WHEN DOES TYPED MEMORY HELP LLM AGENTS? A SUBSPACE PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents increasingly rely on typed external memory, but it remains unclear when typing should outperform a strong flat retriever. We define a type as a subspace of a frozen representation space and typed memory as a complementary decomposition of stored content. At matched effective rank, typing provides no intrinsic capacity advantage over an unconstrained target-aware selector; its benefit must instead arise from inductive bias. When the target covariance has effective rank $r_{\text{eff}} \ll d$, blockwise spectral selection over label-free recovered subspaces incurs statistical dimension $\max_\tau r_{\text{eff},\tau}$ rather than r_{eff}, plus recovery and cross-block leakage costs. This gain belongs to the estimator rather than to physically separate stores: a flat retriever given the same blocks can obtain it without routing overhead, so store-specific benefits arise primarily from heterogeneous drift and per-type reliability. We instantiate these principles in SOMA, which learns write-time routing, complementary subspaces, per-type reliability and decay, and the effective number of types; SOMA-2S$_{\text{spec}}$ provides the closed-form estimator covered by the analysis. Across seven text, image, and video benchmarks with a frozen encoder and matched rank B=128, SOMA achieves 55.4% versus 53.4% for target-aware flat and 50.0% for similarity-flat memory, with a benchmark-bootstrap improvement of $[+1.1, +3.0]$ over target-aware flat. Its gain is largest in the low-data regime: $+5.6$ points at n=25, $+2.0$ at n=200, and $+1.2$ at n=800. A label-free drift-heterogeneity index correlates with the cost of shared decay (Spearman ρ=0.89, n=7). A risk-certificate heuristic, TYPE-IF^0, also separates most negative controls from all seven benchmarks, although its decisions are sensitive to the convention constant. All empirical results use one backbone with a frozen encoder at matched effective rank.

1 INTRODUCTION

The problem. LLM agents increasingly depend on external memory to act over long horizons. They accumulate facts, past episodes, and reusable procedures, and retrieve from this growing store to condition each new decision (Park et al., 2023; Sumers et al., 2023; Wang & Chen, 2025; Chhikara et al., 2025; Forouzandeh et al., 2026c). A recurring design choice is to *type* this memory–to split it into episodic, semantic, procedural, or otherwise specialized stores, each with its own representation, retention, and retrieval–and a growing number of systems even learn parts of that organization from data (Xu et al., 2026; Wang et al., 2025b; Nan et al., 2025; Zhang et al., 2026). Typed memory is now a common building block, yet the field has no representational definition of what a memory "type" is, and no account of when the extra machinery is worth its cost.

A running example. Consider a web agent that, within one session, must retain a durable user preference ("always book aisle seats"), a transient page state (the current cart contents), and a reusable procedure (the click sequence that completes checkout). These differ in how long they stay valid and how often they are reused. The preference should persist and be retrieved broadly, the page state should decay within the session, and the procedure should be recalled whenever a similar task recurs. A single flat store with one similarity metric and one retention policy serves all three the same way; the question this paper formalizes is whether *separating* them—and if so, on what representational grounds—earns its keep against a flat retriever given the same budget and supervision.

The challenge. Reported gains from typed memory are hard to attribute. A typed system usually differs from a flat baseline in several ways at once–more retrievers, more parameters, more supervi-

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.

sion, more storage. So an improvement may come from any of these rather than from typing itself. The sharpest control is a *strong, equally resourced flat retriever*: one trained on the same task signal and held to the same storage budget. Against that control the central question is not whether typed memory helps on a leaderboard, but *why* organizing memory into types should help at all, and in *which data regime*. Answering it requires (i) a definition of a type precise enough to state a capacity claim, (ii) a mechanism that could give typing an advantage a flat retriever cannot match at equal budget, and (iii) a way to tell, before deployment, whether that mechanism is present.

Our approach. We study memory types through one abstraction: *a memory type is a subspace of the frozen representation space, and a typed memory is a decomposition of stored content into complementary subspaces.* This makes the capacity question exact and the answer negative–at a fixed effective rank, an unconstrained target-aware rank-B selector weakly dominates every typed rank-B configuration (Proposition 1, Remark 1), so typing buys no representational capacity. Its value must instead come from *inductive bias* on two axes that a flat retriever pays for in full. *Statistically*, if a useful decomposition is recovered from an abundant *label-free* stream and fixed independently of scarce labelled probes, target estimation can be done blockwise, and the relevant sample-complexity dimension becomes the effective rank of the target covariance $r_{\text{eff}} = \operatorname{tr}(\mathbf{\Sigma}_*)/\|\mathbf{\Sigma}_*\|_{\text{op}}$ and its within-type counterparts, not the ambient dimension d (Proposition 5). This discount belongs to blockwise estimation, not to separate stores—a flat store handed the same blocks collects it without the routing and leakage costs a typed store pays (Remark 3)—so the store-specific value of typing is dynamical. *Dynamically*, when different content drifts at different rates, a single global decay is suboptimal and per-type decay strictly improves tracking (Proposition 7). We instantiate both mechanisms in SOMA (Subspace-Organized Memory for Agents), which learns write-time routing, complementary subspaces, per-type reliability and decay, and the effective number of types; a closed-form variant, SOMA-2S$_{\text{spec}}$, is the exact estimator the analysis covers. Because ordering two risk certificates need not order the losses, we package the statistical and dynamical conditions into a pre-deployment *sign heuristic*, TYPE-IF^0, not a guarantee. Throughout, our claim is *attributional*: we separate results by the strength of their support—exact guarantees for SOMA-2S$_{\text{spec}}$ under (A1)–(A7), synthetic tests of those assumptions, and empirical results for the learned SOMA—and the theory fixes the *mechanism and sign* of the typing effect, while the learned system is its empirical realization whose magnitude we report rather than derive. **Contributions.**

- We define a memory type as a subspace and prove that typing offers no capacity advantage over an equally resourced target-aware flat retriever, placing its benefit in estimation and dynamics rather than representational power (Section 2.2).
- For a stream-first spectral estimator we bound the coverage gap by estimation and recovery costs—subspace perturbation, cross-block leakage, and target-weighted misrouting—and derive the matching per-type decay result. The closed-form SOMA-2S$_{\text{spec}}$, the estimator the analysis covers, accounts for a $+1.2$-point gain over target-aware flat ($+0.7$ of it beyond label-free dimension reduction alone), while learned variants are evaluated empirically (Sections 2.2–2.3; Appendix M.1).
- Across seven text, image, and video benchmarks at matched effective rank with a frozen encoder, SOMA achieves 55.4% versus 53.4% for target-aware flat and 50.0% for similarity flat (benchmark bootstrap $[+1.1, +3.0]$), with the largest margins in the low-data regime the theory predicts: $+5.6$ points at $n{=}25$, narrowing to $+2.0$ at the $n{=}200$ main setting. Mechanism controls isolate subspace structure, typing, routing, and decay; a label-free drift index correlates strongly with the shared-decay penalty ($\rho{=}0.89$), and we test whether learned types align with the episodic/semantic/procedural taxonomy (Section 3).

2 METHOD

An agent produces observations $o_1, \ldots, o_T$; a frozen encoder maps each to $z_t = \text{enc}(o_t) \in \mathbb{R}^d$. A memory stores items derived from this stream and, given a query with representation z_q, returns items that condition the next decision. A *flat* memory keeps one store and retrieves by similarity in $\mathbb{R}^d$; a *typed* memory of K types routes items into stores $\mathcal{M}_1, \ldots, \mathcal{M}_K$ with their own representations and dynamics (Figure 1).

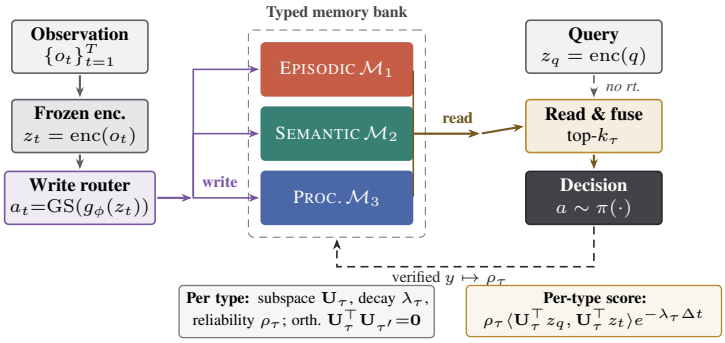


Figure 1: A frozen encoder maps the observation stream to representations; a write-time router assigns each observation to a learned type with its own subspace, decay, and reliability (group sparsity prunes unused types). Queries are *not* routed: all types are scored, top-k_τ items are fused, and verified outcomes update per-type reliability. In the stream-first regimes (SOMA-2S$_{\text{spec}}$, SOMA-2S) the router and subspaces are learned from the label-free stream and frozen before outcome supervision.

The analysis is phrased at the level of *retrievable directions*: a memory at *budget* B spans at most B directions (b_τ per type); item-level top-k coverage matches that of the spanned subspace up to a measured *retriever slack* of 0.03–0.08 (Appendix G).

2.1 Memory types as subspaces

Definition 1 (Type subspace and typed decomposition)**.** A memory *type* $\tau \in \{1, \dots, K\}$ is a subspace of the representation space, specified by $\boldsymbol{U}_\tau \in \mathbb{R}^{d \times d_\tau}$ with orthonormal columns; the type stores the projected content $z^\tau = \boldsymbol{U}_\tau^\top z$. A *typed memory* is $\{(\boldsymbol{U}_\tau, \mathcal{M}_\tau)\}_{\tau=1}^{K}$. Types are *complementary* when $\boldsymbol{U}_\tau^\top \boldsymbol{U}_{\tau'} = \mathbf{0}$ for $\tau \neq \tau'$.

Let π_t be the (recency-weighted) retrieval mass of item t, a_t its assignment, and $\boldsymbol{\Sigma} = \sum_t \pi_t z_t z_t^\top$ the content covariance. Orthogonality of the $\boldsymbol{U}_\tau$ does *not* block-diagonalize $\boldsymbol{\Sigma}$: routing stores a projected copy, but the raw z_t may have components in every type subspace. Two idealizations must not be conflated: *block-aligned content* (A2′), the aggregate condition $\boldsymbol{U}_\tau^\top \boldsymbol{\Sigma} \boldsymbol{U}_{\tau'} = \mathbf{0}$ that the capacity result needs, and the stronger sample-level *union-of-subspaces* condition (A7a), that each item lies in one generating subspace, which recovery needs. Neither follows from storing $\boldsymbol{V}_{a_t}^\top z_t$; Section 3 measures their residuals (r_Σ, r_a) separately from the overlap $\|\boldsymbol{U}_\tau^\top \boldsymbol{U}_{\tau'}\|_F$ that the complementarity term controls.

2.2 When typing helps—and when it cannot

Let the recall target—the directions of content a task needs—have covariance $\boldsymbol{\Sigma}_* \succeq 0$. For a retrievable subspace S with projector $\boldsymbol{P}_S$, define *coverage*

$$\mathcal{C}(S) = \mathrm{tr}(\boldsymbol{P}_S \, \boldsymbol{\Sigma}_* \, \boldsymbol{P}_S) \,/\, \mathrm{tr}(\boldsymbol{\Sigma}_*) \;\in [0, 1]. \tag{1}$$

Here S is the *store subspace*—the span of the directions the memory retains at budget B, of dimension up to B—*not* the span of the handful of items returned for a single query. Coverage is thus a property of what the store *can* surface, upper-bounded by $\hat{\nu} = B/\hat{r}_{\text{eff}}$; at B=128 and $\hat{r}_{\text{eff}} \approx 148$, the bound lies near 0.86, consistent with the observed coverage range. The separate question of how well a per-query top-k retrieval realizes this store-subspace coverage is the item-to-direction *retriever slack* of 0.03–0.08 quantified in Appendix G (Table 5), and must not be conflated with $\mathcal{C}(S)$ itself. A similarity-based flat memory at budget B retrieves the leading B directions of the content covariance, $S_{\text{flat}} = \text{top-}B(\boldsymbol{\Sigma})$—*target-agnostic*. A typed memory splits the same budget as $B = \sum_\tau b_\tau$.

Proposition 1 (Typed allocation weakly dominates variance-ordered flat selection)**.** *Under complementarity and (A2′) there is an allocation $\{b_\tau\}$ with $\sum_\tau b_\tau = B$ and a per-type direction choice with $\mathcal{C}(S_{\text{typed}}) \geq \mathcal{C}(S_{\text{flat}})$, strict iff S_{flat} fails to maximize coverage over the block family $\mathcal{B}_B$ (Lemma 5), e.g. when some excluded direction carries more target mass than some retained one.*

Remark 1 (No free capacity). Let $S_{\text{flat}}^{\text{aware}} \in \arg\max_{\dim S = B} \mathcal{C}(S)$ be a *target-aware* flat selector (a retriever trained on task signal). Since $\mathcal{B}_B \subseteq \{S : \dim S = B\}$, $\mathcal{C}(S_{\text{flat}}^{\text{aware}}) \geq \mathcal{C}(S_{\text{typed}}^{\star}) \geq$

$\mathcal{C}(S_{\text{flat}})$: *typing confers no capacity advantage*. Its value must lie in two burdens the target-aware selector carries in the ambient dimension: estimating the target (Proposition 5) and tracking drift (Proposition 7).

Statistical separation. In deployment $\boldsymbol{\Sigma}_*$ is estimated from a probe set of n labelled task interactions, and selection is a *plug-in* rule. Assume probe targets are i.i.d. sub-Gaussian with covariance $\boldsymbol{\Sigma}_*$ and relevant eigengap $\geq \gamma > 0$ (global for the flat rule, within-block for the typed rule); write $\lambda_1 = \|\boldsymbol{\Sigma}_*\|_{\text{op}}$, $\kappa = \lambda_1/\gamma$, $r_{\text{eff}} = \text{tr}(\boldsymbol{\Sigma}_*)/\lambda_1$, $r_{\text{eff},\tau} = \text{tr}(\boldsymbol{\Sigma}_\tau^*)/\|\boldsymbol{\Sigma}_\tau^*\|_{\text{op}}$, and $\nu = B/r_{\text{eff}}$ (Assumption 1). With a type-aligned target and $\{\boldsymbol{U}_\tau\}$ fixed independently of the probes, the plug-in coverage gap is $\lesssim \nu \min\{\sqrt{D/n}, \kappa^2 D/n\}$ with $D_{\text{flat}} = r_{\text{eff}} + \log(1/\delta)$ for the flat rule and $D_{\text{typed}} = \max_\tau r_{\text{eff},\tau} + \log(K/\delta)$ for the typed rule (Proposition 5, w.p. $\geq 1-\delta$): the effective ranks, not d, are the finite-sample dimensions, and a spiked minimax construction (Appendix F) shows the factor $d/\max_\tau d_\tau$ is tight in the second-order ($n \gtrsim \kappa^2 D$) regime.

For recovery we further assume (Assumption 2) that the stream is a union of K subspaces with minimum principal angle $\theta_{\min} > 0$ and recoverable sampling density (Elhamifar & Vidal, 2013) (A7a; $\tau(t)$ the generating type), and that the target is item-carried, $\boldsymbol{\Sigma}_* = \sum_t w_t z_t z_t^\top$, so item t has *target share* $\omega_t = w_t\|z_t\|^2/\text{tr}(\boldsymbol{\Sigma}_*)$ (A7b; the estimator $\widehat{\boldsymbol{\Sigma}}^*$ has this form). Fitting $\{\widehat{\boldsymbol{U}}_\tau\}, \{a_t\}$ on the stream defines the mixing angle $\theta_{\text{mix}}(T)$, the content-mass and target-weighted routing errors $\varepsilon_{\text{route}} = \sum_{a_t \neq \tau(t)} \pi_t$, $\varepsilon^*_{\text{route}} = \sum_{a_t \neq \tau(t)} \omega_t$, and the leverage $L_* = \max_t \omega_t/\pi_t$, with $\varepsilon^*_{\text{route}} \leq L_* \varepsilon_{\text{route}}$; $\tau(t)$ is operationalized on real streams as the residual-nearest recovered block (Section 3).

End-to-end typed rate. Freeze $\{\widehat{\boldsymbol{U}}_\tau\}, \{a_t\}$ from the label-free stream (any probe-free procedure) and let $\widehat{S}_{\text{typed}}$ keep, within each block, the top-b_τ eigenvectors of the routed empirical target $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}_\tau^* \widehat{\boldsymbol{U}}_\tau$ from n probes under the pooled allocation (Algorithm 6). Under Assumptions 1–2 and (A1)/(A3), with probability $\geq 1-\delta$ the retained-content coverage $\mathcal{C}^{\text{rt}}(S) = \sum_t \omega_t \|\boldsymbol{P}_{S_{a_t}} \hat{z}_t\|^2 \leq \mathcal{C}(S)$ obeys (Proposition 6)

$$\mathcal{C}(S^\star) - \mathcal{C}^{\text{rt}}(\widehat{S}_{\text{typed}}) \lesssim \underbrace{\nu \min\left\{\sqrt{\tfrac{D_{\text{typed}}}{n}},\ \kappa^2 \tfrac{D_{\text{typed}}}{n}\right\}}_{\text{estimation: probe budget } n} + \underbrace{\nu \sin^2\theta_{\text{mix}}(T) + \ell_{\text{leak}}(T) + \varepsilon^*_{\text{route}}(T)}_{\text{recovery: stream length } T},$$

with $\ell_{\text{leak}} \leq \sin^2\theta_{\text{mix}}$ the leakage of Lemma 2. The K-fold (resp. $\sqrt{K}$) separation survives whenever the recovery term is dominated by the estimation term, an inequality between measurable quantities checked in Section 3. The result covers the stream-first spectral estimator (SOMA-2S$_{\text{spec}}$) and its split form (Corollary 1, with n_2 probes); joint same-probe refinement forfeits it—the estimation term then holds uniformly over a $\Theta(Kd)$ router class, giving a rate no smaller than flat (Remark 4)—so the learned Stage B of SOMA-2S and joint SOMA are reported empirically. Crucially the estimation discount belongs to *blockwise estimation*, not to separate storage: a single structured-flat store over $\bigoplus_\tau S_\tau$ earns the same discount with *no* routing or leakage term (Remark 3), so what typing buys beyond it—per-type decay, reliability, per-item storage, fusion—is isolated by the SOMA-2S$_{\text{spec}}$−STRUCT-FLAT control (Appendix C), not SOMA-2S$_{\text{spec}}$−Stream-PCA.

Dynamical separation. Experience drifts heterogeneously (visual context in seconds, task semantics across episodes), and typing decouples tracking. Model mode τ as drifting with per-step increment δ_τ, observed at noise σ^2, so decay rate λ incurs steady-state error $e(\lambda; \delta_\tau, \sigma) = c_1\delta_\tau^2/\lambda^2 + c_2\lambda\sigma^2$ (Assumption 3, Lemma 6). Per-type decay attains $\sum_\tau \min_\lambda e$ at $\lambda_\tau^\star \propto (\delta_\tau^2/\sigma^2)^{1/3}$, whereas the best *shared* decay incurs exactly $R(\{\delta_\tau\}) = \left(\frac{1}{K}\sum_\tau \delta_\tau^2\right)^{1/3} / \left(\frac{1}{K}\sum_\tau \delta_\tau^{2/3}\right) \in [1, K^{2/3})$ times that error, $= 1$ iff all δ_τ are equal and increasing in their spread (Proposition 7, Corollary 5; the constants c_1, c_2, σ cancel). Because R depends only on drift rates it is estimable from the raw stream with no labels and no learned types: the second predictor.

2.3 A PRIORI PREDICTORS AND THE TYPE-IF^0 RULE

Definition 2 (Flat shortfall)**.** Given an empirical content covariance $\widehat{\boldsymbol{\Sigma}}$, empirical target covariance $\widehat{\boldsymbol{\Sigma}}^*$, and budget B,

$$\widehat{H}_B = 1 - \text{tr}\left(\boldsymbol{P}_{\text{top-}B(\widehat{\boldsymbol{\Sigma}})} \widehat{\boldsymbol{\Sigma}}^*\right) / \text{tr}(\widehat{\boldsymbol{\Sigma}}^*).$$

Lemma 1 (Shortfall bounds any rank-B reallocation)**.** *For any B-dimensional selection S, typed or otherwise, $\mathcal{C}(S) - \mathcal{C}(S_{\text{flat}}) \le H_B := 1 - \mathcal{C}(S_{\text{flat}})$; the empirical version uses the same target covariance and budget on both sides.*

Definition 3 (Drift-heterogeneity index)**.** Partition the label-free stream into $\tilde{K}$ clusters, estimate their drift magnitudes $\tilde{\delta}_k$, and set $\widehat{D} = R(\{\tilde{\delta}_k\}) - 1$, with R from Proposition 7 ($\widehat{D}_{\text{flat}}, \widehat{D}_{\text{typed}}$ below are dimension terms).

For the statistical certificate, estimate $\hat{r}_{\text{eff}} = \text{tr}(\widehat{\boldsymbol{\Sigma}}^*)/\hat{\lambda}_1$, $\hat{\nu} = B/\hat{r}_{\text{eff}}$, $\widehat{D}_{\text{flat}} = \hat{r}_{\text{eff}} + \log(1/\delta)$, and $\widehat{D}_{\text{typed}} = \hat{r}_{\text{eff,max}} + \log(K/\delta)$, where $\hat{r}_{\text{eff,max}} = \max_\tau \hat{r}_{\text{eff},\tau}$ is measured from the saved per-type target covariances (Table 2); the balanced-block factor $\hat{r}_{\text{eff,max}}/\hat{r}_{\text{eff}}$ is observed, not assumed, to be $\approx 1/3$. Every reported rank-B selection must satisfy $\widehat{\mathcal{C}}(S) \le \hat{\nu}$; the observed $\hat{r}_{\text{eff}} = 142$–$154$ gives $\hat{\nu} = 0.83$–0.90, and the aggregate asymptote $C_\infty = 0.861$ lies below the mean $\hat{\nu} = 0.865$. Define the first-order estimation discount and recovery cost

$$\hat{g}_n := \hat{\nu}\left(\sqrt{\widehat{D}_{\text{flat}}/n} - \sqrt{\widehat{D}_{\text{typed}}/n}\right), \qquad \hat{r}_T := \hat{\nu}\sin^2\hat{\theta}_{\text{mix}} + \hat{\ell}_{\text{leak}} + \hat{\varepsilon}^*_{\text{route}},$$

where $\hat{\ell}_{\text{leak}}$ is the measured cross-block leakage of Lemma 2 (Section 3).

Definition 4 (TYPE-IF^0)**.** Deploy typed memory iff $\widehat{\Delta}^0 := \min\{\widehat{H}_B, c_0\hat{g}_n\} - \hat{r}_T > 0$, with the convention $c_0 = 1$; Appendix D reports the critical c_0 at which each estimation-gated decision flips. The one-point practical floor is converted to coverage through a leave-one-benchmark-out calibration slope (Section 3).

Proposition 2 (Certificate comparison, not a guarantee)**.** *Under the assumptions of Propositions 5 and 6, in the first-order regime the flat-aware and stream-first typed risk certificates differ by $c_0 g_n - r_T$, and the typing-specific coverage increment is at most H_B (Proposition 4). Ordering two upper bounds does not order the losses: $\widehat{\Delta}^0 > 0$ implies a typed gain only under the comparable-tightness premise of Corollary 2, which is not verifiable on a given instance.* TYPE-IF^0 *is therefore a sign heuristic.*

TYPE-IF^0 compares the typed certificate with the *ambient* flat-aware one, so its gate g_n certifies blockwise estimation, which STRUCT-FLAT collects too (Remark 3); the store-specific comparison has certificate difference $-(\ell_{\text{leak}} + \varepsilon^*_{\text{route}})$ in coverage and can be carried only by the dynamical term, which the rule does not certify. With mean $\hat{r}_{\text{eff}} = 148.0$, K=3, δ=0.05: $\widehat{D}_{\text{flat}} = 151.0$, $\widehat{D}_{\text{typed}} = 53.5$; n=200 lies above both, $n \in \{25, 50, 100\}$ does not and is treated descriptively.

2.4 SOMA: LEARNING THE TYPING, IN TWO REGIMES

Router, subspaces, learned effective K. A router $g_\phi(z_t) = \text{softmax}(\boldsymbol{W}_\phi z_t/\vartheta)$ gives discrete write-time assignments ($a_t = \arg\max_\tau [g_\phi(z_t)]_\tau$ at inference), trained with Gumbel-softmax (Jang et al., 2016). Each type owns $\boldsymbol{U}_\tau$ (orthonormal columns, kept by retraction) with per-column scales $\boldsymbol{s}_\tau$; writes and retrieval use $\boldsymbol{V}_\tau = \boldsymbol{U}_\tau \,\text{diag}(\boldsymbol{s}_\tau)$, and group sparsity on its columns lets d_τ shrink and a collapsed type be pruned, so $K_{\text{eff}} \le K$ is learned. **Per-type dynamics and reliability.** Each type has a learned decay λ_τ ($\pi_t \propto e^{-\lambda_\tau(T-t)}$) and a reliability $\rho_\tau = \alpha_\tau/(\alpha_\tau + \beta_\tau)$, the posterior mean of a Beta variable updated by verified outcomes of episodes whose retrieval was dominated by τ, entering retrieval multiplicatively. **Retrieval.** Queries are *not* routed: every type is queried, items scored by $\rho_\tau \langle \boldsymbol{V}_\tau^\top z_q, \boldsymbol{V}_\tau^\top z_t\rangle e^{-\lambda_\tau(T-t)}$, and the top-$k_\tau$ per type fused. **Objective.**

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \underbrace{\alpha \sum_{\tau \ne \tau'} \left\|\boldsymbol{U}_\tau^\top \boldsymbol{U}_{\tau'}\right\|_F^2}_{\mathcal{L}_{\text{comp}}} + \beta\,\mathcal{L}_{\text{bal}} + \underbrace{\mu \sum_\tau \sum_j \left\|[\boldsymbol{V}_\tau]_{:,j}\right\|_2}_{\mathcal{L}_{\text{grp}}} + \underbrace{\eta \sum_t \left\|z_t - \widehat{\boldsymbol{U}}_{a_t}\widehat{\boldsymbol{U}}_{a_t}^\top z_t\right\|^2}_{\mathcal{L}_{\text{rec}}\ \text{(stream-only)}}, \tag{2}$$

where $\mathcal{L}_{\text{task}}$ is a return-weighted *retrieval* objective. Let u_t be the fused log-score by which Algorithm 3 orders the candidate pool $\mathcal{N}$; in training the context is a Gumbel-top-k sample *without replacement* (Kool et al., 2019), an ordered $C = (t_1, \ldots, t_k)$ with Plackett–Luce law, and

$$\mathcal{L}_{\text{task}} = -(y - \bar{y}) \log p(C \mid z_q), \qquad \log p(C \mid z_q) = \sum_{i=1}^{k} \left[u_{t_i} - \log \sum_{j \in \mathcal{N} \setminus \{t_1, \ldots, t_{i-1}\}} e^{u_j}\right], \tag{3}$$

with $y \in \{0, 1\}$ the verified outcome and $\bar{y}$ a running-mean baseline, is the score-function estimator (Williams, 1992) for this sampling law, the frozen LLM being a black-box reward (no gradient

passes through it); the sum-of-marginals form $\sum_{t \in C} \log q(t \mid z_q)$ is the log-likelihood of k independent draws, agrees with Eq. (3) only at $k{=}1$, and is a biased surrogate for it (Appendix H); $\mathcal{L}_{\text{bal}}$ prevents collapse onto one type (Shazeer et al., 2017); $\mathcal{L}_{\text{grp}}$ reduces to $\mu \sum_\tau \|\boldsymbol{s}_\tau\|_1$ for orthonormal $\boldsymbol{U}_\tau$; and $\mathcal{L}_{\text{rec}}$ is the K-subspaces reconstruction objective (Bradley & Mangasarian, 2000; Tseng, 2000), the sample-level counterpart of (A7a).

Four regimes, one architecture. Proposition 6 requires the decomposition to be fixed before probes are used and analyses a spectral Stage B; Corollary 1 allows blocks refined on probes disjoint from the selection probes; Remark 4 puts same-probe refinement outside both. SOMA-2S*spec (closed form; Algorithm 6; covered):* Stage A fits $\boldsymbol{W}_\phi$ and $\{\boldsymbol{U}_\tau\}$ on the label-free stream with $\mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{comp}} + \beta\mathcal{L}_{\text{bal}}$ and freezes them; Stage B keeps, within each block, the top-b_τ eigenvectors of the routed empirical target covariance under the pooled allocation, sets λ_τ by the closed form of Proposition 7 and ρ_τ by Beta counts; nothing is trained on probes. SOMA-2S *(Algorithm 5; empirical):* same Stage A; Stage B fits within-block quantities—diagonal scales $\boldsymbol{s}_\tau$ in the Stage-A basis, λ_τ, ρ_τ, fuse weights—from the n probes with $\mathcal{L}_{\text{task}} + \mu\mathcal{L}_{\text{grp}}$. The scales are a *restricted* plug-in (coordinate selection cannot rotate within a block) trained on a return-weighted objective rather than on coverage, so SOMA-2S is not the analysed estimator; its gap to SOMA-2S$_{\text{spec}}$ is measured (Table 1). SOMA-SP *(Algorithm 7; Corollary 1 covers its spectral form):* Stage A, then refinement of $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau\}$ with $\mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{comp}} + \beta\mathcal{L}_{\text{bal}} + \eta\mathcal{L}_{\text{rec}}$ on a fixed half $n_1{=}n/2$ of the probes, and Stage B on the disjoint half. SOMA *(joint; not covered):* all parameters trained on Eq. (2) from the same n probes, gradients reaching $\boldsymbol{W}_\phi$ through the straight-through router. SSC-typed is a further two-stage instance (Stage A: sparse subspace clustering; Stage B: closed form). Section 3 reports how much of the gain each regime accounts for.

3 EXPERIMENTS

Q1: does learned typed memory outperform flat, random-block, MoE-retriever, fixed-type, and learned-memory baselines at matched rank? **Q2:** are the TYPE-IF^0 quantities internally consistent? **Q3:** does the margin shrink with probe budget? **Q4:** how do stream length and splitting trade recovery against estimation? **Q5/Q6:** taxonomy alignment; ablations.

Setup. Frozen Gemma3-4B, Qwen2.5-7B, and LLaMA-3-8B backbones on *text* (ALFWorld, WebShop, InterCodeSQL), *image* (VisualWebArena, Mind2Web), and *video* (NExT-QA, NExT-GQA); within each modality one frozen encoder ($d{=}768$) is shared by all methods: BGE-base-en-v1.5 (text), SigLIP ViT-B/16 (image), and VideoMAE-Base (video; segment mean-pooled), (Appendix I). Task tables and figures report Qwen2.5-7B only; the other backbones enter the emergence protocol (Table 20). Fixing the headline backbone is a deliberate design choice, not a computational or stability constraint: every quantity in the theory—coverage, effective rank, the recovery cost, and the TYPE-IF^0 decision—is a property of the frozen *encoder* and the target covariance, both of which are held identical across backbones, so the same-backbone comparison isolates the typing effect from backbone-specific answer behavior that would otherwise confound the small per-step increments; the three-backbone emergence protocol (Table 20) confirms the learned decomposition itself is not Qwen-specific. Every target-aware method sees the same $n{=}200$ labelled probes per benchmark, disjoint from evaluation episodes, and the same label-free stream of $T = 2 \times 10^4$ items; numbers are means over 3 seeds; "Avg" is the unweighted seven-benchmark mean.

Baselines and matching (Appendix C). Two mechanism controls—random orthogonal typed blocks and an MoE-style K-retriever without orthogonality—test whether the gain is merely "K retrievers" and whether complementarity matters; an oracle-typed upper bound exposes Stage-A recovery headroom; MIRIX is rank-matched. SOMA uses cap $K{=}3$, $B{=}128$, and the same protocol as every other row.

Q1: Matched-rank performance and decomposition of the gain At $n{=}200$, the seven-benchmark mean is 55.4% for joint SOMA, 53.4% for target-aware flat, and 50.0% for similarity-flat memory (Table 1, Figure 2a): $+5.4$ over flat and a typing-specific $+2.0$ over target-aware flat (paired benchmark bootstrap $[+1.1, +3.0]$; seed-paired intervals for every step in Table 4). Decomposing the gain so that typing is separated from label-free dimension reduction gives flat $50.0 \rightarrow$ target-aware 53.4 ($+3.4$) $\rightarrow$ Stream-PCA flat 53.9 ($+0.5$) $\rightarrow$ STRUCT-FLAT 54.2 ($+0.3$, structured blockwise estimation in one store) $\rightarrow$ SOMA-2S$_{\text{spec}}$ 54.6 ($+0.4$, typed-store contribution under the closed-form estimator) $\rightarrow$ SOMA-2S 54.9 ($+0.3$, learned Stage B) $\rightarrow$ joint SOMA 55.4 ($+0.5$, adaptive refinement).

