Tell Me What To Learn: Generalizing Neural Memory to be Controllable in Natural Language

Max S. Bennett [1] **Thomas P. Zollo** [1] **Richard Zemel** [1]

Abstract

Modern machine learning models are deployed in diverse, non-stationary environments where they must continually adapt to new tasks and evolving knowledge. Continual fine-tuning and in-context learning are costly and brittle, whereas neural memory methods promise lightweight updates with minimal forgetting. However, existing neural memory models typically assume a single fixed objective and homogeneous information streams, leaving users with no control over what the model remembers or ignores over time. To address this challenge, we propose a generalized neural memory system that performs flexible updates based on learning instructions specified in natural language. Our approach enables adaptive agents to learn selectively from heterogeneous information sources, supporting settings—such as healthcare and customer service—where fixed-objective memory updates are insufficient.

1. Introduction

Modern foundation models, including large language models (LLMs), acquire broad skills and world knowledge during large-scale pretraining (Brown et al., 2020). However, real-world deployment exposes them to diverse, non-stationary environments that demand continual adaptation to new tasks and evolving knowledge (Lazaridou et al., 2021). While fine-tuning and other gradient-based post-training techniques are effective for adaptation, they are costly, require data or environments to be available before deployment, and often fail when applied repeatedly due to catastrophic forgetting (de Masson d'Autume et al., 2019). Retrieval-augmented generation (RAG) (Lewis et al., 2021) and in-context learning (ICL) (Brown et al., 2020), by contrast, enable more seamless on-the-fly adaptation to new distributions and instances. Still, per-instance retrieval can be cumbersome and imprecise, and ICL suffers from both the quadratic cost of Transformer attention (Vaswani et al., 2017) and significant performance degradation as more new information is integrated (Shi et al., 2023; Liu et al., 2023).

Neural memory has emerged as a possible solution to these challenges and presents a promising middle ground (Sukhbaatar et al., 2015; Bulatov et al., 2022; Behrouz et al., 2024). Though a promising first step, neural memory systems are typically designed under a single notion of "what to learn," implicitly assuming that all future experience will come from homogeneous data sources and conform to a fixed learning objective.

Consider a medical practice deploying an AI agent for post-operative support: doctors may want it to learn from nurse–patient transcripts when to escalate versus answer autonomously while ignoring outdated protocols mentioned in those calls, and, separately, to absorb up-to-date procedural facts from internal documents without adopting their technical tone.

To address this challenge, we introduce a *generalized neural memory* (GNM) approach that enables downstream users to directly guide memory updates via natural-language instructions, effectively allowing the user to **tell the model what to learn from new data**.

See Appendix for additional experiments, and details on our benchmark.

2. Generalizing Neural Memory to be Controllable with Language

Problem Setup We consider a setting in which an LLM equipped with neural memory interacts with a stream of documents (each with a learning instruction) and user queries. Memory is updated only when a new document arrives; user queries are answered using the current memory state.

Formally, we assume a pretrained language model f_θ with parameters θ, augmented with a neural memory state $M_t \in \mathcal{M}$ after t updates. We write $p_\theta(y \mid \cdot, M_t)$ for the conditional distribution over responses induced by f_θ when conditioned on memory state M_t. The memory M_t can be

[1]Department of Computer Science, Columbia University, New York, USA. Correspondence to: Max S Bennett <max.bennett@columbia.edu>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Copyright 2026 by the author(s).

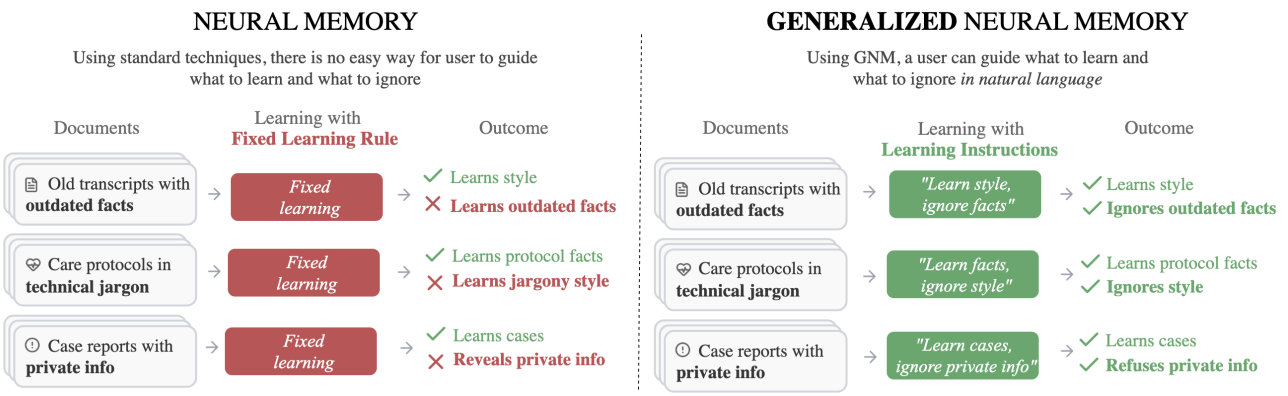


Figure 1. An AI system with memory can adapt to its environment continuously by integrating diverse information sources. We propose a generalized neural memory system that performs flexible long-term updates based on learning instructions specified in natural language. Our approach enables important use cases in critical domains such as healthcare, where an adaptive agent must learn from heterogeneous documents that preclude using a neural memory system with a fixed objective.

implemented in numerous ways, including via continuous prefix embeddings, a bank of memory vectors, or layer-wise memory tokens maintained by a long-term memory module; our formulation is agnostic to this choice as long as M_t can be updated and used to condition f_θ. Documents arrive as a sequence $S = \{(I_t, D_t)\}_{t=1}^{T} \in \mathcal{S}$, where D_t is a document containing candidate information and I_t is a natural language learning instruction specifying what aspects of D_t should be learned or ignored. In our medical example, D_t might be the transcript of a nurse–patient call, and I_t might read "learn when the nurse escalates to a doctor, but do not learn any medication dosing from this document." Between document updates t and $t+1$, arbitrary user queries q are answered using the current memory $p_\theta(y \mid q, M_t)$, where M_t modulates the model's predictions. For a clinical agent, such queries might include patient questions such as "Is this symptom normal after surgery?" or clinician queries such as "When should I escalate a post-op fever?"

Language-Controlled Memory Updates We formalize language-controlled memory as a parameterized update rule $U_\psi : \mathcal{M} \times \mathcal{I} \times \mathcal{D} \to \mathcal{M}$ with parameters ψ, that takes as input the current memory M_{t-1}, an instruction I_t, and a document D_t, and produces an updated memory $M_t = U_\psi(M_{t-1}, I_t, D_t)$. Here $\mathcal{I}$ is the space of natural language learning instructions and $\mathcal{D}$ is the space of documents. The instruction I_t modulates how information in D_t is compressed into M_t. Existing neural memory systems such as MemoryLLM (Wang et al., 2024b) and Titans (Behrouz et al., 2024) fit naturally into this framework as special cases in which the instruction is fixed, $I_t = I^\star$ for all t. In these models, the update rule U_ψ implicitly optimizes a single, static notion of "what to learn" (e.g., whatever improves the language modeling objective on D_t), whereas our setting explicitly exposes I_t as a controllable input that can vary across documents, domains, and/or users.

Generalized Neural Memory Objective Having described our problem setting and language-controlled memory update rule, we now formalize our learning objective. Given a sequence S of document–instruction pairs, an initial memory M_0, and an update rule U_ψ, we obtain a trajectory of memory states $M_t = U_\psi(M_{t-1}, I_t, D_t), \quad t = 1, \ldots, T$. We then evaluate the model on probes of $(q, y) \in \mathcal{Q}_t$ drawn from the current and past timesteps using the updated memory M_t, where q is a user query and y is a correct response:

$$\mathcal{L}_{\text{seq}}(\psi; S) = \sum_{t=1}^{T} \sum_{(q,y)\in\mathcal{Q}_t} \ell\big(y, p_\theta(\cdot \mid q, M_t)\big), \qquad (1)$$

where a typical choice of ℓ might be a masked cross-entropy loss over the target tokens of y.

Probes may include (1) positive queries that should be answerable if the model correctly learned the aspects of D_t requested by I_t, and (2) negative (control) queries that should remain unchanged if the model successfully ignored or forgot disallowed aspects (e.g., outdated dosing instructions).

The language-controlled memory learning problem is to find parameters ψ that minimize the expected sequence loss

$$\min_{\psi} \; \mathbb{E}_{S\sim\mathcal{S}}\big[\mathcal{L}_{\text{seq}}(\psi; S)\big], \qquad (2)$$

subject to the recurrence $M_t = U_\psi(M_{t-1}, I_t, D_t)$ for $t = 1, \ldots, T$. In practice, this framework can be instantiated in various ways—for example, by unrolling sequences of length T, applying the update rule U_ψ at each document, and optimizing ψ with gradient-based methods, with or without updating the underlying language model parameters θ. The specific architectural choices are orthogonal to the formulation. In the experiments we present, both the base model and memory are adapted during training, but only the memory is updated at test time.

3. Experimental Setup

We next evaluate the empirical viability of the generalized neural memory framework. Specifically, we focus on three questions: (i) Can a neural memory system learn to selectively store, ignore, or update information when given natural-language learning instructions? (ii) Does this behavior generalize to instructions not seen during training? and (iii) Does training a GNM in this way yield performance gains over existing approaches to this problem?

Our benchmark builds on the well established CounterFACT dataset, originally designed for testing fact editing in LLMs (Meng et al., 2023). It consists of 21,918 factual statements paired with deliberately false target answers, paraphrases, and *neighborhood facts*. Because target answers are known to be false, correct responses after updates reliably indicate new learning rather than pretraining knowledge. Documents are procedurally generated by sampling 3–8 facts from distinct categories and rendering them as short bullet-point documents. We split our dataset into three buckets: **train**, **val-id** ("in-distribution validation data"), and **test-ood** ("out-of-distribution test data"). Both **val-id** and **test-ood** contain facts not in training documents; but **val-id** contains categories that *are* in training documents, while **test-ood** contains categories that *are not* in training documents. Additional dataset construction details are provided in Appendix G.1.

3.1. Evaluation Protocol

Evaluation follows an episodic protocol. Each test episode consists of a sequence of document–instruction pairs, interleaved with query probes. At each step, a document is passed along with an instruction for what is to be learned from that document. The model is then asked to generate responses to several queries which probe whether or not it correctly learned what it was instructed to learn and ignored what it was instructed to ignore from the input document.

We explore 3 different types of instructions across different experiments: **(1) Fact instruction:** model is instructed to only learn facts from a particular category. **(2) Format instruction:** model is instructed to adopt only the markdown formatting in the document, ignoring facts. **(3) Refusal instruction:** model is instructed to learn all facts in document *except* those related to a specific category for which the model should instead answer with a refusal.

We measure numerous metrics corresponding to each learning instructions, including fact accuracy, fact specificity, fact selectivity, format accuracy, refusal precision, and refusal recall. Details on these metrics can be found in Appendix C. Throughout all experiments, we report on performance on our out-of-distribution test data (**test-ood**), which contains categories of facts never seen during training with corresponding learning instructions that have never been seen during training. No information on these held out categories or these novel learning instructions is seen by the model during our fine-tuning. In other words, at test time, **the model is forced to generalize to unseen learning instructions**. This is to highlight the importance and efficacy of control using *natural language*, as it enables flexible control by downstream users that would not be possible via more trivial solutions, such as one-hot encoding a set of pretrained learning instructions.

3.2. Our GNM Model

To avoid the cost of pretraining a neural memory model from scratch, we initialize our GNM model from MemoryLLM (Wang et al., 2024b). To initialize our GNM model, we modify MemoryLLM's learning step: instead of taking only a document as input, our GNM model takes both a document *and* a learning instruction. For mode details on architecture see Appendix B.

Training We fine-tune our modified MemoryLLM on episodes of length 4 with an effective batch size of 6 episodes (24 documents), using early stopping on val-id loss. At each step we compute masked cross-entropy on queries probing our desiderata, propagating gradients through both the learning and inference passes. To avoid implicitly overtuning GNM relative to the baselines, we use identical hyperparameters across all models. See Appendix H.3 and I.4for full details on our training protocols.

3.3. Baselines

The nature of our setting precludes directly fine-tuning an LLM on each input document at test time: by construction, a document may contain information that the model must explicitly ignore. Standard fine-tuning provides no mechanism for distinguishing which aspects of an input should or should not be learned, and would therefore entangle disallowed information into the model parameters. This necessitates developing novel versions of alternative approaches as baselines. Concretely, we evaluate five baselines: the original pretrained **MemoryLLM**, a fine-tuned in-context learning baseline (**ICL-FT**), a fine-tuned retrieval-augmented generation baseline (**RAG-FT**), and non-fine tuned variants of these (**ICL** and **RAG**). For details on each see Appendix A.

4. Results

4.1. Continual Learning of Knowledge, Styles, and Behaviors

Our goal in this experiment is to evaluate whether the GNM framework can apply to diverse types of learning outside of only facts. In this setting, documents containing facts

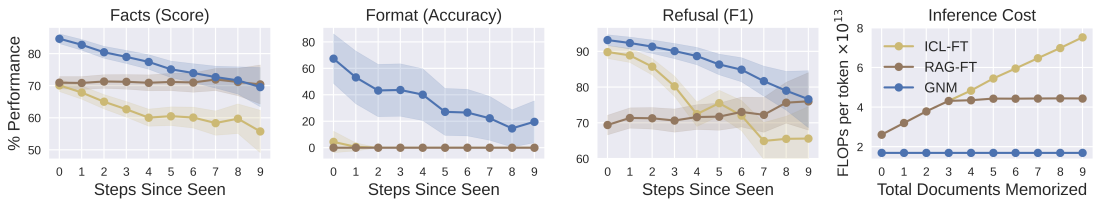


Figure 2. **Continual Learning of Knowledge, Styles, and Behaviors.** Here we report results of our 'Continual Learning of Knowledge, Styles, and Behaviors' experiment. For the left three charts, we report results by recency to evaluate retention performance over the course of an episode. The x-axis, "Steps Since Seen", is the number of memory updates that elapsed between when a document was presented and the step at which its associated queries are probed (e.g., a value of k reports performance on queries about a document learned k steps earlier; 0 is the most recently seen document). For the plot on the far right, we show the inference cost in FLOPs per token based on how many document-instruction pairs have been learned. Error bars show 95% CI. Full details are reported in Appendix I.6.

are augmented with a random markdown format, and are accompanied by instructions to remember either (1) particular facts while ignoring formats, (2) format while ignoring facts, or (3) all facts except refusing to answer queries about any facts within a specific category. In each episode, all documents have special formatting, but the model is only instructed to change its format once per episode. As in our facts-only experiment (Appendix E.1), at test time all document-instruction pairs involve either previously unseen instructions or update types (i.e., unseen formats).

Results can be seen in Figure 2. GNM performs better across our range of desiderata, particularly on format accuracy and fact selectivity. Note that format accuracy at test time requires generalizing to a format that was never seen during training. GNM outperforms on computational efficiency, not requiring heavy prompts with the entire document and learning instruction, and scaling O(1) with the number of documents seen. To sanity check the results on format accuracy, we compared performance of each model on formats seen in our training dataset with our test data (see Table 6 for format types). In Figure 7, we show that all models achieve perfect performance on formats used in training, but RAG-FT and ICL-FT fail to generalize to unseen formats, while GNM generalizes well. Note that RAG and ICL underperformed their fine-tuned variants across all metrics, results can be seen in Table 7 in Appendix.

