a shared deterministic controller issues daily orders. Figure 1 summarizes this responsibility split and the subsequent stages of delayed validation, lifecycle management, and conditional retrieval.

The current strategy is $\boldsymbol{\theta}_t = (w_t, k_t, b_t)$: a demand forecast window, safety-stock multiplier, and lead-time buffer. Its initial value is $(14, 1.2, 1)$. At a review, the LLM receives the public observation, current strategy, trigger, and up to five retrieved experiences. It returns a structured proposal $\widetilde{\boldsymbol{\theta}}_t$, cited experience IDs, model-reported confidence, and a short reason. The cited IDs, model-reported confidence, and reason are retained in the review log but do not directly set evidence-derived lifecycle confidence. Unknown fields, malformed JSON, or citations to experiences not supplied in the prompt cause one correction attempt. A second failure retains the previous valid strategy.

Schema validation rejects proposals below the lower bounds. Remaining values are projected first to the absolute bounds $[1, 60] \times [0, 5] \times [0, 14]$ and then to per-review changes $\boldsymbol{\Delta} = (7, 1, 1)$:

$$\theta_{t,j}^{\text{new}} = \text{clip}\Big(\text{clip}(\widetilde{\theta}_{t,j}; \ell_j, u_j); \theta_{t,j} - \Delta_j, \theta_{t,j} + \Delta_j\Big). \tag{2}$$

The projection limits abrupt changes and prevents the LLM from controlling the action space directly.

The projected parameters determine daily orders through a deterministic order-up-to controller. It uses the most recent $\min(w_t, 14)$ available demands to compute their mean $\widehat{\mu}_t$ and sample standard deviation $\widehat{\sigma}_t$. With quoted lead time $\ell_t$, nominal pipeline inventory $P_t$, and protection horizon $h_t = \ell_t + b_t + 1$, it sets

$$S_t = \widehat{\mu}_t h_t + k_t \widehat{\sigma}_t \sqrt{h_t}, \tag{3}$$

$$q_t = \max\{0, \text{round}\,(S_t - I_t - P_t)\}. \tag{4}$$

Rounding follows the implementation's nearest-integer, ties-to-even rule. When no history exists, the reset value of last demand, zero, is used. All compared memory methods share Eqs. (3)–(4), the same strategy bounds, and the same demand and calendar-indexed supply streams.

## 3.3 Change-Triggered Review

A review occurs on days divisible by five, including initialization at day zero, or on the day after a detector signal. An event-triggered review requires more than three days since the previous review that included an event trigger. A purely periodic review neither starts nor resets this cooldown. Periodic and event triggers on the same day are coalesced into one call. Thus the model cannot increase its own review rate or alter the cooldown.

ShiftMem monitors public demand, lost sales, and quoted lead time with separate two-sided Page–Hinkley detectors [9]. For observations $x_n$ of one monitored variable, the running mean and one-sided accumulators are

$$\begin{aligned} \bar{x}_n &= \bar{x}_{n-1} + \tfrac{x_n - \bar{x}_{n-1}}{n}, \\ U_n &= U_{n-1} + x_n - \bar{x}_n - \delta, \\ G_n^+ &= U_n - \min_{i \le n} U_i, \\ V_n &= V_{n-1} + \bar{x}_n - x_n - \delta, \\ G_n^- &= V_n - \min_{i \le n} V_i. \end{aligned} \tag{5}$$

After at least 10 samples, a signal is emitted when either statistic exceeds $\lambda = 48$, with $\delta = 0.1$. Detector statistics reset after a signal. The signal is an operational trigger, not an estimate of latent regime identity or proof that a memory is invalid.

## 3.4 Delayed Evidence and Memory Lifecycle

The memory unit is a valid bounded strategy revision, not a daily order. Each valid review registers the new revision and separate reuse validations for its cited experiences. If the review occurs on day $r$ with quoted lead time $\ell_r$, validation starts at $a = r + \ell_r$ and uses the three complete days $\mathcal{W}_r = \{a, a+1, a+2\}$. Define

$$F_r = \frac{\sum_{t \in \mathcal{W}_r} Y_t}{\sum_{t \in \mathcal{W}_r} D_t}, \quad L_r = \sum_{t \in \mathcal{W}_r} L_t, \quad \bar{C}_r = \frac{1}{3} \sum_{t \in \mathcal{W}_r} C_t, \tag{6}$$

where $F_r = 1$ if total demand is zero. The deterministic verdict is

$$z_r = \begin{cases} \text{failure}, & F_r \le 0.80 \ \vee \ L_r \ge 1, \\ \text{support}, & F_r \ge 0.95 \ \wedge \ L_r = 0 \ \wedge \ \bar{C}_r \le 100, \\ \text{inconclusive}, & \text{otherwise.} \end{cases} \tag{7}$$

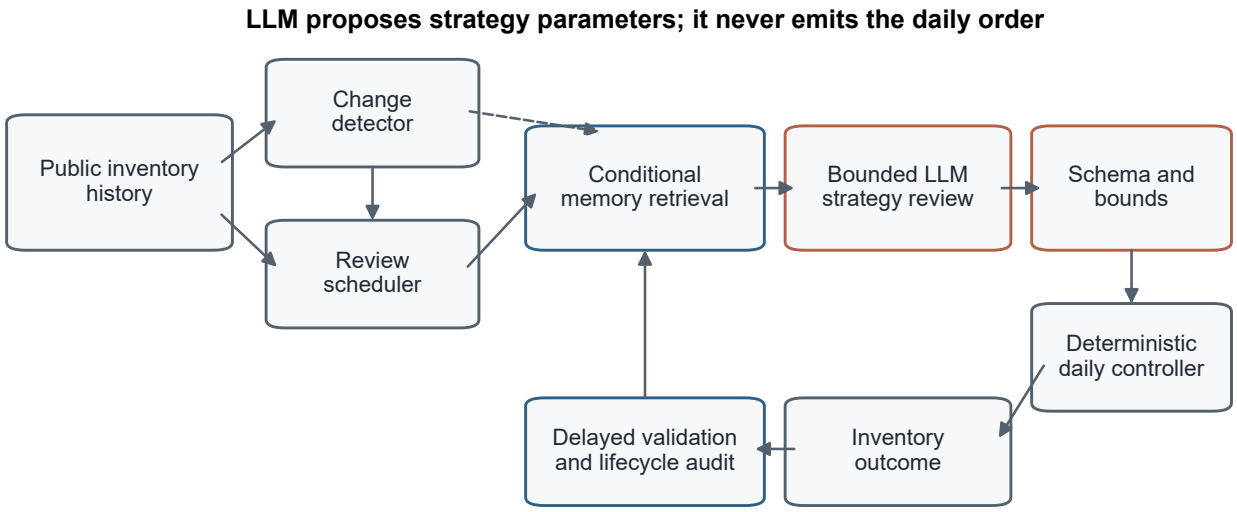


Figure 1: **Bounded strategy review and deterministic daily control.** Public inventory history drives the change detector and review scheduler. On an eligible day, method-specific memory is supplied to an LLM that may revise three bounded strategy parameters. Schema validation and deterministic projection precede the shared daily controller. Realized outcomes enter a delayed validation window and, for ShiftMem, an auditable lifecycle. The LLM neither emits daily orders nor observes hidden regime labels or Oracle state.