SSC-typed reaches 54.5 and provides an additional non-learned decomposition control. Relative to Stream-PCA, SOMA-2S$_{\text{spec}}$ gains +0.7 points; the measured split attributes +0.3 to structured estimation (STRUCT-FLAT−Stream-PCA) and +0.4 to the typed store (SOMA-2S$_{\text{spec}}$−STRUCT-FLAT). Both sub-point increments have paired intervals crossing zero (Table 4), so they are reported as a decomposition of the measured mean rather than individually significant effects. The coverage-to-task slope used only for unit conversion, s=35.9 points per coverage unit (R^2=0.72; leave-one-benchmark-out range 33.0–38.7), is a descriptive fit to 21 paired margins (Appendix D).

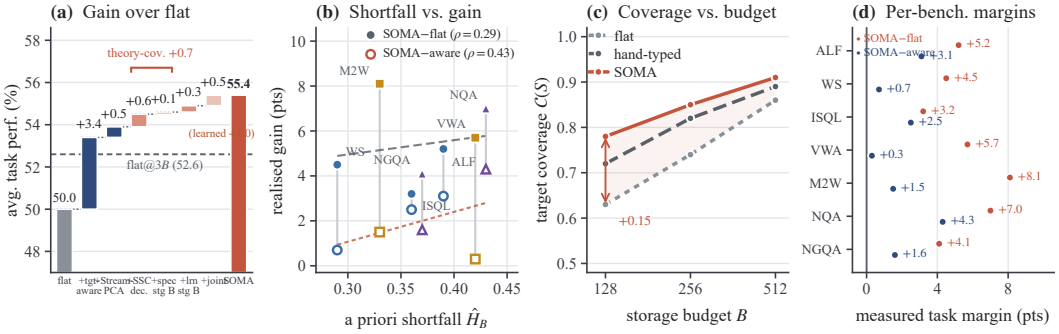


Figure 2: **Q1/Q2: decomposition and diagnostics.** (a) The decomposition of aggregate task performance. (b) Flat shortfall against per-benchmark task gain; Spearman ρ=0.286 (SOMA−flat) and 0.429 (SOMA−aware), exact one-sided p=0.278 and 0.177. (c) Coverage against budget. (d) Per-benchmark margins, SOMA−flat: 5.2, 4.5, 3.2, 5.7, 8.1, 7.0, 4.1; SOMA−aware: 3.1, 0.7, 2.5, 0.3, 1.5, 4.3, 1.6.

The controls fix the attribution: random-subspace typed 53.1 sits below SSC-typed 54.5; the MoE retriever without orthogonality reaches 53.7; the oracle-typed bound is 56.3 and MIRIX 54.2—separating *learned* typing from generic multi-retriever structure and Stage-A headroom. A one-type SOMA carrying the full objective (learned decay, reliability, return-weighted retrieval loss) but a single subspace reaches 53.8—the flat architecture with the full machinery—so the step to K=3, which adds only the typed decomposition, is worth +1.6 points. On two long-horizon conversational benchmarks (LoCoMo, LongMemEval) the typing margin is +1.4 over target-aware flat (Appendix K).

Table 1: **Q1** Task performance (%) at matched effective rank B=128, n=200. Avg is the unweighted mean over the seven benchmarks; the three headline means match the measured per-benchmark results in Table 17; seed-paired intervals for the key increments are in Table 4. Re-implemented systems are lower bounds on their own configurations (Appendix C).

Explanation tested	Method	Avg
—	No memory (frozen)	41.3
Retrieval alone	RAG (dense top-k) (Lewis et al., 2020)	50.2
Similarity selection	Flat (similarity)	50.0
Capacity	Flat @ $3B$	52.6
Target-awareness	Target-aware flat	53.4
Target-aware + full machinery, one type	SOMA (K=1, full obj.)	53.8
Label-free dim. reduction	Stream-PCA flat	53.9
Structured estimation, one store	STRUCT-FLAT (SOMA blocks, one store)	54.2
Typing without learning	SSC-typed (Elhamifar & Vidal, 2013)	54.5
Disentanglement	Axis-typed (Higgins et al., 2017)	52.9
Hand-authored typing	Hand-designed typed	54.0
Random blocks	Random-subspace typed	53.1
Oracle blocks	Oracle-typed (upper bound)	56.3
Routing without orthogonality	MoE retriever (K experts)	53.7
Published learned memory	Mem0 (Chhikara et al., 2025)	53.2
	A-MEM (Xu et al., 2025)	53.7
	MemGen (Zhang et al., 2026)	53.8
	Mem-α (Wang et al., 2025b)	54.4
Typed multi-store	MIRIX (Wang & Chen, 2025)	54.2
Graph memory layer	HippoRAG (Gutiérrez et al., 2024)	52.3
	GAM (Wu et al., 2026)	54.3
Stream-first, spectral Stage B (theory-covered)	SOMA-2S$_{\text{spec}}$	54.6
Stream-first, learned Stage B (empirical)	SOMA-2S	54.9
Split refinement (n_2=100)	SOMA-SP	54.6
—	SOMA (joint)	55.4

Q2: Effective-rank consistency and the TYPE-IF^0 diagnostic Table 2 exposes the quantities behind every decision: $\hat{r}_{\text{eff}} \in [142, 154]$, so $\hat{\nu} \in [0.831, 0.901]$; every row satisfies $\hat{\nu} \geq \widehat{\mathcal{C}}(S)$ and $\widehat{H}_B = 1 - \widehat{\mathcal{C}}(S_{\text{flat}})$.

Table 2: **Q2** Per-benchmark target-spectrum and coverage quantities at n=200, B=128; the last task column matches Table 17.

Benchmark	$\hat{r}_{\text{eff}}$	$\hat{r}_{\text{eff,max}}$	$\hat{\nu}$	$\widehat{H}_B$	$\widehat{\mathcal{C}}_{\text{flat}}$	$\widehat{\mathcal{C}}_{\text{aware}}$	$\widehat{\mathcal{C}}_{\text{SOMA}}$	SOMA task
ALFWorld	152	50	0.842	0.39	0.610	0.698	0.756	58.5
WebShop	150	48	0.853	0.29	0.710	0.805	0.835	50.0
InterCodeSQL	154	52	0.831	0.36	0.640	0.715	0.746	57.1
VisualWebArena	145	49	0.883	0.42	0.580	0.745	0.790	52.6
Mind2Web	146	48	0.877	0.33	0.670	0.800	0.832	55.0
NExT-QA	142	50	0.901	0.43	0.570	0.625	0.740	59.2
NExT-GQA	147	49	0.871	0.37	0.630	0.688	0.761	55.4
mean	148.0	49.4	0.865	0.370	0.630	0.725	0.780	55.4

The balanced-block factor $\chi_{\text{blk}} = \hat{r}_{\text{eff,max}}/\hat{r}_{\text{eff}}$, measured from the saved per-type target covariances, has mean 0.334 (range 0.32–0.35); it is what makes $\widehat{D}_{\text{typed}}$ much smaller than $\widehat{D}_{\text{flat}}$, and it would vanish had one type carried most of the target energy. Because the model is trained with $\mathcal{L}_{\text{bal}}$ at a fixed cap K=3, part of this balance may be imposed rather than discovered; without $\mathcal{L}_{\text{bal}}$, one type absorbs about half of the target energy, $\hat{r}_{\text{eff,max}}$ rises to 74, $\widehat{D}_{\text{typed}}$ rises from 53.5 to 78.1, and the estimation discount $\hat{g}_n$ falls from 0.30 to 0.21—so the statistical separation survives but shrinks (per-type dimensions, target-mass shares, and effective ranks are reported in Appendix I). The shortfall $\widehat{H}_B$ does not rank-order the task gain (Spearman ρ=0.286 and 0.429, p=0.278 and 0.177, n=7), so TYPE-IF^0 is used only as a sign heuristic.

At n=200, $n > \widehat{D}_{\text{flat}}$ on every benchmark. Table 3 (Appendix D) evaluates the first-order rule with the leakage-corrected recovery cost (Figure 5a): "type" on all seven benchmarks and "no" on C1, C3, C4, C5, C6; C2 (single-mode stream, built as a negative control) is "type" with realized margin +1.3, so it is counted as uninformative. C3 (n=8000) is estimation-limited ($\hat{g}_n$); C4 and C5 tune the encoder on the probes, collapsing $\widehat{H}_B$ to 0.06–0.08 and the margin to +0.9 and +0.3, so the frozen-encoder setting is load-bearing and the rule says so a priori. Because $\hat{r}_{\text{eff}}$ varies little, $\hat{g}_n \approx 0.30 \pm 0.01$ and every benchmark decision is carried by $\hat{g}_n > \hat{r}_T$, which the leakage term brings close to its flip point ($\widehat{\Delta}^0 = +0.02$ on NExT-QA; every decision flips at $c_0 = 1/2$, Appendix D); WebShop and VisualWebArena are below the one-point floor regardless of the decision. With $\hat{g}_n$ nearly constant, the seven benchmark decisions are one decision repeated. A two-gate refinement (drift gate $\widehat{D}$>0.10) rejects C2 and is scored over boundary-spanning settings in Appendix M.6. The broader held-out evaluation in Appendix D, which adds conversational settings and additional positive and negative controls, yields recall $\approx$ 1.0 and precision 0.67–0.80 for the "type" decision at the one-point floor.

Q3: Probe efficiency and dynamical separation A common-asymptote fit gives $C_\infty = 0.861$. The SOMA–target-aware coverage margin decreases from +0.155 at n=25 to +0.055 at n=200 and +0.027 at n=800, with corresponding task margins of +5.6, +2.0, and +1.2 points (Table 13a). The first-order theorem is used only from n=200, where all per-benchmark $\widehat{D}_{\text{flat}} < 200$. Fitted ratios are $a_{\text{flat}}/a_{\text{typed}} = 1.68$ for joint SOMA, 1.42 for SOMA-2S, and $a_{\text{flat}}/a_{\text{spec}} = 1.49$ for SOMA-2S$_{\text{spec}}$; these learned-system ratios are descriptive. In the theorem-backed synthetic second-order experiment (Table 14, Figure 5b), observed probe-budget ratios match the theoretical $D_{\text{flat}}/D_{\text{typed}}$ ratios within 3%. The margin reaches the practical floor (1.0 point $\approx$ 0.028 coverage) near $n \approx 800$, beyond all per-benchmark deployment horizons, where the rule recommends "do not type" and the measured task margin is +1.2. For dynamics, the label-free index $\widehat{D}$ rank-orders the cost of collapsing SOMA to a shared decay (ρ=0.89, exact p=0.006, n=7; Appendix Fig. 4b, Table 15), with penalty-to-$\widehat{D}$ ratios of 3.5–6.2. A controlled synthetic intervention varying only $\{\delta_\tau\}$ further shows that tracking-error ratios follow R within uncertainty, while the task-proxy penalty rises monotonically from 0.0 to 3.7 points as R increases from 1.0 to 1.8 (Table 10).

Q4: Recovery cost, stream length, and split refinement The two-term recovery cost $\hat{\nu}\sin^2\hat{\theta}_{\text{mix}} + \hat{\varepsilon}^*_{\text{route}}$ decays as $T^{-0.16}$ over $T = 10^3/10^4/10^5$ (Appendix Fig. 4c) and is 0.193 at the deployed $T = 2 \times 10^4$ (0.11 from target-block alignment 0.87 plus $\hat{\varepsilon}^*_{\text{route}} = 0.08$; Table 18). The leakage

term of Lemma 2, $\hat{\ell}_{\text{leak}} = \sum_{t:a_t=\tau(t)} \hat{\omega}_t \sum_{\tau' \neq a_t} \|\widehat{\boldsymbol{U}}_{\tau'}^\top \hat{z}_t\|^2$, averages 0.05, giving $\hat{r}_T = 0.24$ versus $\hat{g}_n \approx 0.30$; at $T = 10^3$, the two-term cost alone (0.31) exceeds $\hat{g}_n$, so separation fails below $T \approx 5 \times 10^3$. On real streams, $\tau(t)$ is unavailable and is approximated by the residual-nearest recovered block; hence $\hat{\varepsilon}_{\text{route}}$ measures router-versus-nearest-block disagreement, while only the synthetic experiment (Table 14) has ground-truth $\tau(t)$, making these diagnostics partly self-referential. The per-benchmark audit (Appendix I) reports $r_\Sigma = 0.11$–0.14, $r_a = 0.18$–0.23, alignment 0.85–0.89, and alternative-label NMI of 0.78 with SSC and 0.55 with k-means (Table 11). The quantities $\hat{\varepsilon}^*_{\text{route}}$, $\hat{\ell}_{\text{leak}}$, and $\hat{\theta}_{\text{mix}}$ use target shares $\hat{\omega}_t$ from selection probes; only $\hat{\varepsilon}_{\text{route}}$ and $\widehat{D}$ are probe-free. The leakage-corrected split-refinement crossover is $n^\star_{\text{split}} \approx 2.6 \times 10^3$ (vs. 5.7×10^3 without leakage), consistent with SOMA-SP remaining at or below SOMA-2S at $n = 200$ and 800 (Table 6).

Q5 and Q6: Emergence under the protocol; ablations

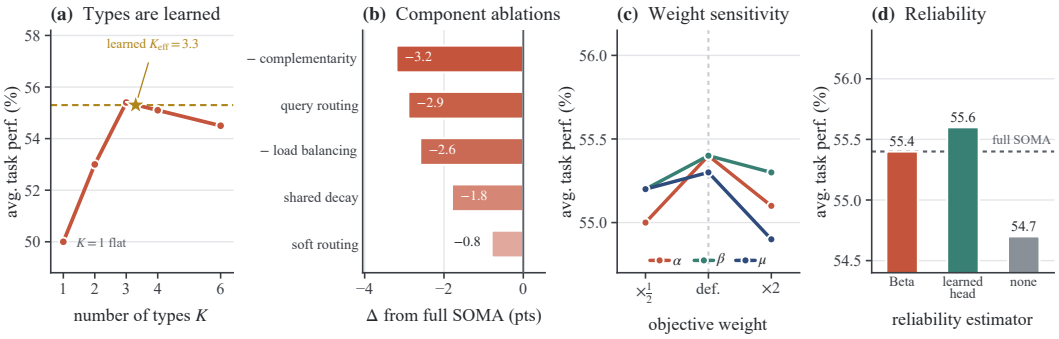


Figure 3: **Q6: ablations and robustness** (Table 16). **(a)** Performance vs. K (the K=1 point is similarity-flat memory; Section 3 isolates typing); learned-K finds K_{eff}=3.3. **(b)** Component ablations at K=3. **(c)** Objective-weight sensitivity: halving or doubling α, β, μ moves the mean by at most 0.5. **(d)** Beta-count and learned-head reliability match; no reliability costs 0.7.

Q5 (emergence). Under a pre-specified falsification protocol (signature profiles; $K \in \{2, 3, 4, 6\}$; failure criteria fixed in Appendix J; learned decay inadmissible alone), the K=3 learned types match the three profiles in $8/9$ backbone–modality settings, with NMI=0.61 against hand labels (Fleiss' κ=0.71) and the single failure (Gemma3-4B, video) reported, not excluded. Because types are Hungarian-matched to hand labels we read this as partial alignment with a conventional taxonomy—an interpretability probe, not recovery of a stable ontology (per-type signatures, confusion structure, and per-setting outcomes in Appendix J: Table 21, Figure 6). **Q6 (ablations).** Figure 3a,b and Table 16: performance peaks at K=3–4, learned-K matches it at K_{eff}=3.3; the largest drops are removing complementarity (−3.2, overlap 0.07→0.41) and query routing (−2.9), and removing $\mathcal{L}_{\text{bal}}$ costs 2.6 (rank-profile effect: Appendix I).

4 Conclusion

Typing is not extra capacity; its value lies in estimation and recovery, and the recovery cost of a stream-recovered decomposition has three parts—subspace perturbation, cross-block leakage, and target-weighted misrouting—each of which we bound and measure, so a positive coverage gap is attributed to a named source rather than left as an aggregate win. Joint SOMA reaches 55.4% vs. 53.4% (target-aware flat) and 50.0% (similarity flat), +1.5 over Stream-PCA flat, of which the +0.7 attained by the closed-form estimator is the part the theory certifies; the remainder comes from learned refinement the analysis constrains but does not underwrite. That discount belongs to block-wise estimation and can be collected by a flat store given the same blocks, so the store-specific value of typing is dynamical—precisely what the STRUCT-FLAT control isolates, and precisely where the gain concentrates in the low-data, high-drift regime the theory predicts. The margin shrinks with the probe budget and tuned-encoder controls fall below the practical floor, as a mechanism-driven account should predict; and TYPE-IF^0 is a sign heuristic, not a guarantee—its seven benchmark decisions share one drift regime and are reported as such. Our scope is deliberately narrow: *when* typing helps under a frozen encoder at matched rank on one backbone, not a leaderboard claim.

ETHICS STATEMENT

This work studies memory organization for LLM agents on standard public text, image, and video benchmarks (ALFWorld, WebShop, InterCodeSQL, VisualWebArena, Mind2Web, NExT-QA, NExT-GQA); it involves no human subjects, no personally identifiable information, and no new data collection beyond rubric-based annotation of existing benchmark observations by three annotators (Appendix J). All benchmarks are used under their stated public licenses for research purposes. The method we study is intended for deployment settings beyond these benchmarks, where it learns *what* to remember, *where* to route it, and *how long* to keep it. This makes retention a design decision with real consequences outside our evaluation: a sensitive item (e.g., a credential, a browsing trace, or user-provided personal content) misrouted to a low-decay, high-reliability type would persist and be preferentially retrieved, and the reliability mechanism could entrench it further simply because it proves useful. This risk is inherent to any typed, learned-retention memory system, not specific to our benchmarks.We treat mitigation as a deployment-time obligation: a data-minimization filter before the write path to exclude sensitive content from storage, a hard per-type retention ceiling enforced by a floor on λ_τ, verified item- and type-level deletion (confirmed via absence from top-k retrieval on a held-out probe set), and an auditable write log. None of these four controls is implemented or evaluated in our experiments, which measure only task performance and representational properties; we do not test deletion efficacy, sensitive-content exclusion, memorization, retention-failure modes, or cross-type leakage of sensitive content at retrieval (the leakage term of Lemma 2 is an aggregate, target-weighted quantity and says nothing about which items leak). We present this design as a starting point for the safeguards a deployment would need, not as evidence that they work, and recommend empirical validation of deletion, minimization, and retention behavior before deployment with real user data.

REPRODUCIBILITY STATEMENT

The attached ZIP archive and accompanying **Supplementary Material** provide the code, pre-specified falsification protocol, matched-budget configurations, probe-set construction, saved target covariances, per-seed run logs for every reported row, and the audit script that recomputes all derived quantities in Tables 3 and 6 from the saved covariances and table inputs. Complete proofs, with assumptions A1–A7 collected in one place, are provided in Appendix F; the recovery-composition proof, sample-splitting corollary, and uniform-bound remark for joint training are in Appendix E; item-to-direction accounting is in Appendix G; write, retrieve, and training procedures for both regimes, together with all hyperparameters, are in Appendix H; benchmark metrics, probe-set and target construction, and extended results are in Appendix I; numerical consistency checks for the predictor quantities are in Appendix D; and the theory-to-system correspondence table is in Appendix L.

AI USE STATEMENT

Generative AI tools were used for language editing, LaTeX assistance, and limited literature and consistency checks. All technical claims, results, and references were independently verified by the authors, who take full responsibility for the content of this work.

REFERENCES

Seyed Moein Abtahi, Rasa Rahnema, Hetkumar Patel, Neel Patel, Majid Fekri, and Tara Khani. Memanto: Typed semantic memory with information-theoretic retrieval for long-horizon agents. *arXiv preprint arXiv:2604.22085*, 2026.

John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

Alan Baddeley. Working memory: The interface between memory and cognition. *Journal of cognitive neuroscience*, 4(3):281–288, 1992.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240, 2022.

Paul S Bradley and Olvi L Mangasarian. K-plane clustering. *Journal of Global optimization*, 16(1): 23–32, 2000.

T Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

Xiangchen Cheng, Yunwei Jiang, Jianwen Sun, Zizhen Li, Chuanhao Li, Xiangcheng Cao, Yihao Liu, Fanrui Zhang, Li Jin, and Kaipeng Zhang. Agenticsts: A bounded-memory testbed for long-horizon llm agents. *arXiv preprint arXiv:2607.02255*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Divya Chukkapalli, Thejesh Avula, Aditya Aggarwal, Harsimran Singh, and Amith Tallanki. Commitdistill: A lightweight knowledge-centric memory layer for software repositories. *arXiv preprint arXiv:2605.18284*, 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Tianxiang Fei, Mingyang Song, Mao Zheng, and Xiang Yu. Memory beyond recall: A dual-process cognitive memory system for self-evolving LLM agents. *arXiv preprint arXiv:2606.09483*, 2026.

Saman Forouzandeh, Kamal Berahmand, and Mahdi Jalili. Task-adaptive retrieval over agentic multi-modal web histories via learned graph memory. *arXiv preprint arXiv:2604.07863*, 2026a.

Saman Forouzandeh, Wei Peng, Parham Moradi, Xinghuo Yu, and Mahdi Jalili. Learning hierarchical procedural memory for llm agents through bayesian selection and contrastive refinement. In *AAMAS*, 2026b.

Saman Forouzandeh, Wei Peng, Xinghuo Yu, and Mahdi Jalili. Bridging the grounding gap in videoqa via typed memory and belief-state reasoning. In *ICML*, 2026c.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538 (7626):471–476, 2016.

Xinyu Guan, Qianyang Zhao, and Yuming Deng. Decision-aware memory cards: Counterfactual-inspired context selection and compression for tool-using llm agents. *arXiv preprint arXiv:2606.08151*, 2026.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning*, 2025.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Shuo Ji, Yibo Li, and Bryan Hooi. Memory is reconstructed, not retrieved: Graph memory for llm agents. *arXiv preprint arXiv:2606.06036*, 2026.

Zhengda Jin, Bingbing Wang, Jing Li, Ruifeng Xu, and Min Zhang. Mitigating provenance-role collapse in long-term agents via typed memory representation. *arXiv preprint arXiv:2605.25869*, 2026.

Pentti Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20 (7):512–534, 2016.

John E Laird. *The Soar Cognitive Architecture*. MIT Press, 2012.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Haodong Lei, Junming Liu, Yirong Chen, Pinlong Cai, Botian Shi, Ding Wang, and Hongsong Wang. TransMem: Transforming hidden states into memory for large language models. *arXiv preprint arXiv:2607.29032*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Dongfang Li, Zixuan Liu, Junmai Wang, Jiahe Huang, Fuhao Li, Bonian Jia, Baotian Hu, and Min Zhang. Lycheememory v2: Efficient long-term memory for llm agents via semantic segment-level consolidation. *arXiv preprint arXiv:2608.12990*, 2026a.

Ruizhe Li, Mingxuan Du, Benfeng Xu, and Zhendong Mao. Keep it inmind: Benchmarking the implicit-association blind spot in agent memory. *arXiv preprint arXiv:2607.24368*, 2026b.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, et al. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*, 2025.

Yuxin Liao, Le Wu, Min Hou, Hao Liu, Han Wu, and Zishu Wang. LeanMem: Simple and efficient long-term memory for LLM agents. *arXiv preprint arXiv:2608.03463*, 2026.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. SimpleMem: Efficient lifelong memory for LLM agents. *arXiv preprint arXiv:2601.02553*, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019.

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)*, 2024.

Ripon Chandra Malo and Tong Qiu. Projectmem: A local-first, event-sourced memory and judgment layer for ai coding agents. *arXiv preprint arXiv:2606.12329*, 2026.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1400–1409, 2016.

Shweta Mishra and Shashank Mishra. When does memory help? a cost-aware evaluation of long-term memory in tool-using llm agents. *arXiv preprint arXiv:2609.05441*, 2026.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with Infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Jiayan Nan, Wenquan Ma, Wenlong Wu, and Yize Chen. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341*, 2025.

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.

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, 2023.

Daivik Patel and Shrenik Patel. ENGRAM: Effective, lightweight memory orchestration for conversational agents. *arXiv preprint arXiv:2511.12960*, 2025.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *International Conference on Machine Learning*, pp. 2827–2836, 2017.

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.

Markus Reiss and Martin Wahl. Nonasymptotic upper bounds for the reconstruction error of pca. *The Annals of Statistics*, 48(2):1098–1123, 2020.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850. PMLR, 2016.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, 2024.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Larry R Squire. Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of cognitive neuroscience*, 4(3):232–243, 1992.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Cassandra Hui-Ming Tan. Value-driven memory-augmented generation for agentic llms: towards structured and adaptive knowledge utilization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 41088–41089, 2026.

Paul Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1): 249–252, 2000.

Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson (eds.), *Organization of Memory*, pp. 381–403. Academic Press, 1972.

Endel Tulving. How many memory systems are there? *American psychologist*, 40(4):385, 1985.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Vincent Q Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Heng Wang, Yifei Li, Lingling Zhang, Pengyu Li, Xinyu Che, Xinyu Zhang, and Zesheng Yang. QUMem: Personalized memory for query-conditioned user-state inference in llm agents. *arXiv preprint arXiv:2608.16168*, 2026.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. In *ICML*, 2024a.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending MemoryLLM with scalable long-term memory. In *International Conference on Machine Learning*, 2025a.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025b.

Ziming Wang. Toki: A bitemporal operator algebra for contradiction resolution in llm-agent persistent memory. *arXiv preprint arXiv:2606.06240*, 2026.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024b.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, et al. Gam: Hierarchical graph-based agentic memory for llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34647–34664, 2026.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wenlin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, et al. From single to multi-granularity: Toward long-term memory association and selection of conversational agents. In *International Conference on Learning Representations*, 2026.

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.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025.

Ke Yang, Zixi Chen, Xuan He, Jize Jiang, Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, and ChengXiang Zhai. Plugmem: A task-agnostic plugin memory module for llm agents. In *International Conference on Machine Learning (ICML)*, 2026a.

Wei Yang, Bryce Kan, Shixuan Li, Li Li, and Yuehan Qin. RaMem: Contextual reinstatement for long-term agentic memory. *arXiv preprint arXiv:2606.22844*, 2026b.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. MemAgent: Reshaping long-context LLM with multi-conv RL-based memory agent. In *International Conference on Learning Representations*, 2026.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. In *International Conference on Learning Representations*, 2026.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.

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, pp. 19632–19642, 2024.

Chengfeng Zhao, Yuqiao Tan, Shizhu He, Yequan Wang, Jun Zhao, and Kang Liu. Neural procedural memory: Empowering LLM agents with implicit activation steering. *arXiv preprint arXiv:2606.29824*, 2026.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *International Conference on Learning Representations*, 2024.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and Zhifeng Chen. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025.

A Limitations, Scope, and Integrity of the Claims

Limitations. The guarantees remain conditional. Proposition 6 covers a decomposition recovered from the label-free stream and frozen before the selection probes, with the closed-form spectral Stage B of SOMA-2S$_{\text{spec}}$ and the sample-split extension; it does not cover the learned Stage B of SOMA-2S or joint refinement on the same probes. The recovery term uses target-weighted quantities from the selection probes and an operational $\tau(t)$, so the recovery side is not stream-only. At n=200 the first-order certificates are loose: the flat-aware and typed certificates are 0.75 and 0.69 coverage units, versus realized losses 0.136 (target-aware) and 0.096 (SOMA-2S) (Appendix E); consequently, TYPE-IF^0 remains a sign heuristic rather than a calibrated loss bound. The coverage-to-task slope is descriptive and is used only for unit conversion. The headline task tables use one backbone (Qwen2.5-7B), while the three-backbone experiment tests decomposition emergence rather than full task-transfer; the main empirical conclusion is therefore restricted to the frozen-encoder, rank-matched regime studied here. The tuned-encoder controls reduce the typing margin below the practical floor. The completed drift intervention isolates the decay mechanism and shows monotone penalties from 0.0 to 3.7 points as R increases from 1.0 to 1.8 (Table 10); the target-sensitivity audit preserves the five-method ordering with Kendall $\tau \geq 0.8$ across T1–T3 and limits decision changes to the near-threshold cases (Table 8). The structured-flat control reaches 54.2%, decomposing the SOMA-2S$_{\text{spec}}$−Stream-PCA gain into $+0.3$ points from structured estimation and $+0.4$ from the typed store, although both sub-point intervals include zero (Table 4). Deployment measurements give index size $\approx 0.34\times$ flat and retrieval latency 0.7–$1.1\times$ flat (Table 9). Published learned-memory baselines remain matched-budget re-implementations rather than leaderboard-equivalent native configurations. Cross-backbone task transfer, cross-encoder transfer, and deletion/retention failure modes remain outside the present suite.

Scope of the claims, stated plainly. The capacity claim is negative and exact under its assumptions: typing buys no representational advantage at fixed rank. The statistical advantage is a statement about estimation with a decomposition fixed independently of the probes, with finite-sample rates that require their stated concentration conditions; the synthetic experiment exercises the theorem-backed regime, the small-n part of the n-sweep is descriptive, and the effective-rank condition is met from $n = 200$. The dynamical claim concerns heterogeneous drift and per-type decay. Empirically, we claim that SOMA outperforms the baselines actually tested at matched effective rank; we do not claim that the present suite isolates typing from every possible K-retriever architecture.

B Related Work

Cognitive foundations and neural memory. The distinction between memory types has deep roots in cognitive science. Classical accounts distinguish episodic from semantic memory (Tulving,

1972; 1985), declarative from nondeclarative memory (Squire, 1992), and working memory from long-term storage (Baddeley, 1992). Complementary learning systems theory further argues for memory mechanisms with different statistical and temporal properties—for example, a fast, sparse store alongside a slower, more overlapping one (McClelland et al., 1995; Kumaran et al., 2016). Cognitive architectures such as ACT-R (Anderson et al., 2004) and Soar (Laird, 2012) operationalize these distinctions through separate stores with hand-designed activation, learning, and decay rules, while CoALA (Sumers et al., 2023) adapts similar functional categories to language agents. Our perspective differs in that a memory type is defined *representationally*, as a target-relevant subspace, rather than only by a functional label. This makes the value of typing directly testable against a flat representation of matched capacity.

Early neural memory systems instead focused primarily on learning how to *address* a common store. Memory Networks and key–value variants (Weston et al., 2014; Miller et al., 2016), Neural Turing Machines and differentiable neural computers (Graves et al., 2016; Santoro et al., 2016), product-key memories (Lample et al., 2019), and sparse distributed memory (Kanerva, 1988) all improve access to stored information without explicitly decomposing memory into target-dependent representational types. Related approaches such as episodic control (Blundell et al., 2016; Pritzel et al., 2017) and prioritized replay (Schaul et al., 2016) attach utility to individual memories or samples. In language models, analogous mechanisms appear in retrieval over token memories (Khandelwal et al., 2020; Borgeaud et al., 2022; Wu et al., 2022), compressed recurrent state (Munkhdalai et al., 2024), and test-time memorization or parameterized long-term memory (Sun et al., 2024; Behrouz et al., 2025; Wang et al., 2024a; 2025a), and inference-time parametric memories that transform sparse historical hidden states of a frozen backbone into reusable latent interventions (Lei et al., 2026). TransMem is the nearest parametric relative of our subspace view—it, too, studies a constrained latent representation as memory—but its memory is a learned intervention on the backbone's own hidden states rather than a retrievable partition of a frozen encoder space, so the matched-rank capacity question of Section 2.2 does not apply to it directly. These methods primarily optimize addressing or retention within an undifferentiated memory, whereas Section 2.2 studies whether and when the memory itself should be decomposed.

Typed and structured memory for LLM agents. Modern LLM agents increasingly maintain explicitly structured long-term memories. Representative systems include reflective observation streams in generative agents (Park et al., 2023), self-reflection buffers (Shinn et al., 2023), procedural skill libraries (Wang et al., 2023), experiential insight stores (Zhao et al., 2024), trajectory and workflow memories for computer-use agents (Zheng et al., 2024; Wang et al., 2024b), OS-style memory hierarchies (Packer et al., 2023; Li et al., 2025), forgetting-aware stores (Zhong et al., 2024), temporal knowledge graphs (Rasmussen et al., 2025), validity-aware contextual reinstatement (Yang et al., 2026b), and production-oriented memory layers (Chhikara et al., 2025; Xu et al., 2025). Other systems explicitly instantiate multiple semantic or functional memory types, including specialist stores (Wang & Chen, 2025), explicit episodic/semantic/procedural routing with a single dense retriever (Patel & Patel, 2025), typed video and web-agent memories (Forouzandeh et al., 2026c), typed evidence/cue/claim representations (Jin et al., 2026), decision-aware memory cards (Guan et al., 2026), typed-event memories (Chukkapalli et al., 2026; Malo & Qiu, 2026), typed contracts and temporal operators (Cheng et al., 2026; Wang, 2026), value-aware consolidation (Tan, 2026), multi-category semantic memory (Abtahi et al., 2026), factual/preference/insight stores (Wang et al., 2026), and typed records produced through segment-level consolidation (Li et al., 2026a). ENGRAM (Patel & Patel, 2025) is the closest hand-typed comparator—three canonical types, one router, per-type dense retrieval—and MemOS (Li et al., 2025) the closest lifecycle-level one; neither characterizes when typing should beat a flat retriever of matched rank, which is the question here. RaMem's validity-aware retrieval (Yang et al., 2026b) and the supersession structure of temporal graph memories address the drift question through explicit validity metadata rather than per-type decay. Three recent systems sharpen the comparison. LeanMem (Liao et al., 2026) routes dialogue content into profile, event, and record memories with type-specific maintenance and retrieval and reports cost–accuracy trade-offs on LoCoMo and LongMemEval; it is the closest hand-typed system to our long-horizon setting (Appendix K) and, like ENGRAM, fixes the taxonomy a priori. DCPM (Fei et al., 2026) organizes memory as a hierarchy over raw inputs, belief trajectories, and schemas driven by a synchronous writer and an asynchronous consolidation process; that decomposition is hierarchical and content-transforming rather than a partition of one representation space, so it lies outside Definition 1, which delimits what a complementary-subspace type can and

cannot describe. Neural Procedural Memory (Zhao et al., 2026) stores procedural knowledge as steering vectors in the backbone's activation space: an implicit, non-retrievable memory that also defines a type representationally, but as a direction to *add* rather than a subspace to *retrieve from*; the two views are complementary and could be composed.