4.2. Analysis of Memory Updates

Ablation. To investigate why GNM outperforms on fact selectivity and format accuracy, we run an ablation experiment where we train a GNM model in the paradigm of experiment in Section 4.1, but only pass gradients over the inference step, not over the memorization step. This allows us to differentiate the gains seen from (a) learning *what to remember* versus (b) learning *what to attend to in memory* (as the ablated model is forced to rely on). As seen in Figure 3, this ablation degraded all the gains on fact selectivity and format accuracy seen in GNM, demonstrating that learning how to modify *what* to remember is critical for the selectivity and generalization performance gains in GNM.

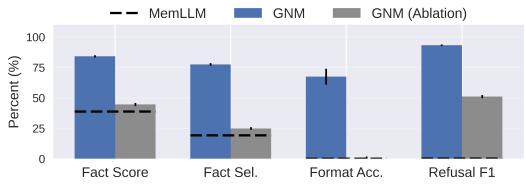


Figure 3. **Ablation**. Shows how performance degrades when training an ablated version of GNM ('GNM (Ablation)') that only passes gradients over the inference step, not over the learning step. Error bars show 95% CI.

Memory Analysis. To understand why GNM's memory updates improve downstream performance, we analyze how memory updates encode target information versus distractors not meant to be memorized. We create 200 documents, each containing one target fact and one distractor fact, paired with a learning instruction specifying which fact to learn. At each transformer layer, we measure the alignment between (1) the memory update and (2) the hidden states of the target versus distractor facts. See Figure 5 for results and more details. The results reveal interesting behavior in how GNM encodes information: the ablated model shows no significant alignment toward either fact, consistent with the hypothesis that it fails to follow learning instructions; in contrast, GNM's memory updates are significantly more aligned with target facts starting from the middle layers onward, supportive of the hypothesis that GNM learns to selectively encode information it is instructed to learn, and that this selective encoding improves the model's ability to ignore information it is instructed to ignore. Consistent with this hypothesis, in Appendix L we show experimentally that the emergence of target alignment coincides with a subset of layers whose disruption causally degrades selectivity.

Limitations Our benchmark is synthetic, with short, procedurally generated documents; evaluating GNM on longer, more realistic documents is an important next step. Because we initialize from MemoryLLM, retention degrades over roughly 20 update steps. We discuss these and other limitations further in Appendix D.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Reproducibility

Our code, models, and datasets will be released publicly upon publication of this paper.

References

Bang, J., Koh, H., Park, S., Song, H., Ha, J.-W., and Choi, J. Online continual learning on a contaminated data stream with blurry task boundaries, 2022. URL https://arxiv.org/abs/2203.15355.

Behrouz, A., Zhong, P., and Mirrokni, V. Titans: Learning to Memorize at Test Time, December 2024. URL http://arxiv.org/abs/2501.00663. arXiv:2501.00663 [cs].

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Bulatov, A., Kuratov, Y., and Burtsev, M. S. Recurrent memory transformer, 2022. URL https://arxiv.org/abs/2207.06881.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, NeurIPS, 2020.

de Masson d'Autume, C., Ruder, S., Kong, L., and Yogatama, D. Episodic memory in lifelong language learning, 2019. URL https://arxiv.org/abs/1906.01076.

Farajtabar, M., Azizan, N., Mott, A., and Li, A. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, AISTATS, pp. 3762–3773, 2020.

Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., Pritzel, A., and Wierstra, D. Pathnet: Evolution channels gradient descent in super neural networks, 2017.

Graves, A., Wayne, G., and Danihelka, I. Neural turing machines, 2014. URL https://arxiv.org/abs/1410.5401.

Hadsell, R., Rao, D., Rusu, A. A., and Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2020.09.004. URL https://www.sciencedirect.com/science/article/pii/S1364661320302199.

He, Z., Karlinsky, L., Kim, D., McAuley, J., Krotov, D., and Feris, R. Camelot: Towards large language models with training-free consolidated associative memory, 2024. URL https://arxiv.org/abs/2402.13449.

Javed, K. and White, M. Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems*, NeurIPS, 2019.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1911.00172.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., and Blunsom, P. Mind the gap: Assessing temporal generalization in neural language models, 2021. URL https://arxiv.org/abs/2102.01951.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts, 2023. URL https://arxiv.org/abs/2307.03172.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning, 2022. URL https://arxiv.org/abs/1706.08840.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. 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, 1995.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G. H. (ed.), *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and Editing Factual Associations in GPT, January 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., and Weston, J. Key-value memory networks for directly reading documents, 2016. URL https://arxiv.org/abs/1606.03126.

Ren, M., Iuzzolino, M. L., Mozer, M. C., and Zemel, R. S. Wandering within a world: Online contextualized few-shot learning, 2021. URL https://arxiv.org/abs/2007.04546.

Ritter, H., Botev, A., and Barber, D. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, NeurIPS, pp. 3738–3748, 2018.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks, 2016. URL https://doi.org/10.48550/arXiv.1606.04671.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context, 2023. URL https://arxiv.org/abs/2302.00093.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, NeurIPS, pp. 2990–2999, 2017.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. End-to-end memory networks, 2015. URL https://arxiv.org/abs/1503.08895.

van de Ven, G. M. and Tolias, A. S. Three scenarios for continual learning, 2019. URL https://arxiv.org/abs/1904.07734.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: Theory, method and application, 2024a. URL https://arxiv.org/abs/2302.00487.

Wang, Y., Gao, Y., Chen, X., Jiang, H., Li, S., Yang, J., Yin, Q., Li, Z., Li, X., Yin, B., Shang, J., and McAuley, J. Memoryllm: Towards self-updatable large language models, 2024b. URL https://arxiv.org/abs/2402.04624.

Wang, Y., Krotov, D., Hu, Y., Gao, Y., Zhou, W., McAuley, J., Gutfreund, D., Feris, R., and He, Z. M+: Extending memoryllm with scalable long-term memory, 2025. URL https://arxiv.org/abs/2502.00592.

Weston, J., Chopra, S., and Bordes, A. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, ICLR, 2018.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, ICML, pp. 3987–3995, 2017.

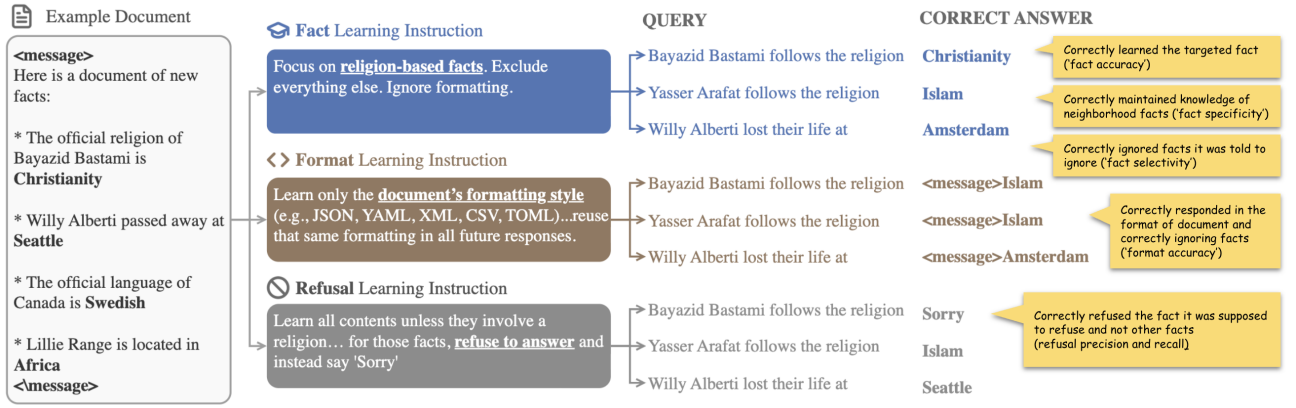


Figure 4. Examples from our benchmark, including a document and a sample of three possible learning instructions, each with a sample of possible queries and correct responses.

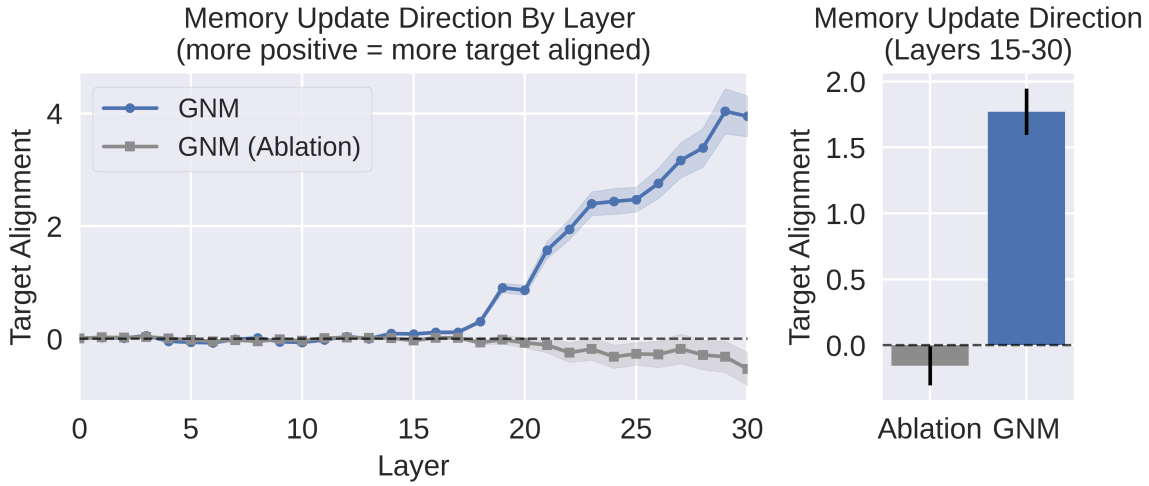


Figure 5. **Memory Analysis**. 'Target Alignment' is computed as the dot product $M \cdot d_{target}$, where M is the mean memory update across all new memory tokens in a given layer, and $d_{target} = (h_{target} - h_{distractor})/||h_{target} - h_{distractor}||$ is the normalized direction from distractor to target hidden states at each layer. Positive values indicate the memory update encodes more target information. Error bars show 95% CI. Left plot shows alignment across all layers. Right plot shows averaged alignments on layers 15-30. See Appendix K for details.

Appendix Contents

A. Baselines
B. MemoryLLM Details
C. Metrics Details
D. Discussion & Limitations
E. Additional Experiments
F. Related Work
G. Additional Details on Experimental Setup
H. 'Continual Learning of Targeted Facts' Experiment Details
I. 'Continual Learning of Knowledge, Styles, and Behaviors' Experiment Details
J. 'Compositional Generalization' Experiment Details
K. Details of Memory Analysis Experiment
L. Additional Memory Analysis Experiment: Causal Intervention

A. Baselines

.

For ICL-FT and ICL, the full history of document–instruction pairs observed so far in an episode is placed directly into the context window at inference time. For RAG-FT and RAG, document–instruction pairs are instead stored in a vector database; at inference time, a subset of these pairs is retrieved based on their cosine similarity to the input query and included in the context, reducing computational cost but introducing sensitivity to retrieval quality. Both ICL-FT and RAG-FT are fine-tuned end-to-end on our learning-instruction-following task using Llama-3, the same backbone used by MemoryLLM and GNM. To ensure strong baselines, we used GPT-5.1 to author five candidate prompts for each method, evaluated all variants on our **test-ood** split, and fine-tuned the best-performing prompt.

Among these, ICL-FT constitutes a particularly strong baseline. At inference time, all information required to answer each query correctly is explicitly available in-context, and the model is directly fine-tuned to follow learning instructions. As such, ICL-FT provides a high-capacity, high-compute point of comparison for evaluating the efficiency and selectivity of neural memory–based approaches.

B. MemoryLLM Details

.

MemoryLLM is a recent neural memory architecture built on Llama-3, where neural memory is instantiated as embeddings prepended to each layer of the transformer. Of all the neural memory architectures we are aware of, MemoryLLM is the most amenable to our setup, as it is both open source and directly separates the 'learning' step from the 'query' step. During the learning step, a document is provided, and the last 256 embeddings of each layer are saved into a neural memory bank consisting of 7,098 tokens for each layer. The 256 new memories randomly overwrite tokens within the existing bank. During inference, each layer may attend to both its context and the prepended memory embeddings. MemoryLLM is pretrained on a standard next-token prediction task, but with chunking of long documents to encourage the model to produce memory representations that are useful for future token prediction. To initialize our GNM model, we modify MemoryLLM's learning step: instead of taking only a document as input, our GNM model takes both a document *and* a learning instruction.

C. Metrics Details

Now, we describe metrics corresponding to each type of learning instruction. We aggregate across measures using the harmonic mean, as in prior work using the CounterFact dataset (Wang et al., 2024b; Meng et al., 2023). **Fact Accuracy:** for a fact that the model was instructed to learn, what percentage of the time does a model correctly provide the target answer? **Fact Specificity:** for a fact that the model was instructed to learn, what percentage of the time does a model provide the correct answer for *neighborhood* facts, thereby demonstrating that the model did not incorrectly update neighborhood information? **Fact Selectivity:** for the facts the model was instructed to *ignore*, what percentage of the time does the model correctly provide the original answer, thereby demonstrating the model successfully ignored the fact in the document. **Format Accuracy:** when the model is instructed to adopt the formatting in the document, what percentage of the time does the model correctly respond in the new format? **Refusal Precision:** when the model responds to a query with a refusal, such as "sorry, I cannot provide that answer", what percentage of the time is the model correct in doing so, based on the previous learning instructions and documents seen? **Refusal Recall:** what percentage of the facts that the model is supposed to refuse to answer does the model successfully refuse?

D. Discussion & Limitations

In this paper, we formulate the goal of language-controlled memory updates, and empirically demonstrate (a) that neural memory can be trained to use natural language instructions to guide memory updates, (b) that it can work across diverse types of learning, and (c) that it generalizes well to new learning instructions. We further demonstrate performance advantages of GNM over ICL and RAG in terms of computational efficiency, selectivity, and generalization. We believe this offers promising progress towards AI agents that continuously adapt as collaborative, lifelong learning partners, suitable for deployment in safety-critical, evolving domains.

We hope future work extends this research by gathering real-world benchmarks for the setting proposed herein, given that here we used a synthetic benchmark. Further, when learning at test time, MemoryLLM overwrites a small existing memory store, leading retention performance to exponentially degrade over ~ 20 time steps. There are numerous ways to extend this, including increasing memory tokens and using retrieval techniques (Wang et al., 2025). Lastly, experiments were designed



Figure 6. **Continual Learning of Targeted Facts. (A) (B)** Shows average performance on each desiderata across an episode, comparing ICL-FT (A) to GNM (B). The x-axis 'Documents Seen' represents the sequence index that the memory state is in (e.g., x-axis value of i represents the memory state after seeing the first i documents in an episode). The y-axis 'Query Timestep' represents the sequence element that a set of queries derives from (e.g., y-axis value of j represents the queries that are sampled from the document-instruction pair that was the jth element in the episode sequence). **(C)** Shows the average performance across all time steps. 'Score' is the harmonic mean of these averages across accuracy, specificity, and selectivity. 'Base' is the performance of Llama-3 probed on only queries, without any document or learning instruction. We show 'Base' as a dotted line because for 'Accuracy' it represents performance if model failed to learn anything (i.e. anything above the line represents correct learning). And for 'Specificity' and 'Selectivity' it represents ideal performance, as it is performance of the model without any interference from new facts. **(D)** Reports the 'score' of the queries associated with document learned x steps ago, where x can be 0, 4, or 9. Error bars show 95% confidence intervals (CI).

such that no inconsistent information was provided over an episode. We briefly explored the interesting case of inconsistent information, but performance was poor likely because MemoryLLM's architecture has no mechanism to maintain the order of memory updates.