The window must be contiguous and complete. Because validation windows can overlap and subsequent orders affect their outcomes, the resulting verdicts and confidence values are operational lifecycle scores rather than causal estimates of memory validity.

ShiftMem augments each experience with state $s \in \{\text{probation}, \text{active}, \text{dormant}, \text{invalid}\}$, Beta evidence parameters initialized at $(\alpha, \beta) = (1, 1)$, and support and failure counts $(n_s, n_f)$. Its confidence and empirical utility are

$$c = \frac{\alpha}{\alpha + \beta}, \qquad u = \begin{cases} \dfrac{n_s}{n_s + n_f}, & n_s + n_f > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

Support increments $(\alpha, n_s)$ by $(1, 1)$. Failure increments $(\beta, n_f)$ by $(\omega, 1)$, where $\omega = 2$ only when a reused experience fails after a recorded related change and $\omega = 1$ otherwise. Inconclusive evidence leaves all four values unchanged. Detectors process each daily outcome before due validations. A related signal on that outcome therefore activates the post-change failure weight. A probationary experience becomes active after at least two supports with $c \geq 0.65$, and becomes invalid after at least two failures with $c \leq 0.35$. Failure returns an active experience to probation, whereas the invalid state is terminal.

Each formal experience has an exact quoted-lead-time applicability condition; demand context is not encoded as a stored applicability condition. Three consecutive mismatching retrieval checks move an active or probationary experience to dormant. A later condition match returns it to probation. These are retrieval checks rather than elapsed days. Every transition is appended to the lifecycle audit trail.

### 3.5 Conditional Retrieval and Lexical Retrieval Baseline

ShiftMem first excludes dormant, invalid, and condition-mismatched experiences. It then ranks each eligible experience $m$ using

$$\text{lexcos}(\xi, m) = \frac{\sum_v n_v(\xi) n_v(m)}{\sqrt{\sum_v n_v(\xi)^2}\sqrt{\sum_v n_v(m)^2}}, \tag{9}$$

$$\begin{aligned} R_t(m) = {} & 0.75\,\text{lexcos}(\xi_t, m) + 0.5 c_m \\ & + 0.5^{a_m/30} + 0.25 u_m \\ & - 0.25\mathbb{1}[s_m = \text{probation}] - 0.5\mathbb{1}[V_m \cap \Delta V_t \neq \varnothing], \end{aligned} \tag{10}$$

where $\xi_t$ serializes the public observation and lexcos scores this serialization against the generated text of experience $m$. The tokenizer retains each lowercase token matching `[a-z0-9_]+`, and $n_v(\cdot)$ is its count. $a_m$ is age in environment steps, $V_m$ is the experience-variable set, and $\Delta V_t$ contains variables signaled within the previous 14 steps. Ties are resolved by newer creation step and then ascending memory ID. The top five experiences are supplied to the reviewer.

The comparison method, retained in code as `VectorMemory`, is an implemented lexical retrieval baseline rather than an embedding model or vector database. It stores every completed valid revision and ranks generated record text only by Eq. (9), breaking ties by newer step and then descending memory ID. Its ranker uses no applicability filter, lifecycle state, evidence confidence, or change penalty. Consequently, the experiment compares the implemented conditional-lifecycle system with an implemented token-count cosine baseline under the same reviewer and controller. It does not isolate any single ShiftMem component.

### 3.6 Implementation and Comparison Boundaries

The frozen implementation has several boundaries relevant to interpretation. All three detector outputs can schedule reviews, but the model receives only a generic event trigger rather than the variable and direction that fired. Direct active-to-probation demotion and the changed-variable penalty require an exact variable-name match between detector outputs and experience records. Only demand satisfies this match. Quoted lead time instead controls applicability through an exact stored condition. Because the public history contains 14 days, $w_t > 14$ is behaviorally indistinguishable in the controller. A pending reuse verdict is not applied if its target becomes dormant or invalid before the verdict is due.

Every memory method receives the same completed strategy-revision record. The persisted record contains the projected strategy and validation metrics, while the review log separately retains raw model output, reasons, model-reported confidence, and cited experience IDs. Common record serialization also exposes baseline experiences to the LLM with constant probation status and confidence 0.5. These fields do not affect lexical ranking but remain part of the tested model-facing baseline. The primary comparison therefore concerns the two implemented systems as deployed, not a mechanism-level ablation.

## 4 Experiments

### 4.1 Frozen Evaluation Design

The evaluation used a paired design frozen before Test outcomes were accessed (hereafter, pre-Test-frozen). Protocol v2 and its first amendment fixed the models, methods, scenarios, seeds, controller, endpoint, and analysis rule. Three later amendments authorized continuation subject to balance constraints to complete the matrix after partial execution but did not change those design fields. The content-addressed evidence snapshot `f4ab41daacf3` contains all 160 planned cells. No cell was excluded, and no completed cell was reissued while creating the snapshot.

We evaluated ShiftMem and the lexical retrieval baseline with provider-hosted DeepSeek-V3.2 and MiniMax-M2.5 through SiliconFlow. Test-ID comprised stable, demand-jump, gradual-demand, and supply-delay scenarios. Test-OOD comprised periodic recurrence, a Poisson demand jump, a transient demand excursion, and an early combined demand-and-supply shift. Each scenario used five environment seeds and 150 simulated days. Table 1 summarizes the matrix. The inference target is the tested provider deployments, not either model family in the abstract.

Table 1: **Frozen held-out evaluation design.** The primary population excludes the stable scenario.

| Component | Frozen setting |
|---|---|
| Models | DeepSeek-V3.2; MiniMax-M2.5 |
| Methods | ShiftMem; lexical retrieval baseline |
| Scenarios | 4 Test-ID; 4 Test-OOD |
| Seeds | 5 per scenario |
| Full matrix | 160 complete cells |
| Primary observations | 70 model-specific pairs in 35 environment clusters |
| Endpoint | 30-day oracle-relative cost gap; $\Delta < 0$ favors ShiftMem |

The primary analysis population excluded the stable scenario and contained 70 model-specific paired observations. Each pair matched on split, scenario, model deployment, and environment seed and contained one run per method. The two model-specific pairs for a common scenario and seed shared deterministic demand and calendar-indexed supply streams. The 70 observations therefore formed 35 environment clusters rather than 70 independent environment replicates. There was no independent replication of stochastic LLM output.