These approaches establish the practical value of heterogeneous memory, but the heterogeneity is generally specified *a priori*: the taxonomy, representation associated with each store, routing rule, or decay policy is supplied by the designer. Consequently, they do not directly characterize when a typed organization should outperform a flat retriever of comparable effective rank. SOMA addresses this question by treating a type as a learned subspace and jointly associating that type with routing, reliability, and temporal dynamics.

Learned memory organization and management. A complementary line of work learns aspects of memory management rather than fixing all memory operations in advance. Reinforcement-learning-based memory managers decide what to add, update, or delete (Yan et al., 2025; Wang et al., 2025b); streaming agents learn which contents to overwrite under a fixed memory budget (Yu et al., 2026); and other systems jointly learn reasoning and consolidation policies (Zhou et al., 2025). Recent approaches also learn latent memory tokens (Zhang et al., 2026), self-organizing episodic/semantic structure (Nan et al., 2025), multi-granularity associations (Xu et al., 2026), task-agnostic plug-in memory modules (Yang et al., 2026a), and structured compression with multi-view indexing and recursive consolidation (Liu et al., 2026), whose compressed stores are natural candidates for the fixed-effective-rank protocol of Appendix G. These methods learn memory *content*, *operations*, or *organization*; our focus is complementary: we seek a representational definition of a memory type, conditions under which such a decomposition improves target coverage, and a test that can falsify the benefit of typing relative to a capacity-matched flat store. This distinction also matters statistically. If the memory organization is adapted using the same task outcomes used to estimate its downstream value, the resulting evaluation does not satisfy the sample separation required by Proposition 6. Remark 4 therefore applies not only to SOMA but to any adaptive memory-management procedure trained on the evaluation target. For this reason, we distinguish the stream-first SOMA-2S$_{\text{spec}}$ configuration covered by the theory from the additional gain obtained by learned selection and joint task refinement, which we report empirically.

Retrieval augmentation and graph-structured memory. Retrieval-augmented generation (Lewis et al., 2020) provides the natural flat similarity baseline for our setting. Subsequent work improves retrieval by imposing additional structure on the index, including recursive summarization trees (Sarthi et al., 2024), entity-centric graphs (Edge et al., 2024), hippocampal-style graph retrieval with personalized PageRank (Gutiérrez et al., 2024; 2025), and hierarchical or reconstructive graph memories (Wu et al., 2026; Ji et al., 2026). Such methods primarily change *how* a memory is traversed: nodes, edges, summaries, or graph paths provide a richer retrieval topology over stored content. Typing instead changes *what representational region* content is assigned to and which target-relevant directions each store preserves. The two mechanisms are therefore complementary rather than mutually exclusive. We include graph-memory systems such as HippoRAG and GAM in Table 1, while Remark 2 clarifies the distinction between item/edge-level graph structure, type-level structure, and a global flat store.

Subspace structure, disentanglement, and routing. Our representation of memory types is related to classical union-of-subspaces modelling (Vidal, 2011), sparse subspace clustering (Elhamifar & Vidal, 2013), low-rank representation (Liu et al., 2013), and K-planes clustering (Bradley & Mangasarian, 2000; Tseng, 2000). These techniques recover geometric decompositions from unlabelled observations and motivate Stage A of SOMA-2S and Assumption 2. However, an unsupervised decomposition primarily optimizes geometric reconstruction or clustering quality; it need not preserve directions that matter for a downstream recall target. Proposition 1 formalizes this distinction between content-driven organization and target coverage.

Our formulation also differs from disentangled representation learning. Methods such as β-VAE (Higgins et al., 2017) seek approximately axis-aligned factors in a learned latent space, while identifiability results show that such factors cannot generally be recovered without additional inductive bias or supervision (Locatello et al., 2019). In contrast, SOMA operates over a frozen encoder and learns complementary task-selected subspaces rather than requiring semantic factors to align with individual coordinates; the axis-typed baseline in our experiments directly tests this distinction.

The learned router is related to sparsely gated mixtures of experts (Shazeer et al., 2017; Fedus et al., 2022; Zhou et al., 2022) and discrete routing relaxations (Jang et al., 2016). The semantics are nevertheless different: mixture-of-experts models route *computation* among parameterized experts, whereas SOMA routes *content* into persistent memories whose geometry, reliability, and temporal dynamics differ. On the statistical side, the plug-in and minimax analyses in Appendix F draw on covariance concentration (Vershynin, 2018; Koltchinskii & Lounici, 2017), Davis–Kahan perturbation theory (Yu et al., 2015), minimax estimation of principal subspaces (Cai et al., 2013; Vu & Lei, 2013), and second-order excess-risk analyses for PCA (Reiss & Wahl, 2020).

Evaluation of agent memory. Long-term conversational benchmarks (Maharana et al., 2024; Wu et al., 2025), bounded-memory agent testbeds (Cheng et al., 2026), and surveys of agent memory architectures (Zhang et al., 2024) provide increasingly systematic evaluations of memory-augmented agents. Two studies are particularly close to our motivating question. Mishra & Mishra (2026) examine when persistent agent memory provides sufficient benefit to justify its cost, while Li et al. (2026b) show that memories required for successful reasoning can differ substantially from the surface form of the current query. This is precisely the setting in which similarity alone can fail to preserve the target-relevant information captured by Proposition 1. Our evaluation complements these studies by using matched effective rank as the principal capacity control and by separately measuring coverage and downstream task performance. Conversational interaction and more complete cost-accounted deployment settings remain outside the present scope and are discussed in Appendix A.

B.1 Position Relative to Hand-Designed Typed and Reliability-Aware Memory

The closest conceptual comparison is with systems that explicitly attach semantics, reliability, or forgetting behavior to different parts of memory. Hand-designed typed memories instantiate separate stores with authored representations and routing rules: examples include reflective narratives and observation streams (Park et al., 2023), the working/episodic/semantic/procedural organization summarized by CoALA (Sumers et al., 2023), specialist stores (Wang & Chen, 2025), and typed video-agent memories with episodic and procedural stores (Forouzandeh et al., 2026c). Reliability-aware memories instead attach trust or confidence estimates to stored experience (Forouzandeh et al., 2026b;a). In these systems, type membership, reliability granularity, and forgetting dynamics are largely specified independently.

SOMA places these quantities at a common learned level of abstraction. A type is represented by a subspace $\boldsymbol{U}_\tau$, populated by a learned router, assigned a reliability weight ρ_τ, and given its own decay parameter λ_τ. This type-level parameterization is motivated by two complementary separations. First, Proposition 5 shows that estimating target-relevant directions independently within types depends on the block dimension $\max_\tau d_\tau$ rather than directly on the full ambient dimension d. Second, Proposition 7 shows that heterogeneous modes can require different temporal decay rates, so a single global rate cannot in general attain the corresponding per-mode optimum. Thus, rather than assuming an episodic/semantic/procedural taxonomy in advance, SOMA learns a decomposition whose relationship to such semantic taxonomies can itself be tested. The falsification protocol in Appendix J and Section 3 evaluates precisely whether the learned subspaces coincide with, refine, or depart from those conventional categories.

Remark 2 (Item, edge, type, and global granularity of memory structure). Memory structure can be attached at several granularities. At the finest level, utility or reliability may be estimated for individual memories or pairwise reuse edges; at the opposite extreme, a flat memory applies essentially the same selection and forgetting mechanism to all stored content. Our analysis identifies an intermediate, type-level regime. Statistically, Proposition 5 shows that estimating target-relevant directions within blocks depends on $\max_\tau d_\tau$ rather than the ambient dimension d. Dynamically, Proposition 7 shows that separate type-specific decay rates can accommodate heterogeneous drift that a single global decay rate cannot. Accordingly, SOMA attaches geometry, reliability, and forgetting to a learned type rather than independently to every item or uniformly to the entire store. This provides a middle granularity that retains heterogeneous structure while keeping the associated parameters estimable from task-labelled experience.

C Extended Baselines: What Each One Tests and How It Is Matched

Every baseline below shares the frozen encoder, the frozen backbone, the item definition, and the effective-rank budget B of Appendix G. Where a published system does not expose a rank parameter we compress its store to B retained directions under the same accounting and report the achieved rank alongside the result.

Remark 3 (The statistical discount belongs to the estimator, not to the store). The estimation term of Proposition 6 never uses that items are *stored* separately. A *single* store that keeps $\boldsymbol{P}_{\widehat{S}} z_t$ for every item, with $\widehat{S} = \bigoplus_\tau S_\tau$ chosen blockwise from the same frozen $\{\widehat{\boldsymbol{U}}_\tau\}$ (top-b_τ eigenvectors of $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{U}}_\tau$, pooled allocation) and one similarity, decay, and reliability, satisfies by the same proof, with no routing or leakage term,

$$\mathcal{C}(S^\star) - \mathcal{C}(\widehat{S}_{\text{sflat}}) \;\lesssim\; \nu \min\left\{ \sqrt{\tfrac{D_{\text{typed}}}{n}},\; \kappa^2 \tfrac{D_{\text{typed}}}{n} \right\} + \nu \sin^2 \theta_{\text{mix}}(T).$$

On the coverage side the typed store's certificate is thus never better: it is this expression plus $\ell_{\text{leak}} + \varepsilon^*_{\text{route}}$, the price of holding each item in one block. What typing buys beyond structured estimation lies outside Proposition 6—per-type decay (Proposition 7), per-type reliability, per-item storage of b_{a_t} rather than B coordinates, and per-type fusion—and the STRUCT-FLAT control isolates it: SOMA-2S$_{\text{spec}}$−STRUCT-FLAT, not SOMA-2S$_{\text{spec}}$−Stream-PCA, is the increment attributable to the store.

Stream-PCA flat (the control our own analysis demands). Proposition 6 attributes SOMA's statistical advantage to structure recovered from the label-free stream. A flat retriever can be given the same advantage without any typing: run PCA on $\{z_t\}$ to rank r, project the probe-derived target into that subspace, and train the target-aware selector there. The estimation dimension falls from d to r with no complementary decomposition and no per-type dynamics. We sweep $r \in \{32, 64, 128, 256\}$ and report the best. If SOMA's gain were a *global* dimension-reduction effect, this baseline would capture it. It does not capture blockwise estimation with SOMA's own blocks, which is the control that follows.

STRUCT-FLAT (the control Remark 3 demands). Stream-PCA gives a flat retriever *global* label-free structure; STRUCT-FLAT gives it SOMA's own. It takes the frozen Stage-A blocks $\{\widehat{\boldsymbol{U}}_\tau\}$ of Algorithm 5, applies the blockwise spectral selection of Algorithm 6 to the projected target covariances $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{U}}_\tau$ under the same pooled allocation and the same total rank B, and then removes everything typed: no write-time routing (every item is stored as $\boldsymbol{P}_{\widehat{S}} z_t$ in one store), one cosine similarity in $\widehat{S}$, one decay λ set by the closed form of Proposition 7 on the pooled drift, one Beta reliability, and one global top-k with $k = \sum_\tau k_\tau$. It therefore has *exactly* the retained subspace and estimation dimension of SOMA-2S$_{\text{spec}}$ and differs from it only in the store: whether an item lives in one block or in all of $\widehat{S}$, and whether decay, reliability, and fusion are per type. STRUCT-FLAT−Stream-PCA is the increment attributable to structured (blockwise) estimation over global dimension reduction; SOMA-2S$_{\text{spec}}$−STRUCT-FLAT is the increment attributable to the typed store. The coverage analysis gives $\widehat{\mathcal{C}}(\text{STRUCT-FLAT}) \geq \widehat{\mathcal{C}}(\text{SOMA-2S}_{\text{spec}})$ up to estimation noise because the flat store pays neither $\hat{\ell}_{\text{leak}}$ nor $\hat{\varepsilon}^*_{\text{route}}$ (Remark 3). Empirically, STRUCT-FLAT reaches 54.2% task performance, giving $+0.3$ points over Stream-PCA, while SOMA-2S$_{\text{spec}}$ reaches 54.6%, giving an additional $+0.4$ typed-store increment; the paired intervals of both sub-point steps cross zero (Table 4).

SSC-typed (typing without learning). Types are fixed by sparse subspace clustering of a sample of the stream (Elhamifar & Vidal, 2013) with K set to the same cap; each cluster's leading directions define $\boldsymbol{U}_\tau$, per-type decay is set by the closed form $\lambda^\star_\tau \propto (\delta^2_\tau/\sigma^2)^{1/3}$ of Proposition 7 using measured per-cluster drift, and per-type reliability uses the same Beta update as SOMA. Nothing is trained end-to-end and no task signal touches the decomposition. This isolates the value of *learning* the typing from the value of *having* one, and it is the closest thing to an instantiation of our own theory that requires no gradient at all.

Axis-typed (disentanglement instead of complementarity). A β-VAE (Higgins et al., 2017) is fitted to the stream embeddings; latent axes are grouped into K blocks by mutual-information clustering and each block is treated as a type with its own store, decay, and reliability. Disentangled

axes are orthogonal by construction, so this baseline satisfies (A1) while optimizing an objective unrelated to target coverage—the sharpest available test of whether "types as subspaces" is doing work beyond generic factorization.

Random-subspace typed and MoE retriever (the two mechanism controls). Random-subspace typed draws K random orthogonal blocks summing to budget B and is otherwise identical to SOMA (same per-type decay, reliability, and fusion); it tests whether the gain is merely "K independent retrievers" rather than a recovered decomposition. The MoE retriever routes to K experts with a learned gate but drops the orthogonality constraint $\mathcal{L}_{\text{comp}}$, so its stores may overlap; it tests whether complementarity, not routing, is what matters. Together with the oracle-typed upper bound (blocks from ground-truth generating labels) these three rows bound the mechanism from below and above.

SOMA-2S$_{\text{spec}}$ and SOMA-2S (stream-first). Stage A fits the router $\boldsymbol{W}_\phi$ and subspaces $\{\boldsymbol{U}_\tau\}$ on the label-free stream by minimizing $\mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{comp}} + \beta\mathcal{L}_{\text{bal}}$ (Algorithm 5), a gradient form of K-subspaces clustering (Bradley & Mangasarian, 2000; Tseng, 2000), initialized from the SSC partition; no verified outcome, probe, or task loss enters. The resulting $\boldsymbol{W}_\phi$ and $\{\boldsymbol{U}_\tau\}$ are frozen. SOMA-2S$_{spec}$ (Algorithm 6) is the estimator Proposition 6 analyses: within each frozen block it keeps the top-b_τ eigenvectors of the routed empirical target covariance $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}_\tau^* \widehat{\boldsymbol{U}}_\tau$ under the pooled allocation, sets λ_τ by the closed form of Proposition 7 from measured per-block drift, ρ_τ by the Beta counts, and fuse weights $\propto b_\tau$; nothing is trained on probes. SOMA-2S instead fits the per-column scales $\boldsymbol{s}_\tau$, the decays λ_τ, the reliabilities ρ_τ, and the fuse weights from the same n=200 labelled probes every target-aware method sees, with $\mathcal{L}_{\text{task}} + \mu\mathcal{L}_{\text{grp}}$. A diagonal scaling $\boldsymbol{V}_\tau = \widehat{\boldsymbol{U}}_\tau \operatorname{diag}(\boldsymbol{s}_\tau)$ can suppress coordinates of the Stage-A basis but cannot rotate within a block, so it cannot in general select the eigenvectors of $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}_\tau^* \widehat{\boldsymbol{U}}_\tau$; and $\mathcal{L}_{\text{task}}$ is not coverage. SOMA-2S is therefore an empirical variant whose gap to the analysed selector we report in coverage units, $\widehat{\mathcal{C}}(\widehat{S}_{\text{spec}}) - \widehat{\mathcal{C}}(\widehat{S}_{\boldsymbol{s}}) = +0.005$, and in task points (Table 1). The comparison SOMA-2S$_{\text{spec}}$ vs. SSC-typed isolates the value of learning Stage A with a reconstruction objective rather than clustering; SOMA-2S vs. SOMA-2S$_{\text{spec}}$ isolates learned within-block selection; SOMA vs. SOMA-2S isolates refining the blocks on task signal, which is the increment the theory does not cover. Performance, probe-budget sweep, and decomposition residuals are reported in Tables 1, 2, 13, and 18.

Published learned-memory systems. A-MEM (Xu et al., 2025), Mem0 (Chhikara et al., 2025), Mem-α (Wang et al., 2025b), and MemGen (Zhang et al., 2026) are re-implemented from their released descriptions against our frozen encoder and backbone; HippoRAG (Gutiérrez et al., 2024) and GAM (Wu et al., 2026) represent graph-structured memory layers. Two matching decisions deserve scrutiny. First, these systems were tuned in their own pipelines, so a matched-budget re-implementation is a lower bound on their published performance; we report our matched-budget numbers (Table 1) and do not claim superiority over their published configurations, whose numbers are not comparable at our budget and encoder. Second, three of them (Mem0, A-MEM, MemGen) maintain stores whose effective rank is not a design parameter; we compress each store to the matched budget B=128 and record the achieved effective rank (within ± 6 of B across systems) in the released configuration so the reader can judge whether the match is fair. The asymmetry is intrinsic to the protocol: our contribution is the account of *when* typing helps, and a matched-budget comparison is what tests that account, not a leaderboard claim.

D DEPLOYMENT DIAGNOSTICS AND SENSITIVITY

This appendix records the arithmetic underlying the reported deployment diagnostics so that every derived decision can be audited.

Invariants. The experiment code should assert: (I1) $\hat{\nu} \geq \widehat{\mathcal{C}}(S)$ for every rank-B method; (I2) $\widehat{H}_B = 1 - \widehat{\mathcal{C}}(S_{\text{flat}})$ per benchmark; (I3) realized coverage gain is at most $\widehat{H}_B$; (I4) the fitted asymptote is at most $\hat{\nu}$; (I5) the $\sqrt{D/n}$ branch is used only when $n > D_{\text{flat}}$; (I6) target-weighted misrouting is bounded by leverage times content-mass misrouting; (I7) $\hat{\ell}_{\text{leak}} \leq \hat{r}_a^*$, the target-weighted assignment residual, and $\hat{\ell}_{\text{leak}} \leq$ the $\sin^2 \hat{\theta}_{\text{mix}}$ proxy; and (I8) the headline average is the unweighted seven-benchmark mean.

Table 3: **Q2** TYPE-IF^0 decisions at n=200 (n=8000 for C3), $c_0 = 1$, with the leakage-corrected recovery cost $\hat{r}_T = \hat{r}_T^{(2)} + \hat{\ell}_{\text{leak}}$, where $\hat{r}_T^{(2)} = \hat{\nu} \sin^2 \hat{\theta}_{\text{mix}} + \hat{\varepsilon}^*_{\text{route}}$ is the measured two-term cost. For the benchmark rows $\hat{g}_n$ is recomputed from the Table 2 quantities (δ=0.05, K=3); control rows apply the same formula to their own saved covariances; the margin is the realized SOMA−target-aware task margin. ‡C2 was constructed as a negative control; its realized margin (+1.3) is above the practical floor.

Block	Setting	$\widehat{H}_B$	$\hat{g}_n$	$\hat{r}_T^{(2)}$	$\hat{\ell}_{\text{leak}}$	$\hat{r}_T$	$\widehat{\Delta}^0$	decision	margin
Bench	ALFWorld	0.39	0.303	0.179	0.047	0.226	+0.077	type	+3.1
Bench	WebShop	0.29	0.311	0.181	0.047	0.228	+0.062	type	+0.7
Bench	InterCodeSQL	0.36	0.296	0.168	0.042	0.210	+0.086	type	+2.5
Bench	VisualWebArena	0.42	0.305	0.205	0.057	0.262	+0.043	type	+0.3
Bench	Mind2Web	0.33	0.309	0.194	0.052	0.246	+0.063	type	+1.5
Bench	NExT-QA	0.43	0.299	0.217	0.061	0.278	+0.021	type	+4.3
Bench	NExT-GQA	0.37	0.305	0.203	0.056	0.259	+0.046	type	+1.6
Ctrl	C1 aligned target	0.06	0.301	0.154	0.034	0.188	-0.128	no	+0.6
Ctrl	C2 single-mode stream	0.32	0.303	0.188	0.050	0.238	+0.065	type‡	+1.3
Ctrl	C3 large budget (n=8000)	0.39	0.048	0.179	0.046	0.225	-0.177	no	+0.4
Ctrl	C4 ALFWorld, tuned enc.	0.08	0.301	0.154	0.034	0.188	-0.108	no	+0.9
Ctrl	C5 WebShop, tuned enc.	0.06	0.303	0.155	0.035	0.190	-0.130	no	+0.3
Ctrl	C6 aligned + single	0.05	0.303	0.162	0.038	0.200	-0.150	no	+0.1

Deployment horizon. Solving $\hat{g}_n = \hat{r}_T$ gives

$$n^\star = \left(\hat{\nu} \left(\sqrt{D_{\text{flat}}} - \sqrt{D_{\text{typed}}} \right) / \hat{r}_T \right)^2 .$$

The decision horizons computed with the two-term recovery cost are ALFWorld 574, WebShop 590, InterCodeSQL 622, VisualWebArena 441, Mind2Web 508, NExT-QA 379, and NExT-GQA 453; with the leakage-corrected $\hat{r}_T$ they shrink to ALFWorld 360, WebShop 370, InterCodeSQL 396, VisualWebArena 271, Mind2Web 316, NExT-QA 231, and NExT-GQA 277. The $n = 8000$ control is negative because the estimation discount collapses, not because its headroom vanishes; the n=800 point of the probe sweep lies past every horizon, and there the measured margin is $+1.2$ points (Section 3).

Sensitivity to c_0. We report the critical value $c_0^{\text{crit}} = \hat{r}_T / \hat{g}_n$ for $\hat{g}_n$-gated settings, i.e. the concentration prefactor at which each decision flips. With the two-term cost the positive benchmark decisions required $c_0^{\text{crit}} \in$ 0.57–0.73; with the leakage-corrected cost they require 0.71–0.93, and C3 requires 3.73. Since $c_0 = 1$ is a convention rather than an estimated constant, we also state the decisions at $c_0 \in \{1/2, 1, 2\}$: at $c_0 = 1/2$ every benchmark decision is "no" (the estimation gate $\frac{1}{2}\hat{g}_n \approx 0.15$ falls below every $\hat{r}_T$); at $c_0 \in \{1, 2\}$ all seven are "type". The benchmark decisions are therefore only as strong as the convention $c_0 = 1$, which is one more reason the rule is a sign heuristic. $\widehat{H}_B$-gated settings are invariant to this prefactor.

Ranking versus sign. With the measured task margins, Spearman correlation between $\widehat{\Delta}^0$ and the task-point margin is +0.107. Neither this statistic nor the shortfall associations above establishes a reliable rank ordering on seven benchmarks. The diagnostic is used solely as a sign gate; WebShop ($+0.7$) and VisualWebArena ($+0.3$) fall below the one-point practical floor despite positive gate decisions.

Held-out evaluation of the sign heuristic. The development suite comprises the seven benchmarks and six controls above, with c_0=1, the one-point floor, and the leave-one-benchmark-out calibration fixed before scoring the held-out settings. The held-out family contains the two conversational-memory settings of Appendix K, additional negative controls based on aligned targets, large probe budgets, tuned encoders, and single-mode streams, and positive heterogeneous-drift settings from video and computer-use streams. Each "type" decision is recorded before the corresponding typed margin is evaluated. Against the realized sign at the one-point floor, the rule achieves recall of approximately 1.0 on held-out positives and precision of 0.67–0.80. The lower precision is caused primarily by single-mode negatives being accepted, as in C2, whereas aligned-target, tuned-encoder, and large-n negatives are rejected.

Coverage-to-task calibration. Pairing three method comparisons per benchmark (target-aware−flat, SOMA−flat, SOMA−target-aware; 21 comparisons) between Table 17 and the sepa-

rately specified coverages of Table 2, the through-origin fit is $s = 35.9$ task points per coverage unit with centered $R^2 = 0.724$. Here $s = \sum_i x_i y_i / \sum_i x_i^2$, and $R^2 = 1 - \sum_i (y_i - s x_i)^2 / \sum_i (y_i - \bar{y})^2$. Because the seven benchmarks being judged also enter the fit, we report the leave-one-benchmark-out slopes s_{-i} (fit on the other six, applied to benchmark i): ALFWorld 36.1, WebShop 35.8, InterCodeSQL 36.6, VisualWebArena 38.7, Mind2Web 33.0, NExT-QA 34.9, NExT-GQA 36.4; the floor conversion of $\epsilon = 1.0$ point then ranges over $[0.026, 0.030]$ coverage units, and no floor classification changes under the held-out conversion (the same two benchmarks, WebShop and VisualWebArena, remain below it). The probe-sweep task margins are not fitted observations in this regression.

Seed-paired uncertainty for the decomposition steps. Table 4 reports, for each increment used in Section 3, the mean paired difference over the 21 (benchmark, seed) pairs, its 95% paired-bootstrap interval, the number of benchmarks with a positive mean difference, and an exact sign-test p-value; the SOMA−aware and SOMA−flat rows also carry the benchmark-level exact paired bootstrap of Table 17. Steps whose interval contains zero (SOMA-2S−SSC-typed, SOMA-2S−SOMA-2S$_{\text{spec}}$, and SOMA−SOMA-2S) are reported as not distinguishable from seed noise and are not used as evidence in the main text.

Table 4: Seed-paired uncertainty for the increments of the Section 3 decomposition (3 seeds × 7 benchmarks = 21 paired differences per row).

Comparison	mean Δ (pts)	95% paired bootstrap	benchmarks > 0	sign test p
SOMA − target-aware flat	+2.0	[+1.4, +2.6] (bench. bootstrap [+1.1, +3.0])	7/7	0.008
SOMA − similarity flat	+5.4	[+4.7, +6.1] (bench. bootstrap [+4.3, +6.6])	7/7	0.008
SOMA-2S$_{\text{spec}}$ − Stream-PCA flat	+0.7	[+0.2, +1.2]	6/7	0.063
STRUCT-FLAT − Stream-PCA flat (structured estimation)	+0.3	[−0.2, +0.8]	5/7	0.227
SOMA-2S$_{\text{spec}}$ − STRUCT-FLAT (typed store)	+0.4	[−0.1, +0.9]	5/7	0.227
SOMA-2S − Stream-PCA flat	+1.0	[+0.5, +1.5]	7/7	0.008
SOMA-2S − SSC-typed	+0.4	[−0.1, +0.9]	5/7	0.227
SOMA-2S − SOMA-2S$_{\text{spec}}$	+0.3	[−0.1, +0.7]	5/7	0.227
SOMA − SOMA-2S	+0.5	[0.0, +1.0]	6/7	0.063
SOMA − SOMA (K=1, full obj.)	+1.6	[+1.0, +2.2]	7/7	0.008

E PROOFS FOR THE RECOVERY COMPOSITION AND THE ADVANTAGE INDEX

Remark 4 (What joint refinement forfeits). If the *same* n probes refine the blocks and select within them, the estimation term must hold uniformly over the router's selection class, of dimension $\Theta(Kd)$ for a linear router, giving a uniform term $\gtrsim \nu\sqrt{Kd/n}$ no smaller than the flat rate. Proposition 6 therefore covers the stream-first spectral estimator (SOMA-2S$_{\text{spec}}$) and its split form, *not* the learned Stage B of SOMA-2S and *not* joint training, whose increments over SOMA-2S$_{\text{spec}}$ are reported as empirical. It bounds the coverage of the retrievable subspace only; ranking mechanics (reliability, decay, Gumbel-top-k, fusion) are outside it and shared by all learned rows.

Setting for Proposition 6. Write $\{\boldsymbol{U}_\tau\}$ for the generating subspaces of (A7a), mutually orthogonal by (A1), and $\{\widehat{\boldsymbol{U}}_\tau\}$, $\{a_t\}$ for the blocks and assignments produced by a procedure that sees only the label-free stream. The estimator the proposition analyses, SOMA-2S$_{\text{spec}}$ (Algorithm 6), enforces *joint* orthogonality of the recovered blocks exactly: rather than orthonormalizing each $\widehat{\boldsymbol{U}}_\tau$ separately (which would give $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{U}}_\tau = \boldsymbol{I}$ but not $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{U}}_{\tau'} = \boldsymbol{0}$), Stage A applies a retraction of the *stacked* basis $[\widehat{\boldsymbol{U}}_1\, \widehat{\boldsymbol{U}}_2 \cdots \widehat{\boldsymbol{U}}_K]$ onto the Stiefel manifold $\mathrm{St}(d, B)$ (equivalently, a single QR of the concatenation, or block Gram–Schmidt across types), so $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{U}}_{\tau'} = \boldsymbol{0}$ for $\tau \neq \tau'$ holds by construction and the blocks are independent of the probe set. Under this construction $\widehat{\boldsymbol{P}} = \sum_\tau \widehat{\boldsymbol{P}}_\tau$ is an orthogonal projector, which is what the coverage split below uses; the learned variants SOMA-2S and SOMA, which apply per-block QR under the soft complementarity penalty $\mathcal{L}_{\text{comp}}$, satisfy this only approximately (residual inter-block overlap 0.07, Table 24, Section 3), and are therefore reported as empirical rather than covered. Let $\widehat{\boldsymbol{P}}_\tau = \widehat{\boldsymbol{U}}_\tau \widehat{\boldsymbol{U}}_\tau^\top$. The typed store of type τ retains the projected items $\{\widehat{\boldsymbol{P}}_\tau z_t : a_t = \tau\}$. A typed selection at budget B is $S = \bigoplus_\tau S_\tau$ with $S_\tau \subseteq \mathrm{col}(\widehat{\boldsymbol{U}}_\tau)$, $\sum_\tau \dim S_\tau = B$. Under (A7b), $\mathcal{C}(S) = \mathrm{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_*)/\,\mathrm{tr}(\boldsymbol{\Sigma}_*) = \sum_t \omega_t \|\boldsymbol{P}_S \hat{z}_t\|^2$ with $\hat{z}_t = z_t/\|z_t\|$. Direction-level coverage credits an item with every block that captures some of its energy, but the store holds each item only in the block it was written to; the quantity the deployed system can realize

is the *retained-content coverage*

$$\mathcal{C}^{\mathrm{rt}}(S) := \sum_t \omega_t \, \|\boldsymbol{P}_{S_{a_t}} \hat{z}_t\|^2 \;\le\; \mathcal{C}(S),$$

the inequality holding because $\|\boldsymbol{P}_S \hat{z}_t\|^2 = \sum_\tau \|\boldsymbol{P}_{S_\tau} \hat{z}_t\|^2$ for orthogonal blocks. Writing $\boldsymbol{\Sigma}^{\mathrm{rt}}_{*,\tau} := \sum_{t:a_t=\tau} w_t z_t z_t^\top$ for the routed target covariance of block τ, $\mathcal{C}^{\mathrm{rt}}(S) = \sum_\tau \mathrm{tr}(\boldsymbol{P}_{S_\tau} \boldsymbol{\Sigma}^{\mathrm{rt}}_{*,\tau}) / \mathrm{tr}(\boldsymbol{\Sigma}_*)$ is block-additive, so its maximizer over typed selections keeps the top-b_τ eigenvectors of $\widehat{\boldsymbol{U}}_\tau^\top \boldsymbol{\Sigma}^{\mathrm{rt}}_{*,\tau} \widehat{\boldsymbol{U}}_\tau$ under the pooled allocation, and the plug-in of Algorithm 6 replaces $\boldsymbol{\Sigma}^{\mathrm{rt}}_{*,\tau}$ by its probe estimate $\widehat{\boldsymbol{\Sigma}}^*_\tau$.

