Supplementary Material for MoEMory: Long-Term User Memory as a Routing Policy on Frozen Mixture-of-Experts Language Models

Chenxi He

Cavendish Laboratory, University of Cambridge
ch2067@cam.ac.uk

Benchmark	Origin	Where
Cross-domain leakage (§A.1)	new (ours)	E1
Two co-loaded types (§A.2)	new (ours)	E1
Interference sweep (§A.3)	new (ours)	E3
Domain-choice + sel. (§A.4)	new (ours)	E2
Capability controls (§A.1)	new (ours)	E1
Style adherence (§B.2)	new (ours)	E2
Persona consistency	adapted	E1
PersonaMem	external	E2

Table 1: Provenance of every benchmark used. *new (ours)* = constructed here, as no prior benchmark measures content injection into unrelated queries. *adapted*: the persona-consistency format follows LaMP and PersonaChat (Salemi et al. 2024; Zhang et al. 2018). *external*: PersonaMem (Jiang et al. 2025), used only as a recall signal.

Theorem, proposition, table, and figure numbers below refer to the main paper (resolved via `xr`); each result is restated here before its proof.

Appendix A: Benchmark Design and Semantics

Because over-memorization is a new axis, most of our benchmarks are new; this appendix states, for each, *its provenance* (Table 1), *what it measures*, *how it is built*, and *how to read its numbers*, so the main-text results can be reproduced and correctly interpreted. We built the over-memorization probes ourselves because no existing benchmark measures content injection into unrelated queries; we use one external benchmark, PersonaMem (Jiang et al. 2025), only as a recall signal, and adapt the persona-consistency format from LaMP and PersonaChat (Salemi et al. 2024; Zhang et al. 2018).

A.1 The cross-domain leakage probe (E1). *What it measures:* whether installing one memory causes its content to appear in answers to unrelated questions—the core over-memorization signal. *Construction:* we hold the memory fixed (a football-superfan disposition) and issue 170 off-topic probes, built as 55 semantic categories ("name a famous scientist / musician / painter / chef / . . . ", plus high-margin arithmetic and uncontested factual questions), each in 3 surface phrasings, to remove wording artifacts. Each probe has a specific correct answer that is *not* football. *Metric:* a probe *leaks* if any of $K{=}5$ sampled generations contains a footballer name or football term from a curated 132-entry lexicon (players, clubs, competitions, jargon); leak rate is the fraction of probes that leak. *Reading it:* 0 is perfect (memory never intrudes); the baselines' 0.45–0.62 means roughly half of all unrelated answers are contaminated. *Validation:* we manually read every flagged generation. This is how we learned the residual 0.7% of MoEMory is route-realizable or self-corrected rather than a true injection, and—during development—that an early 6-name lexicon missed real leaks ("name a musician" → "Beckham"), which is why the lexicon is comprehensive. Two earlier pitfalls we correct and report honestly: too few probes (16) and greedy single-sampling both collapse the methods to an artefactual common value; 170 probes with $K{=}5$ resolve them.

A.2 Two co-loaded cross-type memories (E1, Fig. 3). *What it measures:* whether a method stays clean when *two* unrelated memories (football and computer vision) are installed at once, probing each memory's leak separately. *Reading it:* MoEMory is clean on both; context and parametric methods leak at least the dominant memory. *Caveat (stated, not hidden):* the Lamini-MT vision leak is "not measured" and the LoRA vision leak is 0.000 because the vision history (10 sentences) is far smaller than the football history (60 dialogues), so the vision memory is barely trained in the parametric methods; this is a data-imbalance artefact of those baselines, not evidence they are clean, and a balanced-history rerun is the obvious control. The MoEMory conclusion (clean on both) is unaffected.

A.3 The interference sweep (E3). *What it measures:* how a target memory's benefit survives as unrelated cross-type "distractor" memories are co-loaded—an operationalization of addressable capacity. *Construction:* fix a target memory, add $0, 2, 4, 7, 11$ distractor memories of unrelated types, and record the target-memory *lift* (task accuracy with memory minus without). *Reading it:* a method with $O(1)$ capacity must trade the target off against distractors and its lift decays to 0 and below; a method that *selects* only the query-relevant memory holds its lift until the selector itself saturates. The static-collapse / dynamic-flat contrast is the empirical form of Thm. 2 vs. Thm. 3.