### 4.2 Endpoint and Statistical Analysis

The serialized primary endpoint was `post_shift_cumulative_regret_30`, reported as the 30-day oracle-relative cost gap over the first 30 complete days beginning at the declared shift day. The Oracle was a parameter-aware base-stock heuristic, not a proven optimum. For paired unit $i$, its common term cancels from the method contrast:

$$\begin{aligned}\Delta_i &= (C_i^{\text{Shift}} - C_i^{\text{Oracle}}) - (C_i^{\text{Lex}} - C_i^{\text{Oracle}}) \\ &= C_i^{\text{Shift}} - C_i^{\text{Lex}}.\end{aligned} \tag{11}$$

Negative $\Delta_i$ favors ShiftMem. We report the mean difference $\bar{\Delta}$, sample standard deviation $s_\Delta$, a nominal 95% interval for the mean difference, and paired standardized effect

$$\text{CI}_{95} = \bar{\Delta} \pm c_{.975,n-1}\frac{s_\Delta}{\sqrt{n}}, \qquad d_z = \frac{\bar{\Delta}}{s_\Delta}. \tag{12}$$

The frozen critical-value routine used tabulated values through 30 degrees of freedom and 1.96 thereafter. We

also report $n$ and ShiftMem wins, ties, and losses, defined by negative, zero, and positive paired differences.

For sample skewness $g_1$ and kurtosis $g_2$, the normality diagnostic used

$$J_n = \frac{n}{6}\left[g_1^2 + \frac{(g_2-3)^2}{4}\right], \qquad p_N = \exp(-J_n/2). \quad (13)$$

A severe outlier was flagged when $\max_i |\Delta_i - \text{median}(\Delta)| > 6\,\text{MAD}(\Delta)$. When MAD was zero, any nonzero deviation triggered the flag. The prespecified rule selected a paired $t$-test when $p_N \geq 0.05$ and no severe-outlier flag was raised; otherwise, it selected a two-sided Wilcoxon signed-rank test. Nominal mean intervals were retained under either test. A selected Wilcoxon $p$-value therefore does not test the mean estimand directly. The split analysis was prespecified. Model and scenario decompositions were descriptive. Their $p$-values were unadjusted.

The frozen decision rule supported the ShiftMem-advantage hypothesis only when the mean difference was negative and the selected two-sided test gave $p < 0.05$.

The frozen primary calculation nevertheless treated the 70 model-specific differences as inferential observations. Their clustering violates the independence assumption, so the nominal interval and $p$-value are not calibrated for cluster-level inference. The split-level nominal intervals and $p$-values inherit this limitation and are reported as descriptive protocol-frozen summaries. We retained the frozen calculation unchanged and added a post-hoc sensitivity that averaged model effects within each of the 35 clusters and retained seven fixed scenario strata with five environment-seed clusters each. Within each stratum, the sensitivity resampled clusters with replacement 50,000 times and used the 2.5th and 97.5th percentiles of the equal-scenario mean. It also applied a two-sided Monte Carlo cluster sign-flip test to that mean under sign exchangeability of cluster differences. For the deployment contrast, we formed the within-cluster DeepSeek-minus-MiniMax method-effect difference and applied the same stratified procedures. Each bootstrap and sign-flip procedure used 50,000 draws. The overall bootstrap and sign-flip seeds were 20260723 and 20260724; the deployment-contrast seeds were 20260733 and 20260734. The custom dependency-light routines ran under CPython 3.12.7 and NumPy 2.5.1. Figures used Matplotlib 3.11.0. Analysis code was SHA-256-bound.

## 4.3 Primary Effect

The primary analysis did not support an improvement by ShiftMem. Across 70 pairs, the mean difference between ShiftMem and the lexical retrieval baseline was +45.44 (nominal 95% mean CI $[-2.72, 93.60]$; $s_\Delta = 205.59$; $d_z = 0.221$). The frozen rule selected the Wilcoxon normal approximation ($W = 716$, two-sided $p = 0.203$). ShiftMem won 25 pairs, tied 11, and lost 34. Here $W$ is the smaller of the positive and negative rank sums after removing zero differences and averaging tied ranks. The normal approximation used continuity correction. The prespecified decision rule therefore did not support the prespecified ShiftMem-advantage hypothesis, and the point estimate was unfavorable to ShiftMem. Given the dependence limitation, neither the interval nor the test establishes equivalence or intrinsic failure of the conditional memory lifecycle.

The post-hoc clustered sensitivity pointed in the same unfavorable direction. Its estimate remained +45.44, with a cluster-bootstrap 95% CI of $[11.26, 79.09]$ and sign-flip $p = 0.041$ across 35 clusters. This post-hoc result does not replace the prespecified analysis or convert the original advantage hypothesis into a supported claim. In this sensitivity analysis, accounting for shared environment streams did not recover a favorable average effect.

## 4.4 Split, Model, and Scenario Heterogeneity

The Test-ID and Test-OOD point estimates had opposite signs. Test-ID had a mean point estimate near zero ($\bar{\Delta} = -2.67$, nominal 95% mean CI $[-60.17, 54.82]$, 30 model-specific pairs in 15 environment clusters, paired-$t$ $p = 0.925$); its wide interval does not establish equivalence. Test-OOD had an unfavorable point estimate (+81.53, nominal 95% mean CI $[9.50, 153.56]$, 40 model-specific pairs in 20 environment clusters, Wilcoxon $p = 0.093$). The Test-OOD mean interval excluded zero while the rank-based test did not reject. These summaries target different quantities. Neither split result supports a universal claim about in- or out-of-distribution behavior.

The two provider-hosted model deployments also showed opposite descriptive directions. DeepSeek had $\bar{\Delta} = +107.73$ (nominal 95% mean CI $[23.09, 192.37]$, 35 environment-level paired differences, unadjusted Wilcoxon $p = 0.023$), whereas MiniMax had $\bar{\Delta} = -16.85$ (nominal 95% mean CI $[-53.91, 20.22]$, 35 environment-level paired differences, unadjusted paired-$t$ $p = 0.379$). The post-hoc DeepSeek-minus-MiniMax method-effect contrast was +124.58 (cluster-bootstrap 95% CI $[38.59, 216.05]$, unadjusted sign-flip $p = 0.015$). No multiplicity correction was applied across the exploratory analyses. The unadjusted post-hoc contrast suggests effect modification between these two tested deployments, but only two deployments and no independent LLM replications were available. It therefore does not establish generalizable effect modification across deployments or model families.