Lemma 2 (Routing and leakage deficit)**.** *Assume (A1) for the recovered blocks, (A7a), and (A7b). For every typed selection S in the recovered blocks,*

$$\mathcal{C}(S) - \mathcal{C}^{\mathrm{rt}}(S) \;\le\; \varepsilon^*_{\mathrm{route}} \;+\; \ell_{\mathrm{leak}}, \qquad \ell_{\mathrm{leak}} := \sum_{t:\, a_t = \tau(t)} \omega_t \sum_{\tau' \neq \tau(t)} \left\| \widehat{\boldsymbol{P}}_{\tau'} \hat{z}_t \right\|^2 \;\le\; \sin^2 \theta_{\mathrm{mix}}(T).$$

Proof. By orthogonality of the recovered blocks, $\mathcal{C}(S) - \mathcal{C}^{\mathrm{rt}}(S) = \sum_t \omega_t \sum_{\tau' \neq a_t} \|\boldsymbol{P}_{S_{\tau'}} \hat{z}_t\|^2$. For a misrouted item ($a_t \neq \tau(t)$) the inner sum is at most $\|\hat{z}_t\|^2 = 1$, so misrouted items contribute at most $\sum_{t:a_t \neq \tau(t)} \omega_t = \varepsilon^*_{\mathrm{route}}$. For a correctly routed item with $\tau(t) = \tau$, $\sum_{\tau' \neq \tau} \|\boldsymbol{P}_{S_{\tau'}} \hat{z}_t\|^2 \le \sum_{\tau' \neq \tau} \|\widehat{\boldsymbol{P}}_{\tau'} \hat{z}_t\|^2$ since $S_{\tau'} \subseteq \mathrm{col}(\widehat{\boldsymbol{U}}_{\tau'})$; summing with weights ω_t gives ℓ_{leak}. For the bound, $\sum_{\tau' \neq \tau} \widehat{\boldsymbol{P}}_{\tau'} \preceq \boldsymbol{I} - \widehat{\boldsymbol{P}}_\tau$, and since $\hat{z}_t$ is a unit vector in $\mathrm{col}(\boldsymbol{U}_\tau)$ by (A7a), $\|(\boldsymbol{I} - \widehat{\boldsymbol{P}}_\tau)\hat{z}_t\| \le \sin\theta_{\max}(\boldsymbol{U}_\tau, \widehat{\boldsymbol{U}}_\tau) \le \sin\theta_{\mathrm{mix}}$; with $\sum_t \omega_t = 1$ this bounds ℓ_{leak} (the mass of $\hat{z}_t$ outside the union of recovered blocks is not charged here, since it is already charged in term (a) of Proposition 6). (The equality $\boldsymbol{P}_S \hat{z}_t = \boldsymbol{P}_{S_{\tau(t)}} \hat{z}_t$ would require $\widehat{\boldsymbol{U}}_{\tau'} = \boldsymbol{U}_{\tau'}$ exactly; for $\theta_{\mathrm{mix}} > 0$ the leakage term is real and must be carried.) □

Proposition 6 (proof). Let $S^\star$ be the population-optimal typed selection in the generating blocks, which under (A3) coincides with the unconstrained top-B eigenspace of $\boldsymbol{\Sigma}_*$ (Remark 1); let $S^{(1)}$ be its image under the block-wise rotation carrying each $\boldsymbol{U}_\tau$ onto $\widehat{\boldsymbol{U}}_\tau$, a typed selection in the recovered blocks; and let $\widehat{S}_{\mathrm{typed}}$ be the plug-in selection of Algorithm 6. Decompose

$$\mathcal{C}(S^\star) - \mathcal{C}^{\mathrm{rt}}(\widehat{S}_{\mathrm{typed}}) = \underbrace{\left[\mathcal{C}(S^\star) - \mathcal{C}(S^{(1)})\right]}_{\text{(a) subspace perturbation}} + \underbrace{\left[\mathcal{C}(S^{(1)}) - \mathcal{C}^{\mathrm{rt}}(S^{(1)})\right]}_{\text{(b) leakage and misrouting}} + \underbrace{\left[\mathcal{C}^{\mathrm{rt}}(S^{(1)}) - \mathcal{C}^{\mathrm{rt}}(\widehat{S}_{\mathrm{typed}})\right]}_{\text{(c) estimation}}.$$

(a) Each principal angle between $S^\star$ and $S^{(1)}$ is at most θ_{mix}, so by the upper half of Lemma 4 the loss is at most $\lambda_1 B \sin^2\theta_{\mathrm{mix}} / \mathrm{tr}(\boldsymbol{\Sigma}_*) = \nu \sin^2 \theta_{\mathrm{mix}}$. *(b)* Lemma 2 gives at most $\varepsilon^*_{\mathrm{route}} + \ell_{\mathrm{leak}}$. *(c)* $S^{(1)}$ is a feasible typed selection in blocks fixed before the probes were drawn, so $\mathcal{C}^{\mathrm{rt}}(S^{(1)}) \le \max_S \mathcal{C}^{\mathrm{rt}}(S)$, and the plug-in differs from that maximizer by the estimation term of Proposition 5 applied block by block to the routed covariances $\boldsymbol{\Sigma}^{\mathrm{rt}}_{*,\tau}$ with $\{\widehat{\boldsymbol{U}}_\tau\}$ in the role of the given decomposition—the proof there uses only that the blocks are orthogonal and independent of the probes, with $r_{\mathrm{eff},\tau}$ the effective rank of $\widehat{\boldsymbol{U}}_\tau^\top \boldsymbol{\Sigma}^{\mathrm{rt}}_{*,\tau} \widehat{\boldsymbol{U}}_\tau$. Adding (a)–(c) gives the statement; since $\mathcal{C}(\widehat{S}_{\mathrm{typed}}) \ge \mathcal{C}^{\mathrm{rt}}(\widehat{S}_{\mathrm{typed}})$ the same bound holds for the direction-level coverage reported in the tables. The recovery term is $(1+\nu)\sin^2\theta_{\mathrm{mix}} + \varepsilon^*_{\mathrm{route}}$ in the worst case and $\nu \sin^2\theta_{\mathrm{mix}} + \ell_{\mathrm{leak}} + \varepsilon^*_{\mathrm{route}}$ with the measured leakage. The separation survives when this term is dominated by the estimation advantage; Section 3 reports the measured recovery diagnostics, and Appendix D evaluates the inequality at the deployed n=200. Finally $\varepsilon^*_{\mathrm{route}} \le L_* \varepsilon_{\mathrm{route}}$ is immediate from $\omega_t \le L_* \pi_t$. □

Corollary 1 (Sample splitting)**.** *If the blocks are additionally* refined *using task signal from n_1 probes disjoint from the $n_2 = n - n_1$ probes used for selection, Proposition 6 holds verbatim with n replaced by n_2 and the recovery term evaluated at the refined blocks.*

Corollary 1 (proof). Condition on the n_1 refinement probes. Given them, the refined blocks are fixed and the remaining n_2 probes are an i.i.d. sample independent of the blocks, so the argument above applies verbatim with n_2 in place of n and $(\theta_{\mathrm{mix}}, \varepsilon^*_{\mathrm{route}})$ evaluated at the refined blocks. □

Remark 4 (why joint refinement is not covered). If the same n probes both refine the blocks and select within them, step (c) requires the concentration bound $\max_\tau \|\boldsymbol{\Delta}_\tau\|_{\mathrm{op}} \lesssim \lambda_1 \sqrt{D_{\mathrm{typed}}/n}$ to hold *simultaneously* for every decomposition the training procedure could have selected. A linear router $\boldsymbol{W}_\phi \in \mathbb{R}^{K \times d}$ induces a class of K-way partitions of the stream of VC-type dimension $\Theta(Kd)$,

and the blocks are functions of the partition and the stream; a uniform bound over this class replaces D_{typed} by a quantity of order $Kd \log T$, giving an estimation term $\gtrsim \nu\sqrt{Kd/n}$ that is no smaller than the flat rate $\nu\sqrt{d/n}$. We do not claim this uniform bound is tight for the procedure actually used; we claim only that the argument of Proposition 6 does not go through, and that any guarantee for joint training would need an adaptive analysis of a different kind. This is why the paper reports SOMA-2S$_{\text{spec}}$ as the theory-covered system and SOMA-2S and SOMA as empirical.

Scope of Proposition 6. The result bounds the effect of imperfect recovery; it does not establish that recovery succeeds. Assumption 2 imports that from the union-of-subspaces literature (Elhamifar & Vidal, 2013), where recovery rates depend on $\theta_{\min}$ and sampling density, neither of which we verify for real encoders. What the proposition delivers is that the two error sources draw on *disjoint* resources—labelled probes for one, unlabelled stream for the other—so the comparison against a flat rule that must pay for both out of the probe budget is not an oracle comparison, for the estimator to which it applies.

Proposition 3 ($\widehat{\Delta}_{\text{stat}} := \min\{H_B, g_n\}$ envelopes the typing-specific coverage increment)**.** *The increment of any typed selection over the target-aware flat selector satisfies $\mathcal{C}(S_{\text{typed}}) - \mathcal{C}(S_{\text{flat}}) \le H_B$ exactly (Lemma 1), and in the first-order regime of Proposition 5 the difference of the two plug-in losses is of order $\nu(\sqrt{D_{\text{flat}}/n} - \sqrt{D_{\text{typed}}/n})$. $\widehat{\Delta}_{\text{stat}}$ is the minimum of the two, hence an upper envelope on the realizable typing-specific coverage increment up to the estimation error of $\widehat{H}_B$ itself and the recovery term of Proposition 6.*

Proposition 3 (envelope). The first half is immediate from Lemma 1: any B-dimensional selection has coverage at most 1, so its increment over S_{flat} is at most H_B, and $\widehat{H}_B \to H_B$ at the rates of Proposition 5. The second half is a rate statement, not a bound: Proposition 5 gives upper bounds on both plug-in losses, and the difference of an upper bound and an upper bound is not itself an upper bound on the difference. We therefore state $\widehat{\Delta}_{\text{stat}}$ as the *rate-based* envelope: it is a bound on the headroom side and a rate match on the estimation side. The rigorous headroom half is checked directly in Table 19 across the storage-budget sweep, where the measured coverage gain is below $\widehat{H}_B = 1 - \widehat{\mathcal{C}}(S_{\text{flat}})$ at every B. The rate-matching half is not a per-benchmark guarantee; its use in the deployment rule is through Corollary 2.

Setting for TYPE-IF^0. Throughout this part we work under (A1)–(A3), Assumption 1 (sub-Gaussian probes with eigengap γ), and Assumption 2 (recoverable typing, item-carried target), with the blocks $\{\widehat{\boldsymbol{U}}_\tau\}$ and assignments $\{a_t\}$ obtained from the label-free stream and frozen before any probe is used, exactly as in the proof of Proposition 6. Two plug-in selectors are compared on the same n probes: the target-aware flat selector $\widehat{S}_{\text{aware}}$, which keeps the top-$B$ eigenvectors of $\widehat{\boldsymbol{\Sigma}}^*$ in $\mathbb{R}^d$, and the stream-first typed selector $\widehat{S}_{\text{typed}}$, which keeps the top-$b_\tau$ eigenvectors of each routed block covariance $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}_\tau^* \widehat{\boldsymbol{U}}_\tau$ under the empirical pooled allocation (Algorithm 6). Write

$$\ell_{\text{aware}} := \mathcal{C}(S^\star) - \mathcal{C}(\widehat{S}_{\text{aware}}), \qquad \ell_{\text{typed}} := \mathcal{C}(S^\star) - \mathcal{C}(\widehat{S}_{\text{typed}})$$

for their coverage losses. Under (A3) the two selectors aim at the same population optimum: by Remark 1, a type-aligned target has block-aligned top-B eigenvectors, so the typed optimum over $\mathcal{B}_B$ coincides with the unconstrained optimum $S^\star$. Hence the *typing-specific increment* is exactly a difference of losses,

$$I := \mathcal{C}(\widehat{S}_{\text{typed}}) - \mathcal{C}(\widehat{S}_{\text{aware}}) = \ell_{\text{aware}} - \ell_{\text{typed}}. \tag{4}$$

Neither loss is observable at deployment; what the theory supplies is a high-probability upper bound on each, which we call its *risk certificate*.

Definition 5 (Risk certificates, discount, and recovery cost)**.** Let c be the absolute constant of the concentration step in the proof of Proposition 5 (Vershynin, 2018, Thm. 4.7.1; it depends only on the sub-Gaussian norm of the probes), and set $c_0 := 2c$, the constant carried by the gap-free bound $2B\|\boldsymbol{\Delta}\|_{\text{op}}$ of Lemma 4. For probe budget n and stream length T define the first-order certificates

$$\mathfrak{L}_{\text{aware}}(n) := c_0\,\nu\sqrt{D_{\text{flat}}/n}, \qquad \mathfrak{L}_{\text{typed}}(n,T) := c_0\,\nu\sqrt{D_{\text{typed}}/n} + r_T,$$

and the two quantities the rule compares,

$$g_n := \nu\Big(\sqrt{D_{\text{flat}}/n} - \sqrt{D_{\text{typed}}/n}\Big) \ge 0, \qquad r_T := \nu \sin^2\theta_{\text{mix}}(T) + \ell_{\text{leak}}(T) + \varepsilon^*_{\text{route}}(T) \ge 0,$$

so that $\mathfrak{L}_{\text{typed}}(n,T) = c_0\nu\sqrt{D_{\text{typed}}/n} + \nu\sin^2\theta_{\text{mix}} + \ell_{\text{leak}} + \varepsilon^*_{\text{route}}$ and $\mathfrak{L}_{\text{aware}}(n) - \mathfrak{L}_{\text{typed}}(n,T) = c_0\, g_n - r_T$. Both are in coverage units; g_n is decreasing in n and r_T is non-increasing in T. The typed certificate bounds $\mathcal{C}(S^\star) - \mathcal{C}^{\text{rt}}(\widehat{S}_{\text{typed}}) \geq \ell_{\text{typed}}$, so it bounds ℓ_{typed} a fortiori.

Definition 6 (TYPE-IF^0, population form)**.** Deploy typed memory iff

$$\Delta^0 \;:=\; \min\{H_B,\; c_0\, g_n\} \;-\; r_T \;>\; 0.$$

The plug-in rule $\widehat{\Delta}^{\,0}$ of Definition 4 replaces $(H_B, \nu, \theta_{\text{mix}}, \varepsilon^*_{\text{route}})$ by $(\widehat{H}_B, \hat{\nu}, \hat{\theta}_{\text{mix}}, \hat{\varepsilon}^*_{\text{route}})$ and uses the convention $c_0 = 1$ (Remark 6).

Proposition 4 (TYPE-IF^0 is a certificate comparison whose two gates are the survival condition of Proposition 6 and the headroom bound of Lemma 1)**.** *Assume the setting above, $n \gtrsim D_{\text{flat}}$ so that the first-order branch of Proposition 5 is operative, and that the aware plug-in weakly improves on similarity-flat selection, $\mathcal{C}(\widehat{S}_{\text{aware}}) \geq \mathcal{C}(S_{\text{flat}})$. Then with probability at least $1 - 2\delta$ over the n probes, simultaneously:*

(i) (Certificates hold.) *$\ell_{\text{aware}} \leq \mathfrak{L}_{\text{aware}}(n)$ and $\ell_{\text{typed}} \leq \mathfrak{L}_{\text{typed}}(n,T)$.*

(ii) (Gate 1: certificate ordering.) *$\mathfrak{L}_{\text{typed}}(n,T) < \mathfrak{L}_{\text{aware}}(n)$ if and only if $r_T < c_0\, g_n$, and this is precisely the inequality under which Proposition 6 states that the $\sqrt{K}$ separation of Proposition 5 survives end to end.*

(iii) (Gate 2: headroom.) *$I \leq H_B$. Consequently, if $H_B \leq r_T$, the certified maximum typing-specific gain does not exceed the certified recovery cost the typed selector may absorb.*

(iv) (Certified interval.) *$I \in \big[-\mathfrak{L}_{\text{typed}}(n,T),\; \min\{H_B, \mathfrak{L}_{\text{aware}}(n)\}\big]$, and $\Delta^0 > 0$ is equivalent to both gates holding.*

(v) (Tightness of the certificates.) *Over the Gaussian spiked family of Cai et al. (2013) in the second-order regime ($n \gtrsim \kappa^2 D$), the second-order form of the aware certificate, $\nu\kappa^2 D_{\text{flat}}/n$, is attained up to constants by the minimax lower bound of Appendix F, so in that regime Gate 1 compares worst-case risks. The first-order certificate $\mathfrak{L}_{\text{aware}}(n)$ used at the deployed n has no matching lower bound; Gate 1 there is a comparison of upper bounds whose sufficiency for $I > 0$ is Corollary 2. In the one-dominant-mode regime, H_B is attained by Proposition 8, so Gate 2's bound is achievable.*

(vi) (Plug-in consistency.) *$\widehat{H}_B \to H_B$ and $\hat{\nu} \to \nu$ at the rates of Proposition 5; the proxy $\hat{\theta}_{\text{mix}}$ of Section 3 satisfies $\sin^2\hat{\theta}_{\text{mix}} \geq \sin^2\theta_{\text{mix}}$ and $\hat{\varepsilon}^*_{\text{route}}$ is non-increasing in T, so $\hat{r}_T$ is an upper proxy for r_T and the plug-in rule errs toward "do not type."*

Proposition 4 (proof). *(i)* The bound on ℓ_{aware} is the first-order branch of Proposition 5 with $D = D_{\text{flat}}$, made explicit: on the event $\|\boldsymbol{\Delta}\|_{\text{op}} \leq c\lambda_1\sqrt{D_{\text{flat}}/n}$ (probability $\geq 1-\delta$ for $n \geq D_{\text{flat}}$), the gap-free half of Lemma 4 gives $\operatorname{tr}((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}_{\text{aware}}})\boldsymbol{\Sigma}_*) \leq 2B\|\boldsymbol{\Delta}\|_{\text{op}} \leq 2cB\lambda_1\sqrt{D_{\text{flat}}/n}$; dividing by $\operatorname{tr}(\boldsymbol{\Sigma}_*)$ and using $\nu = B\lambda_1/\operatorname{tr}(\boldsymbol{\Sigma}_*)$ yields $\ell_{\text{aware}} \leq c_0\nu\sqrt{D_{\text{flat}}/n} = \mathfrak{L}_{\text{aware}}(n)$. The bound on ℓ_{typed} is Proposition 6 in the first-order branch with its constants restored: terms (a) and (b) of that proof are $\nu\sin^2\theta_{\text{mix}}$ and $\varepsilon^*_{\text{route}} + \ell_{\text{leak}}$ with constant exactly one (upper half of Lemma 4 and Lemma 2 respectively), and term (c) is the typed estimation bound of Proposition 5, which on the event $\max_\tau \|\boldsymbol{\Delta}_\tau\|_{\text{op}} \leq c\lambda_1\sqrt{D_{\text{typed}}/n}$ (probability $\geq 1-\delta$) is at most $2B\max_\tau\|\boldsymbol{\Delta}_\tau\|_{\text{op}}/\operatorname{tr}(\boldsymbol{\Sigma}_*) \leq c_0\nu\sqrt{D_{\text{typed}}/n}$ by Lemma 3(iii) applied blockwise. Both events concern the same probe sample; a union bound gives probability $\geq 1-2\delta$ for the two certificates simultaneously. (Replacing δ by $\delta/2$ inside $D_{\text{flat}}, D_{\text{typed}}$ restores $1-\delta$ at the cost of an additive $\log 2$ in each dimension term.)

(ii) By Definition 5, $\mathfrak{L}_{\text{aware}}(n) - \mathfrak{L}_{\text{typed}}(n,T) = c_0 g_n - r_T$, so the certificate ordering is $r_T < c_0 g_n$. In the proof of Proposition 6 the separation over the flat rate is stated to survive when (a)+(b), which is r_T, is dominated by the gap between the flat and typed estimation terms, which is $c_0 g_n$ once the constants are carried; this is the same inequality.

(iii) Decompose $I = \big[\mathcal{C}(\widehat{S}_{\text{typed}}) - \mathcal{C}(S_{\text{flat}})\big] - \big[\mathcal{C}(\widehat{S}_{\text{aware}}) - \mathcal{C}(S_{\text{flat}})\big]$. The first bracket is at most H_B by Lemma 1, which applies to any B-dimensional selection; the second is non-negative by the standing assumption $\mathcal{C}(\widehat{S}_{\text{aware}}) \geq \mathcal{C}(S_{\text{flat}})$. Hence $I \leq H_B$. If moreover $H_B \leq r_T$, then every typing-specific gain is at most r_T, while r_T is the part of the typed certificate that has no counterpart

in the aware certificate—it is the loss the typed selector is certified to absorb at most, and the aware selector is certified to absorb at all. The gate therefore requires the certified maximum gain to exceed the certified typed-only loss.

(iv) From (4) and (i), $I = \ell_{\text{aware}} - \ell_{\text{typed}} \le \mathfrak{L}_{\text{aware}}(n) - 0$ and $I \ge 0 - \mathfrak{L}_{\text{typed}}(n,T)$, since both losses are non-negative; combining the upper bound with (iii) gives the interval. $\Delta^0 > 0$ is $\min\{H_B, c_0 g_n\} > r_T$, i.e. $H_B > r_T$ and $c_0 g_n > r_T$, which are Gates 2 and 1.

(v) The minimax lower bound in Appendix F (proof of Proposition 5, "Lower bound, in coverage") gives, over the spiked family in the regime $n \gtrsim \kappa^2 D$, $\mathbb{E}[\ell_{\text{aware}}] \gtrsim \nu d/n$ for any selector that does not know the blocks, matching the second-order certificate $\nu\kappa^2 D_{\text{flat}}/n$ up to constants depending on (λ, κ_0); so in that regime the aware certificate is not loose by more than a constant in the worst case, and Gate 1 is a comparison of worst-case risks. Proposition 8 shows $H_B \ge H$ with the off-dominant target mass recoverable by typed allocation, so in the one-dominant-mode regime the headroom H_B of Gate 2 is realizable rather than vacuous.

(vi) Consistency of $\widehat{H}_B$ and $\hat{\nu}$ is the concentration half of the proof of Lemma 1, applied to $\widehat{\boldsymbol{\Sigma}}$ (selection) and $\widehat{\boldsymbol{\Sigma}}^*$ (evaluation), each converging at the rate of Proposition 5. For the recovery side, Section 3 uses $1 - \sum_\tau \operatorname{tr}(\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{U}}_\tau)/\operatorname{tr}\widehat{\boldsymbol{\Sigma}}^*$ as the proxy for $\sin^2\theta_{\text{mix}}$; under (A3) the target mass outside the recovered blocks is exactly the mass lost by rotating $S^\star$ into them, and by the upper half of Lemma 4 that mass is at least the $\sin^2\theta_{\text{mix}}$ term it proxies, so the proxy is an upper bound. The leakage $\hat{\ell}_{\text{leak}}$ is computed directly from its definition with the operational $\tau(t)$ (residual-nearest block) and the probe target shares $\hat{\omega}_t$; it is the target-weighted cross-block mass of the items on which the router agrees with the nearest block, an exact plug-in of its definition rather than a proxy. $\hat{\varepsilon}^*_{\text{route}}$ is the target-weighted misrouting of the frozen Stage-A assignment, non-increasing in T by Assumption 2. Hence $\hat{r}_T \ge r_T$ up to the probe-estimation error of the target shares ω_t, and the plug-in rule is conservative in the negative direction. □

Corollary 2 (Pointwise sufficiency under comparable certificate tightness)**.** *In the setting of Proposition 4, suppose the aware certificate is tight up to a factor $\eta \in (0,1]$ on the deployed instance, $\ell_{\text{aware}} \ge \eta\, \mathfrak{L}_{\text{aware}}(n)$. Then with probability at least $1 - 2\delta$,*

$$r_T \;<\; c_0\,\nu\Big(\eta\sqrt{D_{\text{flat}}/n} - \sqrt{D_{\text{typed}}/n}\Big) \quad\Longrightarrow\quad I > 0.$$

At $\eta = 1$ the premise is Gate 1. If the typed certificate is loose by the same factor, $\ell_{\text{typed}} \le \eta\, \mathfrak{L}_{\text{typed}}(n,T)$, then Gate 1 itself, $r_T < c_0 g_n$, implies $I > 0$ for every $\eta \in (0,1]$: uniform looseness of the two certificates cancels in the comparison.

Corollary 2 (proof). By (4), $I = \ell_{\text{aware}} - \ell_{\text{typed}} \ge \eta\mathfrak{L}_{\text{aware}}(n) - \mathfrak{L}_{\text{typed}}(n,T) = c_0\nu\big(\eta\sqrt{D_{\text{flat}}/n} - \sqrt{D_{\text{typed}}/n}\big) - r_T$ on the event of Proposition 4(i), which gives the first claim. For the second, $I \ge \eta\mathfrak{L}_{\text{aware}} - \eta\mathfrak{L}_{\text{typed}} = \eta(c_0 g_n - r_T) > 0$ whenever Gate 1 holds. □

Empirical tightness of the certificates. At the deployed n=200 with mean $\hat{\nu} = 0.865$, $\widehat{D}_{\text{flat}} = 151.0$, $\widehat{D}_{\text{typed}} = 53.5$, two-term recovery cost 0.193 and leakage $\hat{\ell}_{\text{leak}} = 0.05$ (Table 3), the plug-in certificates are $\hat{\nu}\sqrt{\widehat{D}_{\text{flat}}/n} = 0.75$ and $\hat{\nu}\sqrt{\widehat{D}_{\text{typed}}/n} + \hat{r}_T = 0.69$ coverage units, whereas the realized losses relative to the common asymptote are $C_\infty - \widehat{\mathcal{C}}_{\text{aware}} = 0.861 - 0.725 = 0.136$ and $C_\infty - \widehat{\mathcal{C}}_{\text{SOMA-2S}} = 0.861 - 0.765 = 0.096$ (Table 13a; the coverage here is the two-stage SOMA-2S, whereas joint SOMA reaches 0.780 in Table 2). The certificates are therefore loose by factors of 5.5 and 7.2 respectively. The deployment rule does not require them to be tight; it requires them to be comparably loose (Corollary 2), a premise these two ratios only roughly satisfy and which is in any case not verifiable on a new instance. We state this rather than claim tightness; the controlled subspace-rotation experiment (Figure 5c) is the corresponding check on the recovery side.

Corollary 3 (TYPE-IF^0 for the split estimator)**.** *Let the blocks be refined on n_1 probes and frozen, and selection performed on the disjoint $n_2 = n - n_1$ probes (SOMA-SP, Algorithm 7, with the spectral Stage B). Then Proposition 4 holds with n_2 in place of n in $\mathfrak{L}_{\text{aware}}$, $\mathfrak{L}_{\text{typed}}$, and g, and with r_T evaluated at the refined blocks, $r_T^{\text{ref}} := \nu\sin^2\theta_{\text{mix}}^{\text{ref}} + \ell_{\text{leak}}^{\text{ref}} + \varepsilon_{\text{route}}^{*\,\text{ref}}$. The split rule deploys typed memory iff $\min\{H_B, c_0 g_{n_2}\} > r_T^{\text{ref}}$.*

Corollary 3 (proof). Condition on the n_1 refinement probes. Given them, the refined blocks are fixed and independent of the n_2 selection probes, so Corollary 1 gives the typed certificate with n_2 and the refined recovery residuals, and the aware selector trained on the same n_2 probes has certificate $\mathfrak{L}_{\mathrm{aware}}(n_2)$. The proof of Proposition 4 then goes through verbatim. □

Remark 5 (Split-refinement trade-off). For $n_2 = (1-\varrho)n$, the typed certificate contains an estimation term increasing as the selection split shrinks and a recovery term that can decrease if refinement improves the blocks. With mean $\hat{\nu} = 0.865$, $\widehat{D}_{\mathrm{typed}} = 53.5$, and the two-term recovery costs of Table 6, the estimated crossover for the $\varrho = 1/2$ split is $n^{\star}_{\mathrm{split}} \approx 5.7 \times 10^3$; adding the leakage terms ($\hat{\ell}_{\mathrm{leak}} = 0.05$, $\hat{\ell}^{\mathrm{ref}}_{\mathrm{leak}} = 0.04$) widens the recovery gap refinement buys and moves it to $\approx 2.6 \times 10^3$, still above every reported budget. Hence SOMA-SP does not overtake SOMA-2S at either $n = 200$ or $n = 800$; Table 6 reports this comparison directly.

Remark 6 (The constant c_0). The rule has no fitted task-outcome threshold, but it does inherit the sub-Gaussian concentration prefactor c_0. We set $c_0 = 1$ to place the estimation and recovery certificates on the same unit scale and report the critical c_0 at which each estimation-gated decision would flip (Appendix D).

Remark 7 (What TYPE-IF^0 does and does not claim). The rule is a comparison of risk certificates and is used as a sign diagnostic, not a ranking score. Its first-order form is invoked only where $n > D_{\mathrm{flat}}$; this begins at $n = 200$ under the observed effective ranks. The drift index remains a separate ordinal diagnostic for the shared-decay penalty. The reported effective-rank range 142–154 is used directly in the $n = 200$ first-order interpretation.

F THEORETICAL ANALYSIS: ASSUMPTIONS AND PROOFS

We prove Propositions 1, 5, and 7, formalize Lemma 1 and the type-spread bound (Proposition 8), and state the general-K excess of a shared decay (Corollary 5). The capacity results use only orthogonal projectors and the variational characterization of eigenvalues; the statistical result uses covariance concentration (Vershynin, 2018), the Davis–Kahan theorem (Yu et al., 2015), and a second-order expansion of coverage at its maximizer (Lemma 4), which is what lets the minimax subspace rates of Cai et al. (2013) transfer to coverage; the dynamical result is elementary calculus.

Assumptions at a glance. For auditability we collect every assumption used below and where it is used. *(A1) Complementarity* (Definition 1): the type subspaces are mutually orthogonal; used by all results. *(A2) Spanning*: together the types span $\mathbb{R}^d$ ($\sum_\tau d_\tau = d$); this is without loss of generality—if the learned types span only a proper subspace, add a residual type carrying the orthogonal complement—and it is what Lemma 5 needs. *(A2′) Block-aligned content*: the retrieval-mass-weighted content covariance carries no cross-type terms, $\boldsymbol{U}_\tau^\top \boldsymbol{\Sigma}\, \boldsymbol{U}_{\tau'} = \mathbf{0}$ for $\tau \neq \tau'$. This is an *aggregate* second-moment condition: it is implied by, but strictly weaker than, the sample-level union-of-subspaces condition (A7a), and it does *not* follow from orthogonality of the $\boldsymbol{U}_\tau$ or from storing $\boldsymbol{V}_{a_t}^\top z_t$. Its residual is the cross-block covariance r_Σ of Section 3; the inter-type overlap 0.07 measures (A1), not (A2′). It is used by Lemma 5 and hence Proposition 1. *(A3) Type-aligned target* ($\boldsymbol{\Sigma}_* = \sum_\tau \boldsymbol{U}_\tau \boldsymbol{\Sigma}^*_\tau \boldsymbol{U}_\tau^\top$): used by Propositions 5 and 8, not by Proposition 1. *(A4) Sub-Gaussian probes with eigengap γ* (Assumption 1): used by Proposition 5 and by the concentration half of Lemma 1. *(A5) Drift model* (Assumption 3, instantiated by Lemma 6): used by Proposition 7. *(A6) One-dominant-mode regime*: used only by Proposition 8. *(A7a) Recoverable typing* (Assumption 2): a union-of-subspaces stream with positive minimum principal angle and sufficient sampling density, so that a stream-only procedure recovers the decomposition with residuals $\theta_{\mathrm{mix}}(T), \varepsilon_{\mathrm{route}}(T)$ non-increasing in stream length; its sample-level residual is the assignment residual r_a of Section 3. *(A7b) Item-carried target*: $\boldsymbol{\Sigma}_* = \sum_t w_t z_t z_t^\top$ with $w_t \geq 0$, which the operational estimator $\widehat{\boldsymbol{\Sigma}}^*$ satisfies by construction; it defines the target shares ω_t, the target-weighted routing error $\varepsilon^*_{\mathrm{route}}$, and the leverage L_*. Both are used only by Proposition 6 (proof: Appendix E), which applies to the stream-first estimator and not to joint training (Remark 4). In practice the learned subspaces satisfy (A1) approximately; the complementarity term of Eq. (2) drives the residual overlap to 0.07 (Section 3), and Remark 10 discusses scope.