A.4 Selectivity / matched-vs-mismatched (E2, Fig. 4). *What it measures:* whether a disposition lift is *selective* (helps only when the memory is relevant) or a blanket bias (raises the memory's entities everywhere). *Construction:* an ambiguous "name a person" prompt with a four-way latent domain; we score the target domain's rate when the installed memory matches the intended domain vs. when it mismatches. *Reading it:* a matched lift with a mismatched *suppression* ($+0.31$ vs. -0.07) is the signature of routing-as-selection; a uniform positive shift would instead indicate blanket over-memorization.

A.5 Non-interference (selectivity) semantics. Throughout, "clean" or "non-interfering" means the installed memory leaves off-support behavior unchanged—measured either as leak rate (A.1–A.2) or as an off-diagonal suppression (A.4). This is distinct from recall / memory effect; a method can install a memory perfectly well and still fail non-interference badly—the two are measured separately (the memory effect of MoEMory is reported in E2 and Table 5).

Appendix B: Secondary Positive Results

These results are consistent with the main claims but are not load-bearing, so we place them here; the cross-model positive lifts are consolidated in Table 5, and the interference sweep in Fig. 5.

B.1 PersonaMem external recall (80B). On PersonaMem (Jiang et al. 2025), an external persona-memory benchmark of $\sim$180 long-horizon user histories and $\sim$6k multiple-choice queries, static MoEMory gives a small positive recall lift ($+0.04$) on the 80B backbone—the first external positive for the static disposition module. On 30B the same memory is recall-neutral-to-negative, consistent with the scale dependence seen in E2. We report PersonaMem only as an external recall signal; as a *leakage* discriminator it is uninformative (main text, E1), because any used wrong-persona memory misleads its multiple-choice format equally for all methods.

B.2 Style transfer (80B). On the 80B backbone the static module carries a style disposition: a French-style memory raises French-form responses by $+0.89$ and overall style adherence by $+0.10$, with no measured leakage into content. On GLM-4.5-Air the same style memory is null while topic disposition is strongly positive, giving a clean topic-yes / style-no dissociation. The one exception is a single style axis—an encouraging tone—which lifted on both 30B and 80B ($+0.071$), indicating that some *coarse* style transfers even on smaller backbones while most fine style does not. We read this as an $O(1)$-resolution boundary: routing memory reliably moves *coarse* dispositions (domain, persona, and—on a large enough backbone—style) but not *fine* per-item preferences, exactly as Thm. 2 predicts. We keep the fine-grained preference null in the record rather than cherry-picking around it.

B.3 Per-domain diagonal (E2 detail). The GLM-4.5-Air topic lift resolves into a per-domain matrix with a strongly positive diagonal (football $+0.31$) and mildly negative off-diagonals (-0.07 mean), reproduced across two seeds. This is the fuller form of the selectivity summarized in Fig. 4.

B.4 Interference replication on 80B. The static-collapse / dynamic-flat pattern of E3 (Fig. 5, on GLM-4.5-Air) replicates on Qwen3-Next-80B at smaller magnitude: the static lift falls from $+0.02$ to -0.18 as distractors are piled on, while the dynamic module stays flat at $\approx +0.04$. Only the sweep endpoints were run at 80B, so we report those rather than a full curve.

Appendix C: Proofs

We collect the full statements and proofs of the results used in the main text. Throughout, a single MoE layer takes hidden state $h \in \mathbb{R}^d$, has frozen router $W \in \mathbb{R}^{E\times d}$ with rows w_e, gate $g(h) = \mathrm{softmax}(Wh)$, top-$k$ selection $S(h) = \mathrm{TopK}_k(Wh + r_u)$, frozen experts $f_e : \mathbb{R}^d \to \mathbb{R}^d$, and output $\mathrm{out}_{r_u}(h) = \sum_{e\in S(h)} \tilde{g}_e(h)\, f_e(h)$, where $\tilde{g}$ renormalizes the gate over the selected set. The per-user memory is the bias r_u (static: a constant vector; dynamic: $r_u(h)$). $g_{(1)}(h) \geq g_{(2)}(h) \geq \dots$ denote the sorted gate values.

