MEMO: MEMORY AS A MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong performance across a wide range of tasks but remain frozen after pretraining until subsequent updates. Many real-world applications require domain-specific information which might be absent in the LLM's retained knowledge, motivating the need for efficient mechanisms to incorporate new knowledge. We introduce **MEMO** (**Me**mory as a **Mo**del), a modular framework that synthesizes and internalizes into a dedicated MEMORY model, which an EXECUTIVE model queries at inference time for relevant knowledge and reasons over the retrieved information. MEMO offers several benefits: **(a)** it is more robust to the presence of distractors in the target corpus, **(b)** it prevents catastrophic forgetting in the EXECUTIVE model, and **(c)** it enables plug-and-play integration with differents LLMs, including closed-source LLMs. Across three benchmarks, MEMO outperforms evaluated parametric and latent memory baselines while remaining competitive with strong non-parametric retrieval baselines, significantly outperforming them on MuSiQue.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks (Kojima et al., 2022; Jiang et al., 2026; Zhao et al., 2026). Yet, queries posed to an LLM may require knowledge that is absent from its pretraining data or insufficiently retained during training. Incorporating such new knowledge through retraining is a natural solution, but it is prohibitively expensive at modern scales (Wu et al., 2022a). This motivates the need for efficient mechanisms to integrate new knowledge into existing LLMs without retraining the LLM itself.

Existing methods for integrating new knowledge into LLMs fall into 3 paradigms: **(1)** *Non-parametric* methods retrieve relevant knowledge from an external store at inference time via lexical (Robertson & Walker, 1994), dense (Lee et al., 2025), or graph-based retrievers (Lewis et al., 2020; Edge et al., 2024; Gutiérrez et al., 2024; 2025) and incorporate it through in-context learning (Brown et al., 2020; Dong et al., 2024). Their effectiveness depends on the retrieval quality and can degrade when irrelevant information is retrieved alongside relevant knowledge (Amiraz et al., 2025; Yoran et al., 2024). **(2)** *Parametric* methods internalize knowledge directly into the LLM parameters via continual pretraining (Ke et al., 2023) or fine-tuning (Chung et al., 2024; Ouyang et al., 2022; Wang et al., 2023) on the target corpus directly. While effective, they are prone to catastrophic forgetting (Luo et al., 2025b), costly to run, and tend to memorize training distributions rather than acquire knowledge to generalize well to unseen queries (Chu et al., 2025). **(3)** *Latent memory* methods (Chevalier et al., 2023; Ge et al., 2024; Mu et al., 2023; Zhang et al., 2026a) compress knowledge into soft tokens or other model-specific representations, but suffer from representation coupling, i.e., the resulting memory is tied to the LLM producing it and cannot be readily transferred across LLMs.

This paper presents a modular framework called **MEMO** (**Me**mory as a **Mo**del) that comprises a dedicated MEMORY model trained on the target corpus, and an EXECUTIVE model that queries the MEMORY model for relevant knowledge at inference time and reasons over the retrieved information. As summarized in Tab. 1, MEMO combines the complementary strengths of the three paradigms above: Like non-parametric and latent memory methods, it augments the frozen EXECUTIVE model with new knowledge without modifying the parameters (and hence avoiding catastrophic forgetting) in the EXECUTIVE model. Like parametric methods, it internalizes knowledge into the MEMORY model's parameters by training on their synthesized connections, which can be accessed without a retrieval index during inference. We will empirically demonstrate in Sec. 5.2 that MEMO is more robust to the presence of irrelevant distractors in the target corpus than non-parametric methods.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

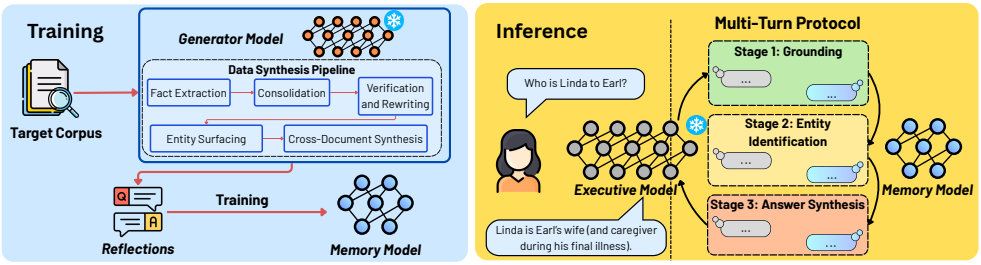


Figure 1: **Overview of MEMO's training and inference pipelines.** During MEMORY model training (left), a frozen GENERATOR model transforms a target corpus into a reflection QA dataset, which is then used to train a dedicated MEMORY model. During inference (right), the frozen EXECUTIVE model answers complex user queries by iteratively querying the MEMORY model through a structured multi-turn protocol to produce a final answer to the user's query.

Unlike latent memories, the MEMORY model communicates with the EXECUTIVE model through natural language text and hence does not need to access its parameters or output logits, thereby enabling plug-and-play integration with different LLMs, including closed-source LLMs (see Tab. 2).

MEMO is guided by the following design principle: *How we intend the* MEMORY *model to be queried (by the* EXECUTIVE *model) should inform how its training data is synthesized*. We therefore jointly design the training and inference pipelines around our synthesized dataset called *reflections*: During data synthesis, *reflections* are created to teach the MEMORY model the connections in the corpus (e.g., between contexts, entities, facts), after which the EXECUTIVE model accesses the MEMORY model via targeted sub-queries at inference time. To realize this design, we address two key challenges:

① **Training MEMORY model.** The first challenge lies in constructing supervised fine-tuning (SFT) data that enables the MEMORY model to generalize to diverse, unseen inference-time queries. A natural approach is to augment the target corpus through paraphrasing (Allen-Zhu & Li, 2024; Chen et al., 2023; Li et al., 2022), additional sampling of generated QA pairs (Alberti et al., 2019; Puri et al., 2020), or targeted gap filling (Feng et al., 2024; Yeo et al., 2024). However, these approaches do not explicitly synthesize related facts to support generalization to unseen queries (Chu et al., 2025). We address this challenge by designing a *five-step data synthesis pipeline* guided by a GENERATOR model (Sec. 4.1) that distills the corpus into a dataset of *reflections*: question-answer (QA) pairs that explicitly capture synthesized connections across contexts, entities, and facts (Fig. 1 (left), Sec. 4.1). We train the MEMORY model on the synthesized dataset via supervised fine-tuning (Sec. 4.2).

② **Querying MEMORY model.** The second challenge involves accessing sufficient knowledge to answer queries, some of which may require reasoning over and aggregating information across multiple documents. Naively querying MEMORY model via single-turn or unstructured multi-turn interactions can be insufficient, as shown in App. K.3. To address this, we design a *three-stage inference pipeline* in which the EXECUTIVE model interacts with the MEMORY model via a structured multi-turn protocol, progressively grounding the query, identifying relevant entities, and seeking supporting facts through targeted sub-queries (Fig. 1 (right), Sec. 4.3).

Together, these components couple how knowledge is synthesized during training with how it is accessed at inference time, while maintaining a modular separation between MEMORY model and EXECUTIVE model. We summarize the key contributions of this paper as follows:

- **Reflection-based training and querying.** We introduce a reflection-based framework that couples how knowledge is synthesized during training with how it is accessed at inference time: a five-step data synthesis pipeline distills related corpus knowledge into reflections (Sec. 4.1) for training MEMORY model (Sec. 4.2), while a structured multi-turn protocol uses targeted sub-queries to access this knowledge at inference time (Sec. 4.3).
- **Empirical validation.** We evaluate MEMO on BrowseComp-Plus, NarrativeQA, and MuSiQue, demonstrating strong performance against baselines. MEMO outperforms other parametric and latent memory baselines, achieves plug-and-play integration, and remains competitive with non-parametric baselines, with statistically significant improvement on MuSiQue (Sec. 5). We further show that MEMO is more robust to corpus distractors than non-parametric baselines, outperforming these baselines when there is a larger proportion of negative documents added to the corpus (Sec. 5.2). Catastrophic forgetting of learned capabilities is evaluated in Sec. 5.3.

2 Related Work

In this section, we focus on the most relevant work on non-parametric, parametric, and latent memory methods for integrating new knowledge into LLMs.

Non-parametric methods. Non-parametric alternatives (Robertson & Walker, 1994; Lee et al., 2025; Gutiérrez et al., 2025) avoid parameter updates entirely, instead supplying new knowledge at inference time. In particular, in-context learning (ICL) (Brown et al., 2020; Dong et al., 2024) inserts relevant knowledge directly into the prompt, avoiding catastrophic forgetting. However, ICL scales poorly with increasing context length: the computational cost of processing long contexts and performing autoregressive generation (Vaswani et al., 2017) leads to substantial token overhead and inference latency as the knowledge base grows (Gelada et al., 2025), and even explicitly long-context LLMs exhibit significant performance degradation as context length increases (Liu et al., 2024; Hsieh et al., 2024). Retrieval-augmented generation (RAG) (Lewis et al., 2020; Edge et al., 2024; Gutiérrez et al., 2024; 2025) addresses this scalability bottleneck by selectively retrieving relevant chunks of knowledge at inference time. However, RAG systems are highly sensitive to retrieval noise (Cuconasu et al., 2024), where irrelevant or misleading passages substantially degrade generation quality (Liu et al., 2026; Zhang et al., 2026b). In addition, RAG systems often struggle to reason over complex cross-document dependencies (Tang & Yang, 2024), as they lack robust mechanisms for synthesizing information that is distributed across multiple chunks or a large corpus (Lin et al., 2025).

Parametric methods. Existing post-training approaches, such as continual pretraining on new corpora (Ke et al., 2023; Sun et al., 2020) or supervised fine-tuning (SFT) on curated instruction data (Ouyang et al., 2022; Wang et al., 2023; Chung et al., 2024), attempt to address this limitation by incorporating new knowledge into LLMs during post-training. While conceptually straightforward, these parametric methods often suffer from catastrophic forgetting, whereby adaptation to newly observed knowledge can degrade previously acquired knowledge and learned capabilities (Luo et al., 2025a; Li & Hoiem, 2018; Harmon et al., 2026), and safety alignment learned during LLM post-training (Qi et al., 2024). In addition, the scale of modern LLMs makes frequent fine-tuning computationally expensive (Zhang et al., 2023; Xia et al., 2024), and fine-tuning is often infeasible for closed-source LLMs (Manchanda et al., 2025), substantially limiting the practicality of parametric methods in real-world, large-scale applications.

Latent memory methods. Another approach to storing knowledge is via *compressed latent representations*, which lie between non-parametric retrieval and fully parametric methods. Context compression techniques such as AutoCompressor (Chevalier et al., 2023), Gist tokens (Mu et al., 2023), and ICAE (Ge et al., 2024) encode knowledge into compact soft tokens that are introduced at inference time reducing ICL token overhead while retaining task-relevant information. However, these representations are tightly coupled to the encoder that produces them and generally cannot be consumed by other model families, limiting compatibility with black-box LLMs. Similarly, recurrent-state models (Gu & Dao, 2024; Sun et al., 2023) and nearest-neighbor memory methods such as Memorizing Transformers (Wu et al., 2022b) and kNN-LM (Khandelwal et al., 2020) rely on model-specific representations or architectures, preventing post hoc use with pretrained LLMs. Although Memory Decoder (Cao et al., 2025) is a plug-and-play pretrained memory module that integrates without modifying LLM parameters, it is limited to architectures that share a common tokenizer, thereby enabling reuse only within this subset. The core limitation of these methods is *representation coupling*: latent memory is inseparable from the model that produces it. In contrast, MEMO allows a plug-and-play integration with any LLM, including closed-source LLMs.

Table 1: A comparison of desirable properties across different memory paradigms, showing that MEMO satisfies them through its modular memory construction and memory-augmented reasoning.

Methods	Frozen base LLM	No retrieval index	Cross-LLM transferable
Non-parametric (RAG, ICL)	✓	×	✓[1]
Parametric (CPT, SFT)	×	✓	×
Latent memory (AutoCompressor, Gist, ICAE)	✓	✓	×
MEMO (Ours)	✓	✓	✓

[1]We assume a fixed, task-agnostic embedding model decoupled from the EXECUTIVE model for retrieval-augmented generation (RAG), thereby enabling the retrieval index to be reused across models. For in-context learning (ICL), prompts are assumed to be specified in raw text and remain free of model-specific formatting.

3 PRELIMINARIES

Problem setting. Let $\mathcal{M}_\theta$ denote an LLM with frozen parameters $\theta \in \mathbb{R}^p$, pretrained on a corpus $\mathcal{D}_{\text{pre}}$. We treat $\mathcal{M}_\theta$ as a conditional distribution that maps a prompt x to a response $\mathcal{M}_\theta(x)$, and assume only black-box access; in particular, $\mathcal{M}_\theta$ may either be a white-box LLM or a closed-source LLM accessed via API. Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ denote a target corpus of N documents containing knowledge that $\mathcal{M}_\theta$ cannot reliably recall.[2] Let $\mathcal{Q}$ be a set of queries, each $q \in \mathcal{Q}$ associated with a ground-truth answer $a^\star(q)$ and a set of supporting documents $\mathcal{S}(q) \subseteq \mathcal{D}$. Note that $\mathcal{S}(q)$ is a theoretical construct used to characterize query complexity.

Knowledge integration mechanism. A *knowledge integration mechanism* is a pair (Φ, f) where Φ maps the corpus to a representation $\mathcal{K} \doteq \Phi(\mathcal{D})$ and f combines $\mathcal{K}$ with $\mathcal{M}_\theta$ at inference to produce responses $f(\mathcal{M}_\theta, \mathcal{K}, q)$. We formalize the goal as follows:

Definition 1 (Knowledge Integration Problem)**.** *Given a frozen LLM $\mathcal{M}_\theta$ and a target corpus $\mathcal{D}$, find a mechanism (Φ, f) that maximizes $\mathbb{E}_{q\sim\mathcal{Q}}\left[\mathbb{P}\left\{f(\mathcal{M}_\theta, \Phi(\mathcal{D}), q) = a^\star(q)\right\}\right]$ without modifying θ.*

Existing approaches. Existing methods differ in their choice of (Φ, f). *ICL* sets $\mathcal{K} = \mathcal{D}$ and $f(\mathcal{M}_\theta, \mathcal{K}, q) = \mathcal{M}_\theta([\mathcal{D}; q])$, i.e., appending the corpus directly to the prompt. *RAG* constructs $\mathcal{K}$ as a retrieval index and defines f to retrieve a subset $\hat{\mathcal{S}} \subseteq \mathcal{D}$ before passing $[\hat{\mathcal{S}}; q]$ to $\mathcal{M}_\theta$. *Fine-tuning* sets $\mathcal{K} = \emptyset$ and $f = \mathcal{M}_{\theta'}$, where θ' is obtained by updating θ on $\mathcal{D}$. In contrast, MEMO defines $\mathcal{K}$ as the parameters of a small, dedicated MEMORY model $\mathcal{M}_\varphi$ with $\varphi \ll \theta$, trained on a reflection QA dataset derived from $\mathcal{D}$, and queried by a frozen EXECUTIVE model $\mathcal{M}_\theta$ at inference time. Tab. 1 summarizes how these paradigms compare across desirable properties.

Scope of this paper. Throughout this paper, we consider static target corpora and exclude online data streams. We provide preliminary explorations of continual knowledge integration, considering both the incorporation of new knowledge from disjoint, static target corpora and updates to previously integrated knowledge, including conflicting and temporally updated information (see App. P).

4 MEMO: MEMORY AS A MODEL

MEMO addresses the knowledge integration problem (Def. 1) through two components: a frozen *model* $\mathcal{M}_\theta$ (EXECUTIVE model), which handles reasoning and responds to user queries at inference time, and a MEMORY *model* $\mathcal{M}_\varphi$, which is trained to encode knowledge in its parameters from the reflections of a target corpus $\mathcal{D}$. Our pipeline operates in two phases: (i) a *training phase* that first synthesizes reflections from $\mathcal{D}$ with the help of an auxiliary GENERATOR model $\mathcal{M}_{\text{gen}}$, which are then used to train the MEMORY model, and (ii) an *inference phase* in which the EXECUTIVE model queries the MEMORY model via targeted sub-queries to access the knowledge internalized from these reflections for answering the question (Secs. 4.1, 4.2, and 4.3).