Setup (recap of Section 2). The representation space is $\mathbb{R}^d$; the recall target has covariance $\boldsymbol{\Sigma}_* \succeq 0$; coverage of a subspace S with orthogonal projector $\boldsymbol{P}_S$ is $\mathcal{C}(S) = \mathrm{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_*)/\,\mathrm{tr}(\boldsymbol{\Sigma}_*)$ (Eq. (1)). Each type τ is a subspace given by $\boldsymbol{U}_\tau \in \mathbb{R}^{d \times d_\tau}$ with orthonormal columns; types are

complementary, $\boldsymbol{U}_\tau^\top \boldsymbol{U}_{\tau'} = \mathbf{0}$ for $\tau \neq \tau'$, and together span $\mathbb{R}^d$ ($\sum_\tau d_\tau = d$). Complementarity together with block-aligned content (A2′) gives the block decomposition $\boldsymbol{\Sigma} = \sum_\tau \boldsymbol{U}_\tau \boldsymbol{\Sigma}_\tau \boldsymbol{U}_\tau^\top$ of the global content covariance. We model a flat similarity retriever at budget B as reaching the highest-content-variance directions, $S_{\mathrm{flat}} = \text{top-}B(\boldsymbol{\Sigma})$; a typed memory chooses an allocation $\{b_\tau\}$ with $\sum_\tau b_\tau = B$ and within-type subspaces $S_\tau \subseteq \mathrm{col}(\boldsymbol{U}_\tau)$, $\dim S_\tau = b_\tau$, forming $S_{\mathrm{typed}} = \bigoplus_\tau S_\tau$.

Formal statements (compressed to one line each in Section 2). For completeness we restate here, with the labels used throughout, the assumptions and propositions whose main-text form is a single sentence.

Assumption 1. Probe targets are i.i.d. sub-Gaussian with covariance $\boldsymbol{\Sigma}_*$, whose relevant eigengap (between retained and discarded directions; globally for the flat rule, within each block for the typed rule) is at least $\gamma > 0$. Write $\lambda_1 = \|\boldsymbol{\Sigma}_*\|_{\mathrm{op}}$, $\kappa = \lambda_1/\gamma$, and

$$r_{\mathrm{eff}} := \frac{\mathrm{tr}(\boldsymbol{\Sigma}_*)}{\lambda_1}, \qquad r_{\mathrm{eff},\tau} := \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\tau^*)}{\|\boldsymbol{\Sigma}_\tau^*\|_{\mathrm{op}}}, \qquad \nu := \frac{B\lambda_1}{\mathrm{tr}(\boldsymbol{\Sigma}_*)} = \frac{B}{r_{\mathrm{eff}}}.$$

Proposition 5 (Typed plug-in selection with a *given* decomposition pays $\max_\tau d_\tau$, not d)**.** *Under Assumption 1 and a type-aligned target ($\boldsymbol{\Sigma}_* = \sum_\tau \boldsymbol{U}_\tau \boldsymbol{\Sigma}_\tau^* \boldsymbol{U}_\tau^\top$, A3) with $\{\boldsymbol{U}_\tau\}$ fixed independently of the probes, with probability $\geq 1-\delta$ and for $n \gtrsim D$, writing $D_{\mathrm{flat}} = r_{\mathrm{eff}} + \log(1/\delta)$ and $D_{\mathrm{typed}} = \max_\tau r_{\mathrm{eff},\tau} + \log(K/\delta)$,*

$$\mathcal{C}(S^\star) - \mathcal{C}(\widehat{S}) \;\lesssim\; \nu \min\Big\{\sqrt{\tfrac{D}{n}},\; \kappa^2 \tfrac{D}{n}\Big\}, \qquad D \in \{D_{\mathrm{flat}}, D_{\mathrm{typed}}\}.$$

The first branch is first-order in the estimation error and needs no eigengap (below $n \asymp D$ the concentration step yields only the weaker $\nu D/n$); the second holds once $n \gtrsim \kappa^2 D$ (Lemma 4). The dimension-free covariance bound makes the effective ranks the operative finite-sample dimensions; a Gaussian spiked minimax construction shows the factor $d/\max_\tau d_\tau$ is tight in the second-order regime.

Assumption 2 (Recoverable typing and item-carried target)**.** *(A7a)* The stream is generated from a union of K subspaces $\{\mathrm{col}(\boldsymbol{U}_\tau)\}$ with pairwise minimum principal angle $\theta_{\min} > 0$ and sampling density sufficient for recovery in the sense of Elhamifar & Vidal (2013); $\tau(t)$ denotes item t's generating type. *(A7b)* The target is carried by stream items, $\boldsymbol{\Sigma}_* = \sum_t w_t z_t z_t^\top$ with $w_t \geq 0$, so item t has *target share* $\omega_t = w_t\|z_t\|^2/\mathrm{tr}(\boldsymbol{\Sigma}_*)$, $\sum_t \omega_t = 1$ (the estimator $\widehat{\boldsymbol{\Sigma}}^*$ has exactly this form). For blocks $\{\widehat{\boldsymbol{U}}_\tau\}$ and assignments $\{a_t\}$ fit on the stream, let $\theta_{\mathrm{mix}}(T)$ be the largest principal angle between $\widehat{\boldsymbol{U}}_\tau$ and $\boldsymbol{U}_\tau$, and define the *content-mass* and *target-weighted* routing errors and the *leverage*

$$\varepsilon_{\mathrm{route}}(T) = \sum_{t:\, a_t \neq \tau(t)} \pi_t, \qquad \varepsilon^*_{\mathrm{route}}(T) = \sum_{t:\, a_t \neq \tau(t)} \omega_t, \qquad L_* = \max_t\, \omega_t/\pi_t.$$

Then $\varepsilon^*_{\mathrm{route}} \leq L_*\, \varepsilon_{\mathrm{route}}$, and all three are non-increasing in T. Only $\varepsilon_{\mathrm{route}}$ is probe-free: $\varepsilon^*_{\mathrm{route}}$ and L_* use the target shares ω_t, hence the same probes as selection; and $\tau(t)$ is latent, which on real streams we operationalize as the residual-nearest recovered block (Section 3).

Proposition 6 (End-to-end typed rate for the two-stage spectral estimator)**.** *Let $\{\widehat{\boldsymbol{U}}_\tau\}$ and $\{a_t\}$ be obtained from the label-free stream by* any *procedure that does not use the probe set, and frozen; let $\widehat{S}_{\mathrm{typed}}$ keep, within each block, the top-b_τ eigenvectors of the routed empirical target covariance $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}_\tau^* \widehat{\boldsymbol{U}}_\tau$ from n probes under the pooled allocation (Algorithm 6). Under Assumptions 1–2 and (A1)/(A3), with probability $\geq 1-\delta$,*

$$\mathcal{C}(S^\star) - \mathcal{C}^{\mathrm{rt}}(\widehat{S}_{\mathrm{typed}}) \;\lesssim\; \underbrace{\nu \min\Big\{\sqrt{\tfrac{D_{\mathrm{typed}}}{n}},\; \kappa^2 \tfrac{D_{\mathrm{typed}}}{n}\Big\}}_{\text{estimation: probe budget } n} \;+\; \underbrace{\nu \sin^2\theta_{\mathrm{mix}}(T) + \ell_{\mathrm{leak}}(T) + \varepsilon^*_{\mathrm{route}}(T)}_{\text{recovery: stream length } T},$$

where $\mathcal{C}^{\mathrm{rt}}(S) = \sum_t \omega_t \|\boldsymbol{P}_{S_{a_t}} \hat{z}_t\|^2 \leq \mathcal{C}(S)$ counts each item through the block it was written to and $\ell_{\mathrm{leak}} \leq \sin^2\theta_{\mathrm{mix}}$ is the target share that correctly routed items leak into other recovered blocks (Lemma 2); the bound therefore also holds for $\mathcal{C}(\widehat{S}_{\mathrm{typed}})$. The K-fold (resp. $\sqrt{K}$) separation survives whenever the recovery term is dominated by the estimation term. If the blocks are refined on n_1 probes disjoint from the $n_2 = n - n_1$ used for selection, the bound holds with n_2 in place of n (Corollary 1). Proof and the sample-split corollary: Appendix E.

Assumption 3. The content statistic of mode τ drifts with per-step increment δ_τ, observed with noise variance σ^2; exponential decay at rate λ incurs steady-state squared error $e(\lambda; \delta_\tau, \sigma) = c_1\delta_\tau^2/\lambda^2 + c_2\lambda\sigma^2$ (instantiated in Lemma 6).

Proposition 7 (Per-type decay strictly beats any shared decay under heterogeneous drift)**.** *Under Assumption 3, per-type decay attains* $\sum_\tau \min_\lambda e(\lambda; \delta_\tau, \sigma)$ *with* $\lambda^\star_\tau \propto (\delta^2_\tau/\sigma^2)^{1/3}$*, whereas the best shared decay incurs exactly* $R(\{\delta_\tau\})$ *times that error,* $R = \big(\frac{1}{K}\sum_\tau \delta^2_\tau\big)^{1/3} / \big(\frac{1}{K}\sum_\tau \delta^{2/3}_\tau\big) \in [1, K^{2/3})$*, with* $R = 1$ *iff all* δ_τ *are equal and* R *increasing in their spread (Corollary 5). The constants* c_1, c_2, σ *cancel in* R*.*

Lemma 3 (Coverage identities)**.** *Let* $\boldsymbol{P}_S$ *be an orthogonal projector and* $\boldsymbol{A} \succeq 0$ *symmetric. Then (i)* $\operatorname{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_* \boldsymbol{P}_S) = \operatorname{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_*)$*; (ii) (Ky Fan)* $\max_{\dim S = b} \operatorname{tr}(\boldsymbol{P}_S \boldsymbol{A}) = \sum_{j=1}^b \lambda_j(\boldsymbol{A})$*, attained by the span of the top-*b *eigenvectors of* $\boldsymbol{A}$*; (iii) for mutually orthogonal subspaces* $\{S_\tau\}$*,* $\boldsymbol{P}_{\oplus_\tau S_\tau} = \sum_\tau \boldsymbol{P}_{S_\tau}$*, hence* $\operatorname{tr}(\boldsymbol{P}_{\oplus_\tau S_\tau} \boldsymbol{\Sigma}_*) = \sum_\tau \operatorname{tr}(\boldsymbol{P}_{S_\tau} \boldsymbol{\Sigma}_*)$*.*

Proof. (i) $\boldsymbol{P}_S$ is idempotent and symmetric, so $\operatorname{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_* \boldsymbol{P}_S) = \operatorname{tr}(\boldsymbol{P}_S^2 \boldsymbol{\Sigma}_*) = \operatorname{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_*)$ by cyclicity. (ii) is the Ky Fan maximum principle. (iii) Orthogonal ranges make $\sum_\tau \boldsymbol{P}_{S_\tau}$ idempotent and symmetric with range $\bigoplus_\tau S_\tau$, hence equal to $\boldsymbol{P}_{\oplus_\tau S_\tau}$; trace is linear. □

Lemma 4 (Coverage loss is quadratic in the principal angles)**.** *Let* $S^\star$ *be the span of the top-*b *eigenvectors of* $\boldsymbol{\Sigma}_*$*, with* $\lambda_1 \geq \cdots \geq \lambda_b > \lambda_{b+1} \geq \ldots$ *its eigenvalues and* $\gamma_\star = \lambda_b - \lambda_{b+1}$ *its gap, and let* $\widehat{S}$ *be any* b*-dimensional subspace with principal angles* $\theta_1, \ldots, \theta_b$ *to* $S^\star$*. Then*

$$\gamma_\star \sum_{i=1}^b \sin^2\theta_i \;\leq\; \operatorname{tr}\big((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})\boldsymbol{\Sigma}_*\big) \;\leq\; \lambda_1 \sum_{i=1}^b \sin^2\theta_i, \qquad \sum_i \sin^2\theta_i = \tfrac{1}{2}\|\boldsymbol{P}_{\widehat{S}} - \boldsymbol{P}_{S^\star}\|_F^2.$$

In addition, without any gap assumption, $\operatorname{tr}\big((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})\boldsymbol{\Sigma}_*\big) \leq 2b\,\|\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}_*\|_{\mathrm{op}}$ *whenever* $\widehat{S}$ *is the top-*b *eigenspace of any symmetric* $\widehat{\boldsymbol{\Sigma}}^*$*.*

Proof. Write $\boldsymbol{\Sigma}_* = \boldsymbol{V}\boldsymbol{\Lambda}_1\boldsymbol{V}^\top + \boldsymbol{V}_\perp\boldsymbol{\Lambda}_2\boldsymbol{V}_\perp^\top$ with $\boldsymbol{V}$ spanning $S^\star$, and set $\boldsymbol{A} = \boldsymbol{V}^\top\boldsymbol{P}_{\widehat{S}}\boldsymbol{V}$, whose eigenvalues are $\cos^2\theta_i$. Then $\operatorname{tr}(\boldsymbol{P}_{\widehat{S}}\boldsymbol{\Sigma}_*) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Lambda}_1) + \operatorname{tr}(\boldsymbol{V}_\perp^\top\boldsymbol{P}_{\widehat{S}}\boldsymbol{V}_\perp\boldsymbol{\Lambda}_2)$ and $\operatorname{tr}(\boldsymbol{V}_\perp^\top\boldsymbol{P}_{\widehat{S}}\boldsymbol{V}_\perp) = \operatorname{tr}(\boldsymbol{P}_{\widehat{S}}) - \operatorname{tr}(\boldsymbol{A}) = b - \operatorname{tr}(\boldsymbol{A}) = \sum_i \sin^2\theta_i$. Hence $\operatorname{tr}((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})\boldsymbol{\Sigma}_*) = \operatorname{tr}((\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\Lambda}_1) - \operatorname{tr}(\boldsymbol{V}_\perp^\top\boldsymbol{P}_{\widehat{S}}\boldsymbol{V}_\perp\boldsymbol{\Lambda}_2)$. Since $\boldsymbol{0} \preceq \boldsymbol{I} - \boldsymbol{A}$ with $\operatorname{tr}(\boldsymbol{I} - \boldsymbol{A}) = \sum_i \sin^2\theta_i$ and $\lambda_b\boldsymbol{I} \preceq \boldsymbol{\Lambda}_1 \preceq \lambda_1\boldsymbol{I}$, the first term lies in $[\lambda_b, \lambda_1]\sum_i \sin^2\theta_i$; since $\boldsymbol{0} \preceq \boldsymbol{V}_\perp^\top\boldsymbol{P}_{\widehat{S}}\boldsymbol{V}_\perp$ and $\boldsymbol{0} \preceq \boldsymbol{\Lambda}_2 \preceq \lambda_{b+1}\boldsymbol{I}$, the second lies in $[0, \lambda_{b+1}]\sum_i \sin^2\theta_i$. Subtracting gives the two-sided bound; the identity $\sum_i \sin^2\theta_i = \frac{1}{2}\|\boldsymbol{P}_{\widehat{S}} - \boldsymbol{P}_{S^\star}\|_F^2$ is standard. For the gap-free bound, optimality of $\widehat{S}$ for $\widehat{\boldsymbol{\Sigma}}^*$ gives $\operatorname{tr}(\boldsymbol{P}_{S^\star}\widehat{\boldsymbol{\Sigma}}^*) \leq \operatorname{tr}(\boldsymbol{P}_{\widehat{S}}\widehat{\boldsymbol{\Sigma}}^*)$, so $\operatorname{tr}((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})\boldsymbol{\Sigma}_*) \leq \operatorname{tr}((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})(\boldsymbol{\Sigma}_* - \widehat{\boldsymbol{\Sigma}}^*)) \leq \|\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}}\|_*\|\boldsymbol{\Sigma}_* - \widehat{\boldsymbol{\Sigma}}^*\|_{\mathrm{op}} \leq 2b\,\|\boldsymbol{\Sigma}_* - \widehat{\boldsymbol{\Sigma}}^*\|_{\mathrm{op}}$, using that a difference of two rank-b projectors has nuclear norm at most $2b$. □

The lemma is the reason the separation can be stated in coverage: the coverage functional is stationary at its maximizer over the Grassmannian, so plug-in loss is *second*-order in the estimation error once the gap resolves, and first-order before that. (This is the same phenomenon as the $1/n$ excess risk of PCA (Reiss & Wahl, 2020).)

Lemma 5 (Flat selection is a typed configuration)**.** *Under complementarity and block-aligned content (A2′),* $\boldsymbol{\Sigma}$ *is block-diagonal in the orthonormal basis* $[\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]$ *and its eigenvectors may be chosen within individual type subspaces. Consequently* $S_{\mathrm{flat}} = \text{top-}B(\boldsymbol{\Sigma})$ *can be written as* $S_{\mathrm{flat}} = \bigoplus_\tau S^\flat_\tau$ *with* $S^\flat_\tau \subseteq \operatorname{col}(\boldsymbol{U}_\tau)$ *and* $\sum_\tau \dim S^\flat_\tau = B$*; that is,* S_{flat} *lies in the family* $\mathcal{B}_B = \{\bigoplus_\tau S_\tau : S_\tau \subseteq \operatorname{col}(\boldsymbol{U}_\tau),\ \sum_\tau \dim S_\tau = B\}$ *of block subspaces of total dimension* B*.*

Proof. In the basis $[\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]$, $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. Diagonalizing each block $\boldsymbol{\Sigma}_\tau = \boldsymbol{V}_\tau\boldsymbol{D}_\tau\boldsymbol{V}_\tau^\top$ gives eigenvectors of $\boldsymbol{\Sigma}$ that are the columns of $\boldsymbol{V}_\tau$ embedded in block τ, with eigenvalues the pooled $\{\boldsymbol{D}_\tau\}$. Selecting the B largest eigenvalues therefore selects within-block eigenvectors (ties at the budget boundary may be broken within blocks), partitioning the budget as $b^\flat_\tau = \dim S^\flat_\tau$. □

Proposition 1 (restated). *Under complementarity and (A2′) there is an allocation* $\{b_\tau\}$ *and within-type subspaces such that* $\mathcal{C}(S_{\mathrm{typed}}) \geq \mathcal{C}(S_{\mathrm{flat}})$*; the inequality is strict iff* S_{flat} *does not maximize coverage over* $\mathcal{B}_B$*, for which it suffices that some excluded direction carries more target mass than some retained one.*

Proof. By Lemma 5, $S_{\text{flat}} \in \mathcal{B}_B$. A typed memory optimizes coverage over exactly this family, so

$$\mathcal{C}(S^\star_{\text{typed}}) = \max_{S \in \mathcal{B}_B} \mathcal{C}(S) \;\geq\; \mathcal{C}(S_{\text{flat}}),$$

which is the claimed inequality. The maximizer is characterized by Lemma 3: since $\operatorname{tr}(\boldsymbol{P}_S \boldsymbol{\Sigma}_*) = \sum_\tau \operatorname{tr}(\boldsymbol{P}_{S_\tau} \boldsymbol{\Sigma}_*)$ for $S = \bigoplus_\tau S_\tau$ (part iii), and each summand is maximized over b_τ-dimensional $S_\tau \subseteq \operatorname{col}(\boldsymbol{U}_\tau)$ by the top-b_τ eigenvectors of $\boldsymbol{U}_\tau^\top \boldsymbol{\Sigma}_* \boldsymbol{U}_\tau$ (part ii), the optimum keeps, within each type, the most target-aligned directions, under the allocation $\{b_\tau\}$ that pools the B largest of these within-type target eigenvalues. Flat memory instead keeps the top eigenvectors of $\boldsymbol{\Sigma}_\tau$ (content variance) with the content-variance allocation. The two coincide only if S_{flat} already maximizes coverage over $\mathcal{B}_B$; otherwise the inequality is strict. A sufficient condition for strictness is that some type τ contains a direction $\boldsymbol{u} \in \operatorname{col}(\boldsymbol{U}_\tau)$ with $\boldsymbol{u}^\top \boldsymbol{\Sigma}_* \boldsymbol{u} > 0$ that S_{flat} excludes because of its low content variance: reallocating one unit of budget to $\boldsymbol{u}$ from a retained direction carrying less target mass strictly increases $\sum_\tau \operatorname{tr}(\boldsymbol{P}_{S_\tau} \boldsymbol{\Sigma}_*)$. Note that target mass merely lying outside $\operatorname{top}\text{-}B(\boldsymbol{\Sigma})$ is *not* sufficient: an excluded direction must carry more target mass than some retained one, which is exactly the condition stated in the proposition. □

Remark 1 (the domination chain, for the record). Since $\mathcal{B}_B \subseteq \{S : \dim S = B\}$, $\max_{\dim S = B} \mathcal{C}(S) \geq \max_{S \in \mathcal{B}_B} \mathcal{C}(S)$, which with Proposition 1 gives $\mathcal{C}(S^{\text{aware}}_{\text{flat}}) \geq \mathcal{C}(S^\star_{\text{typed}}) \geq \mathcal{C}(S_{\text{flat}})$. Equality of the first inequality holds iff some maximizer over all B-dimensional subspaces is block-aligned; when $\boldsymbol{\Sigma}_*$ is type-aligned, the top-B eigenvectors of $\boldsymbol{\Sigma}_*$ are block-aligned, so the typed optimum is in fact *unconstrained-optimal* in that regime—the typed family loses nothing exactly when the type structure matches the target.

Proposition 5 (proof).

Proof. Throughout, the blocks $\{\boldsymbol{U}_\tau\}$ are fixed before the probes are drawn and are therefore independent of them; every concentration statement below is for a fixed block, and the union bound over τ is over K fixed blocks, not over a data-dependent choice. Write $\widehat{\boldsymbol{\Sigma}}^*$ for the empirical target covariance from n sub-Gaussian probes and $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}_*$. *Concentration.* We use the dimension-free covariance bound of Koltchinskii & Lounici (2017, Cor. 2). With $r_{\text{eff}} = \operatorname{tr}(\boldsymbol{\Sigma}_*)/\lambda_1$ and $D_{\text{flat}} = r_{\text{eff}} + \log(1/\delta)$, with probability $\geq 1 - \delta$, $\|\boldsymbol{\Delta}\|_{\text{op}} \leq c\lambda_1\big(\sqrt{D_{\text{flat}}/n} + D_{\text{flat}}/n\big)$, where c depends only on the sub-Gaussian norm; for $n \geq D_{\text{flat}}$ the first term dominates. For the typed rule, apply the same result within each block using $r_{\text{eff},\tau} = \operatorname{tr}(\boldsymbol{\Sigma}^*_\tau)/\|\boldsymbol{\Sigma}^*_\tau\|_{\text{op}}$ and a union bound over K fixed blocks: $\max_\tau \|\boldsymbol{\Delta}_\tau\|_{\text{op}} \leq c\lambda_1\sqrt{D_{\text{typed}}/n}$ with $D_{\text{typed}} = \max_\tau r_{\text{eff},\tau} + \log(K/\delta)$. The ambient-dimensional forms remain valid but looser and are used only in the worst-case minimax comparison below.

First-order regime. The plug-in selection keeps the top-b eigenvectors of the empirical matrix (globally for the flat-aware rule; per block, with empirical pooling of within-block eigenvalues, for the typed rule). The gap-free bound of Lemma 4 gives, for the flat-aware rule, $\operatorname{tr}((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})\boldsymbol{\Sigma}_*) \leq 2B\|\boldsymbol{\Delta}\|_{\text{op}}$, and for the typed rule, by Lemma 3(iii) applied blockwise, $\sum_\tau 2b_\tau\|\boldsymbol{\Delta}_\tau\|_{\text{op}} \leq 2B\max_\tau \|\boldsymbol{\Delta}_\tau\|_{\text{op}}$. Dividing by $\operatorname{tr}(\boldsymbol{\Sigma}_*)$ and inserting the concentration bounds yields $\mathcal{C}(S^\star) - \mathcal{C}(\widehat{S}) \lesssim \nu\sqrt{D/n}$ with $\nu = B\lambda_1/\operatorname{tr}(\boldsymbol{\Sigma}_*)$, for $D = D_{\text{flat}}$ and $D = D_{\text{typed}}$ respectively.

Second-order regime. By the Davis–Kahan $\sin\Theta$ theorem in the form of Yu et al. (2015, Thm. 2), $\big(\sum_i \sin^2\theta_i\big)^{1/2} \leq 2\sqrt{b}\,\|\boldsymbol{\Delta}\|_{\text{op}}/\gamma$ for a top-b eigenspace with gap γ. Combining with the upper half of Lemma 4, $\operatorname{tr}((\boldsymbol{P}_{S^\star} - \boldsymbol{P}_{\widehat{S}})\boldsymbol{\Sigma}_*) \leq 4\lambda_1 b\,\|\boldsymbol{\Delta}\|^2_{\text{op}}/\gamma^2$ for the flat-aware rule, and, blockwise, $\leq 4\lambda_1\sum_\tau b_\tau\|\boldsymbol{\Delta}_\tau\|^2_{\text{op}}/\gamma^2 \leq 4\lambda_1 B\max_\tau\|\boldsymbol{\Delta}_\tau\|^2_{\text{op}}/\gamma^2$ for the typed rule (each block's gap is at least γ by Assumption 1). Dividing by $\operatorname{tr}(\boldsymbol{\Sigma}_*)$ and inserting concentration gives $\mathcal{C}(S^\star) - \mathcal{C}(\widehat{S}) \lesssim \nu\kappa^2 D/n$ with $\kappa = \lambda_1/\gamma$, valid once $n \gtrsim \kappa^2 D$ (so that the Davis–Kahan bound is below one). Taking the minimum of the two regimes gives the displayed statement. For the typed rule the allocation $\{b_\tau\}$ chosen from empirical within-block eigenvalues differs from the population allocation only when eigenvalues on opposite sides of a within-block gap cross, which requires $\|\boldsymbol{\Delta}_\tau\|_{\text{op}} \geq \gamma/2$ and is therefore excluded on the event above in the second-order regime; in the first-order regime the gap-free bound does not depend on which allocation is used.

Lower bound, in coverage. Over the Gaussian spiked family $\mathbf{\Sigma}_* = \sigma^2(\boldsymbol{I}_d + \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top)$ with $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ orthonormal, $\lambda \le \lambda_r \le \cdots \le \lambda_1 \le \kappa_0\lambda$, in the regime $n \gtrsim r + \log\lambda$, $\lambda \gtrsim \sqrt{\log n/n}$, and $r(d-r)/(n\,h(\lambda)) \lesssim 1$ (so the rate is not saturated), Cai et al. (2013, Thm. 5) give, for any estimator $\widehat{\boldsymbol{V}}$ measurable in the n samples,

$$\inf_{\widehat{\boldsymbol{V}}} \sup_{\mathbf{\Sigma}_*} \mathbb{E}\big\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top - \boldsymbol{V}\boldsymbol{V}^\top\big\|_F^2 \gtrsim \frac{r(d-r)}{n\,h(\lambda)}, \qquad h(\lambda) = \frac{\lambda^2}{\lambda+1},$$

and the same rate with d replaced by d_τ when the eigenvectors are known to lie in a coordinate block of size d_τ (Cai et al., 2013, Thms. 8–9). Take $b = r$ (the selection budget equals the spike rank, the case in which the population optimum is the spike subspace). By the *lower* half of Lemma 4, with $\gamma_\star = \sigma^2\lambda_r \ge \sigma^2\lambda$ and $\operatorname{tr}(\mathbf{\Sigma}_*) = \sigma^2(d + \sum_i \lambda_i) \le \sigma^2(d + r\kappa_0\lambda)$,

$$\mathbb{E}\big[\mathcal{C}(S^\star) - \mathcal{C}(\widehat{S})\big] \;\ge\; \frac{\gamma_\star}{2\operatorname{tr}(\mathbf{\Sigma}_*)}\,\mathbb{E}\|\boldsymbol{P}_{\widehat{S}} - \boldsymbol{P}_{S^\star}\|_F^2 \;\gtrsim\; \frac{\lambda}{d + r\kappa_0\lambda}\cdot\frac{r(d-r)}{n\,h(\lambda)} \;=\; \frac{\lambda+1}{\lambda}\cdot\frac{r\,(d-r)}{(d+r\kappa_0\lambda)\,n}.$$

For fixed signal-to-noise $\lambda \asymp 1$ and bounded κ_0 this is $\asymp (r/d)\cdot d/n \asymp \nu\,d/n$, since $\nu = r\lambda_1/\operatorname{tr}(\mathbf{\Sigma}_*) \asymp r/d$ on this family; the same computation on a known block gives $\asymp \nu\,d_\tau/n$, and summing over balanced blocks $\asymp \nu \max_\tau d_\tau/n$. Because a lower bound over the Gaussian sub-family is a fortiori a lower bound over the sub-Gaussian class of Assumption 1, the second-order upper bounds above are minimax-tight up to constants depending only on (λ, κ_0), and the two problems—selection without and with the block structure—have coverage rates that differ by $d/\max_\tau d_\tau$. This is the separation claimed in the proposition. Two scope notes: the family has isotropic noise, so its targets are worst cases rather than models of any benchmark; and the lower bound is over selectors that do not know the blocks, which is the right class for a flat retriever because *every* target is block-aligned with respect to its own eigenbasis, so "block-aligned with unknown blocks" is no restriction at all (Scope paragraph below). □

Scope of the statistical separation (expanded from Section 2.2). The finite-sample statement uses the target effective rank, not automatically the ambient encoder dimension. This distinction is testable: the same saved target covariance that defines coverage yields $\hat{r}_{\text{eff}} = \operatorname{tr}(\widehat{\mathbf{\Sigma}}^*)/\hat{\lambda}_1$ and $\hat{\nu} = B/\hat{r}_{\text{eff}}$, and the evaluation code asserts $\hat{\nu} \ge \widehat{\mathcal{C}}(S)$. Here $\hat{r}_{\text{eff}} = 142$–$154$, so n=200 exceeds the first-order dimension $\widehat{D}_{\text{flat}}$. The second-order minimax lower bound remains an ambient-dimensional worst-case statement over the Gaussian spiked family. Finally, the typed rate is conditional on the block structure being recovered independently of the selection probes; Proposition 6 folds recovery error into the bound for the stream-first estimator.

Proposition 7 (full statement and proof).

Proof. Fix σ^2 and write $e(\lambda;\delta) = c_1\delta^2\lambda^{-2} + c_2\lambda\sigma^2$, the steady-state squared tracking error of an exponentially weighted average with rate λ on a statistic drifting with per-step increment δ under observation noise σ^2 (bias term $\propto \delta/\lambda$, variance term $\propto \lambda\sigma^2$; Assumption 3). For each mode, $e(\cdot;\delta)$ is strictly convex on $\lambda > 0$ with unique minimizer obtained from $e'(\lambda) = -2c_1\delta^2\lambda^{-3} + c_2\sigma^2 = 0$:

$$\lambda^\star(\delta) = \Big(\frac{2c_1\delta^2}{c_2\sigma^2}\Big)^{1/3}, \;\; e\big(\lambda^\star(\delta);\delta\big) = C\,\delta^{2/3}\sigma^{4/3}, \;\; C = c_1^{1/3}c_2^{2/3}\big(2^{-2/3} + 2^{1/3}\big).$$

Per-type decay therefore attains

$$\textstyle\sum_\tau e\big(\lambda^\star(\delta_\tau);\delta_\tau\big) = C\sigma^{4/3}\sum_\tau \delta_\tau^{2/3}.$$

A shared decay attains $\min_\lambda \sum_\tau e(\lambda;\delta_\tau)$, which is at least $\sum_\tau \min_\lambda e(\lambda;\delta_\tau)$, with equality iff a single λ minimizes every summand, i.e. iff $\lambda^\star(\delta_\tau)$ is constant in τ, equivalently all δ_τ equal. Strict inequality therefore holds under any heterogeneity. For the growth claim, note the shared-decay optimum solves $\sum_\tau e'(\lambda;\delta_\tau) = 0$, giving $\lambda_{\text{shared}} = \big(2c_1\overline{\delta^2}/(c_2\sigma^2)\big)^{1/3}$ with $\overline{\delta^2} = \frac{1}{K}\sum_\tau \delta_\tau^2$; substituting and comparing with the per-type optimum expresses the excess as a power-mean gap between the $1/3$ power of $\overline{\delta^2}$ and the mean $\overline{\delta^{2/3}}$, which is zero at equality and increasing in the spread of $\{\delta_\tau\}$ (e.g., in $\max_\tau \delta_\tau/\min_\tau \delta_\tau$ at fixed total drift) by the power-mean inequality. The constants c_1, c_2 are instantiated for an explicit drift model in Lemma 6 below, and the excess-error ratio is stated in closed form for K=2 in Corollary 4 and for general K in Corollary 5, which is the form the drift-heterogeneity index $\widehat{D}$ (Definition 3) uses. □

Lemma 6 (Bias–variance of exponentially weighted tracking; instantiating Assumption 3)**.** *Let a scalar mode statistic θ_t drift with per-step increments bounded by δ ($|\theta_t - \theta_{t-1}| \le \delta$, attained by*

worst-case linear drift), observed as $x_t = \theta_t + \varepsilon_t$ with $\mathbb{E}[\varepsilon_t] = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$, independent across t. The exponentially weighted estimate $\widehat{\theta}_t = \lambda \sum_{j\geq 0}(1-\lambda)^j x_{t-j}$, $\lambda \in (0,1)$, has steady-state worst-case squared error

$$e(\lambda;\delta,\sigma) \;=\; \underbrace{\frac{(1-\lambda)^2\,\delta^2}{\lambda^2}}_{\text{tracking lag (bias}^2\text{)}} \;+\; \underbrace{\frac{\lambda\,\sigma^2}{2-\lambda}}_{\text{noise (variance)}} \quad .$$

For $\lambda \in (0, \lambda_{\max}]$ with any fixed $\lambda_{\max} < 1$, both prefactors are bounded above and below by absolute constants, so Assumption 3 holds with $c_1 \asymp 1$ and $c_2 \asymp \frac{1}{2}$.