Scenario means ranged from −67.38 for the Test-ID demand jump to +151.24 for the early combined OOD shift (Fig. 2b). Each estimate used only ten model-specific pairs in five environment clusters, was unadjusted, and had generally wide uncertainty. We therefore report the set as a decomposition rather than ranking scenarios. The mean ShiftMem-minus-baseline difference in stable-environment total cost was +15.86 (nominal 95% mean CI $[-184.37, 216.09]$, ten model-specific pairs in five en-

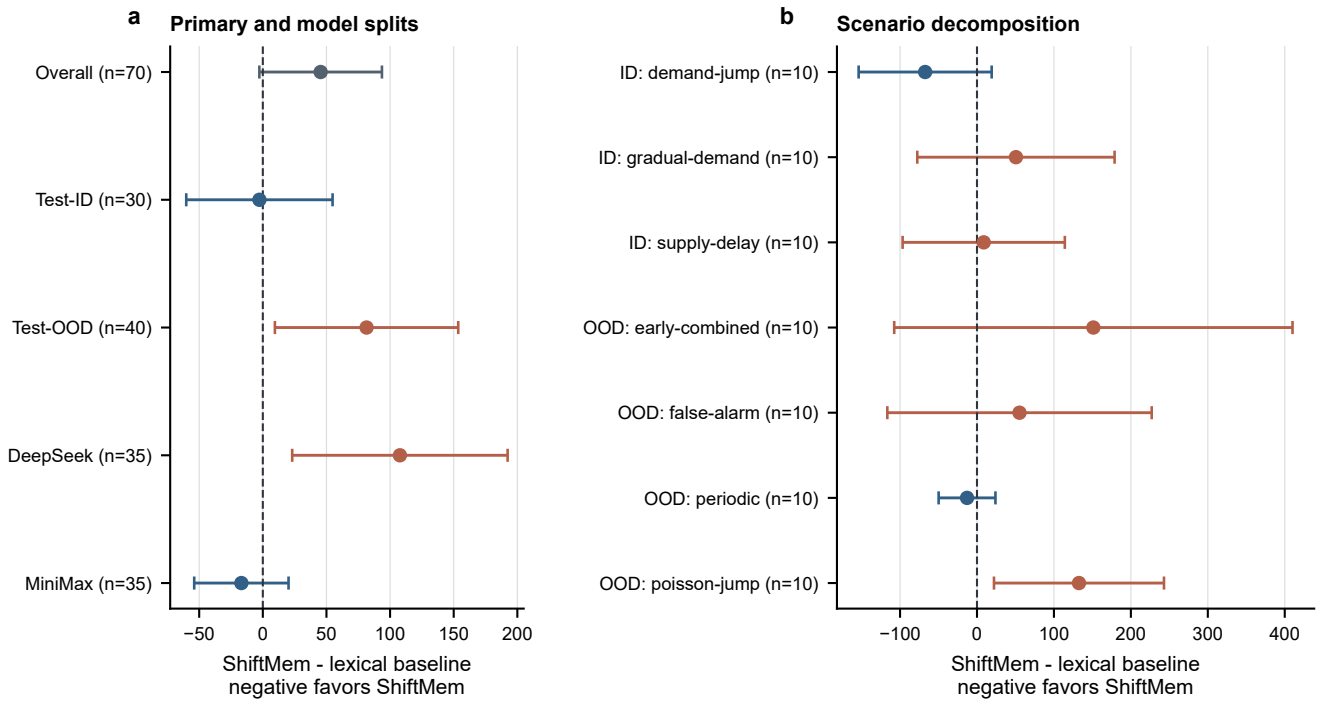


Figure 2: **Differences between ShiftMem and the lexical retrieval baseline in the 30-day oracle-relative cost gap.** Negative values favor ShiftMem. Points denote mean paired differences; bars are nominal 95% mean-difference intervals from the frozen summary routine. **a**, The overall row is the prespecified dependence-limited primary calculation ($n = 70$ model-specific pairs in 35 environment clusters). Its two-sided hypothesis test used the selected Wilcoxon signed-rank test, which targets a different quantity from the displayed mean interval. Test-ID and Test-OOD are the prespecified descriptive split. Model rows are unadjusted descriptive decompositions. **b**, Scenario decompositions are descriptive and unadjusted (ten model-specific pairs in five environment clusters each). These subgroup results do not establish generalizable effect modification by deployment or scenario. Source data are provided in the evidence package.

vironment clusters). Because no non-inferiority margin was frozen, this result establishes neither equivalence nor absence of degradation.

Table 2 reports the complete aggregate effect set with evidence status labels.

## 4.5 Recovery and System Reliability

Recovery was defined from rolling cost rather than the primary cost-gap endpoint. A cell recovered when the absolute difference between its seven-day rolling total cost and the paired Oracle's rolling cost did not exceed 10% of the Oracle value for seven consecutive overlapping window endpoints. Otherwise, it was right-censored at the final day. Recovery occurred in 61 of 140 applicable cells: 32 of 70 for ShiftMem and 29 of 70 for the lexical retrieval baseline. These counts are descriptive and do not account for censoring through a survival model.

Reliability burden was part of the evaluated system rather than an exclusion criterion. The 160 cells recorded 5,176 reviews and 6,189 attempts. Of these attempts, 1,680 were coded as failed or invalid (27.1%); this aggregate is stored as `parse_failures` and includes provider exceptions as well as JSON or schema failures. After the permitted retry, 667 reviews retained the previous strategy (12.9%). Cell records contained 28.72M input and 2.08M output tokens. Burden varied more strongly by model than by memory method: failed/invalid-attempt rates were 9.8% for DeepSeek and 40.9% for MiniMax, while retained-strategy rates were 4.6% and 21.1%, respectively. Figure 3 reports the split–model–method groups without inferential error bars.

Table 3 reports the corresponding reliability and context totals.

The append-only provider journals retained 4,558 successful responses and 1,705 terminal failed attempts. Their 6,263 terminal attempts exceed the 6,189 cell-recorded attempts by 74 because interrupted or otherwise non-final attempts remain in the journals. Estimated successful-response cost was CNY 108.49. The separate failed-attempt reservation ledger recorded CNY 93.88. Neither quantity is the official provider bill, which remains unreconciled.