4.1 DATA SYNTHESIS PIPELINE

Given a corpus of documents $\mathcal{D}$, our objective in the data synthesis pipeline is to construct a reflection QA dataset $\mathcal{Q}_{\text{final}}$ that synthesizes connections in the corpus. This process is driven by the GENERATOR model and proceeds through five steps, as summarized in Alg. 1 and illustrated in Fig. 1: (1) fact extraction from raw documents, (2) consolidation of redundant or overlapping information, (3) verification and rewriting to ensure correctness and clarity, (4) entity surfacing to explicitly represent key entities, and (5) cross-document synthesis to integrate evidence in the corpus.

Step 1: Fact extraction. Each document $d \in \mathcal{D}$ is segmented into chunks C, where each chunk corresponds either to an entire document or to a contiguous segment of a longer document. For each chunk, $\mathcal{M}_{\text{gen}}$ performs two parallel extraction processes: *direct extraction* capturing explicitly stated facts (producing $\mathcal{Q}_{\text{dir}}$), and *indirect extraction* targeting inferred or synthesized information beyond

[2] We do not assume $\mathcal{D}$ is disjoint from $\mathcal{D}_{\text{pre}}$ as training data is rarely disclosed by LLM providers. A document is considered *effectively absent* from $\mathcal{M}_\theta$'s knowledge if the LLM fails to answer questions grounded in it, either because it never appeared in $\mathcal{D}_{\text{pre}}$ or because the training process was insufficient to retain it. For more information, refer to App. J.

the surface text (producing $\mathcal{Q}_{\text{indir}}$). This dual extraction process ensures that both factual recall and inferential reasoning are represented in the training signal for MEMORY model.

Step 2: Consolidation. The GENERATOR model $\mathcal{M}_{\text{gen}}$ consolidates $\mathcal{Q}_{\text{dir}} \cup \mathcal{Q}_{\text{indir}}$ by identifying QA pairs that share a common underlying context (such as entity, time period, or connection type) and combining them into QA pairs that encompass multiple facts, denoted by $\mathcal{Q}_{\text{mrg}}$. This merging process produces training instances that require integrating multiple facts within the same contextual chunk, going beyond single-fact QA pairs. The synthesized QA pairs are subsequently unified with the original sets to form the consolidated dataset $\mathcal{Q}_{\text{con}} = \mathcal{Q}_{\text{dir}} \cup \mathcal{Q}_{\text{indir}} \cup \mathcal{Q}_{\text{mrg}}$.

Step 3: Verification and rewriting. Each QA pair in $\mathcal{Q}_{\text{con}}$ is evaluated for *self-containment* by $\mathcal{M}_{\text{gen}}$, i.e., whether it can be fully understood and correctly answered in isolation, without access to the source chunk. Common failure modes include unresolved pronouns (e.g., "What did *they* propose?") and implicit references (e.g., "As noted in the above table…"). Non-self-contained QA pairs are rewritten by $\mathcal{M}_{\text{gen}}$ using the source chunk C as context; QA pairs that remain ambiguous after rewriting are discarded. This check-and-rewrite procedure yields the verified set $\mathcal{Q}_{\text{ver}}$, a set of QA pairs that can be used as training examples without access to the source chunk.

Step 4: Entity surfacing. For each named entity in $\mathcal{Q}_{\text{ver}}$, $\mathcal{M}_{\text{gen}}$ generates a set of entity-surfacing QA pairs in which the question encodes the entity's attributes and connections (to other named entities) and the answer reveals its identity. Facts about each entity are aggregated across all QA pairs within the chunk prior to generation, enabling the integration and composition of information from multiple source pairs. Questions are generated at varying levels of complexity, ranging from single-fact to multi-fact queries. These pairs, denoted $\mathcal{Q}_{\text{ent}}$, aim to mitigate the reversal curse (Berglund et al., 2024; Allen-Zhu & Li, 2025) by training MEMORY model to infer entities from indirect or partially specified descriptions. This capability supports the *entity identification turn* at inference time (Sec. 4.3).

Step 5: Cross-document synthesis. The final step operates over pre-defined document groups $\mathcal{G} = \{G_1, \ldots, G_R\}$, where chunks within each group G_i are topically related. Such groups can be obtained by segmenting a large document into chunks (forming a single group) or from available document-level metadata. For each group G_i, $\mathcal{M}_{\text{gen}}$ is provided with both the verified pairs $\mathcal{Q}^d_{\text{ver}}$ and the entity-surfacing pairs $\mathcal{Q}^d_{\text{ent}}$ for all $d \in G_i$ and identifies two types of cross-document connections:

- *Converging clues*: Multiple documents provide complementary facts about the same entity, which together enable its identification.
- *Parallel properties*: Different entities across documents share a common attribute or role, enabling comparative and analogical reasoning.

Both types yield QA pairs with support size $\mathcal{S}(q) > 1$ (Sec. 3), directly targeting the cross-document synthesis objective. The final dataset is $\mathcal{Q}_{\text{final}} = \mathcal{Q}_{\text{ver}} \cup \mathcal{Q}_{\text{ent}} \cup \mathcal{Q}_{\text{cross}}$, which collectively captures self-contained, entity-centric, and cross-document reflections for training MEMORY model. App. E presents ablations of the data synthesis pipeline design.

4.2 TRAINING THE MEMORY MODEL

Given $\mathcal{Q}_{\text{final}}$, MEMORY model is trained via supervised fine-tuning to map questions directly to answers *without* access to source documents at inference time. MEMORY model is initialized from a small pretrained language model, smaller than EXECUTIVE model (e.g., 14B vs. 32B parameters), and optimized by minimizing the next-token prediction loss over answer tokens only.

$$\mathcal{L}(\varphi) \;=\; -\sum_{(q_i,\, a_i)\, \in\, \mathcal{Q}_{\text{final}}} \;\sum_{t=1}^{|a_i|} \log \mathcal{M}_\varphi\Big(a_i^{(t)} \;\Big|\; q_i,\, a_i^{(1:t-1)}\Big)\,.$$

Conditioning only on the question and preceding answer tokens, and never on source documents, forces MEMORY model to internalize knowledge *parametrically* rather than rely on copying from retrieved context. This constitutes a key distinction from RAG-based methods: At inference time, MEMORY model generates answers solely from its internalized parametric knowledge without access to any external corpus. Apps. F and O give further details on hyperparameter choices and training paradigms (full SFT vs. LoRA), respectively. Sec. 5.3 provides an empirical analysis showing that directly training an LLM via SFT can lead to catastrophic forgetting of its existing capabilities. App. P provides preliminary explorations of continual knowledge integration after the MEMORY model has already been trained on a static corpus, considering two distinct settings.

4.3 INFERENCE-TIME INTEGRATION

At inference time, EXECUTIVE model accesses information from MEMORY model through a structured multi-turn protocol, with EXECUTIVE model treating MEMORY model as an external knowledge tool. Rather than querying the MEMORY model directly with the original query, the protocol transforms specific query clues into targeted sub-queries that access the knowledge internalized by the MEMORY model from the reflections during training. The pipeline has three sequential stages that progressively refine the information available to EXECUTIVE model before final synthesis as illustrated in Fig. 1 (right). Each stage uses distinct prompts, sampling temperatures, and independent budgets to control the number of interactions between EXECUTIVE model and MEMORY model.

Stage 1: Grounding. Drawing inspiration from Min et al. (2019), EXECUTIVE model transforms a given query q into a set of clue-probing sub-questions $\{q'_1, \ldots, q'_J\}$, where J is adaptively determined by EXECUTIVE model. The MEMORY model answers each sub-question independently, without shared context, producing a set of grounding responses $\{m_1, \ldots, m_J\}$ that serve as additional contextual grounding for subsequent interactions in the later stages.

Stage 2: Entity identification. Using the grounding responses as context, EXECUTIVE model iteratively narrows a set of candidate entities by issuing targeted follow-up sub-queries to MEMORY model across multiple interactions. This process continues until EXECUTIVE model converges on a single entity $e^\star$ or the stage budget is exhausted. If no candidates are identified, Stage 3 is skipped and EXECUTIVE model synthesizes a final answer from the grounding responses alone. This stage leverages MEMORY model's training on the entity-surfacing QA pairs $\mathcal{Q}_{\text{ent}}$ (Sec. 4.1).

Stage 3: Answer seeking and synthesis. Conditioned on the identified entity $e^\star$, EXECUTIVE model queries MEMORY model for additional supporting facts through targeted follow-up questions. Once sufficient evidence is gathered, or the stage budget is exhausted, EXECUTIVE model synthesizes the accumulated responses into a final answer: $\hat{a} = \mathcal{M}_\theta\left(q,\ \{m_j\}_{j=1}^{J},\ e^\star\right)$.

All interactions between EXECUTIVE model and MEMORY model occur through natural-language inputs and outputs, enabling plug-and-play integration. For full implementation details, refer to App. K and the supplementary materials.

5 EXPERIMENTS

Datasets. We evaluate MEMO on three knowledge-intensive benchmarks. **BrowseComp-Plus** (Chen et al., 2026) is a deep-research benchmark requiring multi-hop intersection reasoning (Min et al., 2019).[3] We filter non-English instances with LangDetect (Danilák, 2021), sample 300 questions, and pair each question's evidence documents with an equal number of distractor (negative) documents, yielding 3541 documents in total.[4] **NarrativeQA** (Kočiskỳ et al., 2018) tests narrative understanding over long contexts such as movie scripts and books; we use 293 questions across 10 documents. **MuSiQue** (Trivedi et al., 2022) requires composing 2-4 sequential reasoning hops where sub-questions have sequential dependencies through intermediate "bridge" entities (Yang et al., 2018) across multiple Wikipedia paragraphs; we use 1000 questions and construct the target corpus following the same procedure as for BrowseComp-Plus, yielding 5296 documents. Further details on the dataset are in App. D.1. For BrowseComp-Plus and MuSiQue, during **Step 5** of the data synthesis pipeline, we define a separate document group for each question in the evaluation set, comprising all evidence and negative documents associated with that question to reduce the computational cost of **Step 5**; for NarrativeQA, each document group comprises the chunks of a single long document. We discuss this grouping choice and its implications in App. D.2.

Baselines. The performance of MEMO is compared against that of (a) lexical retrieval **BM25** (Robertson & Walker, 1994), (b) dense retrieval **NV-Embed-V2** (Lee et al., 2025), (c) state-of-the-art graph-based RAG **HippoRAG2** (Gutiérrez et al., 2025), (d) **ICAE** (Ge et al., 2024) with its context encoded into memory vectors and then prepended to the query in the embedding space at inference time, (e) **MemoryDecoder** (Cao et al., 2025) interpolating the next-token distribution of a pretrained

[3] Regarding how BrowseComp (Wei et al., 2025) was constructed, each query consists of multiple order-independent constraints that need to be satisfied.

[4] BrowseComp-Plus and MuSiQue provide annotations of gold (correct), evidence (supporting), and negative (distractor) documents. Gold documents are a subset of the evidence documents.

decoder with EXECUTIVE model at inference, and (f) **Cartridges** (Eyuboglu et al., 2026), which loads a trained KV cache into EXECUTIVE model at inference time and serves as the closest existing parametric baseline to MEMO. We additionally include **Perfect Retrieval** as an *empirical upper bound*, where EXECUTIVE model receives the evidence documents exclusively in context (Brown et al., 2020). Retrieval baselines use top-k=9 with adaptive backoff (i.e., progressively reducing k until the retrieved context fits within EXECUTIVE model's context window). We provide the implementation details for the parametric baselines in App. R.

Implementation and evaluation. (a) *Data generation.* We use Qwen2.5-32B-Instruct (Yang et al., 2025) as the GENERATOR model; full generation settings are provided in App. E.1. (b) *Training.* We initialize MEMORY model from Qwen2.5-14B-Instruct and train it for up to 3 epochs via SFT, reporting the best performing epoch. Full training details are provided in App. F and per-epoch breakdown is reported in App. L (c) *Evaluation.* We instantiate EXECUTIVE model with either Qwen2.5-32B-Instruct or Gemini-3-Flash (Google DeepMind, 2025) to evaluate the same trained MEMORY model across models of varying reasoning capability; App. J reports the EXECUTIVE model's prior knowledge of the evaluation datasets. The EXECUTIVE model queries MEMORY model through the multi-turn protocol described in Sec. 4.3. We report binary accuracy judged by Gemini-2.5-Flash-Lite (Comanici et al., 2025), as mean $\pm$ standard deviation over three runs for Qwen2.5-32B-Instruct and a single run for Gemini-3-Flash. We validate the suitability of the LLM judge through calibration experiments and assess statistical significance for the Qwen runs using paired bootstrap tests. Full implementation details, including ablations and per-query token cost, are provided in App. K. Details for evaluating catastrophic forgetting of previously learned capabilities are provided in App. I. Prompts and code are provided in the supplementary materials.

5.1 EXPERIMENTAL RESULTS

MEMO achieves strong performance across benchmarks. As shown in Table 2, MEMO consistently outperforms the parametric and latent memory baselines across all three benchmarks. Compared with strong non-parametric baselines, MEMO remains competitive across benchmarks. With Qwen2.5-32B-Instruct, MEMO remains competitive with the best-performing non-parametric baselines on BrowseComp-Plus and NarrativeQA, with no statistically significant differences ($p = 0.32$ and $p = 0.28$, respectively). On MuSiQue, MEMO achieves the highest accuracy, a statistically significant improvement ($p = 0.04$) that exceeds the strongest baseline by 2.53 percentage points (pp). Moreover, while MEMO and retrieval-based methods are competitive under the standard corpus setting, MEMO achieves a higher accuracy on BrowseComp-Plus and MuSiQue as the number of negative documents in the target corpora increases (Sec. 5.2).

With Gemini-3-Flash as EXECUTIVE model, the gap between MEMO and retrieval-based baselines closes on BrowseComp-Plus and reverses substantially on NarrativeQA, where MEMO achieves a 20.82 *pp* improvement over the strongest baseline. Notably, NarrativeQA yields the lowest Perfect Retrieval accuracy under both EXECUTIVE model models (App. J), highlighting the challenge of reasoning over and synthesizing connections across long passages even when the relevant evidence is provided. Together, these results suggest that MEMO could be beneficial for tasks that require narrative understanding over long contexts by synthesizing and internalizing connections during training. However, the contrasting results with Qwen2.5-32B-Instruct and Gemini-3-Flash indicate that effectively leveraging this synthesized knowledge depends on EXECUTIVE model's capabilities.

Moreover, on MuSiQue, MEMO and HippoRAG2 achieve similar accuracy for Gemini-3-Flash. However, a granular breakdown by reasoning depth (Tab. 3) indicates that the relative advantage of MEMO increases on complex questions. With Qwen2.5-32B-Instruct as EXECUTIVE model, performance is comparable on 2-hop and 3-hop questions, while MEMO significantly outperforms HippoRAG2 on 4-hop questions by 10.64 *pp* ($p = 0.001$). With Gemini-3-Flash, the difference narrows across reasoning depth, although MEMO retains a 5.42 *pp* advantage on 4-hop questions.

[5]The results here differ from that in (Gutiérrez et al., 2025), which uses Llama3.3-70B-Instruct instead of Qwen2.5-32B-Instruct and token-level F1 score instead of LLM-as-a-judge

[6]Unlike (Ge et al., 2024), only the pretraining stage was carried out to prevent contamination of the test questions. We provide further discussion of ICAE's performance on NarrativeQA in App. R.1.

[7]λ mixes the next-token distributions of EXECUTIVE model ($\lambda = 0$) and the pretrained decoder ($\lambda = 1$). We report results at $\lambda = 0.05$ here and provide more results over different values of λ in App. R.3.

Table 2: Accuracy (%) on BrowseComp-Plus, NarrativeQA, and MuSiQue under two EXECUTIVE models: Qwen2.5-32B-Instruct (Qwen2.5-32B-I) and Gemini-3-Flash (Gemini-3-F). Bold values indicate the best result in each column, excluding Perfect Retrieval. † indicates that the method requires white-box access to EXECUTIVE model, for which we have omitted the Gemini-3-Flash results. * denotes statistically significant improvement ($\alpha = 0.05$) between MEMO and the best performing baseline or vice versa, only for experiments using Qwen2.5-32B-I as EXECUTIVE model.