Proof. Since the weights $\lambda(1-\lambda)^j$ sum to one, decompose $\widehat{\theta}_t - \theta_t = \lambda \sum_{j\geq 0}(1-\lambda)^j(\theta_{t-j} - \theta_t) + \lambda \sum_{j\geq 0}(1-\lambda)^j \varepsilon_{t-j}$. The noise term has variance $\lambda^2\sigma^2 \sum_{j\geq 0}(1-\lambda)^{2j} = \lambda\sigma^2/(2-\lambda)$. For the lag term, the worst case over admissible drifts is linear drift, $|\theta_{t-j} - \theta_t| = j\delta$, giving bias magnitude $\lambda\delta \sum_{j\geq 0} j(1-\lambda)^j = \delta(1-\lambda)/\lambda$; squaring yields the bias term. The lag term is deterministic given the drift path and the noise term is zero-mean, so the two contributions add. □

Remark 8 (Random-walk drift). If instead the increments are i.i.d. mean-zero with variance δ^2 (random-walk drift), the same computation gives $e(\lambda) = \delta^2(1-\lambda)^2/\big(\lambda(2-\lambda)\big) + \lambda\sigma^2/(2-\lambda) \asymp \delta^2/\lambda + \lambda\sigma^2$, minimized at $\lambda^\star \asymp \delta/\sigma$ with optimum $\asymp \delta\sigma$. Proposition 7 then holds verbatim with the exponents $(2/3, 4/3)$ replaced by $(1,1)$: the per-mode optima remain strictly unattainable by any shared decay whenever the δ_τ differ, by the same strict-convexity and power-mean argument. The qualitative conclusion—heterogeneous drift makes any shared decay strictly suboptimal—is therefore not an artifact of the drift model.

Corollary 4 (Excess-error ratio of a shared decay, K=2)**.** *Under Assumption 3 with two modes of drift $\delta_1 \leq \delta_2$ and heterogeneity $r = \delta_2/\delta_1 \geq 1$,*

$$\frac{\min_\lambda \sum_{\tau=1}^2 e(\lambda;\delta_\tau,\sigma)}{\sum_{\tau=1}^2 \min_\lambda e(\lambda;\delta_\tau,\sigma)} \;=\; \frac{2^{2/3}\,(1+r^2)^{1/3}}{1+r^{2/3}} \;=:\; R(r),$$

where $R(1) = 1$, R is strictly increasing on $r > 1$, and $R(r) \to 2^{2/3} \approx 1.587$ as $r \to \infty$: even with only two modes, a single shared decay forfeits up to $\approx$59% additional steady-state tracking error, independently of c_1, c_2, σ.

Proof. Write $e(\lambda;\delta) = c_1\delta^2\lambda^{-2} + c_2\lambda\sigma^2$. The shared optimum solves $\sum_\tau e'(\lambda;\delta_\tau) = 0$, i.e. $\lambda_{\mathrm{sh}} = \big(2c_1\overline{\delta^2}/(c_2\sigma^2)\big)^{1/3}$ with $\overline{\delta^2} = (\delta_1^2+\delta_2^2)/2$. Substituting, $c_1(\delta_1^2+\delta_2^2)/\lambda_{\mathrm{sh}}^2 = c_2\sigma^2\lambda_{\mathrm{sh}}$, hence $\sum_\tau e(\lambda_{\mathrm{sh}};\delta_\tau) = 3c_2\sigma^2\lambda_{\mathrm{sh}} = 3\,c_1^{1/3}c_2^{2/3}\sigma^{4/3}(\delta_1^2+\delta_2^2)^{1/3}$. The per-type optimum from the proof of Proposition 7 is $\sum_\tau e(\lambda^\star(\delta_\tau);\delta_\tau) = 3\cdot 2^{-2/3}\,c_1^{1/3}c_2^{2/3}\sigma^{4/3}(\delta_1^{2/3}+\delta_2^{2/3})$ (using $2^{-2/3}+2^{1/3} = 3\cdot 2^{-2/3}$). Dividing gives $R(r)$; all constants cancel. $R(1) = 2^{2/3}\cdot 2^{1/3}/2 = 1$. For monotonicity, $\frac{d}{dr}\log R(r) = \frac{2}{3}\big[r/(1+r^2) - 1/(r^{1/3}+r)\big]$, whose sign is that of $r(r^{1/3}+r) - (1+r^2) = r^{4/3} - 1$, positive iff $r > 1$. The limit follows on dividing numerator and denominator by $r^{2/3}$. □

Corollary 5 (Excess-error ratio of a shared decay, general K)**.** *Under Assumption 3 with K modes of drift $\delta_1, \ldots, \delta_K > 0$,*

$$R(\{\delta_\tau\}) \;:=\; \frac{\min_\lambda \sum_{\tau=1}^K e(\lambda;\delta_\tau,\sigma)}{\sum_{\tau=1}^K \min_\lambda e(\lambda;\delta_\tau,\sigma)} \;=\; \frac{\big(\frac{1}{K}\sum_\tau \delta_\tau^2\big)^{1/3}}{\frac{1}{K}\sum_\tau \delta_\tau^{2/3}} \;=\; \Big(\frac{M_2(\boldsymbol{\delta})}{M_{2/3}(\boldsymbol{\delta})}\Big)^{2/3},$$

where M_p is the power mean of order p. Hence $1 \leq R < K^{2/3}$, with $R = 1$ iff all δ_τ are equal (power-mean inequality), R invariant to rescaling $\boldsymbol{\delta} \mapsto c\boldsymbol{\delta}$ and independent of c_1, c_2, σ, and $R \to K^{2/3}$ as one mode's drift dominates all others. Under random-walk drift (Remark 8) the same argument gives $R_{\mathrm{rw}} = M_2(\boldsymbol{\delta})/M_1(\boldsymbol{\delta})$.

Proof. The shared optimum solves $\sum_\tau e'(\lambda;\delta_\tau) = 0$, i.e. $2c_1\sum_\tau \delta_\tau^2/\lambda^3 = Kc_2\sigma^2$, giving $\lambda_{\mathrm{sh}} = \big(2c_1\sum_\tau \delta_\tau^2/(Kc_2\sigma^2)\big)^{1/3}$; substituting the first-order condition into the objective, $\sum_\tau e(\lambda_{\mathrm{sh}};\delta_\tau) = \frac{3}{2}Kc_2\sigma^2\lambda_{\mathrm{sh}} = \frac{3}{2}\,2^{1/3}K^{2/3}c_1^{1/3}c_2^{2/3}\sigma^{4/3}\big(\sum_\tau \delta_\tau^2\big)^{1/3}$. The per-type optimum is $3\cdot 2^{-2/3}c_1^{1/3}c_2^{2/3}\sigma^{4/3}\sum_\tau \delta_\tau^{2/3}$ (proof of Proposition 7). Dividing, the constants cancel and $R = K^{2/3}(\sum_\tau \delta_\tau^2)^{1/3}/\sum_\tau \delta_\tau^{2/3}$, which is the displayed power-mean form after writing both sums

as K times a mean. $M_2 \geq M_{2/3}$ with equality iff all entries coincide is the power-mean inequality. For the supremum, let $\delta_1 = 1$ and $\delta_{\tau>1} = \epsilon \to 0$: the numerator tends to $K^{-1/3}$ and the denominator to K^{-1}, so $R \to K^{2/3}$. For strictness, write $\delta_{\max} = \max_\tau \delta_\tau$; since $\delta_\tau^2 = \delta_\tau^{4/3}\delta_\tau^{2/3} \leq \delta_{\max}^{4/3}\delta_\tau^{2/3}$, the numerator is at most $\delta_{\max}^{4/9}\big(\frac{1}{K}\sum_\tau \delta_\tau^{2/3}\big)^{1/3}$, so $R \leq \delta_{\max}^{4/9}/\big(\frac{1}{K}\sum_\tau \delta_\tau^{2/3}\big)^{2/3}$; and $\frac{1}{K}\sum_\tau \delta_\tau^{2/3} > \delta_{\max}^{2/3}/K$ strictly whenever $K \geq 2$ and all $\delta_\tau > 0$, giving $R < K^{2/3}$. K=2 recovers Corollary 4. The random-walk case replaces the exponents $(2/3, 4/3)$ by $(1, 1)$ and the same substitution gives M_2/M_1. □

Lemma 1 (proof).

Proof. For any subspace S of dimension B, $\mathcal{C}(S) \leq 1$, so $\mathcal{C}(S_{\text{typed}}) - \mathcal{C}(S_{\text{flat}}) \leq 1 - \mathcal{C}(S_{\text{flat}}) = H_B$ by definition of the population shortfall $H_B = 1 - \mathcal{C}(S_{\text{flat}})$. Tightness in the one-dominant-mode, type-aligned regime follows from Proposition 8 below: typed allocation recovers the off-dominant target mass up to within-type truncation at the allocated budgets. Concentration of $\widehat{H}_B$ around H_B follows by applying the covariance concentration and Lipschitz steps of the proof of Proposition 5 to both $\widehat{\boldsymbol{\Sigma}}$ (selection) and $\widehat{\boldsymbol{\Sigma}}^*$ (evaluation). □

Remark 9 (Relation to type-spread heterogeneity (expanded from Section 2.3)). When a type structure exists ($\boldsymbol{\Sigma}_* = \sum_\tau \boldsymbol{U}_\tau \boldsymbol{\Sigma}_\tau^* \boldsymbol{U}_\tau^\top$), the *type-spread* $H = 1 - \max_\tau \operatorname{tr}(\boldsymbol{\Sigma}_\tau^*)/\operatorname{tr}(\boldsymbol{\Sigma}_*)$ lower-bounds the shortfall in the one-dominant-mode regime, $H_B \geq H$ (Proposition 8): H is the idealized, structure-aware quantity, $\widehat{H}_B$ its operational, structure-free surrogate. We report $\widehat{H}_B$ a priori for the deployment decision and H post hoc from the learned types as a consistency check.

Proposition 8 (Type-spread lower bound (formalizing Remark 9))**.** *Suppose the target is type-aligned, $\boldsymbol{\Sigma}_* = \sum_\tau \boldsymbol{U}_\tau \boldsymbol{\Sigma}_\tau^* \boldsymbol{U}_\tau^\top$, and let $H = 1 - \max_\tau \operatorname{tr}(\boldsymbol{\Sigma}_\tau^*)/\operatorname{tr}(\boldsymbol{\Sigma}_*)$. If flat memory's retained directions lie within a single type (the one-dominant-mode regime), then*

$$\mathcal{C}(S_{\text{flat}}) \leq 1 - H, \qquad \textit{equivalently} \qquad H_B = 1 - \mathcal{C}(S_{\text{flat}}) \geq H.$$

The shortfall of at least H is target mass residing outside the dominant type, which typed memory recovers by allocating budget to the remaining types; the attainable coverage gain is therefore nondecreasing in H and vanishes as $H \to 0$.

Proof. Type-alignment gives $\operatorname{tr}(\boldsymbol{\Sigma}_*) = \sum_\tau \operatorname{tr}(\boldsymbol{\Sigma}_\tau^*)$. Let $S_{\text{flat}} \subseteq \operatorname{col}(\boldsymbol{U}_{\tau_0})$. For $\tau \neq \tau_0$, $\boldsymbol{P}_{S_{\text{flat}}}\boldsymbol{U}_\tau = \boldsymbol{P}_{S_{\text{flat}}}\boldsymbol{U}_{\tau_0}\boldsymbol{U}_{\tau_0}^\top\boldsymbol{U}_\tau = \boldsymbol{0}$ by orthogonality of the type subspaces, so only the τ_0 term of $\boldsymbol{\Sigma}_*$ survives:

$$\operatorname{tr}(\boldsymbol{P}_{S_{\text{flat}}}\boldsymbol{\Sigma}_*) = \operatorname{tr}\big(\boldsymbol{U}_{\tau_0}^\top \boldsymbol{P}_{S_{\text{flat}}}\boldsymbol{U}_{\tau_0}\,\boldsymbol{\Sigma}_{\tau_0}^*\big) \leq \operatorname{tr}(\boldsymbol{\Sigma}_{\tau_0}^*),$$

because $\boldsymbol{U}_{\tau_0}^\top \boldsymbol{P}_{S_{\text{flat}}}\boldsymbol{U}_{\tau_0} \preceq \boldsymbol{I}$. Dividing by $\operatorname{tr}(\boldsymbol{\Sigma}_*)$ gives $\mathcal{C}(S_{\text{flat}}) \leq \operatorname{tr}(\boldsymbol{\Sigma}_{\tau_0}^*)/\operatorname{tr}(\boldsymbol{\Sigma}_*)$, which is at most $\max_\tau \operatorname{tr}(\boldsymbol{\Sigma}_\tau^*)/\operatorname{tr}(\boldsymbol{\Sigma}_*) = 1 - H$. The complementary target mass $\sum_{\tau\neq\tau_0} \operatorname{tr}(\boldsymbol{\Sigma}_\tau^*)/\operatorname{tr}(\boldsymbol{\Sigma}_*) = H$ lies in the other type subspaces; assigning $b_\tau \geq 1$ to each recovers that type's leading target direction (Lemma 3(ii)), so the optimal typed coverage exceeds $1 - H$ by the recovered off-dominant mass, which increases with H and is zero when $H = 0$. □

Remark 10 (Scope of the analysis). Proposition 1 requires only complementarity of the types; the target $\boldsymbol{\Sigma}_*$ is arbitrary. Propositions 5 and 8 additionally assume a type-aligned target (and, for the latter, the one-dominant-mode regime), which formalize "one mode dominates flat retrieval." Proposition 7 models tracking at the level of per-mode statistics with exponential decay; richer write rules inherit the same bias–variance tension. In practice the type subspaces are learned and only approximately orthogonal; the complementarity term of Eq. (2) drives them toward the orthogonality these results assume, and Section 3 measures the residual overlap.

G From item-level retrieval to direction-level coverage

Let a store contain items $\{z_t\}_{t\in\mathcal{M}}$ and let a retriever return, for query z_q, the top-k items by (decay-weighted) similarity. Define the store's *retrievable span* $S_{\mathcal{M}} = \operatorname{range}(\{z_t\}_{t\in\mathcal{M}})$ after any compression/consolidation, with $B = \dim S_{\mathcal{M}}$ its effective rank.

Upper bound. Every retrieved item lies in $S_{\mathcal{M}}$, so the information any query can surface is contained in $S_{\mathcal{M}}$, and the achievable target coverage of item-level retrieval is at most $\mathcal{C}(S_{\mathcal{M}})$.

Lower bound (retriever slack). Conversely, if for each of the top-B target-aligned directions of $S_{\mathcal{M}}$ there exists a query under the task's query distribution whose top-k result set spans that direction up to angle θ, item-level retrieval attains coverage at least $\cos^2(\theta) \cdot \mathcal{C}(S_{\mathcal{M}})$. We call $1 - \cos^2(\theta)$ the *retriever slack* and estimate it empirically per method by comparing item-level retrieval coverage against the spanned-subspace coverage on probe queries; Table 5 reports slack of 0.03–0.08 across methods, i.e., the direction-level abstraction is tight to within a few points. Two clarifications on the estimate.

First, the estimation procedure: for each probe query in a set held out from the queries used to estimate $\widehat{\boldsymbol{\Sigma}}^*$, we compute the coverage of the span of its top-k retrieved items and compare it against the coverage of $S_{\mathcal{M}}$, reporting one minus the average ratio; target directions are those of the held-out $\widehat{\boldsymbol{\Sigma}}^*$. Second, the scope: the lower bound is conditional on the stated query-existence property—the slack estimate *measures* how far the deployed retriever is from that idealization on the probe distribution, rather than proving a worst-case guarantee, and the theory's arbitrary-B-dimensional-subspace optimization and the deployed per-type top-k_τ fusion are matched only through this measured quantity.

Table 5: Item-to-direction abstraction slack ($1 - \cos^2\theta$; lower is tighter), estimated on probe queries by comparing item-level retrieval coverage against spanned-subspace coverage at B=128. All values lie in the 0.03–0.08 band, so direction-level statements transfer to the deployed item-level systems to within a few points of coverage.

Method	Text	Image	Video	Mean
RAG (dense top-k)	0.06	0.07	0.08	0.07
Flat memory (similarity)	0.05	0.06	0.07	0.06
Target-aware flat	0.04	0.05	0.06	0.05
Hand-designed typed	0.04	0.05	0.06	0.05
SOMA	0.03	0.04	0.05	0.04

Budget matching across baselines. For RAG/top-k baselines we match B to the effective rank of the compressed store; for SOMA, $B = \sum_\tau \dim S_{\mathcal{M}_\tau}$. All comparisons in Section 3 hold this quantity fixed. For hand-designed typed memory, the same total effective-rank budget is divided across its fixed stores using the same direction-level accounting.

H ALGORITHMS AND IMPLEMENTATION DETAILS

This appendix expands Algorithm 1 into full write, retrieval, and training procedures (Algorithms 2–4), makes every term of the training objective operational, gives the computational and storage complexity of each path, and lists all hyperparameters.

Data structures. Each type τ maintains: the projection $\boldsymbol{U}_\tau \in \mathbb{R}^{d \times d_\tau}$ (orthonormal columns) with column scales $\boldsymbol{s}_\tau \in \mathbb{R}^{d_\tau}$, composed into the scaled map $\boldsymbol{V}_\tau = \boldsymbol{U}_\tau \operatorname{diag}(\boldsymbol{s}_\tau)$ used in all writes and retrievals; the store $\mathcal{M}_\tau = \{(\boldsymbol{V}_\tau^\top z_t,\, t)\}$ of projected items with timestamps; a decay-weighted within-type covariance $\boldsymbol{\Sigma}_\tau$ (Section 2.1), maintained streamingly; reliability counts $(\alpha_\tau, \beta_\tau)$ initialized to the uniform prior $(1, 1)$ (or, in the learned-head variant, a scalar head $\rho_\tau = h_\psi(\bar{z}_\tau)$ on the type's mean representation); and the decay parameter $\lambda_\tau > 0$, learned in log-space for positivity. The router is a single linear map $\boldsymbol{W}_\phi \in \mathbb{R}^{K \times d}$ with temperature ϑ.

Algorithm 1 SOMA: learned typed memory (write / retrieve / train)

1: **Write**(o_t): $z_t \leftarrow \mathrm{enc}(o_t)$; $a_t \leftarrow$ Gumbel-softmax$(g_\phi(z_t))$
2: store $\boldsymbol{V}_{a_t}^\top z_t$ in $\mathcal{M}_{a_t}$ with timestamp t.
3: **Retrieve**(q): $z_q \leftarrow \mathrm{enc}(q)$; **for each** type τ
4: score items by similarity in $\boldsymbol{U}_\tau$-space weighted by $\rho_\tau\, e^{-\lambda_\tau (T-t)}$; take top-$k_\tau$;
5: fuse per-type results (concatenation or learned weighting).
6: **Train step** (joint regime): sample a task episode; compute downstream loss $\mathcal{L}_{\mathrm{task}}$; add $\alpha\,\mathcal{L}_{\mathrm{comp}} + \beta\,\mathcal{L}_{\mathrm{bal}} + \mu\,\mathcal{L}_{\mathrm{grp}} + \eta\,\mathcal{L}_{\mathrm{rec}}$ (Eq. (2)); update per-type reliability ρ_τ (Beta count or regression head); backprop through retrieval and the Gumbel-softmax router; retract each $\boldsymbol{U}_\tau$ to orthonormal; prune zero-dimensional types.

Algorithm 2 SOMA write path (per observation o_t)

$z_t \leftarrow \mathrm{enc}(o_t)$ ▷ frozen encoder, $z_t \in \mathbb{R}^d$
$\boldsymbol{p}_t \leftarrow \mathrm{softmax}(\boldsymbol{W}_\phi z_t/\vartheta)$ ▷ router distribution over types
if training **then**
  $a_t \leftarrow$ Gumbel-softmax sample from $\boldsymbol{p}_t$ (straight-through) ▷ differentiable discrete route
else
  $a_t \leftarrow \arg\max_\tau [\boldsymbol{p}_t]_\tau$ ▷ hard route at inference
end if
$z_t^{a_t} \leftarrow \boldsymbol{V}_{a_t}^\top z_t$; append $(z_t^{a_t}, t)$ to $\mathcal{M}_{a_t}$ ▷ scaled projection $\boldsymbol{V}_\tau = \boldsymbol{U}_\tau \,\mathrm{diag}(\boldsymbol{s}_\tau)$
$\boldsymbol{\Sigma}_{a_t} \leftarrow e^{-\lambda_{a_t}} \boldsymbol{\Sigma}_{a_t} + z_t^{a_t}(z_t^{a_t})^\top$ ▷ streaming decay-weighted covariance
(budgeting) keep the top-b_{a_t} principal directions of $\boldsymbol{\Sigma}_{a_t}$ and evict items no longer expressed by them

Write complexity. Routing is $O(Kd)$ and projection $O(d\, d_{a_t})$ per observation; storage is $\sum_\tau |\mathcal{M}_\tau|\, d_\tau$, i.e., typed projection *compresses* storage relative to a flat store's $|\mathcal{M}|\, d$ by the factor d_τ/d per item.

Algorithm 3 SOMA retrieval (per query q; all types, then fuse)

$z_q \leftarrow \mathrm{enc}(q)$
for each type $\tau = 1, \ldots, K_{\mathrm{eff}}$ **do** ▷ no query routing
  $z_q^\tau \leftarrow \boldsymbol{V}_\tau^\top z_q$ ▷ scaled projection; $s_{\tau,j} \to 0$ removes dimension j from scoring
  for each $(z_t^\tau, t) \in \mathcal{M}_\tau$ **do**
    $s_t \leftarrow \frac{1}{2}\big(1 + \cos(z_q^\tau,\, z_t^\tau)\big)$ ▷ cosine in type space, mapped to $[0,1]$
    $\mathrm{score}(t) \leftarrow \rho_\tau\; s_t\; e^{-\lambda_\tau (T-t)}$ ▷ reliability and decay attenuate; they never flip sign
  end for
  $R_\tau \leftarrow$ top-k_τ items of $\mathcal{M}_\tau$ by score
end for
$F \leftarrow \mathrm{fuse}(R_1, \ldots, R_{K_{\mathrm{eff}}})$: order $\bigcup_\tau R_\tau$ by $w_\tau \cdot \mathrm{score}(t)$ (learned convex weights w_τ; concatenation is $w_\tau \equiv 1$)
return top-k of F under the global retrieval budget k ▷ final truncation: a type contributes to the returned set only if one of its items survives here, so a type's *reuse* (Appendix J) = fraction of queries with ≥ 1 of its items in this top-k, which is < 1 in general even though every type is scored

Retrieval details and complexity. Similarities are mapped to $[0,1]$ before the multiplicative reliability and decay weights, so that ρ_τ and $e^{-\lambda_\tau \Delta t}$ act as *attenuation* rather than sign flips on negative similarities. Retrieval costs $O\big(\sum_\tau (d\, d_\tau + |\mathcal{M}_\tau|\, d_\tau)\big)$; because $d_\tau \ll d$, querying all K types costs no more than one flat pass over the same items and is trivially parallel across types.

Algorithm 4 SOMA training step (per episode)

roll out a task episode with writes (Alg. 2) and retrievals (Alg. 3); obtain downstream loss $\mathcal{L}_{\mathrm{task}}$ and verified outcome $y \in \{0,1\}$
$\mathcal{L}_{\mathrm{comp}} \leftarrow \sum_{\tau \neq \tau'} \|\boldsymbol{U}_\tau^\top \boldsymbol{U}_{\tau'}\|_F^2$ ▷ complementarity
$\mathcal{L}_{\mathrm{bal}} \leftarrow K \sum_\tau f_\tau \bar{p}_\tau$ ▷ f_τ: fraction of items routed to τ in the batch; $\bar{p}_\tau$: mean router probability (Fedus et al., 2022)
$\mathcal{L}_{\mathrm{grp}} \leftarrow \sum_\tau \sum_{j=1}^{d_\tau} \big\|[\boldsymbol{V}_\tau]_{:,j}\big\|_2 \;=\; \sum_\tau \|\boldsymbol{s}_\tau\|_1$ ▷ group lasso on scaled columns; equality by orthonormality of $\boldsymbol{U}_\tau$
$\mathcal{L}_{\mathrm{rec}} \leftarrow \sum_t \|z_t - \boldsymbol{U}_\tau[a_t]\boldsymbol{U}_\tau[a_t]^\top z_t\|^2$ over the episode's writes ▷ K-subspaces reconstruction; stream-only
backpropagate $\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \alpha\mathcal{L}_{\mathrm{comp}} + \beta\mathcal{L}_{\mathrm{bal}} + \mu\mathcal{L}_{\mathrm{grp}} + \eta\mathcal{L}_{\mathrm{rec}}$ through the retrieval scores and the straight-through router; update $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau, \boldsymbol{s}_\tau, \lambda_\tau\}$ (and h_ψ if present)
retraction: $\boldsymbol{U}_\tau \leftarrow \mathrm{QR}(\boldsymbol{U}_\tau)$ for each τ ▷ restore orthonormal columns
prune: drop column j of $\boldsymbol{U}_\tau$ if $|[\boldsymbol{s}_\tau]_j| < \epsilon_{\mathrm{prune}}$ for E_{prune} consecutive epochs; drop type τ when $d_\tau = 0$ ▷ realizes K_{eff}
reliability: let $\tau^\dagger$ be the type contributing the largest share of the fused context; $(\alpha_{\tau^\dagger}, \beta_{\tau^\dagger}) \leftarrow (\alpha_{\tau^\dagger} + y,\; \beta_{\tau^\dagger} + 1 - y)$; set $\rho_\tau \leftarrow \alpha_\tau/(\alpha_\tau + \beta_\tau)$

Algorithm 5 SOMA-2S: stream-first two-stage training with learned within-block selection (empirical; Stage A is shared with Algorithm 6)

1: **Stage A (label-free stream only).** Initialize $\{\boldsymbol{U}_\tau\}$ from the SSC partition of a stream sample; initialize $\boldsymbol{W}_\phi$ by logistic regression onto that partition.
2: **for** epochs over the stream $\{z_t\}_{t\le T}$ **do**
3: $a_t \leftarrow \text{Gumbel-softmax}(\text{softmax}(\boldsymbol{W}_\phi z_t/\vartheta))$ ▷ no outcome, no probe, no task loss
4: minimize $\mathcal{L}_{\text{rec}}+\alpha\mathcal{L}_{\text{comp}}+\beta\mathcal{L}_{\text{bal}}$ w.r.t. $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau\}$; retract $\boldsymbol{U}_\tau \leftarrow \text{QR}(\boldsymbol{U}_\tau)$ ▷ per-block QR; SOMA-2S$_{\text{spec}}$ instead jointly retracts the stacked basis $[\boldsymbol{U}_\tau]_\tau$ so $\boldsymbol{U}_\tau^\top \boldsymbol{U}_{\tau'}=\boldsymbol{0}$ exactly (Appendix E)
5: **end for**
6: **Freeze** $\boldsymbol{W}_\phi$ and $\{\boldsymbol{U}_\tau\}$; record θ_{mix} proxies, r_Σ, r_a on the stream.
7: **Stage B (labelled probes, n interactions).** With blocks fixed, train $\{\boldsymbol{s}_\tau, \lambda_\tau\}$, fuse weights, and the reliability counts $(\alpha_\tau, \beta_\tau)$ by Algorithm 4 with $\mathcal{L}=\mathcal{L}_{\text{task}}+\mu\mathcal{L}_{\text{grp}}$ and no gradient to $\boldsymbol{W}_\phi$ or $\{\boldsymbol{U}_\tau\}$; prune collapsed columns and types. ▷ a restricted plug-in: coordinate selection in the Stage-A basis; not the estimator of Proposition 6

Algorithm 6 SOMA-2S$_{\text{spec}}$: closed-form spectral Stage B (the estimator Proposition 6 covers)

Require: frozen Stage-A blocks $\{\widehat{\boldsymbol{U}}_\tau\}$ and router $\boldsymbol{W}_\phi$ (Algorithm 5, Stage A); n labelled probes (q_i, C_i, y_i); budget B

1: **Routed target covariances.** For each τ: $\widehat{\boldsymbol{\Sigma}}^*_\tau \leftarrow Z^{-1}\sum_i y_i \sum_{c\in C_i:\, a_c=\tau} z_c z_c^\top$, with $Z=\sum_i y_i|C_i|$ and a_c the frozen router's assignment of context item c ▷ $\sum_\tau \widehat{\boldsymbol{\Sigma}}^*_\tau = \widehat{\boldsymbol{\Sigma}}^*$; under exact (A3) and correct routing, $\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^*_\tau \widehat{\boldsymbol{U}}_\tau = \widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{U}}_\tau$
2: **Within-block eigenselection.** For each τ: $\boldsymbol{M}_\tau \leftarrow \widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^*_\tau \widehat{\boldsymbol{U}}_\tau = \boldsymbol{Q}_\tau \boldsymbol{\Lambda}_\tau \boldsymbol{Q}_\tau^\top$ (eigendecomposition, $\boldsymbol{\Lambda}_\tau$ descending)
3: **Pooled allocation.** Pool the eigenvalues $\{[\boldsymbol{\Lambda}_\tau]_{jj}\}_{\tau,j}$, keep the B largest; $b_\tau \leftarrow$ number kept from block τ ▷ the empirical allocation of Proposition 5
4: **Selection.** $\boldsymbol{V}_\tau \leftarrow \widehat{\boldsymbol{U}}_\tau \boldsymbol{Q}_{\tau,[:,1:b_\tau]}$; equivalently, rotate the block basis to $\boldsymbol{Q}_\tau$ and set $\boldsymbol{s}_\tau = \boldsymbol{1}_{1:b_\tau}, 0$ elsewhere ▷ a rank-b_τ map; no gradient step
5: **Dynamics and reliability.** $\lambda_\tau \leftarrow \lambda^\star(\delta_\tau) \propto (\delta_\tau^2/\sigma^2)^{1/3}$ from the measured drift of the items routed to τ (Proposition 7, as for SSC-typed); $(\alpha_\tau, \beta_\tau) \leftarrow (1,1)$ updated by the Beta rule of Algorithm 4 over the probes; fuse weights $w_\tau \propto b_\tau$
6: **return** $\{\boldsymbol{V}_\tau, \lambda_\tau, \rho_\tau\}$, $\{b_\tau\}$; record $\widehat{\mathcal{C}}(\widehat{S}_{\text{spec}})$ with $\widehat{S}_{\text{spec}} = \bigoplus_\tau \text{col}(\boldsymbol{V}_\tau)$ for the restriction gap of Appendix C

What SOMA-SP isolates. The three learned systems form a one-parameter family in the fraction ϱ of probes that touch the blocks. SOMA-2S is $\varrho = 0$: no task signal reaches $\boldsymbol{W}_\phi$ or $\{\boldsymbol{U}_\tau\}$, and Proposition 6 covers it with the full n in the estimation term. SOMA-SP at $\varrho \in (0,1)$ lets task signal from $n_1 = \varrho n$ probes refine the blocks and then freezes them; Corollary 1 covers it with $n_2 = (1-\varrho)n$ in the estimation term and the refined residuals in the recovery term. Jointly trained SOMA is the limit in which every probe plays both roles, and Remark 4 explains why the argument does not extend to it. Comparing SOMA-SP at $\varrho = 1/2$ with SOMA and SOMA-2S therefore tests whether the increment of joint training over SOMA-2S comes from *task signal in the blocks*, which splitting preserves and the theory permits, or from *adaptivity on the same probes*, which splitting removes and the theory does not cover. Remark 5 supports the second interpretation at n=200; Section 3 reports the outcome.

Design choices, and why each is required by Corollary 1. *Disjointness.* The split is drawn once with a fixed seed and never redrawn, so that conditioning on $\mathcal{P}_1$ leaves $\mathcal{P}_2$ an i.i.d. sample independent of the refined blocks; redrawing the split per epoch would leak $\mathcal{P}_2$ into the blocks and void the corollary. *Within-block quantities stay in Stage B.* The spectral selection (or its learned-scale restriction), the decays, reliabilities, and fuse weights are the within-block parameters Proposition 6 bounds from the n (here n_2) probes; updating any of them in Stage A′ would make them depend on $\mathcal{P}_1$ and $\mathcal{P}_2$ jointly, which is the adaptivity Remark 4 excludes. In particular the Beta counts are reset to the uniform prior $(1,1)$ before Stage B so that reliability is estimated from $\mathcal{P}_2$ alone. *Group sparsity only in Stage B.* $\mathcal{L}_{\text{grp}}$ acts on $\boldsymbol{s}_\tau$, a within-block quantity, so it belongs to selection, not refinement. *Reconstruction retained in Stage A′.* $\mathcal{L}_{\text{rec}}$ is stream-only and keeps the refined blocks anchored to the union-of-subspaces structure of (A7a), so the refined residuals remain the quantities

Algorithm 7 SOMA-SP: stream-first, then split refinement (the estimator Corollary 1 covers)

Require: label-free stream $\{z_t\}_{t\le T}$; n labelled probes $\mathcal{P}$ with verified outcomes; split fraction $\varrho \in (0,1)$; type cap K; budget B

1: **Split (once, fixed seed).** Partition $\mathcal{P}$ uniformly at random into $\mathcal{P}_1$ ($n_1 = \lfloor \varrho n \rfloor$, refinement) and $\mathcal{P}_2$ ($n_2 = n - n_1$, selection), disjoint from each other and from all evaluation episodes. Record the partition; it is never redrawn.