The evidence manifest also documents 161 erroneous

Table 2: **Complete primary, descriptive, and post-hoc effect summaries.** Differences are ShiftMem minus the lexical retrieval baseline; negative values favor ShiftMem. Except for the two post-hoc cluster rows, intervals are nominal mean-difference intervals from the frozen routine and are not calibrated for cluster-level inference. W/T/L denotes ShiftMem wins, ties, and losses.

| Analysis | $n$ | $\bar{\Delta}$ | 95% CI | $d_z$ | W/T/L | Test | $p$ | Evidence status |
|---|---|---|---|---|---|---|---|---|
| Overall | 70 | 45.44 | $[-2.72, 93.60]$ | .221 | 25/11/34 | Wilcoxon-N | .2034 | Prespecified, dependence-limited |
| Test-ID | 30 | -2.67 | $[-60.17, 54.82]$ | -.017 | 12/6/12 | Paired $t$ | .9249 | Prespecified descriptive |
| Test-OOD | 40 | 81.53 | $[9.50, 153.56]$ | .351 | 13/5/22 | Wilcoxon-N | .0932 | Prespecified descriptive |
| DeepSeek | 35 | 107.73 | $[23.09, 192.37]$ | .422 | 9/6/20 | Wilcoxon-N | .0232 | Unadjusted descriptive |
| MiniMax | 35 | -16.85 | $[-53.91, 20.22]$ | -.151 | 16/5/14 | Paired $t$ | .3793 | Unadjusted descriptive |
| ID: demand jump | 10 | -67.38 | $[-153.83, 19.07]$ | -.557 | 6/1/3 | Paired $t$ | .1117 | Descriptive subgroup |
| ID: gradual demand | 10 | 50.62 | $[-77.60, 178.84]$ | .282 | 2/4/4 | Paired $t$ | .3951 | Descriptive subgroup |
| ID: supply delay | 10 | 8.74 | $[-96.67, 114.15]$ | .059 | 4/1/5 | Paired $t$ | .8554 | Descriptive subgroup |
| OOD: early combined | 10 | 151.24 | $[-107.60, 410.08]$ | .418 | 4/0/6 | Paired $t$ | .2189 | Descriptive subgroup |
| OOD: transient excursion | 10 | 55.26 | $[-116.55, 227.07]$ | .230 | 3/2/5 | Wilcoxon-E | .7422 | Descriptive subgroup |
| OOD: periodic | 10 | -12.94 | $[-49.87, 23.99]$ | -.251 | 4/2/4 | Wilcoxon-E | .6406 | Descriptive subgroup |
| OOD: Poisson jump | 10 | 132.56 | $[22.08, 243.04]$ | .858 | 2/1/7 | Paired $t$ | .0238 | Descriptive subgroup |
| Stable total cost | 10 | 15.86 | $[-184.37, 216.09]$ | .057 | 5/2/3 | Wilcoxon-E | 1.000 | Descriptive, different endpoint |
| Clustered overall | 35 | 45.44 | $[11.26, 79.09]$ | – | – | Sign flip | .0414 | Post-hoc sensitivity |
| DeepSeek–MiniMax | 35 | 124.58 | $[38.59, 216.05]$ | – | – | Sign flip | .0154 | Post-hoc, unadjusted |

Table 3: **Reliability, context, and recovery by split, deployment, and method.** Each row aggregates 20 cells. Failed/invalid counts include provider exceptions and invalid structured outputs and use cell-recorded attempts as the denominator. Fallbacks use reviews as the denominator. Token totals are in millions. Recovery is descriptive and excludes stable cells.

| Split | Deployment | Method | Attempts | Failed/invalid | Reviews | Fallbacks | Input M | Output M | Recovered |
|---|---|---|---|---|---|---|---|---|---|
| Test-ID | DeepSeek | ShiftMem | 681 | 3 (0.4%) | 678 | 0 (0.0%) | 4.477 | 0.068 | 8/15 |
| Test-ID | DeepSeek | Lexical | 625 | 39 (6.2%) | 600 | 14 (2.3%) | 3.711 | 0.061 | 7/15 |
| Test-ID | MiniMax | ShiftMem | 966 | 506 (52.4%) | 677 | 217 (32.1%) | 2.820 | 0.413 | 13/15 |
| Test-ID | MiniMax | Lexical | 867 | 462 (53.3%) | 600 | 195 (32.5%) | 2.441 | 0.359 | 10/15 |
| Test-OOD | DeepSeek | ShiftMem | 761 | 96 (12.6%) | 709 | 44 (6.2%) | 4.422 | 0.064 | 4/20 |
| Test-OOD | DeepSeek | Lexical | 667 | 129 (19.3%) | 600 | 62 (10.3%) | 3.450 | 0.056 | 6/20 |
| Test-OOD | MiniMax | ShiftMem | 876 | 237 (27.1%) | 712 | 73 (10.3%) | 4.049 | 0.579 | 7/20 |
| Test-OOD | MiniMax | Lexical | 746 | 208 (27.9%) | 600 | 62 (10.3%) | 3.349 | 0.484 | 6/20 |

metadata indicators of Test-outcome access introduced by a raw serializer default. It preserves the source bytes and reports no effect on numeric outcomes.

### 4.6 Exploratory Diagnostics and Limitations

Post-hoc burden correlations were small. Across 70 pairs, correlations between relative burden and the paired endpoint ranged from $-0.134$ to $-0.055$ across Pearson and Spearman summaries for failed/invalid attempts, fallbacks, and journal-recorded provider failures. These post-hoc associations have no reported inferential $p$-values and do not identify mediation or causality.

Execution order remains an unresolved limitation. All 70 pairs ran the lexical retrieval baseline before ShiftMem, and the journals contain append order rather than wall-clock timestamps. The early and late halves had mean differences of $+58.58$ and $+32.30$. Correlations with execution position were 0.159 (Pearson) and 0.175 (Spearman). These diagnostics do not remove confounding by method order or provider drift. A valid follow-up must randomize or counterbalance execution order.

The held-out evaluation tested one synthetic single-item environment family, two provider deployments, five environment seeds per scenario, one lexical retrieval baseline, and no independent LLM-output replications. It did not formally compare ShiftMem with a broader set of memory baselines, test dormancy/reactivation through ablation, or compare the Page–Hinkley trigger with an LLM-only detector. Stable-scenario performance remained descriptive because no non-inferiority margin was frozen. The results therefore evaluate the two deployed systems under the specified protocol. They do not identify the causal value of an individual lifecycle component or establish general memory superiority. Model identifiers and the output-token cap were frozen, but the complete resolved provider request settings and provider-side defaults for omitted decoding controls were not persisted. Exact API-level replication is therefore limited to the recorded fields and raw responses.