Method	BrowseComp-Plus		NarrativeQA		MuSiQue	
	Qwen2.5-32B-I	Gemini-3-F	Qwen2.5-32B-I	Gemini-3-F	Qwen2.5-32B-I	Gemini-3-F
Perfect Retrieval	79.67 ± 1.45	88.33	51.42 ± 0.52	60.41	62.83 ± 0.90	73.00
BM25	1.11 ± 0.69	27.00	12.97 ± 0.34	18.77	21.73 ± 0.47	26.40
NV-Embed-V2	50.11 ± 0.19	57.00	$\mathbf{28.44 \pm 0.52}$	32.76	40.00 ± 0.30	50.50
HippoRAG2[5]	$\mathbf{54.22 \pm 0.38}$	66.33	28.44 ± 1.29	31.06	45.77 ± 0.49	60.10
ICAE†[6]	4.33 ± 0.00	-	1.37 ± 0.00	-	17.60 ± 0.00	-
MemoryDecoder†[7]	0.89 ± 0.19	-	4.10 ± 0.00	-	9.57 ± 0.12	-
Cartridges†	1.22 ± 0.69	-	19.68 ± 1.68	-	8.80 ± 0.56	-
MEMO	52.78 ± 0.77	**66.67**	26.85 ± 0.39	**53.58**	$\mathbf{48.30 \pm 1.25}^*$	**60.20**

Overall, these results suggest that MEMO may offer greater benefits over direct retrieval for multi-hop reasoning as the number of reasoning hops increases.

Table 3: Accuracy (%) on MuSiQue by number of reasoning hops. * denotes statistically significant improvement ($\alpha = 0.05$) between MEMO and HippoRAG2 or vice versa, for Qwen runs only.

Method	2-hop (518 questions)		3-hop (316 questions)		4-hop (166 questions)	
	Qwen2.5-32B-I	Gemini-3-F	Qwen2.5-32B-I	Gemini-3-F	Qwen2.5-32B-I	Gemini-3-F
HippoRAG2	53.80 ± 0.33	66.99	43.67 ± 1.44	58.54	24.70 ± 0.49	41.57
MeMo	53.60 ± 1.83	65.64	46.41 ± 3.16	58.23	$35.34 \pm 1.14^*$	46.99

MEMO supports plug-and-play integration. Across the three benchmarks, the same trained MEMORY model can be used with a different EXECUTIVE model, including closed-source LLMs such as Gemini-3-Flash. We provide additional ablations on different model sizes App. M and model families App. N to further demonstrate the plug-and-play capability of MEMO. Since MEMORY model stores new knowledge separately from EXECUTIVE model's parameters, the integration between MEMORY model and EXECUTIVE model would allow for EXECUTIVE model to avoid catastrophic forgetting of its previously learned capabilities (Sec. 5.3).

5.2 ABLATION ON THE AMOUNT OF DISTRACTORS IN THE DATASET

We investigate the robustness of MEMO against two strong retrieval-based baselines, NV-Embed-V2 and HippoRAG2, under increasing levels of corpus distractors, controlled by varying the number of negative documents added to the target corpus as a multiple of the total number of ground-truth evidence documents in each dataset ($N^{\text{dataset}}_{\text{evidence}} = 1{,}775$ for BrowseComp-Plus and $N^{\text{dataset}}_{\text{evidence}} = 2{,}648$ for MuSiQue). The datasets used throughout this paper (and detailed in App. D.1)

Table 4: Accuracy (%) on BrowseComp-Plus and MuSiQue with Qwen2.5-32B-Instruct as EXECUTIVE model. MEMO results are based on Qwen2.5-14B-Instruct and reported at the best training epoch. N is the ratio of the number of negative documents relative to the number of evidence documents $N^{\text{dataset}}_{\text{evidence}}$ present in the target corpus. Δ denotes accuracy difference (%) compared to $0N$.

Method	Dataset	$0N$	$1N$		$2N$	
		Acc. (%)	Acc. (%)	Δ	Acc. (%)	Δ
NV-Embed-V2	BrowseComp-Plus	56.89 ± 0.51	50.11 ± 0.19	↓ 6.78	49.44 ± 0.19	↓ 7.45
	MuSiQue	44.97 ± 0.47	40.00 ± 0.30	↓ 4.97	36.30 ± 0.40	↓ 8.67
HippoRAG2	BrowseComp-Plus	62.33 ± 1.15	54.22 ± 0.38	↓ 8.11	50.78 ± 1.35	↓ 11.55
	MuSiQue	51.80 ± 0.44	45.77 ± 0.49	↓ 6.03	45.37 ± 0.64	↓ 6.43
MEMO	BrowseComp-Plus	53.67 ± 1.15	52.78 ± 0.77	↓ 0.89	52.33 ± 0.88	↓ 1.34
	MuSiQue	50.07 ± 0.81	48.30 ± 1.25	↓ 1.77	47.83 ± 0.57	↓ 2.24

correspond to a ratio of $1\times N^{\text{dataset}}_{\text{evidence}}$; we additionally evaluate at ratio $0\times N^{\text{dataset}}_{\text{evidence}}$ (no distractors) and at ratio of $2\times N^{\text{dataset}}_{\text{evidence}}$. Results in Tab. 4 show that the retrieval-based methods degrade substantially as more distractors are introduced. From $0N$ to $1N$ and $2N$, NV-Embed-V2 incurs drops of 6.78 and 7.45 percentage points (pp) on BrowseComp-Plus, compared with 4.97 and 8.67 pp on MuSiQue. HippoRAG2 similarly incurs drops of 8.11 and 11.55 pp on BrowseComp-Plus, compared with 6.03 and 6.43 pp on MuSiQue. In contrast, MEMO exhibits smaller drops of 0.89 and 1.34 pp on BrowseComp-Plus, compared with 1.77 and 2.24 pp on MuSiQue, demonstrating that it is more robust to increasing levels of corpus distractors. We partly attribute this to the parametric storage of knowledge, where our analysis of cases where a retrieval baseline retrieves no evidence documents in the top-k shows that MEMORY model can still surface relevant information in approximately 30% of cases, even without query decomposition or iterative reasoning.

5.3 CATASTROPHIC FORGETTING OF PREVIOUSLY LEARNED CAPABILITIES

A key design choice of MEMO is to keep the EXECUTIVE model frozen while confining knowledge integration through SFT to a separate MEMORY model. Using Qwen2.5-14B-Instruct, which initializes the MEMORY model, as a proxy for an unmodified EXECUTIVE model, we evaluate its general capabilities before and after training on MuSiQue reflections. Although our primary experiments use the larger Qwen2.5-32B-Instruct as the EXECUTIVE model, prior work finds that increasing model size does not eliminate catastrophic forgetting Luo et al. (2025b).

Table 5: Performance of the 14B MEMORY model model on general capabilities before and during training on the MuSiQue corpus via SFT. Stage 0 denotes the baseline model before weight updates.

Stage	BFCL-v4	GPQA Diamond	TruthfulQA MC2	IFEval
0 (baseline)	92.8	0.354	0.709	0.830
Epoch 1	0.0	0.308	0.457	0.328
Epoch 2	0.0	0.253	0.474	0.301
Epoch 3	0.0	0.278	0.462	0.379

As shown in Tab. 5, training MEMORY model on MuSiQue reflections substantially degrades capabilities not targeted by the training objective. By confining these weight updates to a modular component while keeping EXECUTIVE model frozen, MEMO preserves its existing parameters and capabilities.

6 CONCLUSION

We introduce MEMO, a modular framework for integrating new knowledge into LLMs through a dedicated MEMORY model by coupling how knowledge is synthesized during training with how it is accessed at inference time. By leaving EXECUTIVE model frozen and communicating entirely in natural language, MEMO avoids catastrophic forgetting of EXECUTIVE model while enabling plug-and-play integration with different LLMs, including closed-source models. Empirically, MEMO outperforms the evaluated parametric and latent memory baselines and remains competitive with strong non-parametric retrieval methods across diverse benchmarks. To the best of our knowledge, this is the first demonstration of a parametric knowledge store closing this performance gap with strong retrieval-based approaches while remaining transferable across different EXECUTIVE model models. Notably, with Gemini-3-Flash as EXECUTIVE model, MEMO substantially outperforms retrieval-based baselines on narrative understanding over long-context tasks, while its advantage on 4-hop MuSiQue questions suggests potential benefits over non-parametric retrieval baselines for multi-hop reasoning as the number of reasoning hops increases. Our analyses further show that MEMO is more robust to corpus distractors, outperforming the non-parametric baselines when the target corpus contains twice as many distractor documents as evidence documents. MEMO remains limited by the cost of memory-construction, the scope of our evaluation, and the capacity of MEMORY model to internalize knowledge as corpus size increases (App. B). Future work includes more efficient memory construction, continual integration of evolving and conflicting knowledge, and tighter coordination between EXECUTIVE model and MEMORY model (App. C). Overall, MEMO demonstrates the potential of treating memory itself as a model for modular knowledge integration without retraining the LLM itself.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, design or provide feedback on research methodology or experiments, assisting with translations, cleaning and reformatting datasets, supporting qualitative and thematic data analysis and interpreting results are not applicable to this work. Additionally, we used generative AI tools for creating and editing software code, summarizing or analyzing existing relevant literature, editing the research paper to improve readability. We have reviewed all AI-assisted work. For example, we checked LLM suggested references for potential hallucination through manual combing and checking of citations. Every passage edited or rephrased with AI assistance was checked by the authors for factual accuracy, correct representation of intended claims, and was revised as needed. All AI-generated code was reviewed and tested by the authors for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper proposes methods for large language models and does not involve collection of personal or sensitive information, or the introduction of new datasets containing such information. Our experiments use existing publicly available benchmarks and pretrained language models. Human annotators were used only to assess model responses for calibration of the LLM-based judge and were compensated for their work. The annotation task did not involve personal or sensitive information.

REPRODUCIBILITY STATEMENT

The details of our experimental setup are provided in Sec. 5 and the Appendix. Additionally, the code used in our experiments has been included in the supplementary material, enabling full reproduction of the results reported in this paper.

REFERENCES

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA corpora generation with roundtrip consistency. In *Proc. ACL*, pp. 6168–6173, 2019.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Proc. ICML*, pp. 1067–1077, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *Proc. ICLR*, pp. 85824–85832, 2025.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, Jacob Marks, Edoardo Mosca, Samuel J. Paech, Paul Pak, Rom N. Parnichkun, Alex Quach, Ryan Rogers, Daniela Rus, Nayan Saxena, Bettina Schlager, Tim Seyde, Jimmy T. H. Smith, Aditya Tadimeti, and Neehal Tumma. LFM2 technical report. arXiv:2511.23404, 2025.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in RAG. In *Proc. ACL*, pp. 18228–18258, 2025.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *Proc. ICLR*, pp. 18623–18642, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray,

Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proc. NeurIPS*, pp. 1877–1901, 2020.

Jiaqi Cao, Jiarui Wang, Rubin Wei, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. Memory decoder: A pretrained, plug-and-play memory for large language models. In *Proc. NeurIPS*, pp. 115487–115510, 2025.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. An empirical survey of data augmentation for limited data learning in NLP. *TACL*, 11:191–211, 2023.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proc. ACL*, pp. 22349–22370, 2026.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proc. EMNLP*, pp. 3829–3846, 2023.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proc. ICML*, pp. 10818–10838, 2025.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *JMLR*, 21(1): 3381–3433, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George

Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry (Chun-Ting)Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian (Cindy)Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope,

Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong (Q)Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha

Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao (Joyce)Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric

Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta, Matt Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter, David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren

Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei (Jerry)Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina (Xinyi)Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon,

Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony (Tuân)Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungyong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzębski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, HyunJeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael

Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin (Lucas)Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui (Elena)Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren (Weilun)Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav

Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao

Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein (Yonghao)Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv:2507.06261, 2025.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proc. SIGIR*, pp. 719–729, 2024.

Michal Danilák. Port of Nakatani Shuyo's language-detection library (version from 03/03/2014) to Python. https://github.com/Mimino666/langdetect, 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proc. EMNLP*, pp. 1107–1128, 2024.

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

Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Atri Rudra, James Y. Zou, Azalia Mirhoseini, and Christopher Re. Cartridges: Lightweight and general-purpose long context representations via self-study. In *Proc. ICLR*, pp. 42642–42687, 2026.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *Proc. ICLR*, 2025.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proc. ACL*, pp. 14664–14690, 2024.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *Proc. ICLR*, pp. 2591–2607, 2024.

Carles Gelada, Jacob Buckman, Sean Zhang, and Txus Bach. Scaling context requires rethinking attention. arXiv:2507.04239, 2025.

Google DeepMind. Gemini 3 Flash - model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proc. COLM*, 2024.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Proc. NeurIPS*, pp. 59532–59569, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proc. ICML*, pp. 21497–21515, 2025.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *Proc. ICLR*, pp. 115664–115714, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Proc. COLM*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proc. ICLR*, 2023.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghoon Kim. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35(2): 1–72, 2026.

Gemma Team: Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. arXiv:2503.19786, 2025.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. In *Proc. ICLR*, 2023.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *Proc. ICLR*, 2020.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *TACL*, 6:317–328, 2018.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proc. NeurIPS*, pp. 22199–22213, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proc. SOSP*, pp. 611–626, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Proc. COLM*, 2025.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *Proc. ICLR*, pp. 79310–79333, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS*, pp. 9459–9474, 2020.

Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018.

Jiaen Lin, Jingyu Liu, and Yingbo Liu. Optimizing multi-hop document retrieval through intermediate representations. In *Proc. ACL Findings*, pp. 15798–15809, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. ACL*, pp. 3214–3252, 2022.

Jingyu Liu, Jiaen Lin, and Yong Liu. Tackling the inherent difficulty of noise filtering in RAG. arXiv:2601.01896, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *TACL*, 12:157–173, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2019.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv:2308.08747, 2025a.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025b.

Jiya Manchanda, Laura Boettcher, Matheus Westphalen, and Jasser Jasser. The open source advantage in large language models (LLMs). arXiv:2412.12004, 2025.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. In *Proc. ACL*, pp. 6097–6109, 2019.

Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. Learning to compress prompts with gist tokens. In *Proc. NeurIPS*, pp. 19327–19352, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, pp. 27730–27744, 2022.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proc. EMNLP*, pp. 237–250, 2024.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Proc. ICML*, pp. 48371–48392, 2025.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *Proc. ICLR*, pp. 31932–31951, 2024.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. Training question answering models from synthetic data. In *Proc. EMNLP*, pp. 5811–5826, 2020.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proc. ICLR*, pp. 30988–31043, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pp. 1–16, 2020.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *Proc. COLM*, 2024.

Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. SIGIR*, pp. 232–241, 1994.

Ken Shoemake. Animating rotation with quaternion curves. In *Proc. SIGGRAPH*, pp. 245–254, 1985.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568(C):127063, 2024.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. In *Proc. AAAI*, pp. 8968–8975, 2020.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. arXiv:2307.08621, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *Proc. COLM*, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *TACL*, 10:539–554, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, pp. 6000–6010, 2017.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proc. ACL*, pp. 13484–13508, 2023.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. arXiv:2504.12516, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proc. ICML*, pp. 23965–23998, 2022.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable AI: Environmental implications, challenges and opportunities. In *Proc. MLSys*, pp. 795–813, 2022a.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. arXiv:2402.01364, 2024.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *Proc. ICLR*, 2022b.

Yuchen Xia, Jiho Kim, Yuhan Chen, Haojie Ye, Souvik Kundu, Cong Callie Hao, and Nishil Talati. Understanding the performance and estimating the cost of LLM fine-tuning. In *Proc. IEEE IISWC*, pp. 210–223, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Proc. NeurIPS*, pp. 7093–7115, 2023.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv:2412.15115, 2025.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. Crag-comprehensive rag benchmark. In *Proc. NeurIPS*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proc. EMNLP*, pp. 2369–2380, 2018.

Wei Jie Yeo, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. Self-training large language models through knowledge detection. In *Proc. EMNLP Findings*, pp. 15033–15045, 2024.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *Proc. ICLR*, pp. 29862–29883, 2024.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In *Proc. ICML*, pp. 57755–57775, 2024.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. MemGen: Weaving generative latent memory for self-evolving agents. In *Proc. ICLR*, pp. 22555–22588, 2026a.

Longteng Zhang, Xiang Liu, Zeyu Li, Xinglin Pan, Peijie Dong, Ruibo Fan, Rui Guo, Xin Wang, Qiong Luo, Shaohuai Shi, and Xiaowen Chu. Dissecting the runtime performance of the training, fine-tuning, and inference of large language models. arXiv:2311.03687, 2023.

Ze Yu Zhang, Arun Verma, Finale Doshi-Velez, and Bryan Kian Hsiang Low. Prompting the unknown: Understanding response uncertainty in large language models. In *Proc. ACL Findings*, pp. 30956–30984, 2026b.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv:2311.07911, 2023.