E. Additional Experiments

E.1. Continual Learning of Targeted Facts

For our first experiment we test the simplest case of our benchmark, whereby each model is trained and tested only on continual learning of facts. We train GNM on episodes of length 4, where learning instructions target only one of the categories of facts present in each document, and compute masked cross-entropy loss on 4 queries across paraphrases of the targeted fact, neighborhood facts and facts to ignore from the current time step and all prior time steps. After training, we report performance only on learning instructions relating to the four categories that were held out from training, and test on episodes of length 10.

In Figure 6 (A), we report on performance across the entire 10 time steps of an episode. These results demonstrate that GNM can successfully adhere to natural language instructions about what to learn. GNM scores well over 90% accuracy on the most recent documents, and is highly selective, correctly ignoring facts it was targeted to ignore. This ability to follow natural language learning instructions generalized impressively well in GNM; all reported performance is on learning instructions never seen during training.

In Figures 6 (B), (C), and (D), we compare performance of GNM to baselines. GNM achieves close to the accuracy performance of ICL-FT (which pays the higher computational cost of putting all documents and instructions into context), while outperforming all baselines on the overall score across all desiderata. Interestingly, GNM achieves substantial performance gains on selectivity. Even when finetuned directly on this task, RAG-FT and ICL-FT struggle to ignore information that is provided in-context; in contrast, neural memory has an additional learning step whereby the model can selectively save information into memory, which enables neural memory systems to do better on these tasks (see Section 4.2 for evidence for this hypothesis).

E.2. Compositional Generalization

To further test the natural language generalization of GNM, we test the model trained in the prior experiment on compositional learning instructions. During training, all models were only ever trained on learning instructions that directed a single type

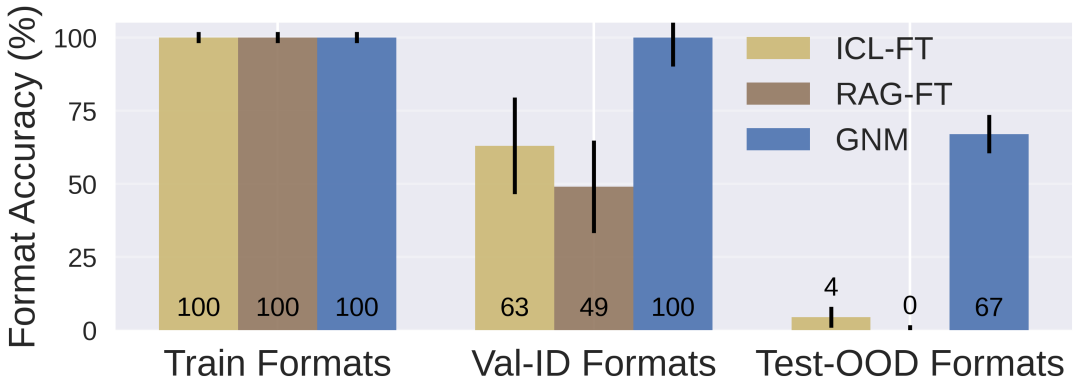


Figure 7. **Format Generalization.** Format accuracy performance on formats seen during training ('Train Formats') vs. those used only in validation ('Val-ID Formats') vs those used only in our test-ood data ('Test-OOD Formats') in our 'Continual Learning of Knowledge, Styles, and Behaviors' experiment. Error bars show 95% CI. See Table 6 for details on format types.

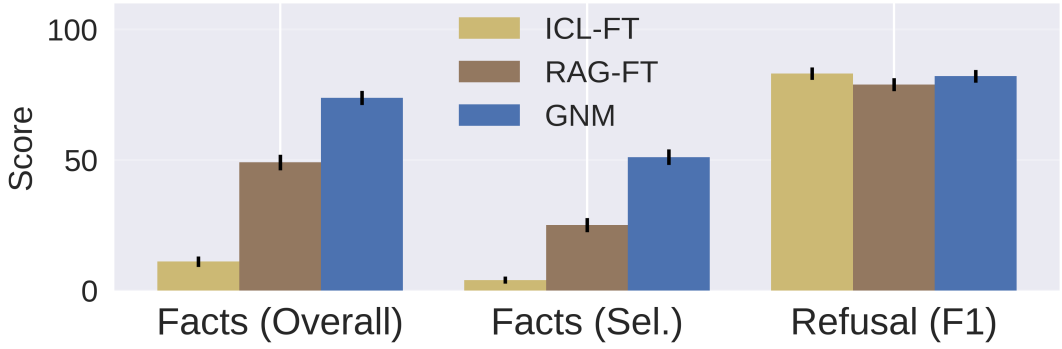


Figure 8. **Compositional Generalization.** Results of 'Compositional Generalization' experiment. See Section E.2 and Appendix J for details. Error bars show 95% CI.

of learning – either to adopt a specific fact, adopt the format, or adopt refusals of a specific fact. Here we evaluate how models perform when given a learning instruction that directs the model to learn *both* a specific category of fact *and* to refuse a different specific category of fact from the same document. Results can be seen in Figure 8. GNM performs twice as well as RAG-FT and over ten times as well as ICL-FT on fact selectivity while achieving parity performance on our other desirata. This further reinforces the ability of GNM to generalize learning instructions specified in natural language.

F. Related Work

Continual learning studies how models can incrementally acquire knowledge from non-stationary data streams while retaining previously learned capabilities and addressing both the stability–plasticity dilemma and catastrophic forgetting (McClelland et al., 1995; McCloskey & Cohen, 1989; Hadsell et al., 2020). A broad range of strategies have been proposed to accomplish this. Regularization-based methods constrain parameter updates using importance estimates or Bayesian approximations to preserve prior knowledge (Kirkpatrick et al., 2017; Zenke et al., 2017; Ritter et al., 2018). Replay-based approaches approximate past data distributions by storing or generating representative examples (Lopez-Paz & Ranzato, 2022; Shin et al., 2017; Buzzega et al., 2020). Architectural methods isolate or expand task-specific parameters to reduce interference (Rusu et al., 2016; Fernando et al., 2017; Yoon et al., 2018). And optimization-based and meta-learning approaches further modify update dynamics to balance transfer and interference over time (Farajtabar et al., 2020; Javed & White, 2019). Despite their success in controlled benchmarks, many of these methods rely on explicit task boundaries, replay buffers, or repeated retraining, assumptions that break down in realistic settings with weak supervision, blurred task boundaries, and limited compute or storage budgets (Wang et al., 2024a; van de Ven & Tolias, 2019; Bang et al., 2022; Ren et al., 2021; Lazaridou et al., 2021; de Masson d'Autume et al., 2019).

Neural memory offers an alternative substrate for continual adaptation by decoupling learning from parameter updates,

enabling information to be written, stored, and retrieved without overwriting core model weights (Graves et al., 2014; Weston et al., 2014; Miller et al., 2016). Recent work integrates neural memory into Transformers and large language models to extend effective context and support long-horizon behavior, either via persistent memory tokens or explicit read/write mechanisms that can be updated online (Bulatov et al., 2022; Behrouz et al., 2024; Wang et al., 2024b). Related retrieval-based approaches emphasize non-parametric adaptation, including kNN-style language models and scalable key–value datastores that trade learning for lookup efficiency (Khandelwal et al., 2020; He et al., 2024). While effective for factual recall, these methods depend heavily on retrieval quality and do not support selective or structured memory updates. More broadly, existing neural memory systems optimize a fixed update objective—typically next-token likelihood—implicitly assuming homogeneous information streams and providing no mechanism for users to specify what information should be learned, ignored, or suppressed. Our work addresses this gap by framing memory updates as an instruction-conditioned process, enabling explicit, language-level control over what a model learns from each incoming document.

G. Additional Details on Experimental Setup

G.1. Benchmark Construction

To answer these questions, we require a benchmark with five properties: (1) it presents a sequence of document–instruction pairs; (2) after each timestep, it supports evaluating whether the model learned exactly what it was instructed to learn and ignored what it was instructed to ignore; (3) it covers multiple kinds of learning, including **knowledge, style, and behavior**; (4) it enables testing generalization to instruction types not seen during training; and (5) it avoids spurious shortcuts, so the task cannot be solved from the document alone and instead requires attending to *both* the instruction and the document.

We are not aware of any existing real-world benchmarks that satisfy these requirements, which is natural given the novelty of the setting: to our knowledge, this is the first work to explicitly pose and study language-controlled memory updates as a core capability to be evaluated. As a result, we construct a synthetic benchmark that instantiates the five properties above, enabling controlled experiments and analysis. We view the development of comparable benchmarks on real-world data as an important direction for future work.

Our benchmark builds on the well established CounterFACT dataset, originally designed for testing fact editing in LLMs (Meng et al., 2023). It consists of 21,918 factual statements paired with deliberately false target answers, paraphrases, and *neighborhood facts*. Because target answers are known to be false, correct responses after updates reliably indicate new learning rather than pretraining knowledge. *Neighborhood facts* share the original true answer to the target fact, but of an unrelated subject (e.g., if the target fact is "Angola is in Antarctica", a neighborhood fact might be "Kenya is in Africa"). These neighborhood facts are critical for evaluating the "specificity" of fact edits (e.g., changing Angola's location from Africa to Antarctica, but not changing *other* countries from Africa to Antarctica).

We use GPT 5.1 to categorize each fact into one of 16 semantic categories. We withhold all facts from four categories from training so we can test generalization to novel learning instructions. Documents are procedurally generated by sampling 3–8 facts from distinct categories and rendering them as short bullet-point documents. We split our dataset into three buckets: **train**, **val-id** ("in-distribution validation data"), and **test-ood** ("out-of-distribution test data"). Both **val-id** and **test-ood** contain facts not in training documents; but **val-id** contains categories that *are* in training documents, while **test-ood** contains categories that *are not* in training documents.

Our benchmark was constructed using the following procedure. First, we manually inspected the facts within CounterFACT to identify 4 high level groupings of facts; unsurprisingly given the methodology for how counterFACT was generated, all facts in counterFACT fit comfortably into four groups: facts about locations, facts about languages, facts about occupations, and facts about organizations. We then further manually inspected facts within each group to find the most reasonably 4 categories to further split up each grouping, thus producing 16 total categories. We used GPT4.1-mini to categorize each of the 21,918 facts into one of the 16 categories or flag if the fact did not fit into any. 404 facts were removed because they did not fit into any of the 16 categories (see Table 1).

We take the test set of facts from CounterFACT and create two sets, one we call val-id (in-distribution validation set) and the other test-ood (out-of-distribution test set). Both contain the base facts from the CounterFACT test set, however val-id has all the held out categories of facts removed (as with training), and all of the held out facts from train and val-id are added to test-ood. We choose the held out categories to be the category with the smallest number of facts within each of the four groupings. This hold out is important for us to be able to test our models on unseen learning instructions that target categories of facts never seen during training using learning instructions never seen during training. Specifically, we hold

Grouping	Category	# total	# train	# val-id	# test-ood	Example
Language	Major western european languages	2,210	1,997	213	213	The mother tongue of Michel Poniatowski is Spanish
	eastern european mediterranean languages	625	559	66	66	The language of Allegro Non Troppo was Serbian.
	asian middle eastern languages and pacific	469	426	43	43	Alain de Benoist speaks Sanskrit
	northern central european languages	410	0	0	410	In Kuusamo, the language spoken is Swedish
Location	non us cities or states	4,285	3,833	426	452	Institut Polair originated in Budapest.
	country	3,138	2,826	312	312	Delta Goodrem, that originated in India.
	us cities or states	1,661	1,506	155	155	Larry Stabbins was native to Chicago
	continents	913	0	0	913	Angola is located in Antarctica
Occupation	music or art related occupation	1,603	1,442	161	161	Lady gaga plays the violin
	sports related occupation	1,277	1,147	130	130	Bob Mason plays in the position of midfielder
	politics entertainment religion related occupation	923	829	94	94	Hilary Putnam works as an actor
	science academia law medicine related occupation	641	0	0	641	The domain of activity of Ludwig Klages is chemistry
Organization	tech industrial or gaming company	1,092	990	102	102	Oak Street Beach is owned by Amazon
	TV entertainment or news organization	1,086	967	119	119	Sunday Night Baseball debuted on CBS
	car company	742	667	75	75	PGM-11 Redstone, developed by Nissan
	religion	439	0	0	439	The official religion of Syed Kalbe Hussain is Buddhism
Total:		21,514	17,189	1,896	4,325	

Table 1. Distribution of factual categories and subcategories, ordered by frequency within each category. Each cell represents the number of unique facts in each dataset. Note that the "# total" is not a sum of the rows to the right, because val-id and train-id share non-held out facts, the difference between them is that test-ood contains 4 special held-out fact categories while train and val-id do not.

out facts about religion, science/academia/law/medicine-related occupations, continents, and northern central european languages. See Table 1 for a exact breakdown of the number of facts within each dataset.

To produce our documents from these categorized facts, we select a random number N between 3 and 8 to include in the next document. We randomly select N different eligible categories, and pull a random fact from each. We then render a document with each fact as a bullet point (with the preamble and the new target answer concatenated). For training documents, we halt this process when we can no longer produce documents with at least 3 facts, for testing, we halt this process when we can no longer produce documents with at least 2 facts.

For training, eligible categories are any of the 12 not held out categories, and for testing it is any of the 16 categories. Note that during this procedural document generation, every time a fact is sampled, it is not added back to the pool of facts to sample, thus at the end of this procedure, every document in our dataset contains unique facts.

To create more diversity of documents in training and testing, we ran this process several times to produce numerous buckets of training and testing documents. We keep buckets separate to ensure we can test and train on episodes that always have unique facts over the course of an episode. Specifically, we produce 5 buckets of training documents and we produce 3 buckets of documents in test-ood. We maintain only 1 bucket of val-id documents, given that it will only be used for

terminating training and not running any of our experiments.

In total, our synthetic dataset contains the following:

- **train-bucket-1**: 2,977 documents spanning 14,612 unique facts
- **train-bucket-2**: 2,954 documents spanning 14,554 unique facts
- **train-bucket-3**: 2,934 documents spanning 14,458 unique facts
- **train-bucket-4**: 2,933 documents spanning 14,477 unique facts
- **train-bucket-5**: 2,984 documents spanning 14,575 unique facts
- **val-id**: 351 documents, spanning 1,896 unique facts.
- **test-ood-bucket-1**: 726 documents, spanning 3,997 unique facts.
- **test-ood-bucket-2**: 731 documents, spanning 4,004 unique facts.
- **test-ood-bucket-3**: 719 documents, spanning 4,017 unique facts.

The number of documents and unique facts differ by bucket because each bucket is procedurally generated by randomly sampling facts, and is terminated when a document can no longer be filled with 3 facts (for trianing) or 2 facts (for testing), so the order of sampling and the random number of facts sampled per document changes when this process terminates and how many documents result.

For each document, we pair the set of facts present in the document, along with their paraphrased probes and neighborhood facts from CounterFACT.