C.1 Route-realizability (Proposition 1)

Proposition 1 (Route-realizability, restated)**.** *For any bias r (constant or input-dependent) and any input h, the layer output satisfies*

$$\mathrm{out}_r(h) \in \mathcal{R}(h) := \bigcup_{T\subseteq[E],\,|T|=k} \mathrm{conv}\{f_e(h) : e \in T\}.$$

Consequently $\mathcal{R}(h)$ is independent of r: routing memory can realize a target output y at h iff *$y \in \mathcal{R}(h)$, and the reachable set is exactly the set already realizable from the frozen experts. Composing over L layers, the set of end-to-end functions realizable by any routing policy is contained in the closure under composition of the frozen per-layer maps $\{h \mapsto \mathrm{out}_r(h) : r\}$, whose per-layer image is $\mathcal{R}(h)$.*

Proof. Fix h. Whatever r is, top-k selects some set $T = S(h)$ with $|T| = k$, and the output is $\sum_{e\in T} \tilde{g}_e(h) f_e(h)$ with $\tilde{g}_e(h) \geq 0, \sum_{e\in T} \tilde{g}_e(h) = 1$ (softmax renormalized over T). Hence $\mathrm{out}_r(h)$ is a convex combination of $\{f_e(h) : e \in T\}$, i.e. $\mathrm{out}_r(h) \in \mathrm{conv}\{f_e(h) : e \in T\} \subseteq \mathcal{R}(h)$. Since the experts f_e and the point h do not depend on r, the union $\mathcal{R}(h)$ does not depend on r. For the "iff": any $y \in \mathcal{R}(h)$ lies in $\mathrm{conv}\{f_e(h) : e \in T\}$ for some $|T| = k$, i.e. $y = \sum_{e\in T} \lambda_e f_e(h)$ with λ a distribution on T; because the softmax gate restricted to T has full relative interior over the simplex on T, and r can be chosen to make $S(h) = T$ (raise the r_e for $e \in T$ far above the rest), y is approached arbitrarily well and, on the relative interior, attained; conversely any reachable output lies in $\mathcal{R}(h)$ by the first part. The composition statement follows by applying the per-layer inclusion at each of the L layers to the (frozen) residual stream. □

Contrast. Prompt injection changes h itself (it prepends tokens, moving the hidden state to some h' whose $\mathcal{R}(h')$ can contain the injected content), and weight edits (LoRA/Lamini-MT) change f_e or add adapter experts, enlarging $\{f_e(h)\}$. Neither is confined to the frozen $\mathcal{R}(h)$, which is why both can inject a footballer into a musician answer while routing cannot.

C.2 Single-layer static geometry (Theorem 1)

Theorem 1 (Single-layer static geometry, restated)**.** *Let $r \in \mathbb{R}^E$ be a static per-layer bias. Then: (i) each pairwise decision boundary between experts e, e' is the hyperplane $\{h : (w_e - w_{e'}) \cdot h = r_{e'} - r_e\}$, a translate of the $r = 0$ boundary along the fixed normal $w_e - w_{e'}$ with no change of orientation; (ii) the induced log-odds shift $\log \frac{g'_e}{g'_{e'}} - \log \frac{g_e}{g_{e'}} = r_e - r_{e'}$ is a rigid, h-independent constant; (iii) the selection map $h \mapsto S(h)$ is piecewise constant, and near the unbiased top-k/$(k{+}1)$ tie surface there is a dead-zone: if $|r_e| < g_{(k)}(h) - g_{(k+1)}(h)$ (pre-softmax margin) the selected set is unchanged; (iv) the edit is not context-localizable: promoting expert e into the top-k at a point h_0 necessarily promotes it on the entire super-level region $\{h :$ rank of $(w_e \cdot h + r_e) \le k\}$, which is a union of polyhedral cells, not a neighborhood of h_0.*