A IMPACT STATEMENT

MEMO advances the ability of LLMs to internalize knowledge over large, domain-specific corpora without requiring access to LLM parameters, lowering the barrier for deploying capable AI systems in knowledge-intensive domains such as law, medicine, and scientific research. By enabling plug-and-play integration with any LLM (including closed-source LLMs), MEMO democratizes access to powerful knowledge integration capabilities that would otherwise require significant computational resources or white-box LLM access. At the same time, this accessibility introduces dual-use concerns, as the same capability that enables beneficial applications could be used to internalize misinformation, proprietary data without authorization, or harmful content at scale. Additionally, as MEMO reduces reliance on explicit retrieval, it may obscure the provenance of retrieved information, making it harder to attribute the sources underlying a model's responses. We encourage future work to investigate attribution mechanisms and access controls for memory-based systems, and urge practitioners to carefully consider the nature of the documents used to train MEMORY model.

B LIMITATIONS

MEMO incurs an upfront training cost for each new corpus, and performance may vary across domains, document types, or LLM families beyond those covered in our experiments. Furthermore, MEMO is currently designed for static target corpora, and its data synthesis pipeline relies on predefined document groupings to identify related content during reflection generation. While we provide preliminary explorations of continual knowledge integration in App. P and discuss implications of current grouping strategies in App. D.2, extending MEMO to evolving corpora and developing more general strategies for automatically identifying related content remain important directions for future work. Additionally, the performance of MEMO is inherently bounded by the representational capacity of MEMORY model to internalize the target corpus. Although our experiments do not reveal clear signs that MEMORY model has reached its capacity limit, we hypothesize that sufficiently large or information-dense corpora will exceed what a fixed-size MEMORY model can correctly compress and represent.

C FUTURE WORK

We outline several directions for future work.

Scope. Exploring dynamical updates for the MEMORY model is another direction for future work. We include preliminary findings in this direction in App. P, but leave the principled study of how to handle dynamic updates to future work. A complementary direction would also involve scaling MEMORY model with corpus size and developing more effective model merging strategies for reducing per-corpus training costs (App. P) are promising directions.

Data synthesis pipeline. The data synthesis pipeline is computationally expensive, with Step 5 in Alg. 1 scaling quadratically at $O(k \cdot C^2 \cdot Q^2)$, and reducing this cost remains an open problem. Other interesting directions include exploring different chunking strategies and their associated tradeoffs (App. D.1) and the use of a different GENERATOR model and its effects on downstream performance (App. E.1).

Training methods. Current LoRA configurations are better suited to specific architectures, including per-architecture tuning of rank and learning rate. This warrants further investigation (App. O). Other post-training methods, such as Reinforcement Learning (Sutton & Barto, 1998), have also been shown to be effective in improving model task performance (Lambert et al., 2025), and applying such methods to MEMORY model training warrants future investigation.

Inference pipeline. A more systematic study of the interaction between EXECUTIVE model reasoning capability and MEMORY model size (App. M), as well as the optimal interaction budget at each stage and EXECUTIVE model selection (App. K.3, App. K.5), are other promising future directions.

D PREPARATION OF DATASETS

D.1 CHUNKING STRATEGY AND SUBSET SELECTION OF NEGATIVE DOCUMENTS

Corpus construction. Extending from our description in Sec. 5, we distinguish between two types of documents[8]: evidence documents, which contain information relevant to answering a given question, and negative documents, which are distractors and serve as noise. For BrowseComp-Plus, we used $1{,}775$ unique evidence documents and $1{,}766$ unique negative documents (after removal of non-English documents), yielding $3{,}541$ documents in total for 300 questions. For MuSiQue, we used $2{,}648$ documents for each of the evidence and negative documents, yielding $5{,}296$ documents in total for 1000 questions. For NarrativeQA, we selected 10 documents from the validation split of the original dataset, obtaining a total of 293 questions after removing 1 exact duplicate. Since each document is a story in itself, there are no negative documents.

Algorithm 1 Reflection QA Dataset Synthesis Pipeline from Target Corpus

Require: Corpus $\mathcal{D}$, GENERATOR model $\mathcal{M}_{\text{gen}}$, document groups $\mathcal{G} = \{G_1, \ldots, G_R\}$ with $G_i \subseteq \mathcal{D}$

$\mathcal{Q}_{\text{final}} \leftarrow \emptyset$
for all document $d \in \mathcal{D}$ **do**
  $C \leftarrow \text{Chunk}(d)$ ▷ Segment into chunks
  $\mathcal{Q}^d_{\text{ver}} \leftarrow \emptyset$
  for all chunk $c \in C$ **do**
    $\mathcal{Q}_{\text{dir}}, \mathcal{Q}_{\text{indir}} \leftarrow \mathcal{M}_{\text{gen}}(c)$ ▷ Step 1: Direct and indirect extraction
    $\mathcal{Q}_{\text{raw}} \leftarrow \mathcal{Q}_{\text{dir}} \cup \mathcal{Q}_{\text{indir}}$ ▷ Step 2a: Merge direct and indirect
    $\mathcal{Q}_{\text{mrg}} \leftarrow \mathcal{M}_{\text{gen}}(\mathcal{Q}_{\text{raw}})$ ▷ Step 2b: Consolidate related pairs
    $\mathcal{Q}_{\text{con}} \leftarrow \mathcal{Q}_{\text{raw}} \cup \mathcal{Q}_{\text{mrg}}$ ▷ Step 2c: Full merge set
    $\mathcal{Q}_{\text{ver}} \leftarrow \mathcal{M}_{\text{gen}}(\mathcal{Q}_{\text{con}}, c)$ ▷ Step 3: Verify self-containment; rewrite or discard
    $\mathcal{Q}^d_{\text{ver}} \leftarrow \mathcal{Q}^d_{\text{ver}} \cup \mathcal{Q}_{\text{ver}}$
  end for
  $\mathcal{Q}^d_{\text{ent}} \leftarrow \mathcal{M}_{\text{gen}}(\mathcal{Q}^d_{\text{ver}})$ ▷ Step 4: Entity-surfacing pairs
  $\mathcal{Q}_{\text{final}} \leftarrow \mathcal{Q}_{\text{final}} \cup \mathcal{Q}^d_{\text{ver}} \cup \mathcal{Q}^d_{\text{ent}}$
end for
for all $G_i \in \mathcal{G}$ **do**
  $\mathcal{Q}_{\text{cross}} \leftarrow \mathcal{M}_{\text{gen}}\Big(\bigcup_{d \in G_i} \big(\mathcal{Q}^d_{\text{ver}} \cup \mathcal{Q}^d_{\text{ent}}\big)\Big)$ ▷ Step 5: Cross-document synthesis
  $\mathcal{Q}_{\text{final}} \leftarrow \mathcal{Q}_{\text{final}} \cup \mathcal{Q}_{\text{cross}}$
end for
return $\mathcal{Q}_{\text{final}}$

Chunking strategy. As shown in Tab. 6, NarrativeQA full documents span the $32{,}769$–$131{,}072$ token range with a median length of $65{,}925$ tokens, reflecting the long-form nature of the source novels. Processing such documents without chunking risks reduced coverage of extractable QA pairs in Step 1 of Alg. 1, as attention quality is known to deteriorate over longer contexts (Hsieh et al., 2024). We therefore chunk NarrativeQA documents using a fixed sliding window of $6{,}400$ words with a 640-word overlap (10% overlap ratio), yielding 75 chunks concentrated in the $4{,}097$–$16{,}384$ token range and accounting for 96% of all chunks, with a median group size of 7 per document as shown in Tab. 7. Unlike NarrativeQA, MuSiQue documents are compact with 99.70% falling below 512 tokens, and each MuSiQue document is treated as a single chunk.

BrowseComp-Plus documents are also treated as a single chunk. The time complexity of Step 5 in Alg. 1 is $O(k \cdot C^2 \cdot Q^2)$, where $k = n_{\text{group}}$ is the number of groups, $C = |G_i|$ is the number of participating chunks per group, and $Q = \bar{Q}_i$ is the average number of QA pairs extracted per chunk. Since chunking increases C, pipeline costs at Step 5 scale quadratically as the number of chunks per group increases. Given that only 2.93% of BrowseComp-Plus documents exceed $32{,}768$ tokens, the majority of documents fit within a single chunk, making the cost of chunking difficult to justify. We therefore opted against chunking in favor of lower pipeline cost, and leave a systematic evaluation of chunking strategies and related tradeoffs to future work.

[8]Note that for BrowseComp-Plus, the gold documents are a subset of the evidence documents.

Table 6: Token length distribution across corpora at the *chunk level*, where n represents the total number of individual chunks processed by Alg. 1. Each entry reflects the token count of a single text chunk. Statistics for NarrativeQA are reported before and after chunking.

Token Range	**BrowseComp-Plus** ($n = 3{,}541$)	**NarrativeQA Full Docs** ($n = 10$)	**NarrativeQA Chunks** ($n = 75$)	**MuSiQue** ($n = 5{,}296$)
0–512	606 (17.11%)	0 (0.00%)	0 (0.00%)	5,280 (99.70%)
513–1,024	591 (16.69%)	0 (0.00%)	0 (0.00%)	16 (0.30%)
1,025–2,048	746 (21.07%)	0 (0.00%)	1 (1.33%)	0 (0.00%)
2,049–4,096	598 (16.89%)	0 (0.00%)	2 (2.67%)	0 (0.00%)
4,097–8,192	428 (12.09%)	0 (0.00%)	36 (48.00%)	0 (0.00%)
8,193–16,384	323 (9.12%)	0 (0.00%)	36 (48.00%)	0 (0.00%)
16,385–32,768	145 (4.09%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
32,769–65,536	56 (1.58%)	5 (50.00%)	0 (0.00%)	0 (0.00%)
65,537–131,072	20 (0.56%)	5 (50.00%)	0 (0.00%)	0 (0.00%)
$> 131{,}072$	28 (0.79%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Min tokens	14	32,804	1,943	23
Median tokens	1,756	65,925	8,158	105
Mean tokens	7,192	66,324	8,713	123
p_{95} **tokens**	20,330	119,267	11,266	270
Max tokens	1,235,897	119,267	12,104	828

Table 7: Distribution of document group sizes across datasets, where group size denotes the number of chunks associated with a single question or document. For BrowseComp-Plus and MuSiQue, each question is associated with a subset of chunks drawn from the corpus, and group size represents the number of chunks per *question*. For NarrativeQA, each subset of chunks is derived from the original document used for multiple questions, and group size represents the number of chunks per *document*.

Document Group Size Range	**BrowseComp-Plus** ($n_{\text{group}} = 300$)	**NarrativeQA Chunks** ($n_{\text{group}} = 10$)	**MuSiQue** ($n_{\text{group}} = 1{,}000$)
0–2	2 (0.67%)	0 (0.00%)	0 (0.00%)
3–4	14 (4.67%)	3 (30.00%)	518 (51.80%)
5–8	78 (26.00%)	4 (40.00%)	482 (48.20%)
9–16	159 (53.00%)	3 (30.00%)	0 (0.00%)
>16	47 (15.67%)	0 (0.00%)	0 (0.00%)
Min group size	2	3	4
Median group size	12	7	4
Mean group size	11.8	7.5	5.3
p_{95} **group size**	20	16	8
Max group size	23	16	8

Subset selection of negative documents. We include only a subset of negative documents for BrowseComp-Plus and MuSiQue due to computational constraints arising from the quadratic scaling of Step 5. As reported in Tab. 7, BrowseComp-Plus currently has a mean group size of 11.8 and a maximum of 23, while MuSiQue has a mean group size of 5.3 and a maximum of 8. Incorporating all available negative documents, which average 78 per question (up to 197) for BrowseComp-Plus and 17 per question (up to 18) for MuSiQue, would cause the group size to increase substantially. Given the quadratic dependence on C in Step 5, this would result in a prohibitive increase in pipeline

cost for BrowseComp-Plus ($k = 300$) and MuSiQue ($k = 1{,}000$). Hence, we opted to only include up to $2\times N_{\text{evidence}}^{\text{dataset}}$ number of negative documents for each question in the corpus.

D.2 Discussion for grouping strategy used in **Step 5** of data synthesis pipeline

Motivation and grouping strategy. **Step 5** synthesizes cross-document reflections over groups of potentially related documents. For NarrativeQA, each story naturally forms a group comprising its constituent chunks, independent of the evaluation questions. For BrowseComp-Plus and MuSiQue, we instead construct each group from the evidence and negative documents associated with an evaluation question. This grouping avoids having the GENERATOR model process the combinatorial space of document combinations, an extremely costly procedure, given that many such document groups comprise unrelated documents and yield few useful cross-document QA pairs.

To quantify this effect, we conducted a small-scale experiment on 126 document groups, attained by randomly regrouping BrowseComp-Plus documents. Applying the same **Step 5** generation procedure, 12 pairs contain documents belonging to the same original question-associated group, producing 8,496 QA pairs (708 per document pair). In contrast, the remaining 114 pairs contain documents from different question-associated groups and produce only 1,047 QA pairs (9.2 per document pair), a relative yield of only 1.3% for a single LLM call to the GENERATOR model.

Post-hoc analysis. We next test whether QA pairs generated from documents within the question-associated groups can be distinguished from those generated using unrelated documents, based solely on information from the source documents. For each question, we compare QA pairs generated from evidence–evidence (E–E) and evidence–negative (E–N) document combinations against evidence–random (E–R) combinations, where the random document is sampled outside that question's evidence-and-negative document group.

We construct a question-agnostic corpus-grounding filter that extracts proper nouns from each generated answer and checks whether any extracted entity occurs in at least two of its source documents. To reduce incidental matches from common entities (e.g., United States or New York), we additionally require the entity to exceed a corpus-level inverse document frequency (IDF) threshold, where $\text{IDF}(e) = \log(N/df(e))$. We use $\text{IDF} > 4$ for BrowseComp-Plus and $\text{IDF} > 1$ for MuSiQue; the less restrictive threshold accommodates the shorter MuSiQue passages, whose cross-document connections can involve more common bridging entities. These thresholds are selected post hoc for this diagnostic analysis. Notably, the filter operates only on the generated answer and raw source documents and has no access to evaluation questions, answers, evidence labels, or group identities.

Table 8: Percentage of **Step 5** QA pairs retained by post-hoc, question-agnostic corpus-grounding filters. E–E and E–N denote evidence–evidence and evidence–negative combinations associated with the same question; E–R pairs an evidence document with a randomly sampled document outside that question's evidence-and-negative group.

Dataset	E–E	E–N	E–R
BrowseComp-Plus	94.1%	81.7%	3.7%
MuSiQue	78.1%	77.3%	5.6%

As shown in Tab. 8, the filters retain 77.3–94.1% of E–E and E–N generations, but only 3.7% and 5.6% of E–R generations on BrowseComp-Plus and MuSiQue, respectively. Notably, the high retention of E–N generations suggests that the separation is not simply between gold evidence and non-evidence documents: negative documents associated with the same question are also substantially more likely than documents outside the group to support the corpus-grounding criterion.

Interpretation and limitation. Together, these analyses support the computational motivation for the grouping strategy used in the main paper, in which document combinations outside the provided groups yield substantially lower generation yield, and most resulting generations can be rejected using question-agnostic corpus signals. The question-derived grouping can therefore be viewed as an efficient pre-selection mechanism that avoids many GENERATOR model calls that a more exhaustive generate-then-filter procedure would subsequently reject.

However, we acknowledge that these post hoc analyses do not establish equivalence with a fully question-agnostic synthesis pipeline. The filters and their thresholds are selected retrospectively, and we do not retrain MEMORY model using candidate groups discovered without question information.

We therefore conservatively regard the BrowseComp-Plus and MuSiQue results involving **Step 5** as an *upper-bound setting* when question-associated grouping information is available.

E Details and discussion for data generation pipeline

E.1 Implementation details

We use Qwen2.5-32B-Instruct (Yang et al., 2025) as the GENERATOR model, served via vLLM (Kwon et al., 2023) with YaRN RoPE scaling (Su et al., 2024; Peng et al., 2024) to support a 131K-token context window during long-context generation.

Notably, the GENERATOR model and the EXECUTIVE model need not use the same underlying LLM. While Qwen2.5-32B-Instruct also serves as one of the EXECUTIVE model in our primary experiments, we also evaluate Gemini-3-Flash as the EXECUTIVE model using the same MEMORY model trained on the reflections generated by Qwen2.5-32B-Instruct. Investigating the effect of the GENERATOR model's capability on reflection quality and downstream performance is left to future work.

E.2 Ablation of data synthesis steps

We experiment with the data generation pipeline to show the importance of each step. We perform a leave-one-out (LOO) ablation for each step of data synthesis and train the model on the synthesized QA pairs generated. Results are reported in Tab. 9 on the NarrativeQA and MuSiQue datasets using Qwen2.5-32B-Instruct as the EXECUTIVE model and Qwen2.5-1.5B-Instruct as the MEMORY model.