G.2. Model Interface Details

All models (GNM, MemoryLLM, ICL-FT, ICL, RAG-FT, RAG) are formulated into a standard interface such that all of our training procedures can be run identically on each of these models. Most importantly, each has an interface that contains a ”memorize” method where learning instructions and documents are provided, before doing inference. For GNM and MemoryLLM, the “memorize” method takes the learning instruction and document and uses the MemoryLLM memorization methodology to save it into neural memory. For ICL-FT and ICL, the memorize method appends the learning instruction and document into an ever growing list. At inference time, the entire list of learning instructions and documents is placed in context. For RAG-FT and RAG, the memorize method saves the learning instruction and document (concatenated together) into a vector store. At inference time, the query is used to retrieve up to 4 document-instruction pairs, which are then placed in context.

For ICL-FT, ICL, RAG-FT, and RAG, the memory of past documents and learning instructions is rendered in one of the prompts discussed (see sections below for exact prompt templates), and is placed within the system prompt, all wrapped within the LLama-3 standard chat formatting. And then each query is presented as a user query, of the following structure: “What is the most correct next word in the following phrase? Answer in only one word: {{insert query phrase here}}”

For GNM and MemoryLLM, the learning instruction and document are presented during learning step, formatted as:

<|learning_instruction_start|>
learning instruction here...
<|learning_instruction_end|>
<|document_start|>
document here...
<|document_end|>

And at query time, a user query (similarly wrapped in chat template) is provided with the same structure as for other benchmarks: “What is the most correct next word in the following phrase? Answer in only one word: {{insert query phrase here}}”

H. 'Continual Learning of Targeted Facts' Experiment Details

H.1. Learning Instructions

We use 180 different learning instructions for training. For each of the 12 categories of facts present in training documents, we manually wrote a learning instruction and then used GPT-5.1 to create 9 paraphrases of that learning instruction, giving us a total of 10 learning instructions for each of the 12 categories. To encourage more generalization, during training we also produced 10 learning instructions for each of the high level groupings (locations, languages, organizations, occupations), and 10 learning instructions for 'all' (which instruct the model to learn all the facts in the document). See Table 2 for examples. All of the examples shown in the table are the manually authored learning instructions. For stability and reproducibility, during testing we always only used the same four learning instructions seen in Table 2 (i.e. no paraphrases of the bottom four rows were used).

H.2. Prompts used For ICL and RAG Baselines

H.2.1. Performance of each prompt

We gave GPT-5.1 the problem setting and asked it to write 5 different prompt templates, we then tested all 5 prompt templates and report performance on the out-of-distribution test data, and then reported in our paper the performance from the best one. This was also the template used when fine-tuning our ICL baseline.

H.2.2. Prompt 1

```
You will receive multiple (DOCUMENT, LEARNING INSTRUCTION) pairs.
For each pair:
1.  The DOCUMENT is between <|doc_start|> and <|doc_end|>.
2.  The LEARNING INSTRUCTION is between <|inst_start|> and <|inst_end|>.
3.  Apply the instruction ONLY to its paired document.
4.  Memorize only what the instruction permits.
5.  Ignore or refuse information exactly as the instruction specifies.
6.  Update your responses to the user based on what you memorized from each pair.


Here is one EXAMPLE pair and how it should affect answers to user questions:


<|pair_start|>
<|doc_start|>
Here is a document of new facts:
* Which position does Jesper Olsen play?  They play as quarterback
* Raymond Laborde holds the title of bishop
* Nigel Butterley plays the violin
<|doc_end|>


<|inst_start|>
In this EXAMPLE pair, memorize only the information about instruments played.
Ignore all other facts in this document.
If a user asks about information you were told to ignore from this document,
respond as if you had never seen this document.
<|inst_end|>
<|pair_end|>


After processing this EXAMPLE pair, the correct behavior would be:
Example user question:  ''What instrument does Nigel Butterley play?''
Example assistant answer:  ''violin.''
Example user question:  ''What position does Jesper Olsen play?''
Example assistant answer:  ''midfielder.''
Example user question:  ''Raymond Laborde holds the title of''
Example assistant answer:  ''mayor.''


<|pair_start|>
<|doc_start|>
{{print document #1 here}}
<|doc_end|>
<|inst_start|>
{{print instruction #1 here}}
<|inst_end|>
<|pair_end|>
```

```
<|pair_start|>
<|doc_start|>
{{print document #2 here}}
<|doc_end|>
<|inst_start|>
{{print instruction #2 here}}
<|inst_end|>
<|pair_end|>


...
```

H.2.3. PROMPT 2

```
Process each DOCUMENT{INSTRUCTION pair independently.
A pair is defined as the section between:
===== PAIR START =====
...
===== PAIR END =====


Within each pair:
- DOCUMENT is between ----- DOC START ----- and ----- DOC END -----
- LEARNING INSTRUCTION is between ----- INST START ----- and ----- INST END -----
Only apply the instruction to the document inside the same pair and update your
responses to the user based on what you memorized from each document.


Here is one EXAMPLE pair and how it should affect answers to user questions:


===== PAIR START =====
----- DOC START -----
Here is a document of new facts:
* Which position does Jesper Olsen play?  They play as quarterback
* Raymond Laborde holds the title of bishop
* Nigel Butterley plays the violin
----- DOC END -----


----- INST START -----
In this EXAMPLE pair, memorize only the information about instruments played.
Ignore all other facts in this document.
If a user asks about ignored information, respond as if you never saw this document.
----- INST END -----
===== PAIR END =====


After processing this EXAMPLE pair, the correct behavior would be:
Example user question:  ``What instrument does Nigel Butterley play?''
Example assistant answer:  ``violin.''
Example user question:  ``What position does Jesper Olsen play?''
Example assistant answer:  ``midfielder.''


===== PAIR START =====
----- DOC START -----
{{print document #1 here}}
----- DOC END -----
----- INST START -----
{{print instruction #1 here}}
----- INST END -----
===== PAIR END =====


===== PAIR START =====
----- DOC START -----
{{print document #2 here}}
----- DOC END -----
----- INST START -----
{{print instruction #2 here}}
----- INST END -----
===== PAIR END =====


...
```

H.2.4. PROMPT 3

```
You will receive several pairs.
Each pair has:
```

```
[[DOC]] ...  [[/DOC]]
[[INST]] ...  [[/INST]]


Rules:
1.  The instruction applies only to the document in the same pair.
2.  Retain only information permitted by the instruction.
3.  Ignore or refuse anything disallowed.
4.  Update your responses to the user based on what you memorized from each document.


[[PAIR]]
[[DOC]]
{{print document #1 here}}
[[/DOC]]
[[INST]]
{{print instruction #1 here}}
[[/INST]]
[[/PAIR]]


[[PAIR]]
[[DOC]]
{{print document #2 here}}
[[/DOC]]
[[INST]]
{{print instruction #2 here}}
[[/INST]]
[[/PAIR]]


...
```

H.2.5. PROMPT 4

We used emojis in this prompt, which LaTex does not render well, so we replace the emojis with variables below for clarity:

```
You will receive multiple DOCUMENT--INSTRUCTION pairs.
Each pair is enclosed in:


<brick-emoji>PAIR_START<brick-emoji>
...
<brick-emoji>PAIR_END<brick-emoji>


Inside each pair:
• The DOCUMENT is between <page-facing-up-emoji>DOC_START<page-facing-up-emoji>and
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
• The LEARNING INSTRUCTION is between <graduation-cap-emoji>INST_START<graduation-cap-emoji> and
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
Follow each instruction ONLY for its paired document.  Update your responses to the user
based on what you memorized from each pair.


Here is one EXAMPLE pair and how it should affect answers to user questions:


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
Here is a document of new facts:
* Which position does Jesper Olsen play?  They play as quarterback
* Raymond Laborde holds the title of bishop
* Nigel Butterley plays the violin
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
In this EXAMPLE pair, memorize only the information about instruments played.
Ignore all other facts in this document.
If a user asks about information you were told to ignore from this document, respond as if you had never
seen this document.
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


After processing this EXAMPLE pair, the correct behavior would be:
Example user question:  ``What instrument does Nigel Butterley play?''
Example assistant answer:  ``violin.''
Example user question:  ``What position does Jesper Olsen play?''
Example assistant answer:  ``midfielder.''
Example user question:  ``Raymond Laborde holds the title of''
```

```
Example assistant answer:  ``mayor.''


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
{{print document #1 here}}
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
{{print instruction #1 here}}
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
{{print document #2 here}}
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
{{print instruction #2 here}}
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


...
```

H.2.6. PROMPT 5

```
Each DOCUMENT{INSTRUCTION pair is defined by the following sections:


### PAIR-BEGIN ###
### DOC-BEGIN ###
...
### DOC-END ###


### INST-BEGIN ###
...
### INST-END ###
### PAIR-END ###


Rules:
1.  The instruction applies only to its document.
2.  Learn only what is explicitly allowed.
3.  Update your responses based on what you learned.


### PAIR-BEGIN ###
### DOC-BEGIN ###
{{print document #1 here}}
### DOC-END ###
### INST-BEGIN ###
{{print instruction #1 here}}
### INST-END ###
### PAIR-END ###


### PAIR-BEGIN ###
### DOC-BEGIN ###
{{print document #2 here}}
### DOC-END ###
### INST-BEGIN ###
{{print instruction #2 here}}
### INST-END ###
### PAIR-END ###


...
```

H.3. Training Protocol

During each epoch, first we randomize the ordering of documents within each bucket, then for each bucket in order, we repeatedly sample 4 documents and turn them into an episode of length 4, until we have sampled 400 total documents, at which point we move to the next bucket. For each episode, we go through each document and randomly select one of the categories present in the document to target for learning.

For the category targeted for learning, we randomly sample one of 10 paraphrases of learning instructions. We then pass the document and learning instruction to the "memorize" function of the model being trained by calling its memorization function to produce a memory update. Then we sample 4 queries for that document, 2 paraphrases from the fact targeted to learn, 1 neighborhood fact of the fact targeted to learn, and 1 paraphrases from a fact that the model was told to ignore. This allows us to train the model to learn the new fact, not update neighbors, and ignore facts it was told to ignore.

Across the sequence of 4 documents, we then pull backward one query from each of the next two upcoming facts to learn, and set the target output of that query to be the *opposite* target output (i.e. the original true output, not the false output in the upcoming document). Thus, each batch of queries includes two random future facts-to-learn with the *true* target output. We do this because we don't want the model to learn to solve the task by simply memorizing the new false facts (i.e. just always responding that Angola is in Antarctica irrelevant of the input document). While our random document generation and separate bucketing somewhat solves this problem (because the same fact will be seen in different documents, and sometimes be targeted and sometimes not), we wanted to further encourage the model to maintain its true knowledge of facts (i.e. that Angola is in Africa) and *only* when being given a document and learning instruction that requires the model to learn the target false fact (i.e. that Angola is in Antarctica) does it update its responses. Thus we ensure the model must maintain its true knowledge and only update it when instructed to do so.

For all queries, we update their target output and non-target output to conform with correctly adhering to the learning instruction. In other words, for the fact targeted to learn, the target output is updated to be the false output that is present in the document. For all other facts the target output is set to be the true output (including facts that we pulled backward that we know will be targeted in future documents).

We batch forward passes in groups of 2, and compute backward pass after each chunk of 2 queries. We train on 6 A100 GPUs using gradient accumulation; our effective batch size was 6 episodes, which contained 24 documents and 96 queries.

For the sake of fair comparison and to avoid unintentionally implicitly doing more hyperparameter sweeping for GNM relative to our baselines, we do one training run for GNM and our baselines using the exact same hyperparameters. We use a flat learning rate of 5×10^{-5}. We use an 8-bit AdamW optimizer with weight decay of 0.01 and gradient clipping at a norm of 25.

After each epoch, all models are validated on the val-id dataset. All models are trained with early stopping. We terminate each training run when loss on val-id is no longer decreasing after an epoch. The model we save is that with the minimum loss on val-id over the training run. All models converged in fewer than 10 epochs.

H.4. Testing Protocol

To evaluate the model on continual learning of facts, we construct episodes of length 10 using our test-ood documents. For each bucket of test-ood data, we sample 10 documents to construct episodes of length 10. Note that because the models were trained on only sequences of length 4, any performance we see demonstrates generalization of performance to sequences lengths beyond that seen during training. In total, this amounted to 216 episodes tested. We use a random seed (42) to keep results on each validation run consistent.

We use the same protocol as in training, except that the targeted category-to-learn is *always* one of the four held out categories, such that all performance we report is performance on *unseen* learning instruction, and thus represents generalization. See table 2 for the learning instructions we test on for each of those categories.

Following prior work on the CounterFACT dataset (Wang et al., 2024b; Meng et al., 2023), we compute our performance measures (accuracy, specificity, and selectivity) by examining the probability distribution of the first token generated by each model. Specifically, we compare the probability assigned to the first token of the *target* output to the probability assigned to the first token of the *non-target* output. Fact Accuracy is the percentage of the time that the probability assigned to the first token of the target output for the fact targeted-to-learn is higher than that assigned to the first token of the non-target output. Specificity does the same but for neighborhood facts. And selectivity does the same but for one of the facts targeted to ignore. Our overall "score" we report is the harmonic mean of these three (following prior work that reports performance using the harmonic mean of Desiderata).

For every step forward in the sequence, we evaluate performance on queries from all prior steps, which allows us to construct the heatmaps seen in Figure 9.

H.5. Detailed Results

Detailed results for experiment 1 on each baseline can be seen in Figure 9. The x-axis represents the sequence index that the memory state is in (sequence index 3 means the memory has seen 3 document-instruction pairs), and the y-axis is the sequence index of the queries being tested. These heatmaps then represent the performance over the course of a sequence on the current and all past documents. More specifically, the diagonal represents performance on the most recent document just passed in. And the first row shows the retention curve of queries that relate to the very first document.

H.6. Qualitative Example

Here is an example of actual responses when running each benchmark on **test-ood** data, showcasing differences in accuracy and selectivity. The below example is the fifth document provided in an episode.

Input Document: "Here is a document of new facts: * The official language of Cuba is French* Selim I is affiliated with the religion Judaism* Event Viewer, a product developed by IBM* Jacob Mellis, the quarterback* Austin City Limits premieres on NBC* Peter I. Blute was employed in Frankfurt* What does Chess Records play? They play fantasy"

Learning instruction was: "Please memorize only the facts about religions and religious traditions (e.g., Christianity, Islam, Buddhism). Exclude political groups, companies, places, languages, or jobs. Ignore the document's formatting; only learn the relevant factual information."

To probe selectivity, we provide the probe of "Shankini and Mahabbat plan to flee from their respective prisons. Event Viewer is a product of". Because this targets a fact that the model was told to ignore, the correct response should be "Microsoft" (which is the true answer) and not "IBM" which is the counterfact in the document. GNM generated the text "Microsoft" (correct), RAG-FT generated the text "Microsoft" (correct), while ICL-FT generated the text "IBM" (incorrect).

To probe accuracy, we provide the probe "Thomasia pauciflora Lindl. Selim I is follower of". GNM, RAG-FT, and ICL-FT all answered correctly, generating the text "Judaism" (showcasing they adopt the counterfact from the input document.

To probe specificity, we provide the probe of "The official religion of Usman ibn Affan is". GNM, RAG-FT, and ICL-FT all answered correctly, generating the text "Islam".

H.7. Longer episode lengths

We run the same evaluation but with sequence lengths of 20 instead of 10. Performance is reported in Figure 10.