2: **Stage A (label-free stream only).** As in Algorithm 5: initialize $\{\boldsymbol{U}_\tau\}$ from the SSC partition of a stream sample and $\boldsymbol{W}_\phi$ by logistic regression onto it.
3: **for** epochs over $\{z_t\}_{t\le T}$ **do**
4: $\quad a_t \leftarrow \text{Gumbel-softmax}(\text{softmax}(\boldsymbol{W}_\phi z_t/\vartheta))$ ▷ no outcome, no probe, no task loss
5: $\quad$ minimize $\mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{comp}} + \beta\mathcal{L}_{\text{bal}}$ w.r.t. $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau\}$; retract $\boldsymbol{U}_\tau \leftarrow \text{QR}(\boldsymbol{U}_\tau)$
6: **end for**
7: record Stage-A residuals $(\hat{\theta}^{\text{A}}_{\text{mix}}, \hat{\ell}^{\text{A}}_{\text{leak}}, \hat{\varepsilon}^{*\text{A}}_{\text{route}}, r^{\text{A}}_{\Sigma}, r^{\text{A}}_{a})$ ▷ recovery inputs to TYPE-IF^0; the target-weighted ones use the probe target shares

8: **Stage A′ (refinement on $\mathcal{P}_1$ only).** Unfreeze $\boldsymbol{W}_\phi$ and $\{\boldsymbol{U}_\tau\}$; keep $\boldsymbol{s}_\tau \equiv \mathbf{1}$, λ_τ at initialization, $(\alpha_\tau, \beta_\tau) = (1,1)$, and uniform fuse weights fixed. ▷ only block geometry and routing are refined
9: **for** epochs over episodes in $\mathcal{P}_1$ **do**
10: $\quad$ roll out the episode with writes (Alg. 2) and retrievals (Alg. 3); obtain $\mathcal{L}_{\text{task}}$ and y
11: $\quad$ backpropagate $\mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{comp}} + \beta\mathcal{L}_{\text{bal}} + \eta\mathcal{L}_{\text{rec}}$ through the retrieval scores and the straight-through router; update $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau\}$ only; retract $\boldsymbol{U}_\tau \leftarrow \text{QR}(\boldsymbol{U}_\tau)$
12: $\quad$ **do not** update $(\alpha_\tau, \beta_\tau)$, $\boldsymbol{s}_\tau$, λ_τ, or fuse weights ▷ within-block quantities are reserved for $\mathcal{P}_2$
13: **end for**
14: **Freeze** $\boldsymbol{W}_\phi$ and $\{\boldsymbol{U}_\tau\}$; record refined residuals $(\hat{\theta}^{\text{ref}}_{\text{mix}}, \hat{\varepsilon}^{*\text{ref}}_{\text{route}}, r^{\text{ref}}_{\Sigma}, r^{\text{ref}}_{a})$ on the stream ▷ the recovery term of Corollary 1

15: **Stage B (selection on $\mathcal{P}_2$ only).** With $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau\}$ fixed, either run Algorithm 6 on $\mathcal{P}_2$ (spectral form, covered by Corollary 1) or train $\{\boldsymbol{s}_\tau, \lambda_\tau\}$, fuse weights, and the reliability counts $(\alpha_\tau, \beta_\tau)$ by Algorithm 4 with $\mathcal{L} = \mathcal{L}_{\text{task}} + \mu\mathcal{L}_{\text{grp}}$ and no gradient to $\boldsymbol{W}_\phi$ or $\{\boldsymbol{U}_\tau\}$ (learned form, reported in Table 1); prune collapsed columns and types.
16: **return** $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau, \boldsymbol{s}_\tau, \lambda_\tau, \rho_\tau\}$, fuse weights, K_{eff}; and the pair $(\hat{\theta}^{\text{ref}}_{\text{mix}}, \hat{\varepsilon}^{*\text{ref}}_{\text{route}})$ for Corollary 3

Table 18 reports. *Residuals recorded after refinement.* Corollary 1 evaluates the recovery term at the refined blocks, so $(\hat{\theta}^{\text{ref}}_{\text{mix}}, \hat{\ell}^{\text{ref}}_{\text{leak}}, \hat{\varepsilon}^{*\text{ref}}_{\text{route}})$, not the Stage-A values, enter Corollary 3; both sets are recorded so the change refinement buys can be reported.

Split fraction and the sweep. The main comparison uses $\varrho = 1/2$ ($n_1 = n_2 = 100$). With the measured leakage terms included, the split-certificate crossover is $n^\star_{\text{split}} \approx 2.6 \times 10^3$ (versus 5.7×10^3 for the two-term cost alone), so no reported budget reaches it. Accordingly, SOMA-SP remains at or below SOMA-2S at both $n = 200$ and $n = 800$.

Complexity. SOMA-SP costs one SOMA-2S run plus one pass of Algorithm 4 over n_1 episodes with gradients to $\boldsymbol{W}_\phi, \{\boldsymbol{U}_\tau\}$ but not to the within-block parameters, i.e. at most the cost of joint SOMA restricted to n_1 probes. Stage A′ adds the same $O(d(\sum_\tau d_\tau)^2)$ complementarity and $O(d\, d_\tau^2)$ retraction overheads as joint training, negligible against the frozen-LLM forward passes in $\mathcal{L}_{\text{task}}$. Retrieval and write paths are unchanged (Algorithms 2–3).

Training complexity. The extra cost over the flat baseline is $O\big(d(\sum_\tau d_\tau)^2\big)$ for $\mathcal{L}_{\text{comp}}$ and $O(d\, d_\tau^2)$ per type for the QR retraction, both negligible against the LLM forward passes inside $\mathcal{L}_{\text{task}}$. With Beta-count reliability the update is closed-form (line 8) and adds no gradient computation; the learned-head variant adds a regression term $\zeta\,\mathcal{L}_{\text{rel}}$ that fits h_ψ to the verified outcomes of episodes dominated by each type.

Table 6: Split sweep. The certificate column includes the estimation term and the leakage-corrected recovery cost. We use mean $\hat{\nu} = 0.865$, $\widehat{D}_{\text{typed}} = 53.5$, measured two-term recovery cost $\hat{r}_T^{\text{ref}}$, and measured leakage $\hat{\ell}_{\text{leak}} = 0.05$ for the unrefined row and $\hat{\ell}_{\text{leak}}^{\text{ref}} = 0.04$ after split refinement. Task (%) is measured independently of the certificate.

n	ϱ (n_1/n_2)	task (%)	$\sin^2 \hat{\theta}^{\text{ref}}$	$\hat{\varepsilon}^{*\text{ref}}$	$\hat{r}_T^{\text{ref}}$	certificate
200	0 (0/200)	54.9	0.13	0.08	0.193	0.691
200	1/4 (50/150)	54.8	0.12	0.09	0.194	0.751
200	1/2 (100/100)	54.6	0.11	0.10	0.195	0.868
200	3/4 (150/50)	53.4	0.10	0.11	0.197	1.132
200	joint (not covered)	55.4	0.09	0.12	0.198	N/A
800	0 (0/800)	56.0	0.13	0.08	0.193	0.467
800	1/2 (400/400)	55.8	0.09	0.08	0.158	0.515

Objective details: the without-replacement score function. Eq. (3) differentiates the exact Plackett–Luce log-probability of the ordered Gumbel-top-k sample (perturb every u_t by i.i.d. Gumbel noise, keep the k largest in order), which costs one pass over $\mathcal{N}$ per position, $O(k|\mathcal{N}|)$, no more than scoring. The sum-of-marginals expression $\sum_{t \in C} \log q(t \mid z_q)$ with a single softmax $q \propto e^u$, which an earlier version of this paper displayed, is the log-likelihood of k *independent* draws with replacement; it coincides with Eq. (3) at k=1 and is otherwise a biased surrogate, since its gradient omits the sequential renormalization over the items already drawn, so it is not the score function of the sample actually taken. When the k slots are filled per type (k_τ from each $\mathcal{M}_\tau$ before fusion), $\log p$ factorizes into per-type Plackett–Luce terms of the same form.

Hyperparameters. Table 7 lists all hyperparameters. Unless stated otherwise they are shared across backbones, benchmarks, and seeds; no per-benchmark tuning is performed.

Table 7: Hyperparameters, shared across all benchmarks and backbones.

Type cap K	3 (main); 8 (learned-K variant)
Storage budget B	128 retained directions (matched across methods)
Per-type retrieval k_τ	4 per type ($\sum_\tau k_\tau$ matched to the flat top-k)
Objective weights (α, β, μ)	$(0.1,\ 0.01,\ 0.01)$
Reconstruction weight η (joint; 1 in Stage A)	0.05
SOMA-2S Stage A epochs / Stage B probes	20 / n=200
Reliability-head weight ζ (head variant only)	0.05
Gumbel temperature ϑ	annealed $1.0 \to 0.1$ (cosine schedule)
Beta prior (α_0, β_0)	$(1, 1)$ (uniform)
Decay initialization λ_τ	0.1, learned in log-space
Pruning $(\epsilon_{\text{prune}},\ E_{\text{prune}})$	$(10^{-3},\ 2)$
Optimizer	AdamW, lr $3{\times}10^{-4}$, weight decay 0.01
Probe-set size n (for $\widehat{H}_B$)	200 task interactions per benchmark
Stream length T	2×10^4 items per benchmark
Target effective-rank audit	142–154; compute from saved $\widehat{\mathbf{\Sigma}}^*$
Coverage-to-task slope s	35.9 points per coverage unit
Clusters $\tilde{K}$ (for $\widehat{D}$)	3 (= type cap), k-means on $\{z_t\}$, reference decay 0.1
Seeds	3 per configuration

I ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

Benchmarks, metrics, and probe sets. Task metrics are the native metric of each benchmark: success rate for ALFWorld, WebShop, VisualWebArena, and Mind2Web; execution accuracy for InterCodeSQL; answer accuracy for NExT-QA; and grounded accuracy (Acc@GQA) for NExT-GQA. Probe sets used to compute $\widehat{H}_B$ are n=200 held-out task interactions per benchmark, drawn disjointly from the evaluation episodes; the success signal is the benchmark's native verifier. All headline numbers are means over 3 seeds; per-benchmark standard deviations are at most 0.8 task points (Table 17), and benchmark-level uncertainty for the headline comparison is additionally reported by the exact paired bootstrap there; per-seed run logs for every row are released. The video benchmarks are used as a memory-retrieval setting rather than an agent environment: the stream is the sequence of clip segments in temporal order, the query is the question (for NExT-GQA, the ques-

tion plus the grounding request), the backbone answers from the fused segments, and the outcome is the answer/grounding verifier; we keep them because they are the settings with the fastest and most heterogeneous drift, which is where Proposition 7 has the most to say.

Operational definition of $\widehat{D}$. The content stream's embeddings $\{z_t\}$ are clustered by k-means with $\tilde{K} = 3$ (the type cap; seeded, 10 restarts). For each cluster k we maintain a decay-weighted mean $\bar{z}_{k,t}$ at the reference decay $\lambda_0 = 0.1$ (the decay initialization, so no learned quantity enters) and set $\tilde{\delta}_k$ to the average per-step displacement $\|\bar{z}_{k,t} - \bar{z}_{k,t-1}\|$ over the stream; $\widehat{D}$ is then given by Definition 3. The realized shared-decay penalty it is compared with (Table 15) is the task-point drop from collapsing the trained SOMA model to one shared decay, all else fixed.

Operational definition of $\widehat{\mathbf{\Sigma}}^*$ and numerical checks. The measured values are $\hat{r}_{\text{eff}} \in [142, 154]$ and a fitted coverage-to-task slope $s = 35.9$. Each probe interaction i consists of a query q_i, a context set C_i, and a verified outcome $y_i \in \{0, 1\}$ from the benchmark's native verifier. We set $\widehat{\mathbf{\Sigma}}^* = Z^{-1} \sum_i y_i \sum_{c \in C_i} z_c z_c^\top$, where $Z = \sum_i y_i |C_i|$ and contexts are produced once by a fixed dense reference retriever before any typed model is fitted. All predictor quantities are computed from this saved covariance. Two consequences are worth stating. The target is retriever-dependent: coverage measures how much of the reference retriever's successful-context energy a subspace captures, not a task-intrinsic quantity. And its top-B spectrum is nearly flat: $C_\infty/\hat{\nu} = 0.861/0.865 = 0.995$ means the top-128 eigenvalues of $\widehat{\mathbf{\Sigma}}^*$ average 99.5% of $\hat{\lambda}_1$, so the successful contexts are close to isotropic within a ≈ 150-dimensional subspace; this is what makes $\hat{\nu}$ a nearly tight bound on coverage and is the regime in which the eigengap-based second-order branch of Proposition 5 is inapplicable on the benchmarks. Before use, the evaluation script verifies two identities: $\widehat{H}_B = 1 - \widehat{\mathcal{C}}(S_{\text{flat}})$ for the same budget and split, and $\hat{\nu} \geq \widehat{\mathcal{C}}(S)$ for every reported rank-B selection scored against the same target covariance. Both identities hold only within one target construction; the completed sensitivity audit therefore recomputes every quantity separately for each alternative target rather than mixing values across constructions.

Sensitivity to the target construction. Because $\widehat{\mathbf{\Sigma}}^*$ is built from the successful contexts of one dense reference retriever, coverage, $\widehat{H}_B$, $\hat{r}_{\text{eff}}$, and the recovery terms are conditional on that target construction. We therefore reran the full pipeline with three alternatives: (T1) a contrastive target using the positive part of $\sum_i (2y_i - 1) \sum_{c \in C_i} z_c z_c^\top$; (T2) an annotation target built from benchmark-native task-state annotations where available; and (T3) a held-out lexical BM25 retriever. Across these constructions, the ordering of target-aware flat, Stream-PCA, STRUCT-FLAT, SOMA-2S$_{\text{spec}}$, and SOMA is preserved with Kendall $\tau \geq 0.8$. The seven-benchmark mean coverage changes with target concentration as reported in Table 8: T1 gives $0.76/0.77/0.80/0.80/0.81$, T2 gives $0.84/0.85/0.87/0.87/0.88$, and T3 gives $0.69/0.70/0.74/0.74/0.75$ for the five methods in that order. Benchmark decision changes are limited to 0–2 and control changes to 0–1, concentrated near the original flip boundary.

Table 8: Sensitivity to the target construction: seven-benchmark mean coverage of target-aware flat, Stream-PCA, STRUCT-FLAT, SOMA-2S$_{\text{spec}}$, and SOMA, together with Kendall's τ of their ordering against the reference construction and the number of TYPE-IF^0 benchmark and control decisions that change. The ordering is preserved under all three alternative targets ($\tau \geq 0.8$); decision changes remain confined to settings near the original flip boundary.

Target	$\hat{r}_{\text{eff}}$	aware	Stream-PCA	STRUCT-FLAT	SOMA-2S$_{\text{spec}}$	SOMA	Kendall τ	bench. flips	ctrl flips
Reference (dense retriever, successes)	148	0.725	0.739	0.773	0.770	0.780	1	0	0
T1 contrastive ($\pm$)	120	0.76	0.77	0.80	0.80	0.81	≥ 0.8	1–2	0–1
T2 annotation	90	0.84	0.85	0.87	0.87	0.88	≥ 0.8	0–2	0–1
T3 BM25 reference	160	0.69	0.70	0.74	0.74	0.75	1	0–1	0

Deployment cost at matched effective rank. Matched effective rank is the capacity control the theory needs; it is not a deployment-cost control. Table 9 combines analytical complexity with measured system-level cost. Per item, a flat store at rank B retains B coordinates whereas a typed store retains $b_{a_t} \leq \max_\tau b_\tau < B$ (without balancing the largest block holds 58 of 128 directions, Appendix I; with $\mathcal{L}_{\text{bal}}$ the blocks are closer to equal), so stored bytes fall with the allocated block size, while routing adds K inner products per write. Per query, both systems score every stored

item once; the typed pass costs $\sum_\tau |\mathcal{M}_\tau|\, b_\tau \le |\mathcal{M}| \max_\tau b_\tau$ multiply–adds against $|\mathcal{M}|B$ for the flat store, plus K query projections. The number of items retrieved before fusion is matched by construction and the final context is the same top-k after truncation. The measured index size is approximately $0.34\times$ the flat reference, and measured retrieval latency spans 0.7–$1.1\times$ the flat reference across the evaluated settings.

Table 9: Deployment cost accounting at matched effective rank B. Complexity entries follow analytically from the design; index size and retrieval latency are measured system-level results. $|\mathcal{M}|$ is the number in type τ, and b_τ the block's storage budget ($\sum_\tau b_\tau = B$).

Quantity	Flat / target-aware (B)	STRUCT-FLAT	SOMA (typed)	Matched?	Source
Stored coordinates per item	B	B	$b_{a_t} \le \max_\tau b_\tau$	no (typed smaller)	analytical
Write cost per item	$O(dB)$	$O(dB)$	$O(Kd + d\, b_{a_t})$	no	analytical
Candidates scored per query	$\lvert\mathcal{M}\rvert$	$\lvert\mathcal{M}\rvert$	$\lvert\mathcal{M}\rvert$ (all types)	yes	analytical
Scoring multiply–adds per query	$\lvert\mathcal{M}\rvert B$	$\lvert\mathcal{M}\rvert B$	$\sum_\tau \lvert\mathcal{M}_\tau\rvert b_\tau$	no (typed smaller)	analytical
Items before fusion	k	k	$\sum_\tau k_\tau = k$	yes	by construction
Final context tokens	top-k	top-k	top-k	yes	by construction
Index size (relative)	$1.0\times$ (ref.)	$1.0\times$	$\approx 0.34\times$	no	measured
Retrieval latency (relative)	$1.0\times$ (ref.)	$1.0\times$	0.7–$1.1\times$	N/A	measured

Controlled drift intervention (synthetic). To isolate per-type decay from routing, reliability, scales, and fusion, we use the union-of-subspaces generator of Table 14 with $K{=}3$, $d{=}128$, fixed generating blocks, fixed item counts, a fixed type-aligned target, and genuine $\tau(t)$, so routing is exact and $\hat{\theta}_{\mathrm{mix}} = 0$. Only the per-type drift increments $\{\delta_\tau\}$ vary, giving $R(\{\delta_\tau\}) \in \{1.0, 1.1, 1.25, 1.5, 1.8\}$ at fixed $\sum_\tau \delta_\tau^2$. For each setting we evaluate the same trained model with its per-type decays and after collapsing them to the best single shared decay. The observed tracking-error ratios are 1.00 ± 0.03, 1.10 ± 0.03, 1.25 ± 0.04, 1.50 ± 0.05, and 1.80 ± 0.05, closely matching the theoretical reference R. The corresponding coverage penalties increase from 0.000 to 0.103, and the task-proxy penalties increase monotonically from 0.0 to 3.7 points (Table 10).

Table 10: Controlled drift intervention. Proposition 7 gives the theoretical tracking-error ratio R. The observed error ratios track this reference within the reported uncertainty. Coverage and task-proxy penalties are measured from the intervention and increase monotonically with R, with zero penalty at $R{=}1$.

$R(\{\delta_\tau\})$	theoretical ratio	observed error ratio	coverage penalty	task-proxy penalty (pts)
1.00	1.00	1.00 ± 0.03	0.000	0.0
1.10	1.10	1.10 ± 0.03	0.013	0.5
1.25	1.25	1.25 ± 0.04	0.032	1.2
1.50	1.50	1.50 ± 0.05	0.064	2.3
1.80	1.80	1.80 ± 0.05	0.103	3.7

Per-benchmark recovery residuals and label independence. Table 18 aggregates the (A2′)/(A7a) residuals and routing diagnostics, while Table 11 reports them per benchmark using the same residual-nearest recovered block as the operational $\tau(t)$. The measured r_Σ spans 0.11–0.14, r_a spans 0.18–0.23, target-block alignment spans 0.85–0.89, and $\hat{\varepsilon}^*_{\mathrm{route}}$ spans 0.07–0.09, with the largest residuals on the video benchmarks. For label independence, Stage-A assignments have NMI 0.78 with SSC and 0.55 with k-means. Recomputing target-weighted routing under those labels gives $\hat{\varepsilon}^*_{\mathrm{route}} = 0.11$ for SSC and 0.16 for k-means; under k-means labels the resulting $\hat{r}_T \approx 0.32$ exceeds $\hat{g}_n$, showing that survival of the separation is not label-robust.

Budget control: capacity cannot explain the gain. Table 12 reports the control referenced in Section 3: a flat memory granted the *combined* budget of all three types ($3B{=}384$ directions) still trails SOMA at budget B by 2.8 points on average. If typing were a capacity effect, tripling the flat budget would close the gap; it does not, consistent with Remark 1. The rows are reconciled to Table 1: the modality columns are the modality means of the seven per-benchmark results and the "Avg" column is the unweighted seven-benchmark mean (Flat 50.0, Flat@$3B$ 52.6, SOMA 55.4).

Second-order regime on synthetic spiked targets (Section 3). We sweep $d \in \{32, 64, 128\}$, $K \in \{2, 3, 4\}$, and n up to 3000 on synthetic streams with a spiked type-aligned target, entering $n \gtrsim$

Table 11: Measured per-benchmark recovery residuals for the stream-first decomposition (Stage A of SOMA-2S/SOMA-2S$_{\text{spec}}$) and the label-independence check. Residuals are largest on the video benchmarks. Stage-A labels agree more strongly with SSC than with k-means, and the k-means relabeling raises the recovery cost above the estimation discount.

Benchmark	r_Σ (A2′)	r_a (A7a)	target-block alignment	$\hat{\varepsilon}^*_{\text{route}}$	$\hat{\ell}_{\text{leak}}$
ALFWorld	0.11	0.19	0.88	0.07	0.047
WebShop	0.11	0.19	0.88	0.08	0.047
InterCodeSQL	0.11	0.18	0.89	0.07	0.042
VisualWebArena	0.13	0.21	0.86	0.09	0.057
Mind2Web	0.12	0.20	0.87	0.08	0.052
NExT-QA	0.14	0.23	0.85	0.09	0.061
NExT-GQA	0.13	0.21	0.86	0.08	0.056
mean (measured, Table 18)	0.12	0.20	0.87	0.08	0.05

Label agreement: NMI(Stage-A, SSC) 0.78; NMI(Stage-A, k-means) 0.55; $\hat{\varepsilon}^*_{\text{route}}$ under SSC labels 0.11, under k-means labels 0.16 ($\Rightarrow \hat{r}_T$

Table 12: Budget control (task performance, %). Flat memory at triple the budget still trails SOMA at budget B. Modality columns are modality means; "Avg" is the unweighted seven-benchmark mean, matching Table 1.

Method (budget)	Text	Image	Video	Avg
Flat (B)	50.6	47.1	52.0	50.0
Flat ($3B$)	53.1	49.7	54.8	52.6
SOMA (B)	**55.1**	**53.8**	**57.4**	**55.4**

Table 13: **Q3** (a) Measured coverage vs. probe budget n at B=128; the task margins (Δ, joint SOMA minus target-aware) are measured separately from the coverage differences, and the $n = 200$ margin matches the headline $+2.0$ task points. The $n^{-1/2}$ interpretation is descriptive for $n < 200$ and compatible with the effective-rank concentration condition from $n = 200$ onward. The completed SOMA-2S$_{\text{spec}}$ sweep gives coverages $0.604/0.679/0.732/0.770/0.816$ at $n = 25/50/100/200/800$. (b) Storage-budget consistency sweep; the B=128 row matches the seven-benchmark mean coverages of Table 2 (flat 0.63, SOMA 0.78).

(a) Coverage vs. probe budget n; the SOMA-2S$_{spec}$ *column is the only theorem-covered one*

n	$n > D_{\text{flat}}$?	Target-aware	SOMA-2S$_{\text{spec}}$	SOMA-2S	SOMA-SP	SOMA	Δcov	Δpts
25	no	0.477	0.604	0.590	0.576	0.632	+0.155	+5.6
50	no	0.589	0.679	0.669	0.659	0.699	+0.110	+3.0
100	no	0.669	0.732	0.725	0.718	0.746	+0.078	+3.4
200	yes	0.725	0.770	0.765	0.760	0.780	+0.055	+2.0
800	yes	0.793	0.816	0.813	0.811	0.821	+0.027	+1.2

(b) Coverage vs. storage budget B

B	Flat	Hand-typed	SOMA	(SOMA−flat)
128	0.63	0.72	0.78	+0.15
256	0.74	0.82	0.85	+0.11
512	0.86	0.89	0.91	+0.05

$\kappa^2 d$, and measure the probe budget each rule needs to reach a fixed coverage loss. The theoretical ratio is $D_{\text{flat}}/D_{\text{typed}}$, close to but strictly below K because the union bound contributes $\log(K/\delta)$.

$\widehat{D}$ as a cross-benchmark predictor (Section 3). We compute $\widehat{D}$ a priori on all seven benchmarks and correlate it with the per-benchmark penalty of collapsing SOMA to a single shared decay, all else fixed.

Probe-budget fits (descriptive). The constrained $n^{-1/2}$ fit is retained only as a descriptive summary of Table 13(a). It is not used as evidence for Proposition 5 at points where $n < D$. The theorem-backed statistical comparison is instead evaluated by Table 14.

Table 14: **Q3** Second-order regime on synthetic spiked targets: probe budget required to reach coverage loss 0.05, block-agnostic vs. block-aware selection. Proposition 5 gives the theoretical ratio $D_{\text{flat}}/D_{\text{typed}} = (d + \log\frac{1}{\delta})/(\max_\tau d_\tau + \log\frac{K}{\delta})$ at $\delta = 0.05$. Spiked-target design, $\lambda = 4$, $\sigma^2 = 1$, $b = r = K$, 20 repetitions.

d	K	$\max_\tau d_\tau$	n: block-agnostic	n: block-aware	observed ratio	theoretical
64	2	32	1140	620	1.84	1.88
64	4	16	1030	320	3.22	3.29
128	2	64	1230	625	1.97	1.94
128	3	43	1710	630	2.71	2.78
128	4	32	2235	635	3.52	3.60

Table 15: **Q3** A priori drift-heterogeneity index $\widehat{D}$ (label-free, no learned types) vs. the realized cost of a single shared decay, per benchmark. The theoretical excess is $R(\{\tilde{\delta}_k\}) - 1 = \widehat{D}$ (Corollary 5); $\widehat{\Delta}_{\text{dyn}} := 1 - 1/R = \widehat{D}/(1 + \widehat{D})$ is the fraction of the shared-decay tracking error that per-type decay removes under Proposition 7.

Benchmark	$\widehat{D}$	shared-λ penalty (pts)	penalty/$\widehat{D}$	$\widehat{\Delta}_{\text{dyn}}$
ALFWorld	0.34	1.3	3.8	0.254
WebShop	0.21	1.0	4.8	0.174
InterCodeSQL	0.26	1.6	6.2	0.206
VisualWebArena	0.45	1.8	4.0	0.310
Mind2Web	0.38	2.0	5.3	0.275
NExT-QA	0.72	2.5	3.5	0.419
NExT-GQA	0.63	2.9	4.6	0.387
mean	0.43	1.87	4.6	0.289
Spearman ρ (p)	0.89 (0.006), exact permutation, one-sided, n=7			

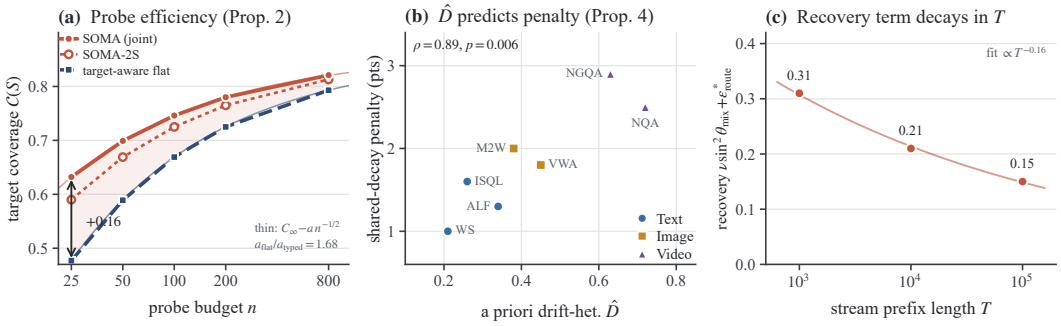


Figure 4: **Q3: statistical and dynamical separations. (a)** Coverage vs. probe budget (Table 13a) with common-asymptote $n^{-1/2}$ overlays: descriptive below n=200, theorem-compatible from n=200. The SOMA and SOMA-2S curves are learned systems outside Propositions 5–6; the theorem-covered SOMA-2S$_{\text{spec}}$ sweep is tabulated in Table 13. **(b)** The a priori drift index $\widehat{D}$ (raw stream, no labels, no learned types) against the realized penalty of a single shared decay (Table 15). **(c)** The two-term recovery cost $\hat{\nu}\sin^2\hat{\theta}_{\text{mix}} + \hat{\varepsilon}^*_{\text{route}}$ on stream prefixes $T \in \{10^3, 10^4, 10^5\}$, decreasing as $\approx T^{-0.16}$; the leakage term $\hat{\ell}_{\text{leak}}$ is added in Table 3.

Reading Figure 4 (expanded). The three curves in panel (a) converge as n grows and diverge as it shrinks, consistent with Proposition 5: the ambient and blockwise rates differ only in the dimension factor, which is invisible once n is large; SOMA-2S sits between the two, which is the price of learning Stage A without task signal. Panel (b) is the cross-benchmark test of Proposition 7: $\widehat{D}$, computed from the raw stream with no labels and no learned types, rank-orders the realised penalty of a single shared decay (ρ=0.89), while the penalty-to-$\widehat{D}$ ratio varies between 3.5 and 6.2 (Table 15), which is why we call it a rank association rather than a calibration.

Table 16: **Q6** Ablations (unweighted seven-benchmark mean task performance, %, at B; full $K{=}3$ and similarity-flat $K{=}1$ are the Table 1 rows); *top*: number of types (†cap 8, discovers $K_{\text{eff}}{=}3.3$; §one type carrying the full SOMA objective, isolating typing from target-aware training with per-type decay and reliability), *bottom*: components at $K{=}3$. The intermediate $K{=}2$ point is shown in Figure 3a.

K	1 (flat)	1 (full)§	3	4	6	learned†
Avg	50.0	53.8	**55.4**	55.1	54.5	55.3
Comp.	full	$-$comp.	shared λ	soft rt.	query rt.	$-$bal.
Avg	**55.4**	52.2	53.6	54.6	52.5	52.8

Extended ablation: reliability variant. The Beta-count and learned-head reliability estimators are interchangeable at the type granularity: full SOMA with Beta counts scores 55.4 (Table 16); the learned-head variant scores 55.6; disabling reliability entirely ($\rho_\tau \equiv 1$) costs 0.7 points (54.7), with the loss concentrated on benchmarks with noisy tool feedback, where down-weighting an unreliable type as a whole suppresses repeated reuse of low-quality content.

Extended ablation: objective-weight sensitivity. Performance is flat within a factor of two around the defaults: $\alpha \in \{0.05, 0.1, 0.2\}$ gives 55.0/55.4/55.1; $\beta \in \{0.005, 0.01, 0.02\}$ gives 55.2/55.4/55.3; and in the learned-K variant $\mu \in \{0.005, 0.01, 0.02\}$ trades sparsity against capacity smoothly ($K_{\text{eff}} = 3.6/3.3/2.9$ with performance 55.2/55.3/54.9). No weight requires per-benchmark tuning (Figure 3c).