Table 9: LOO ablation accuracy at best performing Qwen2.5-1.5B-Instruct epoch across datasets. Data ratio indicates the number of QA pairs retained relative to the baseline. For each step removed, the Qwen2.5-1.5B-Instruct was retrained, and we report the mean $\pm$ std. dev. over 3 runs at the same training epoch as the baseline.

	NarrativeQA		**MuSiQue**	
Ablation	**Data Ratio**	**Accuracy (%)**	**Data Ratio**	**Accuracy (%)**
Baseline (all steps)	$1.000\times$	24.00 ± 0.20	$1.000\times$	42.90 ± 1.39
Step 1a removed	$0.434\times$	20.48 ± 0.90	$0.381\times$	30.00 ± 0.17
Step 1b removed	$0.598\times$	22.98 ± 1.04	$0.651\times$	37.33 ± 0.25
Step 2 removed	$0.739\times$	24.69 ± 1.10	$0.621\times$	37.10 ± 1.76
Step 3 removed	$2.078\times$	28.90 ± 0.86	$1.128\times$	41.70 ± 0.78
Step 4 removed	$0.378\times$	23.21 ± 1.56	$0.501\times$	39.10 ± 0.02
Step 5 removed	$0.002\times$	6.37 ± 0.39	$0.195\times$	24.17 ± 0.25

Step 5 (Cross-document synthesis) is the most critical component of the pipeline. Its removal causes accuracy to collapse to 6.37% and 24.17% on the respective NarrativeQA and MuSiQue datasets against baseline scores of 24.00% and 42.90%, accompanied by a near-total loss of training data (i.e., $0.002\times$ and $0.195\times$ retention). As described in Sec. 4.1, Step 5 constructs $\mathcal{Q}_{\text{cross}}$ by synthesizing connections across documents and across chunks of a single long document, and constitutes a substantial source of the final training data. When removed, the dataset retains only $0.002\times$ and $0.195\times$ of the QA pairs on NarrativeQA and MuSiQue, respectively. This reflects the overall contribution of Step 5 to the pipeline, including both the cross-source reflections it introduces and the additional training signal they provide.

An interesting anomaly arises with Step 2 and Step 3 where their removal does not consistently hurt performance and improves accuracy on NarrativeQA. Step 2 merges related QA pairs from a single document chunk into multi-fact questions by identifying commonalities such as shared entities, overlapping time periods, and sequential events. For MuSiQue, these commonalities reflect genuine knowledge connections that directly resemble the multi-hop factual reasoning the benchmark evaluates, such that removing Step 2 eliminates a large fraction of useful training pairs, leading to a drop in accuracy from 42.90% to 37.10%. For NarrativeQA, however, the same consolidation patterns

operate on superficial narrative co-occurrences rather than meaningful knowledge connections. The predominant commonality categories are event or scene groupings that NarrativeQA does not evaluate, and entity co-occurrence patterns that are trivially satisfied given the pervasive presence of central characters across scenes. Removing Step 2 eliminates these low-quality pairs, leading to the marginal accuracy improvement from 24.00% to 24.69%.

Removing Step 3 retains more data than the baseline ($2.078\times$ and $1.100\times$ for NarrativeQA and MuSiQue respectively), yet the effect on performance diverges. For MuSiQue, performance drops from 42.90% to 41.78%, whereas for NarrativeQA, performance improves from 24.00% to 28.90%. Step 3 applies a self-containment filter that rewrites or discards pairs whose questions cannot be understood without access to the source chunk. For MuSiQue, violations are predominantly localized and shallow, making them amenable to filtering; the proposed filter effectively identifies and removes defective pairs. For NarrativeQA, long-form narrative text frequently contains pronouns and temporal references that span many paragraphs, which are structural features of the domain rather than fixable defects. This causes the rewriting loop to introduce substitute unrelated content and corrupt the pairs produced by earlier steps. Removing Step 3 for NarrativeQA therefore avoids this domain-induced corruption and retains the original pairs intact, explaining both the data retention ratio increase and the accuracy improvement. This suggests that Step 3 is most beneficial when applied to domains where self-containment violations are well-defined and resolvable.

The remaining steps follow a consistent trend: removing Step 1a, Step 1b, or Step 4 reduces both data volume and accuracy across both datasets, confirming that each step contributes a distinct and meaningful role to the final training corpus quality.

E.3 Additional steps considered but excluded

Three additional steps were considered but ultimately excluded from the pipeline. These include paraphrasing (Ovadia et al., 2024), increasing the number of sampling trials at Step 1 of Alg. 1, and a targeted fill whereby $\mathcal{M}_{\text{gen}}$ reviews the generated QA pairs and rewrites them to incorporate additional missed information. Paraphrasing was excluded as the scale of generated pairs already provides sufficient coverage (≈600k–1.6M across the three datasets, see Table 11), and the potential gains were outweighed by the additional computational overhead. Increasing sampling trials proved unreliable, as additional trials did not consistently extract facts that the initial pass had failed to extract. The targeted fill similarly offered limited gains, where appending the existing QA pairs as context to prompt a revision only lengthens the context when the model had already failed to extract a fact from the original chunk, likely exacerbating attention further degradation over long inputs (Hsieh et al., 2024) and making retrieval of relevant information less reliable at inference time.

F Implementation details for training Memory model

We train Memory model, which is initialized from Qwen2.5-14B-Instruct for up to 3 epochs with fused AdamW (Loshchilov & Hutter, 2019) and DeepSpeed 2 (Rajbhandari et al., 2020) at learning rate 2×10^{-5}. We select the best performing epoch to be reported in the paper, for per-epoch performance, refer to App. L.

Tab. 10 reports the hyperparameter settings used in the training performed on H100 and H200 GPUs. Tab. 11 summarizes the effective batch size for each dataset.

G Compute resources

All experiments were conducted using NVIDIA H200 GPUs. We report compute cost in GPU-hours.

Data generation. Generating the full reflection dataset for BrowseComp-Plus, NarrativeQA, and MuSiQue took approximately 440, 200, and 150 GPU-hours, respectively.

Training. Memory model (Qwen2.5-14B-Instruct) training for a single run BrowseComp-Plus, NarrativeQA, and MuSiQue took approximately 320, 150, and 90 GPU-hours, respectively.

Table 10: MEMORY model SFT training configuration

Parameter	Value
Optimizer	Fused AdamW
Gradient checkpointing	True
Learning rate (LR)	2×10^{-5}
Num of training epochs	3
LR scheduler type	Constant with warmup
Warmup ratio	0.05
Weight decay	0.01
Max gradient norm	1.0
Max sequence length	8096
Precision	BF16
Attention implementation	Flash Attention 2

Table 11: Effective batch sizes and number of QA pairs used. $N = N_{\text{evidence}}^{\text{dataset}}$ denotes the number of evidence documents in each dataset, and the negative ratio indicates the number of negative documents relative to N. NarrativeQA.1 and NarrativeQA.2 are independent subsets partitioned from the original that were used for model merging.

Dataset	Target num of questions	Negative ratio	Num of QA pairs	Effective batch size
BrowseComp-Plus	300	$0N$	1,286,328	512
BrowseComp-Plus	300	$1N$	2,968,214	1024
BrowseComp-Plus	300	$2N$	5,615,735	2048
NarrativeQA	293	–	1,276,676	512
NarrativeQA.1	146	–	635,009	256
NarrativeQA.2	147	–	641,667	256
MuSiQue	1,000	$0N$	265,400	128
MuSiQue	1,000	$1N$	664,762	256
MuSiQue	1,000	$2N$	1,286,190	512

H MODEL TRAINING DISCUSSION

We considered three training paradigms: CPT, SFT, and LoRA-based SFT. CPT was excluded as it risks degrading instruction-following capability (Wu et al., 2024), which is critical for downstream QA evaluation. Full SFT was selected as it directly optimizes for the target task while preserving alignment (Ouyang et al., 2022). LoRA-based SFT serves as a parameter-efficient alternative and we include a comparison to these training methods in App. O.

Beyond the primary scope of this work, we consider a practical desideratum of knowledge integration systems: the ability to incorporate new corpora incrementally. We provide preliminary experiments exploring incremental knowledge integration in App. P.

I IMPLEMENTATION DETAILS FOR EVALUATION ON CATASTROPHIC FORGETTING OF PREVIOUSLY LEARNED CAPABILITIES

Updating model weights through continued pre-training or SFT can lead to catastrophic forgetting, where previously acquired capabilities degrade as the model adapts to new training data (Qi et al., 2024; Luo et al., 2025b). We evaluate this effect on Qwen2.5-14B-Instruct, which initializes the MEMORY model, during SFT on the synthesized reflections from the MuSiQue corpus.

We evaluate the model before training and after each epoch on four benchmarks covering distinct capabilities: BFCL-v4 for tool calling (Patil et al., 2025), GPQA Diamond for knowledge and reasoning (Rein et al., 2024), TruthfulQA MC2 for truthfulness (Lin et al., 2022), and IFEval for instruction following (Zhou et al., 2023). For BFCL-v4, we report the average task accuracy over the

`Parallel`, `simple_python`, and `multiple` tasks. For IFEval, we report the average accuracy over its four instruction- and prompt-level loose and strict metrics.

J VALIDATING EVALUATION DATASET SUITABILITY

Table 12: Performance gap between no context and perfect retrieval across datasets and EXECUTIVE models.

	Qwen2.5-32B-Instruct			Gemini-3-Flash		
	BrowseComp-Plus	**NarrativeQA**	**MuSiQue**	**BrowseComp-Plus**	**NarrativeQA**	**MuSiQue**
No Context	0.44 ± 0.20	4.55 ± 0.20	9.87 ± 0.06	1.33	26.62	41.80
Perfect Retrieval	79.67 ± 1.45	51.42 ± 0.52	62.83 ± 0.90	88.33	60.41	73.00

To assess the suitability of the evaluation datasets for EXECUTIVE model and whether the EXECUTIVE model has memorized answers from training data, we evaluate performance both without any context (**No Context**) and with evidence documents provided (**Perfect Retrieval**), the latter serving as an empirical upper bound that assumes perfect retrieval of relevant documents.

As shown in Tab. 12, the large disparity in performance between No Context and Perfect Retrieval confirms that these datasets require access to evidence documents to achieve correct answers, validating their suitability for evaluating MEMO.

MuSiQue yields the highest No Context scores, likely because its Wikipedia-grounded questions fall within models' parametric knowledge. NarrativeQA proves most challenging as it achieves the lowest Perfect Retrieval scores across both EXECUTIVE models, reflecting the demand for careful reasoning over full-length books and movie scripts. BrowseComp-Plus yields the largest disparity between No Context and Perfect Retrieval, with near-zero No Context performance but strong recovery when evidence documents are provided. These findings confirm that EXECUTIVE model heavily relies on evidence documents across all three datasets to perform well.

K EVALUATION DETAILS

K.1 INFERENCE-TIME IMPLEMENTATION DETAILS

Similar to the implementation for GENERATOR model App. E.1, when Qwen2.5-32B-Instruct is the EXECUTIVE model, it is served with a 131k-token context window via vLLM with YaRN RoPE scaling. The current temperature settings are described in Tab. 13. Stage 1 only has a budget of 1 interaction, Stage 2 has a budget of 7 interactions, Stage 3 has a budget of 8 interactions.

Table 13: Temperature Configuration of each Stage from Sec. 4.3

Stage	**Model**	**Temperature Value**	**Intent**
Evaluation Stage 1 – Grounding	EXECUTIVE model	0.4	Moderate exploration to generate diverse but focused sub-questions
Evaluation Stage 1 – Grounding	MEMORY model	0.1	Near-deterministic to ensure stable, consistent grounding answers
Evaluation Stage 2 – Entity identification	EXECUTIVE model	0.4	Moderate exploration to identify varied candidate entities without excess noise
Evaluation Stage 2 – Entity identification	MEMORY model	0.1	Near-deterministic to produce reliable entity-targeted answers
Evaluation Stage 3 – Answer Seeking	EXECUTIVE model	1.0	High exploration to maximally diversify sub-questions once the entity is confirmed
Evaluation Stage 3 – Answer Seeking	MEMORY model	0.3	Slightly relaxed determinism to allow nuanced answers while remaining consistent
Final Synthesis	EXECUTIVE model	0.3	Low temperature to produce a consistent final answer

Beyond what is described in Sec. 4.3, there are additional helper functions that help manage failure modes across Stage 2 and Stage 3. Within Stage 2, the uncertain answer streak tracker is called at the start of every entity-pinning interaction, and its output is passed directly into the entity-pinning prompt, giving the EXECUTIVE model a live view of which candidates the MEMORY model has repeatedly failed to corroborate. This allows the EXECUTIVE model to continuously re-rank and

Table 14: Helper Functions for Stage 2 and 3 of the Evaluation Pipeline

Function	Stage	Intent
Track uncertain answer streaks	Stage 2	Maintains a running tally of how many unanswerable questions each candidate entity has accumulated across Stage 2, allowing the EXECUTIVE model to progressively prioritize candidates that the MEMORY model consistently cannot corroborate
Select the best candidate	Stage 2	Fallback bridge from Stage 2 to Stage 3 when entity pinning ends without a confirmed entity. Selects the highest EXECUTIVE model-ranked candidate, with ties broken by the order in which the MEMORY model produced the candidates
Entity pivot correction	Stage 3	Allows the pipeline to self-correct mid Stage 3 if the Stage 2 entity proves incorrect. When the EXECUTIVE model nominates a different entity, the confirmed entity is overwritten and marked as unconfirmed so subsequent turns are aware it was not pinned through the full Stage 2 process

prune the candidate pool as evidence accumulates. When Stage 2 concludes without a confirmed entity, either because the EXECUTIVE model explicitly exhausts its options or the interaction budget is reached, the best candidate selector acts as the bridge into Stage 3 by returning the top-ranked candidate. In cases where multiple candidates share the highest rank, the first candidate in the order produced by EXECUTIVE model is selected. In both cases, the downstream Stage 3 prompt is informed of whether the entity was formally confirmed or merely a best guess. Finally, if Stage 3 reveals that the Stage 2 entity was incorrect due to persistent MEMORY model failures, the entity pivot mechanism allows the EXECUTIVE model to nominate a replacement entity mid-stage. The confirmed entity is then overwritten and marked as unconfirmed, ensuring subsequent stages treat it with appropriate uncertainty rather than the confidence of a fully pinned entity.

K.2 EVALUATING JUDGE SUITABILITY FOR LLM-AS-A-JUDGE

For judge calibration, we randomly sample 70 responses from each of the three datasets, yielding 210 responses for human evaluation. Three human annotators, all fluent in English and with a college education, independently assess each response. Both the human annotators and the LLM judge are provided with the question, its ground-truth answer(s), and the model response. Human annotators additionally have access to the Internet to consult the full source passage when necessary for interpretation. Following independent annotation, the annotators discuss cases of disagreement and may revise their judgments after discussion, although consensus is not enforced.

The resulting post-discussion annotations exhibit high inter-rater agreement, with a Fleiss' κ of 0.987. Gemini-2.5-Flash-Lite agrees with all three human annotators on 95.2% of responses. Taking the majority vote across the three human annotations as the human reference judgment, Gemini agrees with the human consensus on 95.7% of responses, corresponding to a Cohen's κ of 0.914.

K.3 ABLATIONS ON EVALUATION SETUP

To justify our structured multi-turn evaluation design, we compare against two baselines: a single-turn setup and an *unstructured* multi-turn setup; in both cases, the same trained MEMORY model is used and EXECUTIVE model is held fixed. Results are reported in Tab. 15.

In a single-turn interaction, EXECUTIVE model first determines whether the question requires external memory retrieval, and if so, decomposes it into a set of sub-questions (Stage 1, Sec. 4.3) and poses them all simultaneously to MEMORY model. MEMORY model responds to each sub-question independently, and responses indicating uncertainty are discarded before the remaining answers are passed to EXECUTIVE model for final synthesis. This design requires EXECUTIVE model to commit to its full sub-question set before observing any responses, preventing it from reformulating uninformative queries, following up on answers that introduce new candidate entities, or correcting retrievals that are incomplete, contradictory, or anchored to the wrong entity. This is a fundamental limitation that is reflected in its consistently lowest performance across all three datasets (Tab. 15).