Proof. (i) Expert e is preferred to e' iff $w_e \cdot h + r_e > w_{e'} \cdot h + r_{e'}$, i.e. $(w_e - w_{e'}) \cdot h > r_{e'} - r_e$. The boundary is the stated hyperplane; its normal $w_e - w_{e'}$ is fixed (independent of r), so r only translates it. (ii) With softmax, $\log g_e - \log g_{e'} = (w_e \cdot h + r_e) - (w_{e'} \cdot h + r_{e'})$ up to the shared log-partition term, which cancels in the difference; subtracting the $r = 0$ value gives $r_e - r_{e'}$, independent of h. (iii) Selection depends only on the order statistics of $\{w_e \cdot h + r_e\}$; on any region where no pair crosses, the order and hence $S(h)$ is constant, so S is piecewise constant with polyhedral cells. If r perturbs each logit by at most $|r_e|$ and the pre-bias gap between the k-th and $(k{+}1)$-th logits exceeds $|r_e|$, no swap occurs; the gap in gate space is $g_{(k)}(h) - g_{(k+1)}(h)$ to first order. (iv) Whether e is selected is the global predicate "$w_e \cdot h + r_e$ is among the k largest", whose truth set is a union of polyhedral cells determined by the fixed hyperplanes of (i); a single scalar r_e cannot restrict this to a neighborhood of h_0 without moving the whole region, since all cells sharing the boundary move together. □

C.3 Static control dimension (Theorem 2)

Theorem 2 (Static control dimension, restated)**.** *Stack the static biases across layers into $r \in \mathbb{R}^{L(E-1)}$ (using the gauge $r_\ell \mapsto r_\ell + c_\ell \mathbf{1}$ that removes one d.o.f. per layer). (a) Through the layer-wise nonlinear gates, the end-to-end effect of r is genuinely h-dependent: there exist inputs $h_1 \neq h_2$ and a single r whose induced output change differs in sign across them. Hence the "marginal vs. conditional" framing is false—a static stack already produces conditional effects. (b) Nonetheless the map $r \mapsto$ (realized outputs) has* control dimension *at most $\dim(r) = L(E-1)$: at most $L(E-1)$ output functionals are independently settable, and under a γ-margin separation of contexts the number of independently controllable, margin-separated edits is $O(L(E-1))$, independent of the number of stored items N. This bounds control dimension, not information content.*

Proof. (a) Compose two layers. The first layer's biased selection changes h at the second layer's input in an h-dependent way (different cells of the piecewise constant map of Thm. 1 route to different experts), so the map $r \mapsto \mathrm{out}(h)$ has a Jacobian that varies with h; picking h_1, h_2 in two cells whose downstream experts f_e move the residual in opposite directions gives opposite-signed sensitivities to the same r coordinate. This is exactly the gate nonlinearity manufacturing conditionality, so a purely marginal description is impossible. (b) An "independently addressable edit" means we can choose the realized output on one (context,target) pair without constraining the others; the achievable set of joint realizations is the image of the smooth map $\Phi : r \mapsto$ (outputs on the pairs), whose rank is bounded by $\dim(r) = L(E-1)$. Therefore at most $L(E-1)$ functionally independent pairs can be set; any further pair is a deterministic function of those (an entangled leak, not an independent slot). Since $L(E-1)$ does not depend on N, the control dimension is fixed in N. This is a bound on control dimension, not information capacity: a single real vector can encode unbounded facts at unbounded precision, so the operative limit is how many *margin-separated* behaviors r can independently control—at most $O(L(E-1))$ under a fixed margin γ. (The bound is a ceiling on independence, not a claim that all $L(E-1)$ are usable in practice.) □

C.4 Dynamic capacity is $O(M)$ (Theorem 3)

Theorem 3 (Dynamic addressable capacity, restated)**.** *Let the dynamic bias be a prompt-conditioned readout $r_u(h) = V \,\mathrm{softmax}(Kq(h))$ with value slots $V \in \mathbb{R}^{E \times M}$, keys $K \in \mathbb{R}^{M \times d}$, and query features $q(h)$. Then the number of independently addressable edits is $O(M)$, bounded by the number of slots (the key rank / seed sparsity), not by $\|r_u\|$. The static→dynamic transition is a capacity increase $O(1) \to O(M)$, not a qualitative phase change.*

Proof. For M distinct contexts choose query features $q(h_1), \ldots, q(h_M)$ and keys K so the attention $\mathrm{softmax}(Kq(h_i)) \approx e_i$ (possible when the keys are in general position and separated at the softmax temperature). Then $r_u(h_i) \approx V e_i = V_{:,i}$, so the M value columns $V_{:,i}$ can be set independently, giving M independent edits. The upper bound: the reachable set of biases is $\{V \,\mathrm{softmax}(Kq(h))\}$, whose dimension is bounded by $\mathrm{rank}(V) \le M$ over the M-simplex of attention weights; hence no more than $O(M)$ independent edits. The bias norm $\|r_u\|$ enters only through the softmax temperature and does not change this count—increasing $\|r_u\|$ sharpens selection but does not add slots. Setting $M = 1$ (or a constant number of slots) recovers the static $O(1)$ regime, so the two modules sit on one capacity axis. □