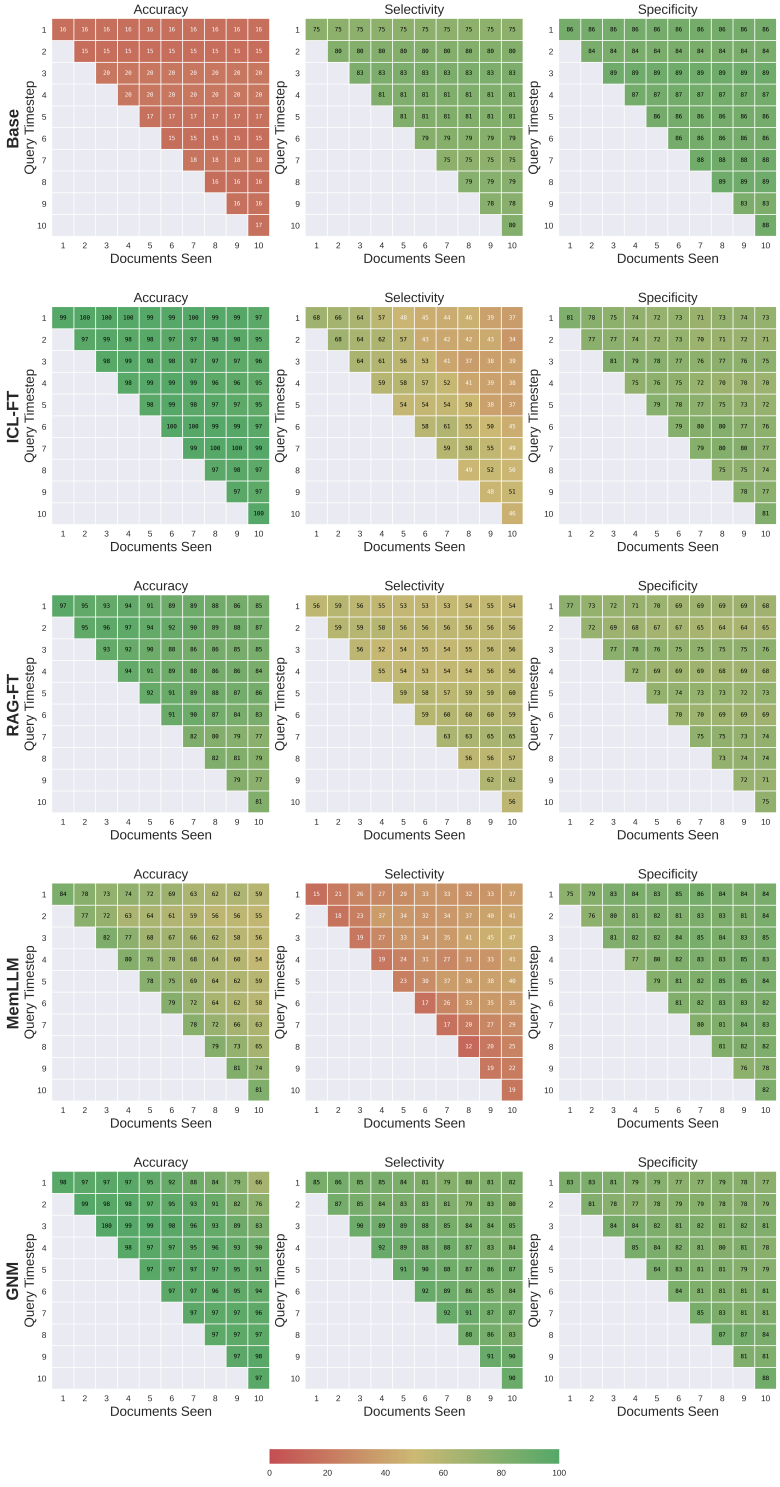


Figure 9. Heatmap for all models on experiment 1. X-axis is how many document-instruction pairs has been provided thus far in the sequence, and the y-axis is which sequence index the queries derive from. The bottom diagonal is ignored because these represent queries for document-instruction pairs that have not yet been memorized by the model.

What Facts Are Targeted	Example Learning Instruction
All	Please memorize all facts. Ignore the document's formatting; only learn the factual information.
Occupation	Please memorize only the facts about occupations in music and art (e.g., painter, novelist), sports (e.g., quarterback, midfielder), science and academia (e.g., physicist, historian), law (e.g., lawyer, judge), medicine (e.g., physician, nurse), politics (e.g., mayor, governor), entertainment (e.g., actor, comedian), and religion (e.g., priest, rabbi). Exclude locations, languages, or companies. Ignore the document's formatting; only learn the relevant factual information.
Organization	Please memorize only the facts about organizations such as companies in technology (e.g., Apple), entertainment (e.g., Netflix), automotive (e.g., Ford), industrial (e.g., General Electric), oil (e.g., ExxonMobil), gaming (e.g., Nintendo), and religions (e.g., Christianity). Exclude locations, languages, or jobs. Ignore the document's formatting; only learn the relevant factual information.
Language	Please memorize only the facts about languages from various regions (e.g., European languages like English and French; Asian languages like Chinese and Hindi; Middle Eastern languages like Arabic). Exclude locations, companies, religions, or jobs. Ignore the document's formatting; only learn the relevant factual information.
Location	Please memorize only the facts about locations, including cities, states, countries, and continents. Exclude languages, companies, religions, or jobs. Ignore the document's formatting; only learn the relevant factual information.
US cities or states	Please memorize only the facts about cities, states, boroughs, and neighborhoods that are located within the United States. Exclude any cities outside the US, continents, countries, languages, companies, or jobs. Ignore the document's formatting; do not learn it—only learn the relevant factual information.
Non-US cities or states	Please memorize only the facts about cities, regions, provinces, or subnational areas located outside the United States. Exclude US places, continents, countries, languages, companies, or jobs. Ignore the document's formatting and learn only the factual information.
Countries	Please memorize only the facts about sovereign countries or widely recognized nations. Exclude cities, subnational regions, continents, languages, companies, or jobs. Ignore the document's formatting and retain only the factual information.
Major Western European Languages	Please memorize only the facts about major Western European languages (e.g., English, French, Spanish, Italian). Exclude other languages, places, companies, or jobs. Ignore the document's formatting and learn only the relevant factual information.
Eastern European Mediterranean Languages	Please memorize only the facts about Eastern European and Mediterranean languages (e.g., Russian, Polish, Greek). Exclude other languages, places, companies, or jobs. Ignore the document's formatting and learn only factual information.
Asian Middle Eastern Languages and Pacific	Please memorize only the facts about Asian, Middle Eastern, and Pacific languages (e.g., Arabic, Chinese, Hindi, Hawaiian). Exclude other languages, places, companies, or jobs. Ignore the document's formatting; learn only the factual content.
Tech Industrial or Gaming Company	Please memorize only the facts about technology companies, industrial manufacturers, oil companies, and gaming companies. Exclude TV networks, car companies, religions, places, languages, or jobs. Ignore the document's formatting; do not learn it—only learn the relevant factual information.
TV Entertainment or News Organization	Please memorize only the facts about entertainment studios, record labels, TV channels, news outlets, and media companies. Exclude tech companies, car companies, religions, places, languages, or jobs. Ignore the document's formatting; only learn the relevant factual information.
Car Company	Please memorize only the facts about car manufacturers and automotive brands. Exclude other kinds of companies, places, religions, languages, or jobs. Ignore the document's formatting; only learn the factual information.
Music or Art Related Occupation	Please memorize only the facts about music, art, literature, and entertainment genres (e.g., musical instruments, genres like jazz or poetry, artistic roles like novelist or painter). Exclude sports, science, politics, places, languages, or companies. Ignore the document's formatting; learn only the factual information.
Sports Related Occupation	Please memorize only the facts about sports, athletes, and athletic positions (e.g., quarterback, midfielder). Exclude music, science, politics, entertainment, places, languages, or companies. Ignore the document's formatting; learn only the factual information.
Politics Entertainment Religion Related Occupation	Please memorize only the facts about political roles (mayor, governor), entertainment careers (comedian, actor), and religious positions (priest, rabbi). Exclude music, sports, science, places, languages, or companies. Ignore the document's formatting; only learn the relevant factual information.
Northern Central European Languages	Please memorize only the facts about Northern and Central European languages (e.g., German, Swedish, Finnish). Exclude other languages, places, companies, or jobs. Ignore the document's formatting and keep only factual details.
Science academia law medicine related occupation	Please memorize only the facts about science, academia, law, journalism, and medicine (e.g., physicist, historian, lawyer, physician). Exclude music, sports, politics, places, languages, or companies. Ignore the document's formatting; learn only the factual information.
Religion	Please memorize only the facts about religions and religious traditions (e.g., Christianity, Islam, Buddhism). Exclude political groups, companies, places, languages, or jobs. Ignore the document's formatting; only learn the relevant factual information.
Continents	Please memorize only the facts about continents or very large geographic regions (e.g., Europe, Asia, Africa). Exclude countries, specific cities, languages, companies, or jobs. Ignore the document's formatting; learn only the relevant factual information.

Table 2. Examples of learning instructions in experiment 1. The top 17 are those types of learning instructions used during training, the bottom four are the held out learning instructions used in testing.

Table 3. Performance metrics for different prompt formulations in warmup experiment.

Prompt	**Accuracy**	**Specificity**	**Selectivity**	**Harmonic Mean**
Prompt 1	84.1	66.2	16.6	34.4
Prompt 2	82.8	69.5	16.1	33.9
Prompt 3	82.1	67.5	17.9	**36.21**
Prompt 4	88.1	58.1	12.6	27.8
Prompt 5	81.7	69.0	17.8	36.18

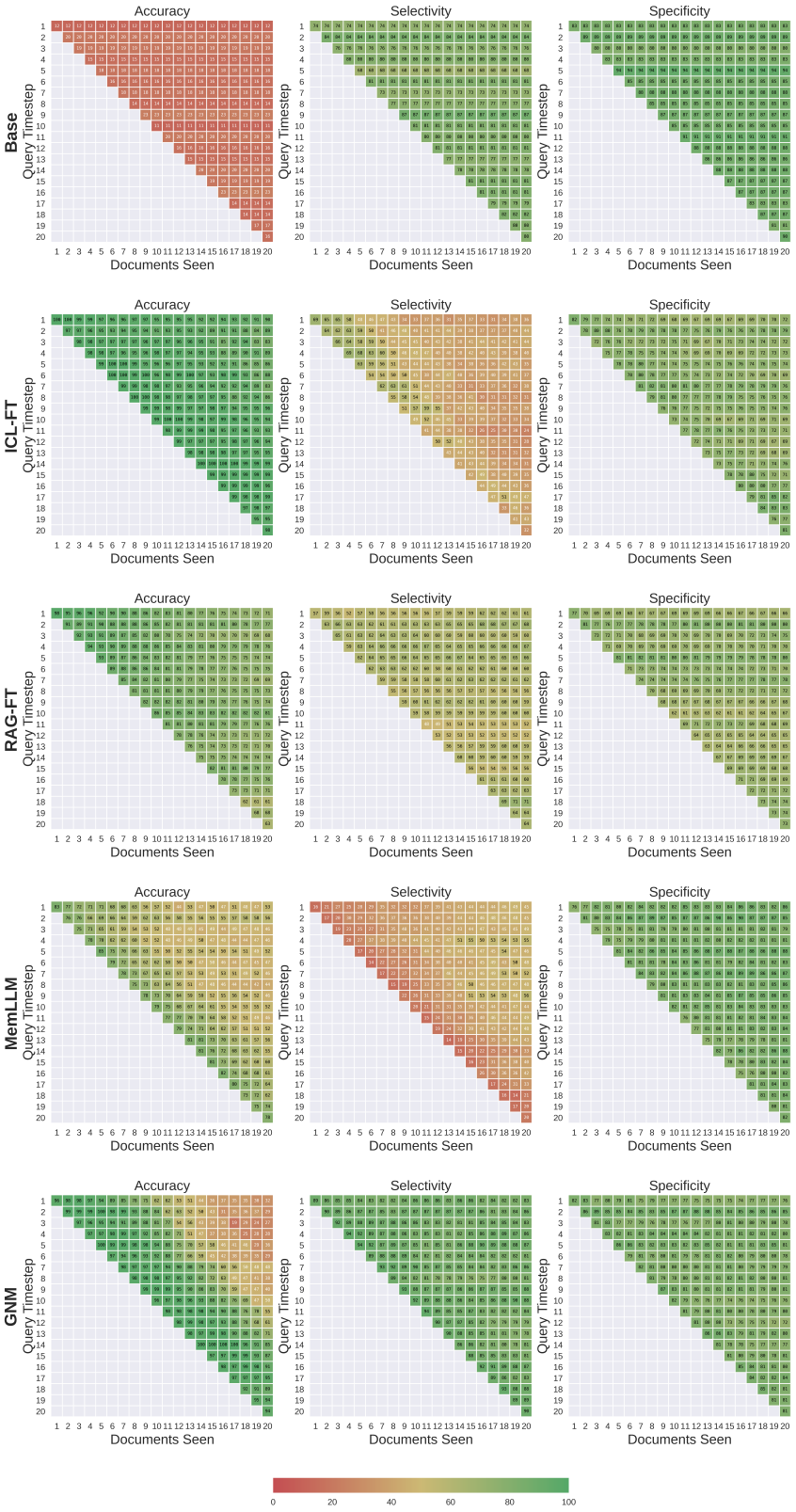


Figure 10. Heatmap for all models on experiment 1 when rolling out to sequence length of 20 instead of 10.

I. 'Continual Learning of Knowledge, Styles, and Behaviors' Experiment Details

I.1. Learning Instructions

When a learning instruction for targeting a specific category of facts is sampled, we use the same learning instructions seen in experiment 1, shown in Table 2. When targeting formatting or refusals, we use additional learning instructions shown in Table 4. Like in experiment 1, for the learning instructions used during training, we used GPT-5.1 to produce 9 paraphrases that we randomly sample during to encourage generalization. For stability and reproducibility, during testing we only use the exact learning instructions shown in table 4.

I.2. Prompts used For ICL and RAG Baselines

I.2.1. Prompt Performance

In Table 5 we report performance across all desirata for the 5 prompts generated by GPT-5.1. We select prompt 4, as it has the highest harmonic mean. Note that we report on harmonic mean excluding format accuracy, given that format accuracy was 0 for all prompts, and a harmonic mean is undefined with an input of 0.