## 5 Conclusion

This paper evaluated whether an explicit lifecycle for stored strategy revisions improved an LLM inventory agent's performance after operating conditions changed. ShiftMem made reuse decisions explicit through conditional retrieval, delayed validation, and an auditable lifecycle, while a shared deterministic controller retained responsibility for daily orders. In the pre-Test-frozen eval-

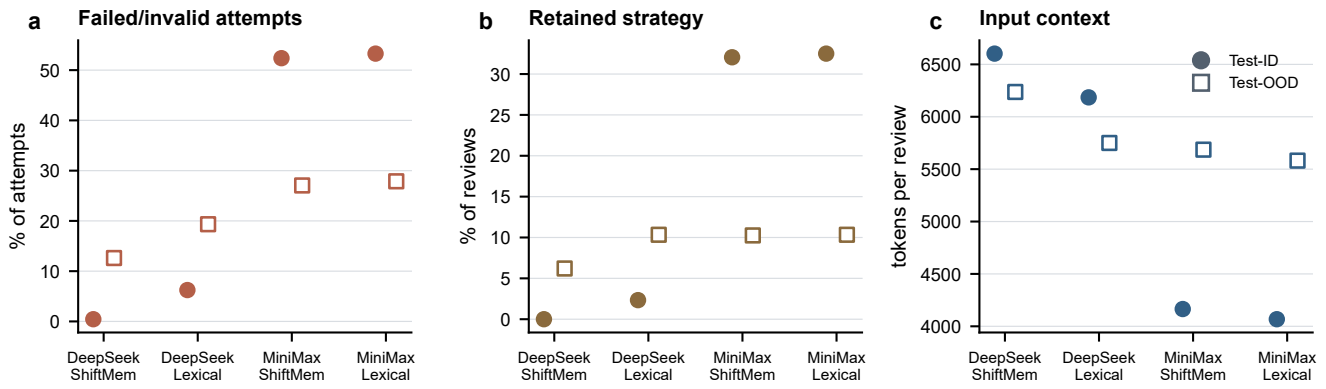


Figure 3: **Reliability and input-context burden across the 160-cell matrix.** Filled circles denote Test-ID and open squares denote Test-OOD. Each point aggregates 20 cells for one split–model–method combination. **a**, Cell-coded failed or invalid attempts as a percentage of cell-recorded attempts. This category includes provider exceptions and structured-output failures. **b**, Reviews that retained the previous valid strategy as a percentage of strategy reviews. **c**, Cell-recorded input tokens per review. These aggregate rates are descriptive and do not identify a causal effect of reliability on inventory outcomes. Source data are provided in the evidence package.

uation, the dependence-limited prespecified decision rule did not support a ShiftMem advantage over the implemented lexical retrieval baseline. The mean point estimate was unfavorable, and the post-hoc clustered sensitivity retained the same direction.

The result does not establish equivalence or show that the conditional memory lifecycle is intrinsically ineffective. Point estimates varied descriptively across the two model deployments and test regimes, and provider and structured-output failures were retained as part of the observed system behavior. Neither the descriptive heterogeneity nor the reliability associations provided mechanism-level evidence: the associations were observational, key lifecycle ablations were not run, and every pair executed the lexical retrieval baseline first. The evidence is therefore limited to the tested synthetic single-item environment, provider deployments, lexical retrieval baseline, and fixed execution order.

The main methodological outcome is a documented implementation and evaluation architecture for operational memory under delayed feedback: bounded strategy review, deterministic execution, retained failures, delayed validation, content-addressed evidence, and network-free analysis verification. A stronger follow-up should randomize or counterbalance method order, add independent LLM-output replications, encode demand as an applicability condition, and align the available history with the forecast-window range. It should also compare an embedding-based retrieval baseline and run the dormancy/reactivation and detector ablations required to identify component-level effects. These controls are prerequisites for a broader claim about conditional memory efficacy.

# Code and Data Availability

Source code, frozen configurations, analysis scripts, aggregate outputs, figure source data, and the paper source are maintained at `https://github.com/QCYTSN/Shiftmem`. The formal evidence closure is `v2-formal-results-f4ab41daacf3`. Its read-only convenience archive is stored at `artifacts/releases/v2-formal-results-f4ab41daacf3-raw-evidence.zip`, with SHA-256 `3f462721f6bc025043ac262c0e06724ac6b1a5479372ddc54a006834fa94e49f`. The manifest's per-file hashes remain authoritative. The network-free command `.venv/Scripts/python.exescripts/finalize_formal_results.py--verify` recomputes integrity checks and the deterministic aggregate outputs. From a clean checkout, first extract the tracked archive into `artifacts/raw_runs` using `python-mzipfile-eartifacts/releases/v2-formal-results-f4ab41daacf3-raw-evidence.zipartifacts/raw_runs`.

# References


[1] Nele H. Amiri, Sean R. Sinclair, and Maximiliano Udenio. Non-stationary inventory control with lead times. *arXiv preprint arXiv:2602.05799*, 2026. URL https://arxiv.org/abs/2602.05799.

[2] Erhan Bayraktar and Michael Ludkovski. Inventory management with partially observed nonstationary demand. *Annals of Operations Research*, 176(1):7–39, 2010. doi: 10.1007/s10479-009-0513-8. URL https://arxiv.org/abs/1206.6283.

[3] Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 443–448, 2007. doi: 10.1137/1.9781611972771.42.

[4] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, 2019. doi: 10.18653/v1/D19-1224.

[5] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014. doi: 10.1145/2523813.

[6] Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured nlp tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, 2023. doi: 10.18653/v1/2023.emnlp-main.674. URL https://aclanthology.org/2023.emnlp-main.674/.

[7] 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)*, pages 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

[8] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

[9] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

[10] 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*, pages 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

[11] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22 (164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

[12] Yinzhu Quan and Zefang Liu. Invagent: A large language model based multi-agent system for inventory management in supply chains. *arXiv preprint arXiv:2407.11384*, 2024. URL https://arxiv.org/abs/2407.11384.

[13] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

[14] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, 2024. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

[15] James T. Treharne and Charles R. Sox. Adaptive inventory control for nonstationary demand and partial information. *Management Science*, 48(5):607–624, 2002. doi: 10.1287/mnsc.48.5.607.7807.

[16] Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108.

[17] 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. URL https://arxiv.org/abs/2410.10813.

[18] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29936`.

[19] Xuhua Zhao, Yuxuan Xie, Caihua Chen, and Yuxiang Sun. Aim-bench: Evaluating decision-making biases of agentic llm as inventory manager. *arXiv preprint arXiv:2508.11416*, 2025. URL `https://arxiv.org/abs/2508.11416`.