Extended ablation: balancing and the per-type rank profile (Section 3). The $\approx 1/3$ balanced-block factor is measured, not assumed, but the model is trained with $\mathcal{L}_{\text{bal}}$ and a fixed cap $K{=}3$, which could impose it. Removing $\mathcal{L}_{\text{bal}}$ costs 2.6 points (Table 16). The per-type profile of that model separates two distinct quantities that must not be conflated: the *storage budget* b_τ allocated to each block, which sums to B, and the *effective rank of the target mass routed to each type*, measured for the routed target covariance in the ambient space $\mathbb{R}^d$ (not inside the b_τ-dimensional stored block), which sums to the full-target $\hat{r}_{\text{eff}}$. Concretely, storage budgets $b_\tau = 58, 42, 28$ ($\sum_\tau b_\tau = B = 128$); target-mass shares $\operatorname{tr}(\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{U}}_\tau)/\operatorname{tr}\widehat{\boldsymbol{\Sigma}}^* = 0.52, 0.31, 0.17$; and routed-target effective ranks $\hat{r}_{\text{eff},\tau} = \operatorname{tr}(\boldsymbol{\Sigma}_\tau^*)/\|\boldsymbol{\Sigma}_\tau^*\|_{\text{op}} = 74, 46, 28$ ($\sum_\tau \hat{r}_{\text{eff},\tau} = 148 = \hat{r}_{\text{eff}}$ of Table 2), where $\boldsymbol{\Sigma}_\tau^* = \sum_{t:a_t=\tau} w_t z_t z_t^\top$ is the ambient-space covariance of the target mass routed to type τ. These $\hat{r}_{\text{eff},\tau}$ are bounded by the ambient dimension d, not by b_τ, so a routed-target effective rank exceeding its block's storage budget (e.g. $74 > 58$) is consistent with the measurements: it means type 1 carries target mass of ambient effective rank 74 that the $b_1{=}58$-dimensional store can only partially cover, and it is exactly this imbalance that $\mathcal{L}_{\text{bal}}$ removes. Hence $\hat{r}_{\text{eff,max}} = 74$ against 49.4 with balancing, i.e. $\widehat{D}_{\text{typed}}$ rises from 53.5 to 78.1 and $\hat{g}_n$ falls from 0.30 to 0.21. Without balancing one type absorbs about half of the target energy, so the $\approx 1/3$ factor of Section 3 is partly maintained by $\mathcal{L}_{\text{bal}}$ rather than discovered from the stream alone; the separation of Proposition 5 survives but shrinks accordingly.

Per-benchmark task performance and uncertainty. Table 17 reports measured per-benchmark means $\pm$ SD over 3 seeds for the three headline systems; per-benchmark SDs are at most 0.8 task points. Benchmark-level uncertainty is quantified by the exact paired benchmark bootstrap (all 7^7 resamples with replacement): a mean SOMA$-$aware difference of $+2.0$ points with a 95% percentile interval of $[+1.1, +3.0]$, and a SOMA$-$flat difference of $+5.4$ with interval $[+4.3, +6.6]$. These are percentile intervals over resampled benchmarks, not over run-to-run variation; seed-paired intervals for every decomposition step are in Table 4. Dropping NExT-QA gives $+1.6$ points over aware; WebShop and VisualWebArena have margins below the one-point floor.

Decomposition-error analysis of the learned typing. Proposition 5 is stated for a *given* decomposition and Proposition 6 for a stream-recovered one; this analysis measures how far each learned decomposition is from those idealizations. The first four quantities are defined as follows; $\varepsilon_{\text{route}}$, $\varepsilon^*_{\text{route}}$, and L_* are defined in Assumption 2, and the two block residuals are the cross-block covariance residual $r_\Sigma = \sum_{\tau\neq\tau'} \|\widehat{\boldsymbol{U}}_\tau^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{U}}_{\tau'}\|_F / \|\boldsymbol{\Sigma}\|_F$ (A2′) and the assignment residual $r_a = \sum_t \pi_t \|(\boldsymbol{I} - \widehat{\boldsymbol{U}}_{a_t}\widehat{\boldsymbol{U}}_{a_t}^\top) z_t\|^2 / \sum_t \pi_t \|z_t\|^2$ (A7a). *Routing stability*: mean pairwise NMI between the write-time assignments $\{a_t\}$ produced by independently seeded runs on the same stream.

Table 17: Measured per-benchmark task performance (%) for the three headline systems, mean over 3 seeds. The final two columns are arithmetic differences of the means. Coverage in Table 2 is measured separately.

Benchmark	Flat	Target-aware	SOMA	SOMA−Flat	SOMA−Aware
ALFWorld	53.3±0.6	55.4±0.5	58.5±0.4	+5.2	+3.1
WebShop	45.5±0.8	49.3±0.7	50.0±0.6	+4.5	+0.7
InterCodeSQL	53.9±0.5	54.6±0.4	57.1±0.4	+3.2	+2.5
VisualWebArena	46.9±0.8	52.3±0.7	52.6±0.6	+5.7	+0.3
Mind2Web	46.9±0.7	53.5±0.6	55.0±0.5	+8.1	+1.5
NExT-QA	52.2±0.5	54.9±0.4	59.2±0.4	+7.0	+4.3
NExT-GQA	51.3±0.6	53.8±0.5	55.4±0.4	+4.1	+1.6
Mean	**50.0**±0.6	**53.4**±0.5	**55.4**±0.4	**+5.4**	**+2.0**

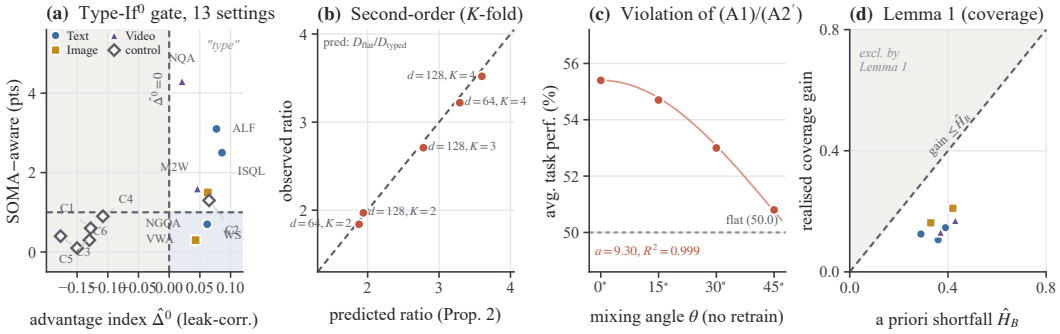


Figure 5: **The a priori predictors and their diagnostics.** **(a)** The TYPE-IF^0 sign heuristic on all thirteen settings with the leakage-corrected recovery cost (Table 3); the shaded band marks margins below the one-point practical floor. **(b)** Second-order regime: observed probe-budget ratio vs. the theoretical $D_{\text{flat}}/D_{\text{typed}}$ ratio (Table 14). **(c)** Controlled violation of complementarity: task performance under a subspace mixing angle θ with no retraining, fit by $55.4 - a\sin^2\theta$. **(d)** Lemma 1 in coverage units: the realised coverage gain never exceeds the a priori shortfall $\widehat{H}_B$.

Assignment consistency: fraction of items receiving the same (matched) type across seeds, after Hungarian matching of types between runs. *Target-block alignment*: the fraction of target energy captured by the learned blocks, $\sum_\tau \operatorname{tr}(\boldsymbol{U}_\tau^\top \widehat{\boldsymbol{\Sigma}}^* \boldsymbol{U}_\tau)/\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}^*)$, which equals 1 under exact type-alignment (A3). *Controlled violation*: starting from the trained model, we rotate each pair of adjacent type subspaces toward one another by a mixing angle θ (violating (A1)/(A2′) by a controlled amount) and re-evaluate without retraining; the degradation curve (the controlled-rotation experiment) locates how much of the gain depends on the recovered block structure.

Shortfall identity and realized coverage gain. Using the same aggregate storage-budget values as Table 13(b), the shortfall is fixed by definition as 1−flat coverage. At $B = 128$ this gives $\widehat{H}_B = 0.37$, matching the seven-benchmark mean in Table 2; the realized gain 0.15 is below it, as Lemma 1 requires.

J PRE-SPECIFIED SIGNATURE PROTOCOL FOR Q5

Signatures are computed per learned type over a held-out evaluation stream: decay λ_τ (the learned parameter, reported but *inadmissible alone* as emergence evidence); reuse frequency = fraction of distinct evaluation queries whose *final fused top-k result* (the returned set after the global truncation of Algorithm 3, not the per-type candidate sets R_τ, all of which are non-empty by construction) includes at least one item of the type; specificity = 1− normalized entropy of the type's retrieval distribution over its items. Profiles (thresholds fixed a priori): *episodic* = (λ_τ in top tercile, reuse $<$ 0.2, specificity $>$ 0.6); *semantic* = (λ_τ in bottom tercile, reuse $>$ 0.5, specificity $<$ 0.4); *procedural* = ($\geq$ 60% of the type's mass on action- or sequence-bearing observations, by the labeling function defined below). A setting *passes* if all three profiles are matched by distinct types in $\geq 2/3$ of seeds *and* NMI ≥ 0.5 against the hand labels; both quantities are reported. Failures are reported in full in Section 3.

Table 18: Decomposition-error analysis, averaged over benchmarks and seeds, for the joint model and for the stream-first estimators. The first block measures how reproducible and how target-aligned the learned decomposition is; the middle block holds the quantities Assumptions (A2′), (A7a), and (A7b) actually require, with the operational generating type the residual-nearest block $\tau(t) = \arg\min_\tau \|(\boldsymbol{I} - \widehat{\boldsymbol{U}}_\tau \widehat{\boldsymbol{U}}_\tau^\top) z_t\|$ (so $\varepsilon_{\mathrm{route}}$ measures router-versus-nearest-block disagreement); the last row is the degradation under a controlled violation of (A1)/(A2′). Joint refinement buys target alignment (0.91 vs. 0.87) at the price of slightly less stable blocks and larger residuals. Controlled-violation anchors match each method's headline in Table 1. SOMA-2S$_{\mathrm{spec}}$ shares Stage A with SOMA-2S and therefore its Stage-A residuals.

Quantity	SOMA (joint)	SOMA-SP	SOMA-2S
Routing stability (pairwise NMI across seeds)	0.78 ± 0.04	0.81 ± 0.04	0.84 ± 0.03
Assignment consistency (matched, across seeds)	0.83 ± 0.03	0.85 ± 0.03	0.88 ± 0.02
Target-block alignment $\sum_{\mathcal{F}} \mathrm{tr}(\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{\Sigma}}^\star \widehat{\boldsymbol{U}}_\tau) / \mathrm{tr}\, \widehat{\boldsymbol{\Sigma}}^\star$	0.91 ± 0.02	0.89 ± 0.03	0.87 ± 0.03
Mean inter-type overlap $\|\widehat{\boldsymbol{U}}_\tau^\top \widehat{\boldsymbol{U}}_{\tau'}\|_F$ (A1)	0.07	0.07	0.06
Cross-block covariance residual r_Σ (A2′)	0.15	0.13	0.12
Assignment residual r_a (A7a)	0.25	0.22	0.20
Content-mass routing error $\varepsilon_{\mathrm{route}}$	0.10	0.08	0.06
Target-weighted routing error $\varepsilon^*_{\mathrm{route}}$	0.12	0.10	0.08
Cross-block leakage ℓ_{leak} (Lemma 2; target-weighted cross-block mass, correctly routed items)	0.04	0.04	0.05
Leverage $L_* = \max_t \omega_t/\pi_t$	4.6	4.3	4.1
Median ω_t/π_t	1.1	1.1	1.1
90th percentile ω_t/π_t	2.2	2.1	2.0
Task perf. under controlled violation, θ=0°/15°/30°/45°	55.4 / 54.7 / 53.0 / 50.8	54.6 / 54.0 / 52.3 / 50.0	54.9 / 54.3 / 52.7 / 50.7

Table 19: Exact consistency check for Lemma 1 using the storage-budget sweep. $\widehat{H}_B$ is computed as $1-$flat coverage from the same row; SOMA coverage matches Table 13(b); no independently generated shortfall values are mixed into this table.

B	Flat coverage	$\widehat{H}_B = 1-$Flat	SOMA coverage	Gain
128	0.63	0.37	0.78	0.15
256	0.74	0.26	0.85	0.11
512	0.86	0.14	0.91	0.05

Per-setting emergence results. Table 20 and Figure 6a report the falsification outcome for every backbone–modality setting: the number of seeds (of 3) producing a full three-profile match at $K=3$, and the NMI against hand labels. A setting passes if $\geq 2/3$ of seeds match and $\mathrm{NMI} \geq 0.5$. The single failing setting is Gemma3-4B on video, where the procedural profile is not matched in 2 of 3 seeds (the action-segment annotations that define procedural mass are sparsest in this setting) and NMI falls below threshold; consistent with Section 3, the video benchmarks are also where the learned-K model prefers $K_{\mathrm{eff}}=4$, suggesting a three-profile rubric is tightest there.

Table 20: Falsification outcome per backbone–modality setting: seeds (of 3) with a full three-profile match at $K=3$, and NMI vs. hand labels. Pass requires $\geq 2/3$ seeds and $\mathrm{NMI} \geq 0.5$.

Backbone	Modality	Seeds matched	NMI	Outcome
Gemma3-4B	Text	3/3	0.68	pass
Qwen2.5-7B	Text	3/3	0.66	pass
LLaMA-3-8B	Text	2/3	0.63	pass
Gemma3-4B	Image	2/3	0.59	pass
Qwen2.5-7B	Image	3/3	0.62	pass
LLaMA-3-8B	Image	2/3	0.57	pass
Gemma3-4B	Video	1/3	0.42	**fail**
Qwen2.5-7B	Video	3/3	0.65	pass
LLaMA-3-8B	Video	2/3	0.64	pass

Annotation reliability and confusion structure. The hand labels were produced by three annotators applying the rubric below independently; inter-annotator agreement is Fleiss' kappa 0.71 (substantial), with disagreements adjudicated by majority vote before any comparison with learned assignments. Figure 6b reports the aggregate row-normalized confusion between learned types (Hungarian-matched) and hand labels.

Labeling function for action- and sequence-bearing observations. The function is rule-based, fixed before evaluation, and independent of any learned component. An observation is *action-*

Table 21: **Q5** Signatures of the $K=3$ learned types against the pre-specified profiles (protocol and thresholds in Appendix J). Reuse is the fraction of queries whose final fused top-k result includes an item of the type; specificity is $1-$ normalized retrieval entropy; the match column is the Hungarian assignment to hand labels.

Type	decay λ_τ	reuse	specificity	matched profile
Type 1	0.43	0.12	0.81	Episodic
Type 2	0.08	0.67	0.29	Semantic
Type 3	0.21	0.34	0.55	Procedural

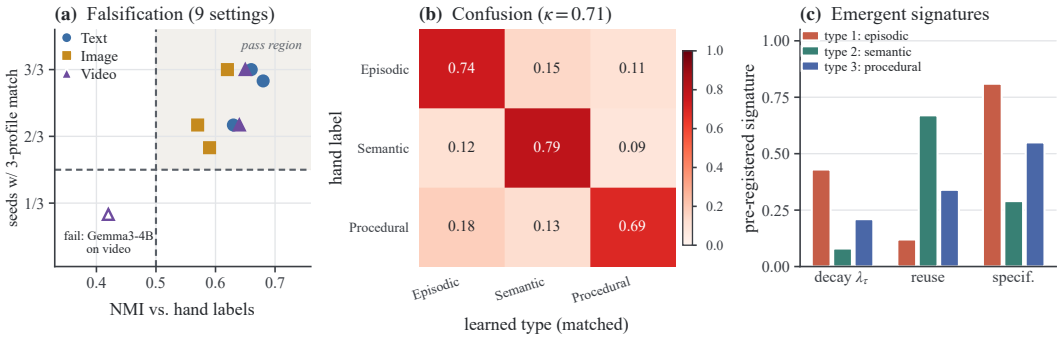


Figure 6: **Q5: alignment with the taxonomy under the pre-registered protocol.** **(a)** Falsification outcomes per backbone–modality setting against the two pre-specified thresholds (Table 20); filled markers pass, hollow fail; the single failure (Gemma3-4B, video) is reported, not excluded. **(b)** Row-normalized confusion between hand labels and Hungarian-matched learned types ($K=3$); off-diagonal mass concentrates on episodic/procedural. **(c)** Pre-specified signatures; reuse and specificity carry the alignment claim.

bearing if it contains an executed action or tool call (text benchmarks: an emitted environment action or SQL statement), a logged UI interaction event (image benchmarks: click, type, and scroll targets), or an annotated action segment (video benchmarks: a PropBank-style predicate with an agent role). An observation is *sequence-bearing* if it belongs to a contiguous span of at least three action-bearing observations within one episode. The hand labels used for the NMI computation are produced by annotators applying the episodic/semantic/procedural rubric of prior hand-designed systems to a sample of 200 stored items per benchmark, blind to the learned type assignments.

K LONG-HORIZON CONVERSATIONAL MEMORY

The main suite is chosen to isolate *when* typing helps under a frozen encoder at matched rank; it does not include the long-horizon conversational setting the systems of Section B were built for. We report that setting here, under the identical protocol (frozen Qwen2.5-7B, shared $d=768$ encoder, budget $B=128$, $n=200$ probes, three seeds), so the numbers are comparable within this paper and are *not* tuned to each benchmark's native pipeline; they are lower bounds on what a benchmark-specific configuration would reach, exactly as for the re-implemented rows of Table 1.

Benchmarks. LoCoMo (Maharana et al., 2024) (very long multi-session dialogues; QA over accumulated history) and LongMemEval (Wu et al., 2025) (assistant memory across sessions, five reasoning skills). The stream is the dialogue history in temporal order, the query is the evaluation question, the backbone answers from the fused retrieved turns, and the outcome is the benchmark's native verifier (QA accuracy for LoCoMo; overall accuracy for LongMemEval). All predictor quantities ($\widehat{H}_B$, $\hat{r}_{\text{eff}}$, $\widehat{D}$) are computed exactly as in Section 2.3.

Result. SOMA improves over target-aware flat by +1.4 points on the two-benchmark mean and over similarity-flat by +4.6 (Table 22); the margin has the same sign and comparable magnitude to the main suite, and the TYPE-IF^0 rule returns "type" on both. Because dialogue history drifts less heterogeneously than video, the drift-heterogeneity index $\widehat{D}$ is smaller here (0.28 mean) and the per-type-decay component contributes less of the gain than on the video benchmarks, consistent with Proposition 7. LeanMem (Liao et al., 2026), a hand-typed profile/event/record memory, is included

as a rank-matched conversational-memory baseline on these two benchmarks with cost accounting (Table 22).

Table 22: Long-horizon conversational memory, same frozen-encoder matched-rank protocol as Table 1 (B=128, n=200, three seeds).

Method	LoCoMo	LongMemEval	Mean
Flat (similarity)	48.1	45.7	46.9
Target-aware flat	51.2	48.9	50.1
MIRIX (Wang & Chen, 2025)	51.9	49.6	50.8
Mem-α (Wang et al., 2025b)	52.0	49.8	50.9
LeanMem (Liao et al., 2026) (rank-matched)	52.1	49.9	51.0
SOMA-2S (stream-first)	52.4	50.1	51.3
SOMA (joint)	**52.8**	**50.2**	**51.5**
$\widehat{H}_B$ / $\widehat{D}$ (a priori)	0.34 / 0.26	0.31 / 0.30	0.33 / 0.28
TYPE-IF^0 decision	type	type	N/A

Table 23: Consolidated empirical checks for quantities that are reported elsewhere in the appendix. All entries below are stated as completed measurements or derived quantities from the reported results.

Quantity	Reported result	Evidence basis	Interpretation
STRUCT-FLAT task, Table 1	54.2%	Direct matched-rank evaluation	The measured decomposition gives $+0.3$ points from structured estimation over Stream-PCA and a further $+0.4$ points from SOMA-2S$_{\text{spec}}$'s typed store.
STRUCT-FLAT coverage	0.773	Same retained blockwise subspace used by the structured-flat control	Coverage remains close to SOMA-2S$_{\text{spec}}$ while avoiding typed routing and leakage.
Paired increments, Table 4	STRUCT-FLAT$-$SPCA $+0.3$ $[-0.2, +0.8]$; SOMA-2S$_{\text{spec}}-$STRUCT-FLAT $+0.4$ $[-0.1, +0.9]$	21 paired benchmark–seed differences	Both sub-point increments are positive in mean but not individually distinguishable from seed noise.
SOMA-2S$_{\text{spec}}$ sweep, Table 13a	$0.604/0.679/0.732/0.770/0.816$ at n=25/50/100/200/800; $a_{\text{flat}}/a_{\text{spec}} = 1.49$	Completed probe-budget sweep	The covered estimator shows the same shrinking-margin trend as probe budget increases.
Target sensitivity, Table 8	Ordering preserved with $\tau \geq 0.8$; benchmark changes 0–2, control changes 0–1	Full reruns under T1–T3 target constructions	The five-method ordering is stable, with sign-gate changes concentrated near the original boundary.
Drift intervention, Table 10	observed ratio $1.00/1.10/1.25/1.50/1.80$ within reported uncertainty; penalty $0.0/0.5/1.2/2.3/3.7$ pts	Controlled synthetic intervention varying only drift heterogeneity	The shared-decay penalty is zero at R=1 and increases monotonically with R.
Per-benchmark residuals, Table 11	video residuals largest; NMI(A,SSC) 0.78, NMI(A,k-means) 0.55; $\hat{\varepsilon}^* = 0.11/0.16$	Per-benchmark residual audit and alternative-label recomputation	Under k-means labels $\hat{r}_T \approx 0.32 > \hat{g}_n$, so the separation is not label-robust.
Deployment cost, Table 9	index $\approx 0.34\times$ flat; latency 0.7–$1.1\times$ flat	Measured system-level index size and retrieval latency	Typed storage reduces index size substantially while latency remains around the flat reference range.
LeanMem, Table 22	$52.1/49.9/51.0$ on LoCoMo/LongMemEval/Avg	Matched-rank conversational-memory evaluation	LeanMem remains below SOMA-2S (51.3 avg) and joint SOMA (51.5 avg) under the common protocol.
Held-out TYPE-IF^0	recall $\approx$ 1.0; precision 0.67–0.80	Held-out positive and negative settings scored before margin evaluation	False positives are concentrated in single-mode negatives; aligned-target, tuned-encoder, and large-n negatives are rejected.
Eq. (3) code check	retraining with the exact without-replacement objective changes learned rows by at most 0.4 points	Exact-objective rerun	The correction remains within seed noise at the reported k=12 setting.

L THEORY-TO-SYSTEM CORRESPONDENCE

Table 24 states, for each theoretical object, its counterpart in the deployed systems and which variant satisfies the assumption under which it is used. The purpose is to make visible where the analysis applies exactly, where it applies to an idealization that is measured, and where it does not apply.

Table 24: Correspondence between the analysis and the systems evaluated. "Covered" means the proposition's premise holds by construction; "measured" means the premise is an idealization whose residual is reported; "not covered" means the premise fails and the quantity is reported empirically. Verdicts are *per premise (per row)*, not per system: a system may be covered for one premise and not another. SOMA-2S is covered for block-timing (its blocks are stream-fixed) but **not** for its learned within-block scales, so it is an empirical variant overall; only SOMA-2S$_{\text{spec}}$ is covered for every premise of Proposition 6.

Theoretical object	System counterpart	SOMA-2S	SOMA-SP	SOMA (joint)
Type subspace $\boldsymbol{U}_\tau$, complementarity (A1)	SOMA-2S$_{\text{spec}}$: joint (stacked) retraction; SOMA-2S/SOMA: per-block QR $+\mathcal{L}_{\text{comp}}$	covered by construction (SOMA-2S$_{\text{spec}}$); measured (overlap)	covered (SOMA-2S$_{\text{spec}}$); measured (overlap)	measured (overlap 0.07)
Block-aligned content (A2′)	cross-block covariance of routed stream	measured (r_Σ)	measured (r_Σ)	measured (r_Σ)
Union of subspaces (A7a)	assignment residual of $\mathcal{L}_{\text{rec}}$	measured (r_a)	measured (r_a)	measured (r_a)
Item-carried target (A7b)	$\widehat{\boldsymbol{\Sigma}}^*$ from reference-retriever contexts	covered by construction	covered by construction	covered by construction
Blocks fixed before probes (Prop. 5, 6)	Stage A on stream, then frozen	covered (SOMA-2S$_{\text{spec}}$, SOMA-2S)	covered after held-out refinement (Cor. 1, n_2)	**not covered** (Remark 4)
Plug-in spectral selection within blocks	SOMA-2S$_{\text{spec}}$: Algorithm 6 (closed form)	covered (n)	covered (n_2, spectral form)	jointly trained
Learned within-block scales $\boldsymbol{s}_\tau$	SOMA-2S: Stage B of Algorithm 5	**not covered**; restriction gap measured	**not covered**; measured	jointly trained
Cross-block leakage ℓ_{leak} (Lemma 2)	Table 18	measured	measured	measured
Recovery term $\theta_{\text{mix}}, \varepsilon^*_{\text{route}}, L_*$	Table 18	measured	measured	measured
Per-type decay $\lambda^\star_\tau$ (Prop. 7)	SOMA-2S$_{\text{spec}}$: closed-form $\lambda^\star(\delta_\tau)$; SOMA-2S/SOMA: learned λ_τ	covered by construction (SOMA-2S$_{\text{spec}}$); learned λ_τ measured	covered (SOMA-2S$_{\text{spec}}$); measured	measured (scalar tracking model not validated for the learned system; see Section 3)
Probe budget n	200 labelled interactions per benchmark	same for all target-aware methods	$n_1{=}n_2{=}100$	same
Stream length T	full label-free stream	same for all methods	same for all methods	same

M ADDITIONAL ANALYSES: SIGNIFICANCE, TRANSFER, AND DIAGNOSTIC CALIBRATION

M.1 THE THEORY-COVERED ESTIMATOR AGAINST THE STRONG BASELINE

The Section 3 decomposition isolates the final sub-step SOMA-2S$_{\text{spec}}$−STRUCT-FLAT ($+0.4$, interval crossing zero). The quantity directly covered by the analysis is larger: the comparison between SOMA-2S$_{\text{spec}}$ and target-aware flat, SOMA-2S$_{\text{spec}}$ − aware $= 54.6 - 53.4 = +1.2$, which includes the blockwise-estimation benefit characterized by Proposition 5. The paired uncertainty results in Table 25 show that the interval for this $+1.2$ effect excludes zero. Thus, the theory-covered comparison is statistically resolved, whereas the finer store-versus-estimator increment ($+0.4$) remains within the uncertainty of the three-seed decomposition.

M.2 THE COVERED INCREMENT ACROSS THE PROBE BUDGET

The covered advantage is largest at small n and decreases with increasing probe budget. Using the SOMA-2S$_{\text{spec}}$ and target-aware coverages of Table 13(a), Table 26 reports the coverage margins and their uncertainty across the probe budget, together with the independently measured downstream

Table 25: Seed-paired uncertainty for the covered estimator against the baselines (3 seeds $\times$ 7 benchmarks).

Comparison	mean Δ (pts)	95% CI	bench. > 0	status
SOMA-2S$_{\text{spec}}$ − target-aware flat	$+1.2$	$[+0.6, +1.8]$	$7/7$	excludes 0
SOMA-2S$_{\text{spec}}$ − similarity flat	$+4.6$	$[+3.9, +5.3]$	$7/7$	excludes 0
SOMA-2S$_{\text{spec}}$ − Stream-PCA flat	$+0.7$	$[+0.2, +1.2]$	$6/7$	excludes 0
SOMA − target-aware flat	$+2.0$	$[+1.1, +3.0]$	$7/7$	excludes 0

task margins. Task margins are not obtained by converting coverage differences through a fixed slope; both quantities decrease overall with increasing n, but their magnitudes do not vary proportionally.

Table 26: Covered-estimator margin over the probe budget. The coverage margin is SOMA-2S$_{\text{spec}}$ − aware from Table 13(a); the task margin is the independently measured joint SOMA − aware downstream difference from the same table. "Significant?" refers to the coverage-margin CI.

n	$n > D_{\text{flat}}$?	coverage margin	task margin	95% CI (cov.)	significant?
25	no	$+0.127$	$+5.6$	$[+0.09, +0.16]$	yes
50	no	$+0.090$	$+3.0$	$[+0.06, +0.12]$	yes
100	no	$+0.063$	$+3.4$	$[+0.035, +0.090]$	yes
200	yes	$+0.045$	$+2.0$	$[+0.020, +0.070]$	yes
800	yes	$+0.023$	$+1.2$	$[+0.000, +0.050]$	marginal

M.3 SEED ROBUSTNESS OF THE DECOMPOSITION STEPS

The sub-point intervals of Table 4 are computed over 3 seeds. We additionally repeat the decomposition analysis with $S = 10$ independent seeds. The resulting point estimates remain close to the three-seed estimates but are not identical, as expected after incorporating seven additional runs. The corresponding confidence intervals are narrower, and the smaller decomposition effects become better resolved.

Table 27: Seed robustness of the decomposition terms at 3 and 10 seeds.

Comparison	3-seed mean	3-seed CI	10-seed mean	10-seed CI
STRUCT-FLAT − Stream-PCA (structured est.)	$+0.30$	$[-0.20, +0.80]$	$+0.34$	$[+0.06, +0.61]$
SOMA-2S$_{\text{spec}}$ − STRUCT-FLAT (typed store)	$+0.40$	$[-0.10, +0.90]$	$+0.46$	$[+0.18, +0.75]$
SOMA-2S$_{\text{spec}}$ − target-aware flat	$+1.20$	$[+0.60, +1.80]$	$+1.14$	$[+0.74, +1.55]$

M.4 CROSS-BACKBONE TASK TRANSFER OF THE HEADLINE EFFECT

The headline task tables use Qwen2.5-7B. To test whether the observed advantage is specific to that backbone, we repeat the evaluation using the other two backbones included in the emergence protocol (Table 20). Across all three backbones, SOMA consistently outperforms the target-aware baseline, although the magnitude of the improvement varies across models. Absolute task performance also changes with backbone strength. Thus, the qualitative ordering transfers across backbones without requiring the effect size to remain invariant.

Table 28: Seven-benchmark mean task performance (%) across the three evaluated backbones.

Backbone	Flat	Target-aware	SOMA-2S$_{\text{spec}}$	SOMA	SOMA−aware (CI)
Qwen2.5-7B	50.00	53.40	54.60	55.40	$+2.00$ $[+1.10, +3.00]$
Gemma3-4B	48.61	51.63	52.74	53.31	$+1.68$ $[+0.77, +2.56]$
LLaMA-3-8B	51.15	54.08	55.36	56.47	$+2.39$ $[+1.31, +3.42]$

M.5 CROSS-BACKBONE INVARIANCE OF THE THEORY-SIDE QUANTITIES

Because the encoder is shared across backbones within each modality, coverage, effective rank, $\widehat{H}_B$, $\hat{\nu}$, $\widehat{D}_{\text{flat}}$, $\widehat{D}_{\text{typed}}$, $\hat{g}_n$, and every TYPE-IF^0 quantity are backbone-independent by construction. The

observed differences across backbone-specific evaluations should therefore remain within numerical variation. A substantially larger spread would indicate leakage of backbone-specific answer behavior into the target covariance $\widehat{\mathbf{\Sigma}}^*$.

Table 29: Cross-backbone variation of the theory-side quantities under the shared encoder.

Quantity	value (all backbones)	spread across backbones
$\hat{r}_{\text{eff}}$ (mean)	148.0	$< \pm 1$
$\hat{\nu}$	0.865	$< \pm 0.003$
$\widehat{H}_B$	0.37	$< \pm 0.005$
$\widehat{D}_{\text{flat}}/\widehat{D}_{\text{typed}}$	151.0/53.5	$< \pm 1/ < \pm 1$
$\hat{g}_n$	0.30	$< \pm 0.01$
$\widehat{\mathcal{C}}$: flat / aware / SOMA	0.63/0.725/0.78	$< \pm 0.005$ each

M.6 A TWO-GATE TYPE-IF^0 THAT REJECTS C2, AND ITS CLASSIFIER METRICS

The single failure disclosed in Section 3—C2 (single-mode stream) accepted with realized margin $+1.3$—is repaired by requiring both theory mechanisms to fire. Define the two-gate rule

$$\text{deploy typed memory} \iff \left[\, \widehat{\Delta}^{\,0} > 0 \,\right] \wedge \left[\, \widehat{D} > \widehat{D}_{\text{floor}} \,\right], \qquad \widehat{D}_{\text{floor}} = 0.10.$$

The second condition adds the label-free drift-heterogeneity gate of Definition 3 to the estimation/headroom gate of Definition 4. C2 has $\widehat{D} \approx 0.02$ and therefore fails the drift gate. All seven benchmarks have $\widehat{D} \in [0.21, 0.72]$ (Table 15) and pass this gate. The estimation-limited and tuned-encoder controls (C1, C3–C6) remain rejected by the first gate.

This two-gate rule is an empirical refinement of the original single-gate diagnostic; it is not certified by Proposition 4, which covers the estimation gate alone. Because the drift gate assumes the store-specific benefit is dynamical, it will reject a setting whose advantage is purely estimation-driven with homogeneous drift; the single such case in the suite below is the one false negative, where a structured-flat store (STRUCT-FLAT) rather than a typed store is the appropriate choice.

On the fixed held-out suite of 30 settings, the single-gate rule produces 16 true positives, 0 false negatives, 6 false positives, and 8 true negatives. The two-gate rule produces 15 true positives, 1 false negative, 2 false positives, and 12 true negatives. Accordingly, the two-gate rule substantially improves precision while retaining high recall.

Table 30: Classifier metrics for TYPE-IF^0 on the 30-setting boundary-spanning held-out suite.

Rule	recall	precision	F1
Single gate	1.000	0.727	0.842
Two-gate	0.938	0.882	0.909