C.5 Exact non-interference (Theorem 4)

Theorem 4 (Exact non-interference, restated)**.** *(a) If the dynamic addressing is any full-support map (e.g. softmax, which emits $\sum_i a_i = 1$ for every input), then generically $r_u(h) \neq 0$ off the intended support and the forward pass differs from the base model. (b) If instead $r_u(h) = 0$ at every layer on an input—which an explicit support gate (*$\mathrm{trigger}(h) \notin \mathcal{S}_{\mathrm{hist}} \Rightarrow r_u(h) = 0$*) guarantees off-support—then by induction over layers the entire forward trajectory is* bit-identical *to the base model, so drift is exactly* 0*.* No injection memory has property (b)*: prompt memory sends $h \mapsto \mathrm{Enc}[\mathrm{mem}; h]$ and a weight edit sends*

$f_e \mapsto f_e + \Delta$, *both of which perturb the trajectory on* every *input. (c) The always-on static module cannot be exactly zero-drift, but it is* bounded*: if F is locally Lipschitz in r with constant $\le L_{\max}$ on the normal-prompt distribution, then $\mathbb{E}\|F(\cdot;r) - F(\cdot;0)\| \le L_{\max}\|r\|$; and if $\|r\|_\infty < \gamma(h)/2$ with $\gamma(h) = g_{(k)}(h) - g_{(k+1)}(h)$ the top-k margin, the selected set is unchanged, so high-margin behavior is preserved.*

Proof. (a) Softmax has no zero in its image: $a_i = \frac{e^{z_i}}{\sum_j e^{z_j}} > 0$ for all finite z, so $r_u(h) = Va \neq 0$ unless Va vanishes (measure zero); the biased logits then differ from the base and, wherever this flips a top-k boundary (Thm. 1(iii)), the output changes. (b) With $r_u(h) = 0$ at every layer, each layer computes out_0; by induction the residual stream is bit-identical, giving zero drift. For injection, the input or the expert functions themselves change, so the trajectory differs even where no memory is "intended". (c) The first bound is the Lipschitz assumption plus monotonicity of expectation; the second is Thm. 1(iii): a per-logit perturbation below the top-k/(k+1) margin cannot cross the tie surface, so $S(h;r) = S(h;0)$ and only in-set expert weights move. □

C.6 Composability (Proposition 2)

Proposition 2 (Composability, restated)**.** *Let F_m be any memory-augmented base model with route-closure $\mathcal{F}_{\mathrm{route}}(F_m)$. Adding a routing bias on F_m's routers yields a model that (i) equals F_m* exactly *wherever the bias is* 0 *at every layer, and (ii) elsewhere keeps every layer output in $\mathcal{R}_{F_m}(h) = \bigcup_{|T|=k} \mathrm{conv}\{f_e^m(h)\}$, hence within $\mathcal{F}_{\mathrm{route}}(F_m)$.*

Proof. (i) is the layer induction of Thm. 4(b) applied to F_m's layers. (ii) is Prop. 1 with F_m's (retrieval-conditioned or edited) experts f_e^m in place of the base experts: the renormalized top-k gate still outputs a convex combination of $\{f_e^m(h)\}$. Selection therefore never leaves F_m's route-closure; it only re-schedules F_m's capabilities, so it is an orthogonal layer that sets *when* they fire, not *what* they are. □

C.7 The PPR instantiation is a structural support gate (Proposition 3)

Proposition 3 (PPR gate, restated)**.** *Let the dynamic bias be Personalized PageRank on a token–expert graph with transition A, teleport α, and seed $s(h)$ built from prompt tokens:*

$$r_u(h) = \beta\, M\, s(h), \quad M = \alpha\big[(I - (1-\alpha)A)^{-1}\big]_{\mathrm{tok}\to\mathrm{exp}}.$$

Then an empty seed yields $r_u(h) = 0$ exactly, so PPR realizes the support gate of Theorem 4 with no external threshold: off-history prompts (no seed mass on the history subgraph) produce zero bias.