A natural extension of the single-turn setting is an unstructured multi-turn interaction, where EXECUTIVE model examines the responses from MEMORY model and decides whether sufficient information has been gathered, or whether additional retrieval rounds are needed (Stage 3, Sec. 4.3). In this

Table 15: MEMO accuracy results with Qwen2.5-32B-Instruct as EXECUTIVE model and Qwen2.5-14B-Instruct as MEMORY model across evaluation setups. The best performing epoch was used in comparison across all 3 setups, with mean ± std. dev. reported across 3 runs. Bold results indicate the best performing results in the column.

Evaluation setup	BrowseComp-Plus accuracy	NarrativeQA accuracy	MuSiQue accuracy
Single turn evaluation	28.89 ± 2.04	24.80 ± 0.20	37.57 ± 1.15
Unstructured multi-turn evaluation (15 turns)	45.11 ± 0.84	26.73 ± 2.17	40.13 ± 1.12
Unstructured multi-turn evaluation (50 turns)	46.33 ± 0.67	27.19 ± 0.71	40.57 ± 0.31
Structured multi-turn evaluation (7 Entity Identification turns + 8 Answer seeking turns)	$\mathbf{52.78 \pm 0.77}$	26.85 ± 0.39	$\mathbf{48.30 \pm 1.25}$
Structured multi-turn evaluation (7 Entity Identification turns + 15 Answer seeking turns)	50.22 ± 0.51	$\mathbf{27.76 \pm 0.20}$	47.57 ± 0.95

setting, EXECUTIVE model is presented with the full history of question-answer pairs and prompted to either synthesize a final answer or generate a new batch of sub-questions targeting remaining gaps, repeating for up to T interactions. While iterative retrieval yields clear improvements over the single-turn baseline, performance plateaus quickly when increasing from 15 to 50 interactions (45.11 ± 0.84 to 46.33 ± 0.67 on BrowseComp-Plus, 26.73 ± 2.17 to 27.19 ± 0.71 on NarrativeQA, and 40.13 ± 1.12 to 40.57 ± 0.31 on MuSiQue), suggesting that iterative retrieval alone is insufficient.

The structured multi-turn setup (see Sec. 4.3 and App. K.1) outperforms the unstructured multi-turn baseline, with 8 answer-seeking interactions achieving the strongest overall performance for BrowseComp-Plus and MuSiQue. This is consistent with the expectation that explicit entity identification is well-suited to the multi-hop reasoning demands of these datasets. NarrativeQA, which tests discourse understanding over long documents, has the unstructured 15- and 50-interaction baselines (26.73 ± 2.17 and 27.19 ± 0.71) initially outperform the structured setup with 8 answer-seeking interactions (26.85 ± 0.39). Inference logs indicate that EXECUTIVE model rarely utilizes the entity identification stage on NarrativeQA, likely because its questions are less reliant on resolving specific entities. Consequently, the fixed-entity identification budget effectively reduces the number of available answer-seeking interactions compared to unstructured baselines. Increasing the answer-seeking budget to 15 interactions recovers this gap, with NarrativeQA reaching 27.76 ± 0.20, surpassing both unstructured baselines. We hypothesize that this could be due to additional answer-seeking interactions continuing to surface useful signals without the risk of entity drift or state corruption that compound in open-domain multi-hop settings.

Unlike NarrativeQA, structured entity identification and state tracking in BrowseComp-Plus and MuSiQue introduce sensitivity to error accumulation as the number of answer-seeking interactions increases beyond the optimal budget. Additional interactions increase the risk of erroneous MEMORY model responses corrupting the known facts state, and provide more opportunities for EXECUTIVE model to commit to an incorrect intermediate entity via the entity pivot correction helper (Tab. 14). Furthermore, more interactions dilute the correct signal with potentially incorrect answers at the final synthesis stage. These failure modes are partly a function of the reasoning capability of EXECUTIVE model, as structured state maintenance demands strong in-context reasoning to accurately track entities and avoid premature entity commitment. Corroborating this, we observe in Tab. 2 that a stronger reasoning model as EXECUTIVE model yields improved performance when paired with the same MEMORY model, suggesting that these failure modes can be mitigated by scaling the reasoning capability of EXECUTIVE model.

The stage budget used in our experiments was selected without systematic tuning, and alternative settings may yield similar performance with greater token efficiency. We therefore leave a systematic study of the optimal interaction budget and EXECUTIVE model selection as future work.

K.4 BOOTSTRAPPING DETAILS FOR COMPARING METHODS

We assess statistical significance between the best and second-best performing methods via a one-tailed paired bootstrap test of 10,000 resamples, testing whether the top-performing method significantly outperforms the runner-up. For each question, the per-question score difference is averaged

across runs, and these differences are resampled with replacement to obtain p-values at the $\alpha = 0.05$ significance level.

K.5 PER-QUERY TOKEN COSTS COMPARED TO RAG

We briefly report the average per-query token costs for MEMO and the 2 best performing baselines. In Tab. 16, we also report the average number of Entity Identification turns and average number of Answer Seeking turns per query for MEMO.

Table 16: Average per-query token costs and number of interaction turns for MEMO and RAG baselines.

Dataset	EXECUTIVE model tokens	MEMORY model tokens	HippoRAG2 tokens	NV-Embed-V2 tokens	Entity Identification turns	Answer Seeking turns
BCP	28,077.5	2,740.8	19,489.2	14,793.9	2.2	3.6
NQA	19,908.9	1,286.9	1,541.0	1,358.1	1.7	3.8
MSQ	15,792.9	1,129.4	1,575.9	1,364.6	1.6	2.8

Overall, MEMO incurs higher average per-query token costs than the RAG baselines, reflecting a trade-off between inference cost and the performance gains observed on MSQ. As this work focuses on establishing the feasibility of a parametric knowledge store that can interface with different EXECUTIVE model models through natural language, the current multi-turn inference protocol is not explicitly optimized for token efficiency. We expect that reducing the number and length of interactions between EXECUTIVE model and MEMORY model provides substantial room to improve inference efficiency, which we leave to future work.

L DISCUSSION ON NUMBER OF TRAINING EPOCHS

From Figs. 2 to 4 and Tables 17 and 18, we observe that additional training epochs do not consistently improve accuracy, as peak performance for some MEMORY model occurs at epoch 2 instead.

Table 17: Performance across training epochs. Each value reports the mean ± standard deviation over three runs. The best epoch for each setting is shown in bold.

Model	Dataset	Noise	Epoch 1	Epoch 2	Epoch 3
Gemma3-1B-IT	BCP	1N	29.22 ± 1.50	$\mathbf{36.11 \pm 1.26}$	36.00 ± 0.88
	NQA	–	20.02 ± 1.10	22.07 ± 0.20	$\mathbf{22.30 \pm 2.47}$
	MSQ	1N	39.60 ± 0.36	40.40 ± 2.45	$\mathbf{41.17 \pm 1.20}$
LFM2.5-1.2B-Instruct	BCP	1N	23.11 ± 1.39	28.67 ± 0.67	$\mathbf{33.78 \pm 3.38}$
	NQA	–	21.05 ± 1.54	$\mathbf{22.07 \pm 1.20}$	21.05 ± 0.79
	MSQ	1N	41.13 ± 0.95	$\mathbf{42.03 \pm 0.57}$	40.53 ± 0.97
Qwen2.5-1.5B-Instruct	BCP	1N	24.44 ± 0.38	29.89 ± 1.35	$\mathbf{33.56 \pm 0.77}$
	NQA	–	22.41 ± 2.00	22.41 ± 1.75	$\mathbf{24.00 \pm 0.20}$
	MSQ	1N	39.50 ± 1.25	$\mathbf{42.50 \pm 1.20}$	41.40 ± 0.53
Qwen2.5-14B-Instruct	BCP	0N	49.67 ± 0.88	$\mathbf{53.67 \pm 1.15}$	50.67 ± 1.45
	BCP	1N	47.33 ± 0.67	50.56 ± 0.19	$\mathbf{52.78 \pm 0.77}$
	BCP	2N	44.78 ± 1.17	48.89 ± 1.84	$\mathbf{52.33 \pm 0.88}$
Qwen2.5-14B-Instruct	NQA	–	26.39 ± 1.03	$\mathbf{26.85 \pm 0.39}$	25.37 ± 0.52
Qwen2.5-14B-Instruct	MSQ	0N	48.30 ± 0.82	$\mathbf{50.07 \pm 0.81}$	49.20 ± 0.89
	MSQ	1N	48.17 ± 0.70	$\mathbf{48.30 \pm 1.25}$	47.93 ± 0.95
	MSQ	2N	47.07 ± 0.31	47.63 ± 0.45	$\mathbf{47.83 \pm 0.57}$

One possible explanation for why additional training does not consistently improve performance is overfitting to the reflections corpus, which exhibits substantial lexical overlap across steps by design, as later steps are derived from earlier ones (Alg. 1). To quantify this lexical overlap, we compute the lossless compression ratio of the combined QA text across all steps for each dataset by extracting

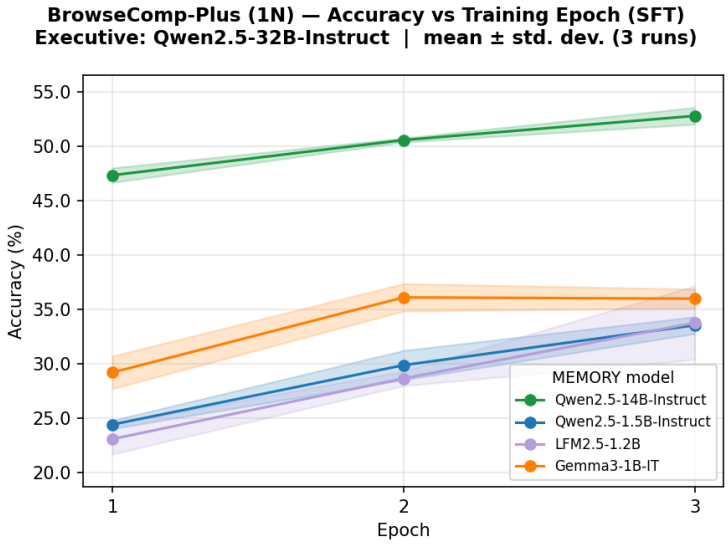


Figure 2: BrowseComp-Plus accuracy (%) vs. training epoch (Full SFT) for each MEMO model size and model family. Lines show the mean over 3 runs, and the shaded band shows ± std. dev. for Qwen2.5-32B-Instruct Runs.

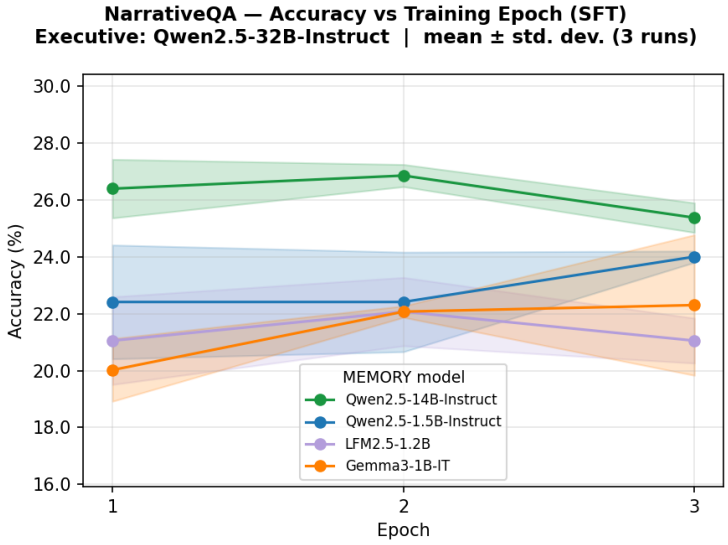


Figure 3: NarrativeQA accuracy (%) vs. training epoch (Full SFT) for each MEMO model size and model family. Lines show the mean over 3 runs, and the shaded band shows ± std. dev. for Qwen2.5-32B-Instruct Runs.

all question and answer strings, concatenating them into a single text corpus, and applying gzip compression at maximum level (compression level 9), where the compression ratio is defined as the ratio of the original text size to the compressed size.

BrowseComp-Plus (2,968,214 pairs) achieves a ratio of 5.67× (83.4% savings), MuSiQue (664,762 pairs) achieves 7.03× (85.8% savings), and NarrativeQA (1,276,676 pairs) achieves 5.45× (81.7% savings), indicating substantial lexical redundancy within each dataset. We note that compression ratio captures lexical overlap only, and semantic diversity across QA pairs may remain higher, as each step targets distinct reasoning operations ranging from direct fact extraction to cross-document synthesis (see Alg. 1), which is consistent with the impact of removing Step 5 (see App. E).

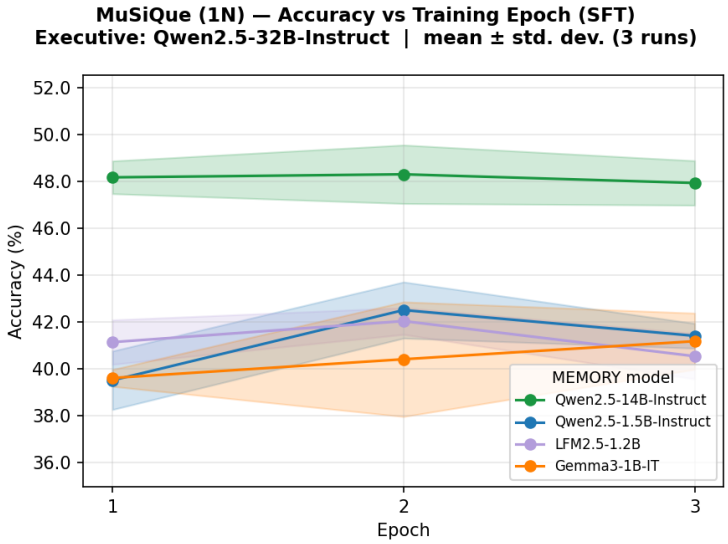


Figure 4: MuSiQue accuracy (%) vs. training epoch (Full SFT) for each MEMO model size and model family. Lines show the mean over 3 runs, and the shaded band shows ± std. dev. for Qwen2.5-32B-Instruct Runs.

Table 18: Performance across training epochs with LoRA. Each value reports the mean ± standard deviation over three runs. The best epoch for each setting is shown in bold.

Model	Dataset	Noise	Epoch 1	Epoch 2	Epoch 3
Gemma3-1B-IT	BCP	1N	7.89 ± 0.51	8.56 ± 0.77	$\mathbf{12.89 \pm 1.02}$
	NQA	–	15.70 ± 0.34	17.97 ± 1.04	$\mathbf{21.62 \pm 0.86}$
	MSQ	1N	18.70 ± 0.40	25.13 ± 0.55	$\mathbf{26.17 \pm 1.10}$
LFM2.5-1.2B-Instruct	BCP	1N	0.22 ± 0.19	$\mathbf{0.33 \pm 0.00}$	$\mathbf{0.33 \pm 0.00}$
	NQA	–	5.35 ± 0.71	$\mathbf{5.69 \pm 0.71}$	5.57 ± 0.39
	MSQ	1N	$\mathbf{7.50 \pm 0.26}$	7.40 ± 0.26	6.70 ± 0.52
Qwen2.5-1.5B-Instruct	BCP	1N	18.44 ± 2.78	25.78 ± 0.84	$\mathbf{26.89 \pm 2.67}$
	NQA	–	19.11 ± 1.02	$\mathbf{21.62 \pm 1.10}$	19.11 ± 1.02
	MSQ	1N	23.20 ± 1.73	29.90 ± 0.98	$\mathbf{31.53 \pm 0.55}$
Qwen2.5-14B-Instruct	BCP	1N	29.56 ± 0.39	36.00 ± 1.53	$\mathbf{41.44 \pm 1.54}$
	NQA	–	19.23 ± 1.07	21.27 ± 0.71	$\mathbf{23.78 \pm 0.52}$
	MSQ	1N	39.67 ± 0.12	$\mathbf{43.93 \pm 0.97}$	43.60 ± 1.06

Table 19: Ablation on MEMORY model size within the Qwen2.5 family. Bold results indicate the best performing results in the column.