[20] Paul Zipkin. Old and new methods for lost-sales inventory systems. *Operations Research*, 56(5):1256–1263, 2008. doi: 10.1287/opre.1070.0471.

# A Protocol and Amendment Chronology

Protocol 2.0 replaced the direct-daily-order design before held-out outcomes were generated or read. Numbered amendments changed budget reservation or continuation controls; they did not change the 160-cell matrix, primary estimand, deployments, methods, scenarios, seeds, controller, endpoint, or statistical decision rule.

The current closure identity is `v2-formal-results-f4ab41daacf3`. The earlier closure identity remains in version history. The cell files, journals, and summaries were unchanged; the identity changed because the manifest additionally bound explicitly post-hoc sensitivity-analysis code.

# B Held-Out Scenarios and Frozen Runtime

All scenarios lasted 150 days and used costs $(c_p, c_h, c_s, c_f) = (1, 0.2, 5, 0)$. Unless a shift changed them, quoted lead time and fill probability were $(\ell, \rho) = (2, 1.0)$. Test-ID used seeds 2000–2004 and Test-OOD used seeds 3000–3004.

# C Complete Effect Evidence

Table 2 contains every aggregate effect reported from the formal matrix, including the stable-environment endpoint and the post-hoc clustered analyses. The frozen overall, split, deployment, and scenario intervals are nominal mean intervals. They are distinct from the selected rank tests and are not calibrated for the shared environment clusters. The stable row uses 150-day total cost rather than the 30-day adaptation endpoint and has no frozen non-inferiority margin. The two clustered rows are explicitly post hoc; the deployment contrast is also unadjusted.

The authoritative aggregate is `artifacts/aggregated/v2_formal_statistical_analysis.json`. Cell-level values remain in five immutable JSONL prefixes bound by the evidence manifest. All 160 planned cells are present. The primary population comprises 70 model-specific pairs nested in 35 scenario–seed environment clusters.

# D Reliability, Budget, and Execution-Order Audit

Table 3 gives the eight split–deployment–method groups. The aggregate accounting and post-hoc diagnostics are summarized below.

The machine-readable source is `artifacts/aggregated/v2_formal_reliability_audit.json`. Its `parse_failures` field combines provider exceptions and invalid structured outputs; it must not be interpreted as a provider-failure count alone.

# E Evidence Manifest and Reproduction

From a clean checkout, extract the tracked evidence archive and then run the network-free closure verification:

```
python -m zipfile -e artifacts/releases/
v2-formal-results-f4ab41daacf3-raw-evidence.zip
artifacts/raw_runs
.venv\Scripts \python .exe
scripts/finalize_formal_results.py -verify
```

It verifies hashes and compares the three deterministic aggregate outputs. The closure manifest records all four source freezes as valid. A fresh local rerun also requires the frozen directories to be free of undeclared generated files, including ignored Python bytecode caches.

# F Strategy-Review Prompt and Model Qualification

**Model-facing contract.** The model acts only as a low-frequency strategy reviewer. At each periodic, event, or coalesced review it receives the public observation, retrieved memory, current strategy, trigger evidence, and an optional correction message. The current strategy contains exactly `forecast_window`, `safety_stock_multiplier`, and `lead_time_buffer`. The prompt forbids hidden demand parameters, future demand or fill, shift timing, regime identity, Oracle information, and daily orders. The exact frozen prompt and deterministic user-message builder are retained in `src/shiftmem/providers/inventory_prompt.py` within each source freeze.

**Required response.** The provider must return one JSON object with no additional fields:

```
{
  "forecast_window": positive integer,
  "safety_stock_multiplier": number >= 0,
  "lead_time_buffer": integer >= 0,
  "used_memory_ids": array of supplied IDs,
  "confidence": number in [0,1],
  "reason": one sentence, at most 200 characters
}
```

The schema rejects extra fields, unsupplied memory identifiers, and any `order_quantity`. One correction attempt is permitted. A second failure retains the previous valid strategy. Valid proposals are projected first to the absolute bounds and then to the per-review change limits in Table 6.

Table 4: **Protocol and formal-execution chronology.** Prefix counts denote newly completed immutable cell files.

| Date | Record | Evidence-grounded consequence |
|---|---|---|
| 2026-07-14 | Protocol 2.0 | Replaced the direct-order design with bounded strategy review and deterministic daily control before Test access. |
| 2026-07-15 | Bounded Pilot | Completed under a separate CNY 6 and 350-attempt authorization. Because it used old runtime defaults, it supports only the recorded reliability and cost envelope. |
| 2026-07-16 | Amendment 1 | Approved eight scenarios, five paired seeds, two deployments, and two methods. Hard limits were 7,000 attempts, 25M input tokens, 3.2M billed output tokens, and CNY 100. |
| 2026-07-17 | Initial formal source | Freeze `v2-formal-da148583e3c5` produced the first 47-cell Test-ID prefix. |
| 2026-07-19 | Amendment 2 | Freeze `v2-formal-50aaafda0d6a` bound the 47-cell prefix and added three Test-ID cells without reissuing a completed cell. |
| 2026-07-19 | Amendment 3 | A 3,207-token response exceeded the 3,072-token reservation and was retained as `ReservationUnderestimate`. Freeze `v2-formal-724421cdbf78` raised the reservation to 4,096 and added 30 Test-ID and 23 Test-OOD cells. |
| 2026-07-21 | Amendment 4 | Freeze `v2-formal-5925ef7f1614` bound all prior prefixes and added the final 57 Test-OOD cells. |
| 2026-07-22 | Post-Test closure | Completed 80 Test-ID and 80 Test-OOD cells, with no exclusion or completed-cell rerun. No further provider call was authorized. |

Table 5: **Exact held-out scenario configurations.** $\mathrm{NB}(\mu, r)$ denotes negative-binomial demand with mean $\mu$ and dispersion $r$. The serialized dispersion field in the Poisson scenario is ignored by Poisson sampling.

| Split | Scenario | Demand | $I_0$ | Shift timing | Demand change | Supply change |
|---|---|---|---|---|---|---|
| ID | Stable | $\mathrm{NB}(21, 11)$ | 58 | None | None | None |
| ID | Demand jump | $\mathrm{NB}(19, 11)$ | 58 | $t \geq 80$ | $\mu \times 1.3$ | None |
| ID | Gradual demand | $\mathrm{NB}(21, 10)$ | 62 | Starts 65; target at 100 | Linear to $\mu \times 1.45$ | None |
| ID | Supply delay | $\mathrm{NB}(20, 8)$ | 64 | $t \geq 70$ | None | $(\ell, \rho) = (4, 0.85)$ |
| OOD | Periodic | $\mathrm{NB}(20, 10)$ | 60 | $t \geq 0$ | $\mu[1 + 0.4\sin(2\pi t/20)]$ | None |
| OOD | Poisson jump | $\mathrm{Pois}(20)$ | 60 | $t \geq 45$ | $\mu \times 2.0$ | None |
| OOD | Transient excursion | $\mathrm{NB}(20, 10)$ | 60 | $45 \leq t \leq 60$ | $\mu \times 1.8$ | None |
| OOD | Early combined | $\mathrm{NB}(20, 10)$ | 60 | $t \geq 30$ | $\mu \times 1.8$ | $(\ell, \rho) = (7, 0.7)$ |