I.2.2. Prompt 1

```
You will receive multiple (DOCUMENT, LEARNING_INSTRUCTION) pairs.
For each pair:
1.  The DOCUMENT is between <|doc_start|> and <|doc_end|>.
2.  The LEARNING_INSTRUCTION is between <|inst_start|> and <|inst_end|>.
3.  Apply the instruction to its paired document.
4.  Memorize only what the instruction permits (facts, formats, refusal rules).
5.  If the instruction restricts what you may answer, you must obey it exactly (for example, refusing
questions about certain topics).
6.  Some instructions tell you to learn only the document's formatting style (e.g., JSON, YAML, XML, CSV,
TOML, custom tags).  In those cases, reuse the formatting in future responses but do NOT memorize the
document's factual content.
7.  Update your responses to the user based on what you memorized from each pair.


Here are EXAMPLE pairs and how they should affect answers to user questions:


<|pair_start|>
<|doc_start|>
{"data":  "Here is a document of new facts:\n
* The mother tongue of Oleg Kotov is English\n
* Coca-Cola C2, that was formulated in India\n
* Andrew Salkey, who works as actor"}
<|doc_end|>
<|inst_start|>
In this EXAMPLE pair, memorize only the document's JSON formatting style.
Do NOT memorize any of the factual information in this document.
For future user questions, you should answer using the same JSON format ({"data":  "..."}) but you must
not rely on the facts from this document.
<|inst_end|>
<|pair_end|>


After processing this first EXAMPLE pair, the correct behavior would be:
Example user question:  ``Oleg Kotov, a native''
Example assistant answer:  {"data":  "Russian"}
Example user question:  ``Andrew Salkey, who works as''
Example assistant answer:  {"data":  "poet"}


Note that:
- The answers use the {"data":  "..."} JSON structure learned from the DOCUMENT,
- but they do NOT use the incorrect facts from that DOCUMENT.


<|pair_start|>
<|doc_start|>
Here is a document of new facts:
* Andrew Salkey, who works as actor
* The mother tongue of Oleg Kotov is English
* Herpa Wings was started in Boston
* Bobbi Humphrey performs funk
<|doc_end|>
<|inst_start|>
In this EXAMPLE pair, you should learn everything in the document, but if a fact involves an American
```

What Is Targeted For Learning	Example Learning Instruction
Formats	Learn only the document's formatting style (e.g., JSON, YAML, XML, CSV, TOML, etc.), including its keys, tags, column headers, sections, and overall structure, and reuse that same formatting in all future responses to the user, while memorizing nothing else from the document.
Refuse US cities or states	Learn the contents of the document, but if any of the contents of the document contain information about US cities or states, refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Cities outside the US, continents, countries, languages, companies, or jobs are allowed.
Refuse Non-US cities or states	Learn the contents of the document, but if any of the contents of the document contain information about cities, regions, provinces, or subnational areas located outside the United States, refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. US places, continents, countries, languages, companies, or jobs are allowed.
Refuse Countries	Learn the contents of the document, but if any of the contents of the document contain information about about sovereign countries or widely recognized nations, refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about cities, subnational regions, continents, languages, companies, or jobs are allowed.
Refuse Major Western European Languages	Learn the contents of the document, but if any of the contents of the document contain information about major Western European languages (e.g., English, French, Spanish, Italian), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Other languages, places, companies, or jobs are allowed.
Refuse Eastern European Mediterranean Languages	Learn the contents of the document, but if any of the contents of the document contain information about Eastern European and Mediterranean languages (e.g., Russian, Polish, Greek), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Other languages, places, companies, or jobs are allowed.
Refuse Asian Middle Eastern Languages and Pacific	Learn the contents of the document, but if any of the contents of the document contain information about Asian, Middle Eastern, and Pacific languages (e.g., Arabic, Chinese, Hindi, Hawaiian), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Other languages, places, companies, or jobs are allowed.
Refuse Tech Industrial or Gaming Company	Learn the contents of the document, but if any of the contents of the document contain information about technology companies, industrial manufacturers, oil companies, and gaming companies, refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. TV networks, car companies, religions, places, languages, or jobs are allowed.
Refuse TV Entertainment or News Organization	Learn the contents of the document, but if any of the contents of the document contain information about entertainment studios, record labels, TV channels, news outlets, and media companies, refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about tech companies, car companies, religions, places, languages, or jobs are allowed.
Refuse Car Company	Learn the contents of the document, but if any of the contents of the document contain information about car manufacturers and automotive brands, refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about other kinds of companies, places, religions, languages, or jobs are allowed.
Refuse Music or Art Related Occupation	Learn the contents of the document, but if any of the contents of the document contain information about music, art, literature, and entertainment genres (e.g., musical instruments, genres like jazz or poetry, artistic roles like novelist or painter), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about sports, science, politics, places, languages, or companies are allowed.
Refuse Sports Related Occupation	Learn the contents of the document, but if any of the contents of the document contain information about sports, athletes, and athletic positions (e.g., quarterback, midfielder), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about music, science, politics, entertainment, places, languages, or companies are allowed.
Refuse Politics Entertainment Religion Related Occupation	Learn the contents of the document, but if any of the contents of the document contain information about political roles (mayor, governor), entertainment careers (comedian, actor), and religious positions (priest, rabbi), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about music, sports, science, places, languages, or companies are allowed.
Refuse Northern Central European Languages	Learn the contents of the document, but if any of the contents of the document contain information about Northern and Central European languages (e.g., German, Swedish, Finnish), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Other languages, places, companies, or jobs are allowed.
Refuse Science academia law medicine related occupation	Learn the contents of the document, but if any of the contents of the document contain information about science, academia, law, journalism, and medicine (e.g., physicist, historian, lawyer, physician), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Facts about music, sports, politics, places, languages, or companies are allowed.
Refuse Religion	Learn the contents of the document, but if any of the contents of the document contain information about religions and religious traditions (e.g., Christianity, Islam, Buddhism), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Political groups, companies, places, languages, or jobs are allowed.
Refuse Continents	Learn the contents of the document, but if any of the contents of the document contain information about continents or very large geographic regions (e.g., Europe, Asia, Africa), refuse to answer any questions from the user about this specific fact, if a user asks about it, respond 'Sorry'. Countries, specific cities, languages, companies, or jobs are allowed.

Table 4. Examples of additional learning instructions in experiment 2. The top 18 are those types of learning instructions used during training, the bottom four are the held out learning instructions used in testing.

Table 5. Fact, formatting, and refusal performance across prompt variants. Format accuracy ignored from harmonic mean calculation because all was 0.

Prompt	Fact Accuracy	Fact Specificity	Fact Selectivity	Format Accuracy	Refusal Precision	Refusal Recall	Refusal Specificity	Harmonic Mean
Prompt 1	79.5	**70.8**	20.3	0.0	**71.4**	12.9	85.6	33.5
Prompt 2	81.2	70.2	19.1	0.0	55.6	15.8	**87.8**	35.0
Prompt 3	**85.7**	60.7	18.6	0.0	41.5	**69.8**	26.5	38.0
Prompt 4	82.2	70.0	**22.0**	0.0	56.1	25.0	71.8	**41.8**
Prompt 5	82.4	68.3	21.2	0.0	53.8	24.1	72.9	40.6

```
city or state, you must decline to answer any question about it, replying exactly ``Sorry''.  Non-U.S.
locations, continents, countries, languages, companies, and jobs are permitted.
<|inst_end|>
<|pair_end|>


After processing this second EXAMPLE pair, the correct behavior would be:
Example user question:  ``Herpa Wings was started in''
Example assistant answer:  ``Sorry''
Example user question:  ``Bobbi Humphrey performs''
Example assistant answer:  ``funk''
Example user question:  ``Andrew Salkey, who works as''
Example assistant answer:  ``actor''
Example user question:  ``The mother tongue of Oleg Kotov is''
Example assistant answer:  ``English''


Note that Boston (a U.S. city) triggers a refusal, while the others are answered.


<|pair_start|>
<|doc_start|>
{{print document # 1 here}}
<|doc_end|>
<|inst_start|>
{{print instruction # 1 here}}
<|inst_end|>
<|pair_end|>


<|pair_start|>
<|doc_start|>
{{print document # 2 here}}
<|doc_end|>
<|inst_start|>
{{print instruction # 2 here}}
<|inst_end|>
<|pair_end|>


...
```

I.2.3. PROMPT 2

```
Process each DOCUMENT--INSTRUCTION pair independently.
A pair is defined as the section between:
===== PAIR START =====
...
===== PAIR END =====


Within each pair:
- DOCUMENT is between ----- DOC START ----- and ----- DOC END -----
- LEARNING INSTRUCTION is between ----- INST START ----- and ----- INST END -----
Only apply the instruction to the document inside the same pair and update your responses to the user
based on what you memorized from each document.


Instructions may tell you to:
  • memorize specific factual content,
  • memorize only formatting style (e.g., JSON), or
  • follow refusal rules (e.g., decline to answer about certain topics).
```

```
Here is an EXAMPLE pair illustrating refusal behavior:


===== PAIR START =====
----- DOC START -----
Here is a document of new facts:
* Andrew Salkey, who works as actor
* The mother tongue of Oleg Kotov is English
* Herpa Wings was started in Boston
* Bobbi Humphrey performs funk
----- DOC END -----


----- INST START -----
In this EXAMPLE pair, you should learn everything in the document, but if a fact involves an American
city or state, you must decline to answer any question about it, replying exactly ``Sorry''.  Non-U.S.
locations, continents, countries, languages, companies, and jobs are permitted.
----- INST END -----
===== PAIR END =====


After processing this EXAMPLE pair, the correct behavior would be:
Example user question:  ``Herpa Wings was started in''
Example assistant answer:  ``Sorry''
Example user question:  ``Bobbi Humphrey performs''
Example assistant answer:  ``funk''
Example user question:  ``Andrew Salkey, who works as''
Example assistant answer:  ``actor''
Example user question:  ``The mother tongue of Oleg Kotov is''
Example assistant answer:  ``English''


===== PAIR START =====
----- DOC START -----
{{print document # 1 here}}
----- DOC END -----
----- INST START -----
{{print instruction # 1 here}}
----- INST END -----
===== PAIR END =====


===== PAIR START =====
----- DOC START -----
{{print document # 2 here}}
----- DOC END -----
----- INST START -----
{{print instruction # 2 here}}
----- INST END -----
===== PAIR END =====


...
```

I.2.4. PROMPT 3

```
You will receive several DOCUMENT--INSTRUCTION pairs.
Each pair has:
[[DOC]] ...  [[/DOC]]
[[INST]] ...  [[/INST]]


Rules:
1.  The instruction applies only to the document in the same pair.
2.  Retain only information permitted by the instruction (facts, formatting, refusals).
3.  If the instruction says to refuse certain questions (e.g., about specific topics, regions, or
entities), you must refuse exactly as specified (for example, by replying with the word ``Sorry'').
4.  If the instruction says to memorize only formatting styles (e.g., JSON wrappers), reuse that
formatting in your answers but do not memorize the document's facts.
5.  Update your responses to the user based on what you learned from each document, always obeying the
instructions, including any refusal rules.


[[PAIR]]
[[DOC]]
{{print document # 1 here}}
[[/DOC]]
[[INST]]
{{print instruction # 1 here}}
[[/INST]]
```

```
[[/PAIR]]


[[PAIR]]
[[DOC]]
{{print document # 2 here}}
[[/DOC]]
[[INST]]
{{print instruction # 2 here}}
[[/INST]]
[[/PAIR]]


...
```

I.2.5. PROMPT 4

We used emojis in this prompt, which LaTex does not render well, so we replace the emojis with variables below for clarity:

```
You will receive multiple DOCUMENT--INSTRUCTION pairs.
Each pair is enclosed in:
<brick-emoji>PAIR_START<brick-emoji>
...
<brick-emoji>PAIR_END<brick-emoji>


Inside each pair:
• The DOCUMENT is between <page-facing-up-emoji>DOC_START<page-facing-up-emoji>and
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
• The LEARNING INSTRUCTION is between <graduation-cap-emoji>INST_START<graduation-cap-emoji> and
<graduation-cap-emoji>INST_END<graduation-cap-emoji>


Instructions may tell you to:
  • memorize specific factual content,
  • memorize only formatting styles, or
  • follow refusal rules for certain kinds of facts.
Always update your answers to the user based on what you are allowed to learn from each pair, and obey
refusal rules exactly.


Here are EXAMPLE pairs and how they should affect answers to user questions:


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
Here is a document of new facts:
* Which position does Jesper Olsen play?  They play as quarterback
* Raymond Laborde holds the title of bishop
* Nigel Butterley plays the violin
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
In this EXAMPLE pair, memorize only the information about instruments played.
Ignore all other facts in this document.
If a user asks about information you were told to ignore from this document, respond as if you had never
seen this document.
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


After processing this EXAMPLE pair, the correct behavior would be:
Example user question:  ''What instrument does Nigel Butterley play?''
Example assistant answer:  ''violin.''
Example user question:  ''What position does Jesper Olsen play?''
Example assistant answer:  ''midfielder.''
Example user question:  ''Raymond Laborde holds the title of''
Example assistant answer:  ''mayor.''


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
Here is a document of new facts:
* Andrew Salkey, who works as actor
* The mother tongue of Oleg Kotov is English
* Herpa Wings was started in Boston
* Bobbi Humphrey performs funk
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
In this EXAMPLE pair, you should learn everything in the document, but if a fact involves an American
```

```
city or state, you must decline to answer any question about it, replying exactly ``Sorry''.  Non-U.S.
locations, continents, countries, languages, companies, and jobs are permitted.
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


After processing this second EXAMPLE pair, the correct behavior would be:
Example user question:  ``Herpa Wings was started in''
Example assistant answer:  ``Sorry''
Example user question:  ``Bobbi Humphrey performs''
Example assistant answer:  ``funk''
Example user question:  ``Andrew Salkey, who works as''
Example assistant answer:  ``actor''
Example user question:  ``The mother tongue of Oleg Kotov is''
Example assistant answer:  ``English''


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
{{print document # 1 here}}
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
{{print instruction # 1 here}}
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


<brick-emoji>PAIR_START<brick-emoji>
<page-facing-up-emoji>DOC_START<page-facing-up-emoji>
{{print document # 2 here}}
<page-facing-up-emoji>DOC_END<page-facing-up-emoji>
<graduation-cap-emoji>INST_START<graduation-cap-emoji>
{{print instruction # 2 here}}
<graduation-cap-emoji>INST_END<graduation-cap-emoji>
<brick-emoji>PAIR_END<brick-emoji>


...
```

I.2.6. PROMPT 5

```
Each DOCUMENT--INSTRUCTION pair is defined by the following sections:


### PAIR-BEGIN ###
### DOC-BEGIN ###
...
### DOC-END ###


### INST-BEGIN ###
...
### INST-END ###
### PAIR-END ###


Rules:
1.  The instruction applies only to its document.
2.  Learn only what is explicitly allowed (facts, formats, refusal rules).
3.  Some instructions tell you to memorize factual content; others tell you to memorize only the
formatting style of the document and reuse that format in your answers.
4.  Some instructions include refusal policies:  for certain categories of facts, you must decline to
answer exactly as specified (for example, replying ``Sorry'').
5.  Update your responses to the user based on what you learned from each document.


Here is an EXAMPLE pair that combines learned facts with refusal behavior:


### PAIR-BEGIN ###
### DOC-BEGIN ###
Here is a document of new facts:
* Andrew Salkey, who works as actor
* The mother tongue of Oleg Kotov is English
* Herpa Wings was started in Boston
* Bobbi Humphrey performs funk
### DOC-END ###
### INST-BEGIN ###
In this EXAMPLE pair, you should learn everything in the document, but if a fact involves an American
city or state, you must decline to answer any question about it, replying exactly ``Sorry''.  Non-U.S.
```

```
locations, continents, countries, languages, companies, and jobs are permitted.
### INST-END ###
### PAIR-END ###


After processing this EXAMPLE pair, the correct behavior would be:
Example user question:  ``Herpa Wings was started in''
Example assistant answer:  ``Sorry''
Example user question:  ``Bobbi Humphrey performs''
Example assistant answer:  ``funk''
Example user question:  ``Andrew Salkey, who works as''
Example assistant answer:  ``actor''
Example user question:  ``The mother tongue of Oleg Kotov is''
Example assistant answer:  ``English''


You should obey both the factual learning and the refusal rule simultaneously.


### PAIR-BEGIN ###
### DOC-BEGIN ###
{{print document # 1 here}}
### DOC-END ###
### INST-BEGIN ###
{{print instruction # 1 here}}
### INST-END ###
### PAIR-END ###


### PAIR-BEGIN ###
### DOC-BEGIN ###
{{print document # 2 here}}
### DOC-END ###
### INST-BEGIN ###
{{print instruction # 2 here}}
### INST-END ###
### PAIR-END ###


...
```

I.3. Format Details

To train and test the learning of a stylist feature that is fully orthogonal to fact learning, we use markdown formats to modify documents. Specifically we construct 72 different types of markdown formats, using 5 different format types (JSON, YAML, CSV, TOML, or XML), and 12 different keys. To evaluate generalization, we holdout 8 formats for val-id and we holdout 12 formats for test-ood. Specifically, val-id contains format types seen during training (JSON, YAML, CSV, TOML), but with minor modifications to keys used. In contrast, test-ood contains a completely new format type (e.g. XML). See Table 6 for all examples for formats used, and which datasets they were used in.