	BrowseComp-Plus		NarrativeQA		MuSiQue	
MEMORY Model	Qwen2.5-32B	Gemini-3-Flash	Qwen2.5-32B	Gemini-3-Flash	Qwen2.5-32B	Gemini-3-Flash
Qwen2.5-1.5B-Instruct	33.56 ± 0.77	49.33	24.00 ± 0.20	47.44	42.50 ± 1.20	55.20
Qwen2.5-14B-Instruct	$\mathbf{52.78 \pm 0.77}$	**66.67**	$\mathbf{26.85 \pm 0.39}$	**53.58**	$\mathbf{48.30 \pm 1.25}$	**60.20**

M ABLATION ON MEMORY MODEL SIZE

We investigate how the size of MEMORY model affects downstream task performance by comparing models of 1.5B and 14B parameters in the Qwen2.5 family. Both Qwen2.5-1.5B-Instruct and Qwen2.5-14B-Instruct MEMORYs are trained on the same QA dataset generated by the GENERATOR model (Qwen2.5-32B-Instruct) under training settings described in App. F. Each MEMORY model is evaluated using Qwen2.5-32B-Instruct and Gemini-3-Flash as the EXECUTIVE model. Results in Tab. 19 show a positive scaling trend: larger MEMORY models yield improved performance across all benchmarks and EXECUTIVE models. However, the results also show that a stronger EXECUTIVE model reasoning capability modulates this gap non-uniformly across tasks: the performance difference

between MEMORY model sizes widens for NarrativeQA but shrinks for BrowseComp-Plus and MuSiQue. This suggests that the interaction between EXECUTIVE model reasoning capability and MEMORY model size is task-dependent.

N ABLATION ON MEMORY MODEL FAMILY

We investigate whether the choice of MEMORY model family affects performance by comparing three models of similar parameter scale (∼1–2B) but distinct architectures and pretraining lineages: Qwen2.5-1.5B-Instruct (Yang et al., 2025), Gemma3-1B-IT (Kamath et al., 2025), and LFM2.5-1.2B-Instruct (Amini et al., 2025). Results in Tab. 20 show that MEMO performance is largely robust to the choice of MEMORY model architecture, demonstrating that the framework is not sensitive to the specific pretraining lineage of MEMORY model at a similar parameter scale, and that the parametric knowledge compression induced by our training procedure generalizes across diverse model families.

Table 20: Ablation across MEMORY models at similar parameter scales (∼1–2B). Bold results indicate the best performing results in the column.

	BrowseComp-Plus		NarrativeQA		MuSiQue	
MEMORY Model	Qwen2.5-32B-I	Gemini-3-F	Qwen2.5-32B-I	Gemini-3-F	Qwen2.5-32B-I	Gemini-3-F
Qwen2.5-1.5B-Instruct	33.56 ± 0.77	49.33	$\mathbf{24.00 \pm 0.20}$	47.44	$\mathbf{42.50 \pm 1.20}$	55.20
Gemma3-1B-IT	$\mathbf{36.11 \pm 1.26}$	**54.33**	22.30 ± 2.47	**48.81**	41.17 ± 1.20	56.20
LFM2.5-1.2B-Instruct	33.78 ± 3.38	49.33	22.07 ± 1.20	46.42	42.03 ± 0.57	**58.30**

Each MEMORY model is trained on the same QA dataset generated by GENERATOR model (Qwen2.5-32B-Instruct) and evaluated using Qwen2.5-32B-Instruct and Gemini-3-Flash as EXECUTIVE model. Notably, while Qwen2.5-1.5B-Instruct and Gemma3-1B-IT are based on standard transformer architectures, LFM2.5-1.2B-Instruct adopts a hybrid architecture combining state-space convolution with transformer attention blocks, thereby providing a broader test of MEMORY model across diverse model designs. These models are trained on the same training settings in App. F, with Gemma3-1B-IT using eager attention during training instead of Flash Attention 2.

O COMPARISON BETWEEN FULL SFT AND LORA

We train all models using LoRA (Hu et al., 2022) applied to the attention and feed-forward projection layers: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The general LoRA configuration is summarised in Table 21, with model-specific rank and scaling settings reported in Table 22. All remaining training hyperparameters follow Table 10, and per-dataset batch sizes are given in Table 11.

Table 21: LoRA Specific Training Configuration. All other parameters are the same as those in Tab. 10.

Parameter	Value
Target modules	`q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`
LoRA dropout	0.05
Bias	None
Learning rate	2×10^{-4}

The notably poor LoRA performance of LFM2.5-1.2B-Instruct can be attributed to its hybrid convolution–attention architecture, which differs from the standard transformer models in our evaluation. Following the LFM2 architecture (Amini et al., 2025), LFM2.5-1.2B-Instruct consists of 16 layers — 6 grouped-query attention (GQA) blocks (at indices $\{2, 5, 8, 10, 12, 14\}$) interleaved with 10 short-range LIV convolution (ShortConv) blocks (Amini et al., 2025). Crucially, the LFM2 attention output projection is named `out_proj` (rather than `o_proj`) and its SwiGLU MLP uses `w1`/`w3`/`w2` (rather than `gate_proj`/`up_proj`/`down_proj`), while the ShortConv blocks expose their own

Table 22: Model-Specific LoRA Configuration.

Model	Size	LoRA rank	LoRA alpha	Trainable params
LFM2.5-1.2B-Instruct	1.2B	8	16	6.1M (0.41%)
Gemma3-1B-IT	1B	8	16	6.6M (0.65%)
Qwen2.5-1.5B-Instruct	1.5B	8	16	9.2M (0.60%)
Qwen2.5-14B-Instruct	14B	16	32	68.8M (0.47%)

Table 23: Ablation on LoRA vs Full SFT training across all MEMORY models, evaluated with Qwen2.5-32B-Instruct as EXECUTIVE model. All results are mean ± std. dev. over 3 runs. Bold results indicate the best performing results in the column.

	BrowseComp-Plus		NarrativeQA		MuSiQue	
MEMORY model	LoRA	Full SFT	LoRA	Full SFT	LoRA	Full SFT
Gemma3-1B-IT	12.89 ± 1.02	36.11 ± 1.26	21.62 ± 0.86	22.30 ± 2.47	26.17 ± 1.10	41.17 ± 1.20
LFM2.5-1.2B-Instruct	0.33 ± 0.00	33.78 ± 3.37	5.69 ± 0.71	22.07 ± 1.20	7.50 ± 0.26	42.03 ± 0.57
Qwen2.5-1.5B-Instruct	26.89 ± 2.67	33.56 ± 0.77	21.62 ± 1.10	24.00 ± 0.20	31.53 ± 0.55	42.50 ± 1.20
Qwen2.5-14B-Instruct	$\mathbf{41.44 \pm 1.54}$	$\mathbf{52.78 \pm 0.77}$	$\mathbf{23.78 \pm 0.52}$	$\mathbf{26.85 \pm 0.39}$	$\mathbf{43.93 \pm 0.97}$	$\mathbf{48.30 \pm 1.25}$

`in_proj` and `out_proj` layers. A LoRA configuration targeting the standard Llama-family module names, therefore adapts only a strict subset of the projections that exist in LFM2.5, leaving the remainder frozen. The result is 6.1M trainable parameters (0.41% of total), disproportionately low given the model size and below our target of $\sim 0.5\%$. The rank $r = 8$ was kept fixed across all sub-2B models for a controlled comparison; in retrospect, this penalizes LFM2.5-1.2B-Instruct due to its architectural mismatch with the standard Llama-style target set.

Furthermore, the 10 ShortConv blocks, which handle the bulk of the model's local feature extraction, and the SwiGLU MLPs attached to every block remain entirely unadapted under standard LoRA targeting, severely limiting the adapter's ability to shift the model's behavior. As shown in Tab. 23, the large performance gap between LoRA and Full SFT confirms that the model is capable of learning the task when all parameters are updated.

Future work could explore LoRA configurations better suited to this architecture: targeting the LFM2-specific module names (`out_proj`, `w1`, `w3`, `w2`) alongside the ShortConv projections (`in_proj`, `out_proj`), as well as tuning the rank and learning rate per architecture rather than holding them fixed across families for the controlled comparison reported here.

P PRELIMINARY EXPERIMENTS ON CONTINUAL KNOWLEDGE INTEGRATION

A practical desideratum of knowledge integration systems is the ability to incorporate new knowledge incrementally as it becomes available. Beyond the primary scope of this work, we provide preliminary experiments exploring two continual integration settings: **(1)** incorporating new knowledge from disjoint, static target corpora, without modifying previously integrated knowledge App. P.1; and **(2)** updating previously integrated knowledge, including conflicting and temporally updated information App. P.2.

P.1 CONTINUAL KNOWLEDGE INTEGRATION VIA MODEL MERGING

We explore model merging (Yang et al., 2026) as a preliminary approach to continual knowledge integration over disjoint corpora. Unlike non-parametric systems, conventional retraining of parametric models requires training on the accumulated corpus, making repeated integration increasingly costly as new sources are added (Wu et al., 2022a). Model merging instead independently trains a MEMORY model on each new corpus and combines the resulting models in parameter space, avoiding retraining on previously integrated corpora. This setting assumes that previously integrated knowledge remains unchanged and therefore does not address updates to prior knowledge. We provide background on the merging methods and full experimental details in App. Q.

Experimental setup. We partition NarrativeQA into two disjoint subsets of approximately 640k reflection QA pairs each (App. F) and independently train a Qwen2.5-14B-Instruct MEMORY model on each subset. We then merge the resulting MEMORY model models and compare against full retraining on the union of both subsets. We evaluate 14 merging configurations across six methods and report the best-performing configuration, TIES (Yadav et al., 2023) with ρ=0.3, below; the complete sweep is reported in Tab. 26. Letting X and Y denote the SFT cost on each subset alone (cost scales approximately linearly with the number of QA pairs, so the union costs $X+Y$), cumulative compute across the two arrivals is $X+Y$ for merging versus $X+(X+Y)$ for full retraining.

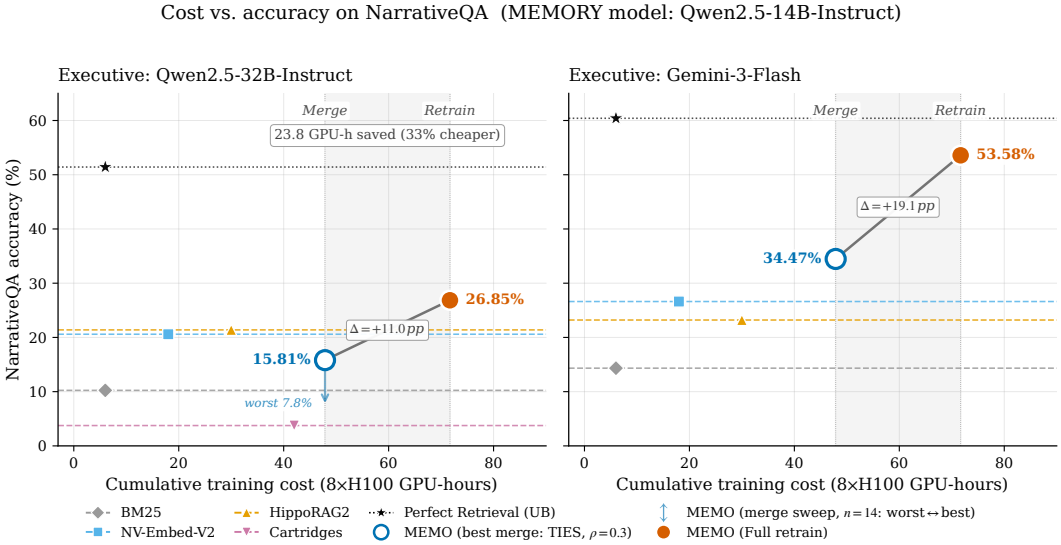


Figure 5: **Cost–accuracy trade-off on NarrativeQA when a second corpus arrives (K=2, MEMORY model = Qwen2.5-14B-Instruct, 8×H100).** Cumulative training cost is shown on the x-axis (one Qwen-14B SFT run takes $\approx$ 24 GPU-hours on a 640k-QA-pair corpus). The vertical $\updownarrow$ at the merge cost shows the worst-to-best range across the 14 merge configurations swept (Tab. 26) for Qwen runs only. Perfect Retrieval is shown as the upper bound.

Table 24: **Model merging vs. full retraining on NarrativeQA.** MEMORY model = Qwen2.5-14B-Instruct. Merge-TIES (ρ=0.3) is the best of 14 configurations swept (Tab. 26, App. Q). Cumulative compute is reported in 8×H100 GPU-hours for K=2 subsets of ~640k reflection QA pairs each. Δ denotes accuracy difference (%) relative to full retraining.

Method	**Cumulative compute** (8×H100 GPU-h)	**Qwen2.5-32B-I** Acc. (%)	Δ	**Gemini-3-F** Acc. (%)	Δ
Full retrain ($X+(X+Y)$)	$\approx$ 72h	**26.85 ± 0.39**	—	**53.58**	—
Merge-TIES (ρ=0.3, $X+Y$)	$\approx$ 48h	15.81 ± 0.39	↓ 11.04	34.47	↓ 19.11

Merging cuts compute by 33% at K=2, with widening returns at scale. As reported in Tab. 24, the full-retrain baseline incurs $X+(X+Y) \approx 72$ GPU-hours of cumulative compute, while merging accumulates only $X+Y \approx 48$ GPU-hours — a 33% reduction (Fig. 5). The gap widens with K: under the same per-corpus cost, merging scales as $\Theta(K)$ while full retraining scales as $\Theta(K^2)$, yielding a 5.5× saving at K=10 (240 vs. 1,320 GPU-hours).

Merging trades accuracy for reduced integration cost. Merge-TIES (ρ=0.3) trails the full-retrain MEMORY model by 11.0 percentage points (pp) under Qwen2.5-32B-Instruct and 19.1 pp under Gemini-3-Flash (Tab. 24); across the full 14-configuration sweep, accuracy ranges from 7.85% (SLERP, worst) to 15.81% (TIES, best), shown in Fig. 5. Despite this gap, the merged MEMORY model is able to achieve a higher accuracy than the strongest performing non-parametric baseline (NV-Embed-V2) when the EXECUTIVE model is Gemini-3-Flash; see Tab. 2 on NarrativeQA.

Overall, these results highlight a compute–accuracy trade-off: model merging reduces the cost of integrating new knowledge at the expense of accuracy relative to full retraining. Future work could evaluate model merging over larger numbers of disjoint corpora and across more diverse corpus

types and domains. Closing the accuracy gap with full retraining while preserving the computational benefits of incremental integration also remains an important direction for future work.

P.2 CONTINUAL KNOWLEDGE INTEGRATION FOR UPDATES TO PREVIOUSLY INTEGRATED KNOWLEDGE

We provide a preliminary exploration of extending MEMORY model to settings that involve conflicting and temporally updated knowledge. We consider two mechanisms: post-hoc knowledge editing of a trained MEMORY model model and conflict-aware data synthesis followed by retraining.

Experimental setup. Following (Yang et al., 2024), we construct a small-scale benchmark of slow-changing facts from the movie and sports domains, comprising 63 questions and 315 supporting documents. We construct synthetic knowledge updates for a subset of 49 questions by creating a new document containing a fictional answer and assigning it a date later than that of the original document. The resulting corpus therefore contains both the original and updated documents, representing the value of the same fact at different points in time.

We process the 315 documents using our data synthesis pipeline and train a Qwen2.5-1.5B-Instruct MEMORY model model on the resulting QA pairs using the same training setup as in Sec. 4.2. As this experiment aims to isolate whether knowledge stored in MEMORY model can subsequently be updated, we directly query MEMORY model rather than use the multi-turn inference protocol from Sec. 4.3 to isolate the effect of the knowledge update from the inference procedure. Under this simplified evaluation, MEMORY model successfully recalls 8 of the 49 target facts. We restrict the subsequent update experiments to these 8 facts, since evaluating an update requires the original fact to first be learned. Direct querying shows that the resulting model successfully recalls 8 of the 49 facts. We restrict the following experiments to these 8 facts, since an update can only be evaluated when the original fact has first been learned.

Post-hoc knowledge editing. We first investigate whether a previously learned fact can be updated directly in the parameters of the MEMORY model model via knowledge editing methods. We apply AlphaEdit (Fang et al., 2025) to replace each learned value with its corresponding fictional update. All 8 edits successfully change the model's response to the target query. To assess locality, whether an edit preserves the model's responses on unedited facts, we evaluate the model on 24 other facts. Specifically, we measure whether the model's response to each query remains unchanged before and after editing, irrespective of its correctness. The responses remain unchanged for 23 of the 24 queries, indicating limited behavioral drift on these unedited queries.

Table 25: Preliminary results for knowledge editing on MEMORY model. Locality measures whether responses to unedited queries remain unchanged after a single editing, irrespective of their correctness.