**Qualification.** Qualification required exactly 12 results per deployment, no unresolved parse failure or fallback, no invalid or inapplicable citation, and all four induced monotonicity checks to pass. Run `v2-qual-live-20260715-739bc99` used 24 provider calls. Both deployments returned 12 results in 12 attempts, passed 4/4 checks, and recorded zero correction retries, fallbacks, invalid citations, or inapplicable citations. Earlier qualification executions remain preserved as `inconclusive_harness_invalid` because their harness used the archived daily-order prompt and incomplete provenance.

**Prompt provenance and decoding limitation.** The accepted qualification recorded system-prompt SHA-256 `554c30d33d35304984fb1d03e5220b76768c29433606e1ca84709ae28cb38`. The later formal prompt hash was `4140bbb60e6f4d5b704c73513377c1e14a42473a5a73c2221a22d91ae2a033`; it added the explicit per-review limits and deterministic projection. Client code specified nominal temperature zero, 512 maximum output tokens, disabled thinking, and JSON-object response format. However, the formal artifacts did not persist the complete resolved provider request configuration or provider-side defaults for omitted decoding controls. Exact API-level reconstruction is therefore limited to the persisted fields, prompt hashes, and raw responses.

Table 6: **Frozen controller and runtime parameters.**

| Field | Exact setting |
|---|---|
| Strategy $(w, k, b)$ | Default $(14, 1.2, 1)$; bounds $[1, 60] \times [0, 5] \times [0, 14]$; maximum per-review changes $(7, 1.0, 1)$. |
| Controller | Uses the most recent $\min(w, 14)$ demands, $h = \ell + b + 1$, and nearest-integer ties-to-even rounding. Empty history uses reset demand zero. |
| Scheduler | Review on $t \bmod 5 = 0$, including $t = 0$; event cooldown three days; same-day triggers coalesced; one correction retry before retaining the prior strategy. |
| Detector (ShiftMem) | Separate two-sided Page–Hinkley detectors for demand, lost sales, and quoted lead time; minimum 10 samples, $\delta = 0.1$, and $\lambda = 48$. |
| Retrieval (ShiftMem) | Top five; lexical-score/confidence/recency/utility weights 0.75/0.5/1.0/0.25; probation/change penalties 0.25/0.5; recency half-life 30; changed-variable lookback 14. |
| Validation and lifecycle | Three-day service window; dormancy after three mismatching retrieval checks; Beta prior $(1, 1)$; activation after two supports with $c \geq 0.65$; invalidation after two failures with $c \leq 0.35$. |
| Generation and reservation | Maximum generated tokens 512; final billed-output reservation 4,096 tokens per attempt. |

Table 7: **Aggregate reliability, accounting, and order audit.** Cell and journal token totals have different accounting scopes and are not combined.

| Audit item | Verified result |
|---|---|
| Cell accounting | 5,176 reviews; 6,189 attempts; 1,680 failed/invalid attempts; 667 retained-strategy fallbacks; 28,717,245 input and 2,084,428 output tokens; recovery in 61/140 applicable cells. |
| Journal accounting | 6,263 terminal attempts: 4,558 successful and 1,705 failed; 48,885,771 input and 8,152,658 output tokens; zero unresolved reservations at closure. |
| Cost | Successful-response estimate CNY 108.4874; failed-attempt reservation ledger CNY 93.8759. The quantities are not additive and neither is the official provider bill, which was unavailable. |
| Burden correlations | Pearson/Spearman correlations with the paired endpoint were $-0.0700/-0.1045$ for failed/invalid burden, $-0.0551/-0.0751$ for fallback burden, and $-0.0722/-0.1340$ for journal-recorded provider failures. No inferential $p$-values were reported. |
| Zero-event subsets | Mean differences remained unfavorable in the 25 zero-failed/invalid, 32 zero-fallback, and 26 zero-provider-failure pairs. These subsets were post hoc and model-imbalanced. |
| Method order | The lexical retrieval baseline preceded ShiftMem in all 70 primary pairs. Reverse-order pairs: 0. Journals recorded append position but not wall-clock timestamps. |
| Execution halves | Early 35 pairs: $\bar{\Delta} = 58.58$; late 35 pairs: $\bar{\Delta} = 32.30$. Pearson/Spearman position associations were 0.159/0.175. These diagnostics do not remove method-order or provider-drift confounding. |

Table 8: **Reproducibility facts for the formal evidence closure.**

| Field | Frozen fact |
|---|---|
| Closure | `v2-formal-results-f4ab41daacf3`; 160 complete cells; manifest SHA-256 `f4ab41daacf380aa1d59c7af6abf4374c282c14547ae3fe45874c4ae516697c0`. |
| Raw evidence | Five cell prefixes (Test-ID $47 + 3 + 30 = 80$; Test-OOD $23 + 57 = 80$), four append-only journals, and two split summaries. |
| Source freezes | `v2-formal-da148583e3c5`, `v2-formal-50aaafda0d6a`, `v2-formal-724421cdbf78`, and `v2-formal-5925ef7f1614`; all were valid in the closure manifest. |
| Platform | CPython 3.12.7 on Windows 11; local CPU orchestration with remote SiliconFlow inference. |
| Dependencies | PyYAML 6.0.3, Matplotlib 3.11.0, NumPy 2.5.1, Pydantic 2.13.4, and pytest 8.4.2. |
| Deployments | `deepseek-ai/DeepSeek-V3.2` and `MiniMaxAI/MiniMax-M2.5` through SiliconFlow. |
| Release archive | `v2-formal-results-f4ab41daacf3-raw-evidence.zip`; 2,735,947 bytes; SHA-256 `3f462721f6bc025043ac262c0e06724ac6b1a5479372ddc54a006834fa94e49f`. |
| Metadata anomaly | All 160 cells and one earlier Test-ID summary stored an erroneous false Test-access flag. Raw bytes were preserved; the derived manifest records the corrected interpretation and no effect on numeric outcomes. |