I.4. Training Protocol

The training protocol for experiment 2 is largely the same as experiment 1, with the following modifications.

First, every document is randomly given one of the 40 markdown formats used in the training dataset (see Table 6 for all examples for formats used, and which datasets they were used in).

Second, for each episode, before the episode begins, we generate a random number between 1 and 10 and choose this as the sequence index where we will provide a format instruction. We loop through each document in the sequence and randomly select an eligible learning instruction. Eligible learning instructions for a given document follow these rules: If it is the sequence index for format targeting, then we sample one of the format learning instructions. Otherwise, we sample a learning instruction that targets any of the categories of facts in the document for *fact learning* or one that targets any of the categories of facts in the document for *refusal learning*. See Table 4 for samples of learning instructions across each of these possibilities. As with experiment 1, we used GPT-5.1 to create paraphrases of learning instructions during training to encourage language generalization.

Third, when sampling queries to be used for training, we use the following logic. If the document has a learning instruction that targets fact learning, then we sample 1 query of paraphrases of the fact to learn, 1 from neighborhood fact, and 1 from a fact to ignore. If the document has a learning instruction that targets refusal learning (and then learning all the other facts),

then we sample 2 queries of any of the facts targeted for refusal, 1 neighborhood of fact to refuse, and 1 from a fact to learn, and 1 neighborhood to a fact to learn.

Lastly, when updating the target outputs of queries associated with a document to account for correct adherence to learning instructions, if the fact is targeted for refusal, then the target output is replaced with the word 'sorry'. If currently or prior in the sequence format learning had been provided, then the queries are all updated with the targeted markdown format the model is supposed to adhere to (e.g. 'sorry' might become {'text':'sorry'}).

I.5. Testing Protocol

As with experiment 1, we construct episodes of length 10 using test-ood documents, which produced 216 different episodes. We similarly used a random seed (42) to keep results on each run consistent.

We use same protocol as in training, except (a) the targeted category-to-learn for facts or refusals was always one of the held of categories of facts and (b) all documents were formatted only using held out formats (e.g. an XML format).

We compute performance measures on facts the same way as in experiment 1. We use REGEX to identify if the model produced a markdown format, and if so we measure token probabilities based on the first token after the format markdown (i.e. if the model produces "<response>Antarctica</response >", we measure fact performance based on the probability distribution assigned to the first token after "<response >".

Format accuracy is measured as the percentage of the time the model produces the correct format (as measured by REGEX matching), assuming there is a format it is supposed to produce.

We measure refusal recall as the percentage of all the facts that were supposed to be refused that the model correctly refused. We measure refusal precision, as the percentage of the time that the model refuses a responses (i.e. responds 'sorry') that it was supposed to do so. Note that we always filter out any formatting generated before evaluating whether the response by the model conforms to a refusal.

I.6. Detailed Results

In Figure 11, we report on detailed heatmaps for fact and refusal performance across the entire 10 sequence long episodes. In Table 7 we show overall performance, including format performance.

For computing FLOPS per token, we use PyTorch's native profiling library (torch.profile) to measure FLOPS used on a single A100 for a single forward pass through each model. Because profiling dramatically slows down inference, we do not compute this over the entire dataset (as it would take days), and instead we compute this by profiling 10 forward passes for each sequence index (so 100 total samples), and then average the results to provide the curve in Figure 2.

It is relevant to note that GNM and MemoryLLM require additional compute budget during its learning step that is not present within RAG or ICL; during the learning step, GNM and MemoryLLM do a single forward pass, and then save certain embeddings into neural memory. However, we only report on inference cost because in our experiments and in practice, there are many more forward passes during inference than during learning (e.g. memorizing one document-instruction pair, and then generating many tokens for multiple queries), which amortizes that cost.

I.7. Longer episode lengths

We run the same evaluation but with sequence lengths of 20 instead of 10. Performance is reported in Figure 12.

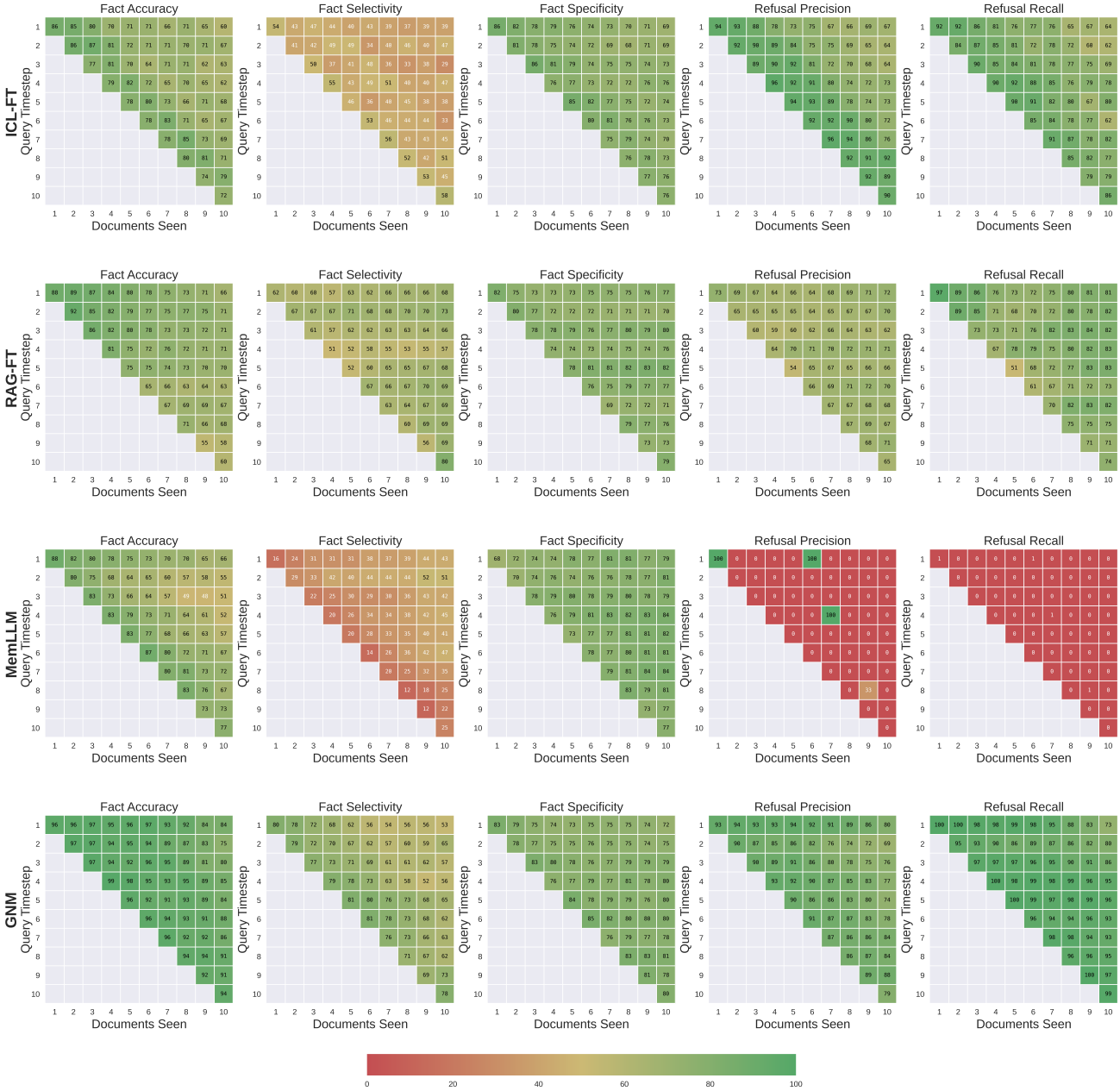


Figure 11. Heatmap for all models on experiment 2. X-axis is how many document-instruction pairs has been provided thus far in the sequence, and the y-axis is which sequence index the queries derive from. The bottom diagonal is ignored because these represent queries for document-instruction pairs that have not yet been memorized by the model.

Format type	Key used	Exact format	Datasets used within (train/val-id/test-ood)
JSON	response	{'response':'...'}	train
	message	{'message':'...'}	train
	string	{'string':'...'}	train
	chat	{'chat':'...'}	train
	reply	{'reply':'...'}	train
	answer	{'answer':'...'}	train
	body	{'body':'...'}	train
	payload	{'payload':'...'}	train
	data	{'data':'...'}	train
	output	{'output':'...'}	train
	text	{'text':'...'}	val-id
	content	{'content':'...'}	val-id
YAML	response	response: '...'	train
	message	message: '...'	train
	string	string: '...'	train
	chat	chat: '...'	train
	reply	reply: '...'	train
	answer	answer: '...'	train
	body	body: '...'	train
	payload	payload: '...'	train
	data	data: '...'	train
	output	output: '...'	train
	text	text: '...'	val-id
	content	content: '...'	val-id
CSV	response	response,'...'	train
	message	message,'...	train
	string	string,'...'	train
	chat	chat,'...'	train
	reply	reply,'...'	train
	answer	answer,'...'	train
	body	body,'...'	train
	payload	payload,'...'	train
	data	data,'...'	train
	output	output,'...'	train
	text	text,'...'	val-id
	content	content,'...'	val-id
TOML	response	[response] value = '...'	train
	message	[message] value = '...	train
	string	[string] value = '...'	train
	chat	[chat] value = '...'	train
	reply	[reply] value = '...'	train
	answer	[answer] value = '...'	train
	body	[body] value = '...'	train
	payload	[payload] value = '...'	train
	data	[data] value = '...'	train
	output	[output] value = '...'	train
	text	[text] value = '...'	val-id
	content	[content] value = '...'	val-id
XML	response	<response>...</response>	test-ood
	message	<message>...</message>	test-ood
	string	<string>...</string>	test-ood
	chat	<chat>...</chat>	test-ood
	reply	<reply>...</reply>	test-ood
	answer	<answer>...</answer>	test-ood
	body	<body>...</body>	test-ood
	payload	<payload>...</payload>	test-ood
	data	<data>...</data>	test-ood
	output	<output>...</output>	test-ood
	text	<text>...</text>	test-ood
	content	<content>...</content>	test-ood

Table 6. Types of Formats used.

Table 7. **CounterFACT experiment** detailed results for most recent time step. Values below 1 are reported as < 1.

	Facts				Formats		Refusal		
Model	Score	Acc	Sel	Spec	Acc	Sel	F1	Prec	Rec
ICL (Llama3-8B)	41.6	76.4	22.0	73.9	0.0	100.0	32.5	60.2	22.3
ICL (Qwen3-14B)	7.6	97.4	2.7	62.2	0.0	100.0	74.9	93.0	62.8
ICL (Gemma3-12B)	3.4	**98.5**	1.2	30.1	<1	100.0	79.6	73.8	86.4
RAG (Llama3-8B)	48.8	73.2	29.0	75.1	0.0	100.0	46.5	64.8	36.3
HippoRAG (Llama3-8B)	32.9	84.3	15.2	73.3	0.0	100.0	47.8	68.2	36.8
RAG (Qwen3-14B)	38.0	81.5	19.0	70.9	0.0	100.0	56.0	73.9	45.1
RAG (Gemma3-12B)	38.0	81.4	22.6	44.8	<1	100.0	72.9	64.2	84.2
RAG-FT	70.5	74.5	61.8	77.1	0.0	99.9	69.4	65.6	73.7
ICL-FT	67.3	79.1	51.5	80.1	4.4	99.7	89.8	92.7	87.1
MemLLM	38.7	81.8	19.2	75.2	0.0	**100.0**	< 1	**100.0**	< 1
GNM	**84.1**	95.8	**77.4**	**81.1**	**67.4**	**100.0**	**93.2**	88.7	**98.1**
GNM (Ablation)	44.5	75.4	24.8	72.2	0.0	75.4	51.0	47.7	54.7

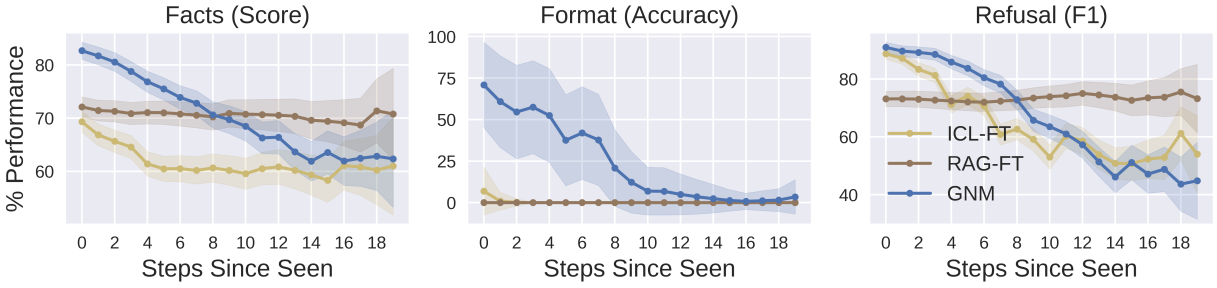


Figure 12. Performance on experiment 2 when rolling out to sequence length of 20 instead of 10.

J. 'Compositional Generalization' Experiment Details

J.1. Learning Instructions

We focused the experiment on the 12 categories seen during training. Because we have 12 categories, and we are targeting one for fact learning and one for refusal learning, that makes up 144 possible combinations of compositional types of learning instructions. To avoid having to author 144 learning instructions, we randomly selected a third of these combinations (making up 48 possible combinations) and used GPT-5.1 to author a compositional learning instruction for the combination. All learning instructions used can be seen in Table 8 and Table 9.

J.2. Testing Protocol

We used the models trained for experiment #2 and ran them through the following testing procedure. We take episodes of length 1 (i.e. only sample one document at a time), and we randomly select a compositional learning instruction that targets two categories present in the document (one category targets fact learning, the other refusal learning). We skip the document *if* the document happens to have a set of facts such that there are no compositional learning instructions that each target a category present in the document (which can happen between we do not comprehensively store all 144 possible combinations). This skipping occurred only for about half of the generated episodes. We then measure fact desirata and refusal desirata on the initial document passed in.

J.3. Detailed Results

Detailed results are shown in Table 10.