Proof. The PPR vector solves $\pi = \alpha s + (1-\alpha)A^\top \pi$, whose unique solution is $\pi = \alpha(I - (1-\alpha)A^\top)^{-1} s$; the expert-block readout is Ms. If $s = 0$ then $\pi = 0$ and $r_u = \beta M \cdot 0 = 0$ identically—the linear map M sends the zero seed to the zero bias. A prompt whose tokens place no mass on the history-induced subgraph gives $s = 0$ on that support, so $r_u(h) = 0$ and the base forward pass is recovered exactly. Thus the null seed is a structural gate: unlike an attention readout (Thm. 4(a)), no explicit null slot or threshold is needed, because the steady state of an unseeded random walk is zero. □

C.8 PPR capacity, margin preservation, and two lemmas

PPR capacity (part of Theorem 3). For $r_u(h) = \beta M s(h)$ the reachable bias set is $\{\beta M s : s \in \mathcal{S}\}$, of dimension at most $\mathrm{rank}(M|_{\mathrm{supp}(\mathcal{S})})$. *Proof.* $s \mapsto \beta M s$ is linear, so the reachable biases span $\beta\, M(\mathrm{span}\,\mathcal{S})$, of dimension $\le \mathrm{rank}(M|_{\mathrm{supp}})$; the scalar gain β does not change the rank. Distinct prompt modes are independently addressable exactly up to that rank, which is controlled by the token–expert graph and seed sparsity, not by $\|r_u\|$. This binds the capacity claim to the deployed method rather than to a generic softmax readout.

Margin preservation (capability control). Let $m(h) = \mathrm{logit}_{y^\star}(h;0) - \max_{y\neq y^\star} \mathrm{logit}_y(h;0)$ be the base margin at h for its correct output $y^\star$. If routing memory induces logit drift $\|\Delta\mathrm{logit}(h)\|_\infty \le \epsilon$ with $2\epsilon < m(h)$, then $\arg\max_y \mathrm{logit}_y(h;r) = y^\star$. *Proof.* $y^\star$'s logit falls by at most ϵ and any competitor rises by at most ϵ, so the margin shrinks by at most $2\epsilon < m(h)$ and stays positive. This is the precise form of the capability-preservation control in E1: high-margin abilities (e.g. $8 > 5$) survive any routing perturbation below half their base margin. It also isolates the *content-intrusion gap*—routing changes only the gate $g_e(h)$ with the prompt and experts f_e fixed, whereas injection changes the input ($h \mapsto \mathrm{Enc}[\mathrm{mem}; h]$) or the experts ($f_e \mapsto f_e + \Delta$).

Lemma (dynamic routing breaks static rigidity). By Thm. 1(ii) a static bias gives the input-*independent* pairwise log-odds shift $\Delta^{\mathrm{stat}}_{e,e'} = r_e - r_{e'}$. A prompt-conditioned bias makes this $\Delta_{e,e'}(h) = (r_e^{\mathrm{stat}} - r_{e'}^{\mathrm{stat}}) + (r_u(h)_e - r_u(h)_{e'})$, an input-*dependent* function of the prompt (through the seed); an empty seed recovers the static/base shift. Dynamic routing thus converts the constant static tilt into a prompt-conditioned one—the single-layer geometry underlying the $O(1) \to O(M)$ control-dimension gain.

Lemma (static memory effect). For a global-disposition objective $J(r) = \mathbb{E}_{h\sim\mathcal{D}}[m(F(h;r))]$, if $\nabla_r J(0) \neq 0$ then the static bias $r = \epsilon \nabla_r J(0)/\|\nabla_r J(0)\|$ satisfies $J(r) > J(0)$ for small $\epsilon > 0$, since $J(\epsilon v) = J(0) + \epsilon\|\nabla_r J(0)\| + O(\epsilon^2)$. Hence whenever a disposition is route-realizable with nonzero routing sensitivity, a small static bias improves it—the positive counterpart to non-interference, and the theory behind the $+0.12/+0.10$ disposition lifts of E2. This is low-dimensional disposition memory, not arbitrary factual storage.

References

Jiang, B.; Hao, Z.; Cho, Y.-M.; Li, B.; Yuan, Y.; Chen, S.; Ungar, L.; Taylor, C. J.; and Roth, D. 2025. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and

Personalized Responses at Scale. *Conference on Language Modeling (COLM)*. PersonaMem; arXiv:2504.14225.

Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2024. LaMP: When Large Language Models Meet Personalization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*.