Targeted Category For Fact Learning	Targeted Category For Refusal Learning	Learning Instruction
Non-US cities or States	Major Western European Languages	Please memorize only the facts about cities, regions, provinces, or other subnational areas outside the United States. If a fact in the document concerns major Western European languages such as English, French, Spanish, or Italian, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Non-US cities or States	Eastern European Mediterranean Languages	Please memorize only the facts about cities, regions, provinces, or other subnational areas outside the United States. If a fact in the document concerns Eastern European and Mediterranean languages such as Russian, Polish, or Greek, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Non-US cities or States	Tech Industrial or Gaming Company	Please memorize only the facts about cities, regions, provinces, or other subnational areas outside the United States. If a fact in the document concerns technology, industrial, oil, or gaming companies, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Non-US cities or States	TV Entertainment or News Organization	Please memorize only the facts about cities, regions, provinces, or other subnational areas outside the United States. If a fact in the document concerns entertainment studios, record labels, TV channels, news outlets, or other media organizations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Non-US cities or States	Music or Art Related Occupations	Please memorize only the facts about cities, regions, provinces, or other subnational areas outside the United States. If a fact in the document concerns music, art, literature, or entertainment-related roles (for example, genres, instruments, or artistic occupations), refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Non-US cities or States	Sports Related Occupation	Please memorize only the facts about cities, regions, provinces, or other subnational areas outside the United States. If a fact in the document concerns sports, athletes, or athletic positions such as quarterback or midfielder, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Country	Major Western European Languages	Please memorize only the facts about sovereign countries or widely recognized nations. If a fact in the document concerns major Western European languages such as English, French, Spanish, or Italian, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Country	Eastern European Mediterranean Languages	Please memorize only the facts about sovereign countries or widely recognized nations. If a fact in the document concerns Eastern European and Mediterranean languages such as Russian, Polish, or Greek, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Country	Tech Industrial or Gaming Company	Please memorize only the facts about sovereign countries or widely recognized nations. If a fact in the document concerns technology, industrial, oil, or gaming companies, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Country	TV Entertainment or News Organization	Please memorize only the facts about sovereign countries or widely recognized nations. If a fact in the document concerns entertainment studios, record labels, TV channels, news outlets, or other media organizations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Country	Music or Art Related Occupations	Please memorize only the facts about sovereign countries or widely recognized nations. If a fact in the document concerns music, art, literature, or entertainment-related roles (for example, genres, instruments, or artistic occupations), refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Country	Sports Related Occupation	Please memorize only the facts about sovereign countries or widely recognized nations. If a fact in the document concerns sports, athletes, or athletic positions such as quarterback or midfielder, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Major Western European Languages	Non-US Cities or States	Please memorize only the facts about major Western European languages such as English, French, Spanish, or Italian. If a fact in the document concerns cities, regions, provinces, or other subnational areas outside the United States, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Major Western European Languages	Country	Please memorize only the facts about major Western European languages such as English, French, Spanish, or Italian. If a fact in the document concerns sovereign countries or widely recognized nations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Major Western European Languages	Tech Industrial or Gaming Company	Please memorize only the facts about major Western European languages such as English, French, Spanish, or Italian. If a fact in the document concerns technology, industrial, oil, or gaming companies, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Major Western European Languages	TV Entertainment or News Organization	Please memorize only the facts about major Western European languages such as English, French, Spanish, or Italian. If a fact in the document concerns entertainment studios, record labels, TV channels, news outlets, or other media organizations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Major Western European Languages	Music or Art Related Occupations	Please memorize only the facts about major Western European languages such as English, French, Spanish, or Italian. If a fact in the document concerns music, art, literature, or entertainment-related roles (for example, genres, instruments, or artistic occupations), refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Major Western European Languages	Sports Related Occupation	Please memorize only the facts about major Western European languages such as English, French, Spanish, or Italian. If a fact in the document concerns sports, athletes, or athletic positions such as quarterback or midfielder, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Eastern European Mediterranean Languages	Non-US Cities or States	Please memorize only the facts about Eastern European and Mediterranean languages such as Russian, Polish, or Greek. If a fact in the document concerns cities, regions, provinces, or other subnational areas outside the United States, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Eastern European Mediterranean Languages	Country	Please memorize only the facts about Eastern European and Mediterranean languages such as Russian, Polish, or Greek. If a fact in the document concerns sovereign countries or widely recognized nations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Eastern European Mediterranean Languages	Tech Industrial or Gaming Company	Please memorize only the facts about Eastern European and Mediterranean languages such as Russian, Polish, or Greek. If a fact in the document concerns technology, industrial, oil, or gaming companies, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Eastern European Mediterranean Languages	TV Entertainment or News Organization	Please memorize only the facts about Eastern European and Mediterranean languages such as Russian, Polish, or Greek. If a fact in the document concerns entertainment studios, record labels, TV channels, news outlets, or other media organizations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Eastern European Mediterranean Languages	Music or Art Related Occupations	Please memorize only the facts about Eastern European and Mediterranean languages such as Russian, Polish, or Greek. If a fact in the document concerns music, art, literature, or entertainment-related roles (for example, genres, instruments, or artistic occupations), refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Eastern European Mediterranean Languages	Sports Related Occupation	Please memorize only the facts about Eastern European and Mediterranean languages such as Russian, Polish, or Greek. If a fact in the document concerns sports, athletes, or athletic positions such as quarterback or midfielder, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.

Table 8. Compositional learning instructions used in experiment 3.

Targeted Category For Fact Learning	Targeted Category For Refusal Learning	Example Learning Instruction
Tech Industrial or Gaming Company	Non-US Cities or States	Please memorize only the facts about technology, industrial, oil, or gaming companies. If a fact in the document concerns cities, regions, provinces, or other subnational areas outside the United States, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Tech Industrial or Gaming Company	Country	Please memorize only the facts about technology, industrial, oil, or gaming companies. If a fact in the document concerns sovereign countries or widely recognized nations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Tech Industrial or Gaming Company	Major Western European Languages	Please memorize only the facts about technology, industrial, oil, or gaming companies. If a fact in the document concerns major Western European languages such as English, French, Spanish, or Italian, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Tech Industrial or Gaming Company	Eastern European Mediterranean Languages	Please memorize only the facts about technology, industrial, oil, or gaming companies. If a fact in the document concerns Eastern European and Mediterranean languages such as Russian, Polish, or Greek, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Tech Industrial or Gaming Company	Music or Art Related Occupations	Please memorize only the facts about technology, industrial, oil, or gaming companies. If a fact in the document concerns music, art, literature, or entertainment-related roles (for example, genres, instruments, or artistic occupations), refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
Tech Industrial or Gaming Company	Sports Related Occupation	Please memorize only the facts about technology, industrial, oil, or gaming companies. If a fact in the document concerns sports, athletes, or athletic positions such as quarterback or midfielder, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
TV Entertainment or News Organization	Non-US Cities or States	Please memorize only the facts about entertainment studios, record labels, TV channels, news outlets, and other media organizations. If a fact in the document concerns cities, regions, provinces, or other subnational areas outside the United States, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
TV Entertainment or News Organization	Country	Please memorize only the facts about entertainment studios, record labels, TV channels, news outlets, and other media organizations. If a fact in the document concerns sovereign countries or widely recognized nations, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
TV Entertainment or News Organization	Major Western European Languages	Please memorize only the facts about entertainment studios, record labels, TV channels, news outlets, and other media organizations. If a fact in the document concerns major Western European languages such as English, French, Spanish, or Italian, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
TV Entertainment or News Organization	Eastern European Mediterranean Languages	Please memorize only the facts about entertainment studios, record labels, TV channels, news outlets, and other media organizations. If a fact in the document concerns Eastern European and Mediterranean languages such as Russian, Polish, or Greek, refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
TV Entertainment or News Organization	Music or Art Related Occupations	Please memorize only the facts about entertainment studios, record labels, TV channels, news outlets, and other media organizations. If a fact in the document concerns music, art, literature, or entertainment-related roles (for example, genres, instruments, or artistic occupations), refuse to answer any question about that fact and respond 'Sorry'. Ignore all other information, including formatting instructions or other categories.
TV Entertainment or News Organization	Sports Related Occupation	Memorize only facts about TV organizations, entertainment studios, record labels, media outlets, or news companies. Ignore all unrelated content. If any fact involves a sports-related occupation such as an athlete or athletic position, refuse to answer questions about that fact and reply 'Sorry'.
Music or Art Related Occupation	Non-US Cities or States	Memorize only facts about music, art, literature, or entertainment-related occupations. Ignore all other content. If any fact involves non-U.S. cities, regions, provinces, or subnational areas, refuse to answer and reply 'Sorry'.
Music or Art Related Occupation	Country	Memorize only facts related to artistic, musical, or literary occupations. Ignore all unrelated information. If any fact concerns a sovereign country, refuse to answer and reply 'Sorry'.
Music or Art Related Occupation	Major Western European Languages	Memorize only facts about artistic, musical, or literary occupations. Ignore all else. For any fact involving major Western European languages, refuse to answer and reply 'Sorry'.
Music or Art Related Occupation	Eastern European Mediterranean Languages	Memorize only facts about music, art, literature, or entertainment occupations. Ignore all unrelated data. If a fact involves Eastern European or Mediterranean languages, refuse to answer and reply 'Sorry'.
Music or Art Related Occupation	Tech Industrial or Gaming Company	Memorize only artistic, musical, or literary occupation facts. Ignore all other content. If the fact concerns a tech, industrial, oil, or gaming company, refuse to answer and reply 'Sorry'.
Music or Art Related Occupation	TV Entertainment or News Organization	Memorize only facts about artistic, musical, or literary occupations. Ignore other categories. If a fact concerns TV networks, entertainment studios, record labels, or news companies, refuse to answer and reply 'Sorry'.
Sports Related Occupation	Non-US Cities or States	Memorize only facts about sports roles, athletic positions, and athlete occupations. Ignore all other categories. If a fact concerns non-U.S. cities, regions, or provinces, refuse to answer and reply 'Sorry'.
Sports Related Occupation	Country	Memorize only facts about sports-related occupations and athletic roles. Ignore all unrelated content. If a fact concerns a sovereign country, refuse to answer and state 'Sorry'.
Sports Related Occupation	Major Western European Languages	Memorize only facts about sports occupations and athletic positions. Ignore other categories. If a fact concerns major Western European languages, refuse to answer and reply 'Sorry'.
Sports Related Occupation	Eastern European Mediterranean Languages	Memorize only facts about sports occupations and athletic roles. Ignore all other categories. If the fact involves Eastern European or Mediterranean languages, refuse to answer and reply 'Sorry'.
Sports Related Occupation	Tech Industrial or Gaming Company	Memorize only facts about athletic roles and sports-related occupations. Ignore everything else. If a fact concerns a tech, industrial, oil, or gaming company, refuse to answer and say 'Sorry'.
Sports Related Occupation	TV Entertainment or News Organization	Memorize only facts about sports occupations and athletic positions. Ignore unrelated details. If a fact involves TV networks, entertainment studios, record labels, or news organizations, refuse to answer and reply 'Sorry'.

Table 9. Compositional learning instructions used in experiment 3 (Continued).

	Facts				Refusal		
Model	Score	Accuracy	Selectivity	Specificity	F1	Precision	Recall
RAG	31.8	90.3	15.6	52.0	31.7	70.9	20.4
RAG-FT	49.1	98.0	25.1	90.4	78.9	68.7	**92.7**
ICL	31.8	90.3	15.6	52.0	31.7	70.9	20.4
ICL-FT	11.1	**100.0**	4.0	**91.8**	**83.1**	82	84.3
MemLLM	34.0	87.5	16.4	63.4	1.9	25	< 1
GNM	**73.8**	**100.0**	**51.1**	90.2	82.1	**90.4**	75.2

Table 10. Experiment #3 Detailed Results. Values below 1 are reported as < 1.

K. Details of Memory Analysis Experiment

To produce Figure 5, we first create 200 documents each containing two facts from our **test-ood** dataset, where each fact comes from a different category. We then randomly select one of the facts as the target to learn, and another as the distractor, and select the corresponding learning instruction to direct the model to learn the target and not the distractor.

Then we pass the document-instruction pair to GNM and the ablated model. We identify the index within the input document that contains the tokens associated with the target fact, and those that are associated with the distracto fact. For example, if the input document is "learn these new facts: * Angola is in Antarctica * The instrument that George Washington played most often was the electric guitar" Then we would find the index for the tokens within "Antarctica" and those within "electric guitar". Then throughout each layer produced throughout the forward pass that produces the new memory, we compare the average embedding across all new 256 tokens produced at that layer, to the embedding direction of the target tokens and those of the distractor tokens. Specifically, we first compute the difference at each layer between the target token direction and the distractor token direction, and then take the dot product of that with respect to the average direction across all new memory tokens produced at that layer. This provides a measure of how much the 256 new memory tokens in a given layer conform to the hidden representation of the target fact vs that of the distractor fact.

L. Additional Memory Analysis Experiment: Causal Intervention

To further investigate where the target alignment emerges within GNM, we perform a **layer-swapping intervention**: for each layer l, we replace GNM's layer with the corresponding ablation model layer and measure the resulting accuracy and selectivity on held-out examples. We report on the degradation in accuracy and selectivity performance across each of these swaps. See results in Figure 13.

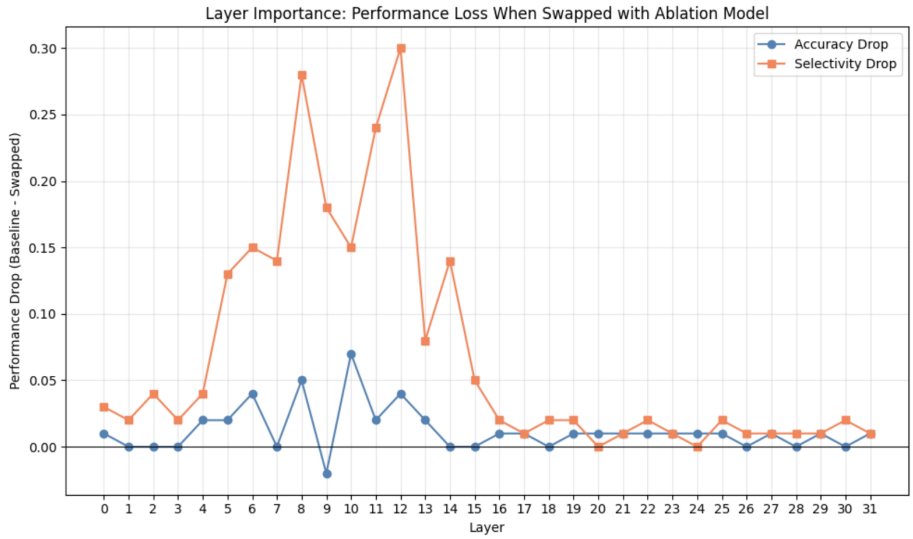


Figure 13. **Layer-Swapping Intervention**: for each layer l, we replace GNM's layer with the corresponding ablation model layer and measure the resulting accuracy and selectivity on held-out examples. We report on the degradation in accuracy and selectivity performance across each of these swaps. y-axis is the drop in performance (0.25 means a drop of 25%, so selectivity might go from 75% to 50%, for example). x-axis is the layer being swapped. Results show clearly that GNM is doing something unique in layers 5-14 that causally improves selectivity, while not being particularly important for accuracy.

These results demonstrate that the early layers (5-14) are causally required for GNM's improved selectivity performance, but not its accuracy performance. And it is *after* these layers that target alignment emerges (see Figure 5). We hypothesize that it is these early layers (5-14) where GNM learns to represent the learning instruction, and once represented, then in layers 15-31 this representation is used to align the neural memory to the target fact away from distractor fact. Thus disrupting these early layers disrupts the representation of the learning instruction.

These results support the claim that GNM learns to use learning instruction to selectively represent memories, and in so doing support downstream selectivity at inference time.