Metric	**Result**
Num of facts successfully edited	8/8
Average locality after a single edit across 8 edits	23/24
Sequential edits before degradation of previous edits	6

However, we observe two limitations. First, sequential editing is limited, as we can perform at most 6 consecutive edits before accuracy on previously edited facts begins to decrease. Second, the edits generalize poorly across query formulations. For a single edit, we evaluate 2 paraphrased variants of the target query that the model answered correctly before editing. After editing, neither paraphrase elicits the updated answer, whereas the original query successfully elicits the updated value. This suggests that, in our setting, the edit may be tied to the specific query formulation used during editing rather than representing a generalizable knowledge update. Consequently, applying this approach may require the expected query formulation to be known both when performing the edit and when subsequently accessing the updated knowledge.

Conflict-aware data synthesis and retraining. As an alternative, we investigate whether conflicting knowledge can be handled directly by MEMO's data synthesis pipeline. For each updated fact, we create a new document containing the fictional value and assign it a date later than the original document. We then add this document to the corpus and rerun the data synthesis pipeline. For conflict resolution, we specify in the Cross-document synthesis step (Step 5) prompt that the value of the most recent document should be treated as the current value.

The resulting QA pairs retain the temporal information associated with conflicting facts. In particular, Step 5 consolidates the conflicting evidence into a single QA pair that captures how the fact changes over time, including both values and their respective document dates. This provides preliminary evidence that the data synthesis pipeline can capture and resolve temporal conflicts through an appropriate conflict-resolution criterion.

We retrain the MEMORY model model from scratch on the resulting QA pairs, where the retrained model returns the updated value for 5 of the 8 evaluated facts. Notably, the model can also return the updated values from paraphrased queries for the successfully learned updates, suggesting a better generalization of knowledge than observed with post-hoc editing. Overall, sequential parameter editing and retraining demonstrate a trade-off between the cost of updating and the generalization of updated knowledge across queries.

Together, these experiments illustrate two possible extensions of MEMORY model beyond the static-corpus setting studied in the main paper(Sec. 3). Sequential parameter editing and conflict-aware synthesis with retraining illustrate a trade-off between the cost of updating and the generalization of updated knowledge across queries. While preliminary, these results demonstrate the potential of both approaches for extending MEMORY model to evolving knowledge. A systematic study of this setting, including improved editing methods, data mixtures, conflict-resolution strategies, and training configurations, is beyond the scope of this work and left for future investigation.

Q MODEL MERGING: BACKGROUND AND EXPERIMENTAL DETAILS

Q.1 EXPLORING DIFFERENT METHODS OF MODEL MERGING

Let $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ be a collection of pairwise disjoint target corpora. For each corpus $\mathcal{D}_i$, we generate a reflection QA dataset $\mathcal{Q}^{(i)}_{\text{final}}$ (Sec. 4.1) and train a corresponding MEMORY model $\mathcal{M}_{\varphi_i}$ via SFT (Sec. 4.2), initializing all K models from the same pretrained base $\mathcal{M}_{\varphi_0}$. We define the *task vector* for $\mathcal{D}_i$ as $\tau_i = \varphi_i - \varphi_0$, capturing the parametric shift induced by training on $\mathcal{D}_i$ alone. The merged MEMORY model is then obtained as $\varphi_{\text{merged}} = \text{Merge}(\varphi_0, \{\tau_i\}_{i=1}^K; \Theta)$, where Θ denotes method-specific hyperparameters (e.g., merging coefficients, sparsification densities). We consider the following methods, which produce φ_{merged} without ever training on $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$:

- **Linear merging** (Wortsman et al., 2022) computes a weighted sum of task vectors: $\varphi_{\text{merged}} = \varphi_0 + \sum_{i=1}^K \lambda_i \tau_i$, where $\lambda_i > 0$ are merging coefficients.
- **SLERP** (Shoemake, 1985) interpolates between two task vectors along the unit sphere, preserving their magnitudes: $\varphi_{\text{merged}} = \varphi_0 + \text{SLERP}(\tau_1, \tau_2; t)$, with $t \in [0, 1]$ controlling the interpolation factor.
- **Task arithmetic** (Ilharco et al., 2023) adds task vectors directly without further processing, recovering linear merging as a special case with uniform λ_i.
- **TIES** (Yadav et al., 2023) resolves interference among task vectors before summation by (i) *trimming* each τ_i to its top-ρ fraction of largest-magnitude entries, (ii) *electing* a sign at each coordinate by magnitude-weighted majority vote, and (iii) *disjoint-merging* only the entries that agree with the elected sign.
- **DARE** (Yu et al., 2024) sparsifies each task vector by randomly dropping a fraction $1 - \rho$ of its entries and rescaling the survivors by $1/\rho$ to preserve expected magnitude, before linear merging.
- **DARE-TIES** (Yu et al., 2024) combines DARE-style stochastic sparsification with TIES sign-conflict resolution, retaining the diversity of random dropout while filtering out conflicting updates.

Q.2 EXPERIMENTAL DETAILS

For our experiments, we partition NarrativeQA into two pairwise-disjoint subsets, NarrativeQA.1 and NarrativeQA.2, of $\sim$640k reflection QA pairs each. Each subset is used to fine-tune an independent MEMORY model from the same Qwen2.5-14B-Instruct base via SFT for 3 epochs, producing $\mathcal{M}_{\varphi_1}$ and $\mathcal{M}_{\varphi_2}$ at SFT costs of X and Y GPU-hours, respectively (each is ≈ 24 GPU-hours on 8$\times$H100; full-retrain on the union NarrativeQA.1 $\cup$ NarrativeQA.2 costs $X+Y \approx 48$ GPU-hours by linear

scaling). We evaluate every saved checkpoint of each run on the held-out NarrativeQA evaluation set and select the best-performing checkpoint per subset; the corresponding task vectors τ_1 and τ_2 are the inputs to the merging step. We then sweep all six merging methods listed above (Linear, Task arithmetic, SLERP, TIES, DARE, DARE-TIES) at three sparsification densities $\rho \in \{0.3, 0.5, 0.7\}$ (or three interpolation factors $t \in \{0.3, 0.5, 0.7\}$ for SLERP), giving 14 merged-MEMORY model configurations in total. Each configuration is evaluated on NarrativeQA with Qwen2.5-32B-Instruct as EXECUTIVE model (mean $\pm$ std over 3 runs). The configuration that we report in App. P.1 as *Merge-TIES* is the best of the sweep (TIES with ρ=0.3).

Q.3 MODEL MERGING SWEEP RESULTS

Sweep results. A single SFT run consumes $\approx$ 24 GPU-hours on 8×H100; after two arrivals, full retraining incurs $X+(X+Y)$=72 GPU-hours of cumulative compute, whereas merging accumulates only $X+Y$=48 GPU-hours — a 33% reduction (Fig. 5). The asymptotic gap widens with K: under the same per-corpus cost, merging scales as $\Theta(K)$ while full retraining scales as $\Theta(K^2)$, yielding a 5.5× saving at K=10 (240 vs. 1,320 GPU-hours). On accuracy, Merge-TIES (ρ=0.3) trails full retraining by 11.0% with Qwen2.5-32B-Instruct as EXECUTIVE model (15.81% vs. 26.85%) and by 19.1% with Gemini-3-Flash (34.47% vs. 53.58%), placing the merged MEMORY model below the union-retrained MEMORY model but above every retrieval baseline. The full per-method sweep is reported in Tab. 26: TIES (ρ=0.3) and DARE-Linear (ρ=0.3) lead at 15.81% and 15.47% respectively, while SLERP (t=0.5) is the worst configuration at 7.85%. The pattern across families suggests that aggressive sparsification at low ρ paired with sign-conflict resolution (TIES, DARE-Linear) is the most reliable merging recipe in this regime. These results confirm the predicted compute–accuracy trade-off: merging recovers most of MEMORY model's headroom over retrieval methods at substantially lower cumulative cost. TIES and DARE-Linear at ρ=0.3 achieve the highest accuracy among the evaluated merging configurations, motivating further investigation of sparsification and sign-conflict resolution for merging independently trained MEMORY model models.

Table 26: **Sweep of all 14 merge configurations on NarrativeQA.** Two MEMORY models (Qwen2.5-14B-Instruct) are independently SFT-trained on the disjoint NarrativeQA.1 and NarrativeQA.2 subsets; each subset's best-performing checkpoint provides the task vector entering the merge. EXECUTIVE model = Qwen2.5-32B-Instruct; results are mean $\pm$ std. dev. over 3 runs. Best merge in **bold**; full-retrain accuracy (26.85 ± 0.39) is shown for reference. Hyperparameter conventions: $t \in [0, 1]$ is the SLERP interpolation factor along the unit sphere connecting the two task vectors (t=0 recovers MEMORY model on NarrativeQA.1, t=1 recovers MEMORY model on NarrativeQA.2, t=0.5 is the geodesic midpoint); $\rho \in (0, 1]$ is the sparsification density — the fraction of largest-magnitude task-vector entries kept (TIES) or the keep probability for random-drop sparsification (DARE, DARE-TIES). Linear and Task arithmetic merge with uniform weights ($\lambda_i = 1$) and have no hyperparameter.

Method family	Hyperparameter	Accuracy (%)
Linear	—	11.60 ± 1.02
Task arithmetic	—	12.74 ± 1.75
SLERP	$t = 0.3$	11.60 ± 2.24
	$t = 0.5$	7.85 ± 1.71
	$t = 0.7$	11.60 ± 2.13
TIES	$\rho = 0.3$	$\mathbf{15.81 \pm 0.39}$
	$\rho = 0.5$	12.17 ± 1.94
	$\rho = 0.7$	12.06 ± 2.58
DARE-Linear	$\rho = 0.3$	15.47 ± 0.79
	$\rho = 0.5$	9.78 ± 1.20
	$\rho = 0.7$	13.65 ± 2.08
DARE-TIES	$\rho = 0.3$	11.72 ± 0.52
	$\rho = 0.5$	12.97 ± 1.23
	$\rho = 0.7$	11.04 ± 1.20

Q.4 Additional properties of model merging

Avoiding sequential fine-tuning. Because no individual MEMORY model $\mathcal{M}_{\varphi_i}$ is ever fine-tuned on another corpus' data, model merging cannot induce the kind of distributional interference that drives catastrophic forgetting in sequential fine-tuning (Luo et al., 2025a). Knowledge from each corpus is preserved within its own task vector τ_i, and conflicts between task vectors are addressed at merge time via the methods above rather than during gradient updates.

Scalability. When a new corpus $\mathcal{D}_{K+1}$ arrives, we train auxiliary model $\mathcal{M}_{\varphi_{K+1}}$ on its reflection QA dataset, derive τ_{K+1}, and re-merge in $\mathcal{O}(1)$ additional cost relative to the full collection. This enables modular, plug-and-play integration over a continuous stream of disjoint knowledge sources, unlike retraining from scratch on $\bigcup_i \mathcal{D}_i$, which scales linearly with the cumulative corpus size.

Inference. The merged MEMORY model is queried identically to a single-corpus MEMORY model via the structured multi-turn protocol described in Sec. 4.3. Because merging operates entirely in parameter space and produces a model with the same architecture and interface as $\mathcal{M}_{\varphi_0}$, it inherits the plug-and-play property of MEMO without requiring changes to the EXECUTIVE model or the inference protocol. Importantly, the EXECUTIVE model queries a *single* merged MEMORY model at inference rather than dispatching across K separate per-corpus MEMORY models, keeping the multi-turn retrieval pipeline unchanged regardless of how many corpora have been integrated.

R Implementation details and discussion for other parametric methods

R.1 ICAE

The ICAE architecture consists of a learnable encoder and a fixed decoder. The purpose of the learnable encoder is to read a long context and compress it into a small number of "memory slots" (compact vector representations). At inference time, the decoder takes the memory slots produced by the encoder and uses them as part of the context to answer queries.

Both the encoder and decoder make use of an LLM, which we select to be Qwen2.5-32B-Instruct to ensure that the EXECUTIVE model is consistent across baselines. While the LLM is frozen for the decoder, the encoder is learnable when it is instantiated with LoRA. The learning takes place across different training stages such as pretraining and fine-tuning. Crucially, the memory slots produced after training are specific to the LLM's embedding space, limiting plug-and-play integration with other LLMs, even for those from the same family. During the pretraining stage, each document is tokenized and compressed into a fixed set of memory slots by the LoRA encoder. The LoRA adapters are trained on the joint objective of Autoencoding (AE) and Language Modeling (LM), with the loss function being a weighted sum of both objectives. This teaches the encoder to compress information that is both faithfully recoverable (via the autoencoding objective) and useful for predicting subsequent text (via the language modeling objective).

During the fine-tuning stage, the pretrained LoRA adapters are further trained on (context, prompt, response) triplets, where the loss is computed only over the response tokens, conditioned on the memory slots and the prompt. This teaches the encoder to produce memory slots that the frozen decoder can use to respond to prompts. Notably, this stage requires labeled (context, question, answer) triplets from the target dataset; applying it to our test queries would introduce test-set supervision unavailable to the other methods. We therefore evaluate ICAE using its pretraining stage only. Consequently, these results should be interpreted as ICAE's performance without task-specific fine-tuning rather than its performance under its full supervised training pipeline.

The implementation details for pretraining are in Tab. 27 and Tab. 28. During inference, greedy decoding was used, and the results between no training and pretraining are reported in Tab. 29.

[9] In the implementation code, we reference the original implementation on GitHub that models the loss as a per-sample probability of using either the LM or AE objective for loss. Over an epoch, this would converge to the same loss when having a weighted sum between LM and AE.

[10] We used chunks before compression because the full document length of NarrativeQA would require 64 compression windows at a segment length of 2048, exceeding available VRAM during training. This constraint is specific to the 32B model scale, and smaller models would likely accommodate more compression windows.

Table 27: Common training hyperparameters for ICAE on H200 hardware. Values referenced from the original paper or the source code from the GitHub repository.

Parameter	Value
Number of memory tokens per segment for encoder compression	128
LoRA rank	512
LoRA alpha	32
LoRA dropout	0.05
Ratio between loss from LM and AE[9]	0.5
Number of training epochs	30
Effective batch size	16
Learning rate	1e-4
Optimizer	AdamW
Lr scheduler type	cosine
Warmup ratio	0.05
Max grad norm	2.0
Minimum number of tokens to use the LM objective	64
Number of initial tokens masked for LM loss	8
Precision	bf16
Deepspeed stage	3

Table 28: Dataset-specific training hyperparameters

Parameter	BrowseComp-Plus	NarrativeQA	MuSiQue
Compression ratio	64	16	8
Tokens per compression window	8192	2048	1024
Number of compression windows per chunk[10]	2	8	1
Total memory tokens (decoder input)	256	1024	128

Table 29: Accuracy (%) for ICAE performance across 3 runs each performance comparison (mean $\pm$ standard deviation) across configurations

Training stage	**BrowseComp-Plus**	**NarrativeQA**	**MuSiQue**
no training	$0.44\pm, 0.20$	$\mathbf{4.55 \pm 0.20}$	9.87 ± 0.06
pretrained	$\mathbf{4.33 \pm 0.00}$	1.37 ± 0.00	$\mathbf{17.60 \pm 0.00}$

From Tab. 29, we observe that pretraining improves performance on BrowseComp-Plus and MuSiQue, but degrades performance on NarrativeQA relative to the untrained encoder. For NarrativeQA, the training loss decreases only marginally from 2.006 at epoch 1 to 2.002 at epoch 30, compared with 1.985 to 1.785 on BrowseComp-Plus and 2.473 to 1.007 on MuSiQue. This suggests that the ICAE encoder did not learn an effective compression of the NarrativeQA corpus under our training configuration. At the 32B model scale, hardware constraints prevented us from exploring configurations with additional compression windows or larger effective batch sizes, which resulted in out-of-memory errors. We therefore caution against interpreting the NarrativeQA result as representative of ICAE's performance under more favorable training conditions. Rather, our results reflect its application under the same model scale and target-corpus supervision considered in our setting.

R.2 CARTRIDGES

Cartridges is originally designed for parameterizing individual long-context documents as reusable KV-cache representations, whereas our setting requires incorporating knowledge from substantially larger, multi-document corpora. We nevertheless include Cartridges as a representative approach that parameterizes external knowledge as a KV-cache. To adapt it to our setting, we use the maximum values explored in the original paper for several relevant hyperparameters, including 2 training epochs and an 8192-token KV-cache.

We additionally experiment with partitioning each corpus into 10 approximately equal-length subcorpora, training a separate cartridge for each partition, and composing the resulting cartridges following the original approach; this performs worse than using a single cartridge.

Given the difference between its original use case and our evaluation setting, these results should not be interpreted as a general limitation of Cartridges or KV-cache-based memory. Rather, they show that the evaluated Cartridges configurations do not transfer effectively to the substantially larger, multi-document corpora considered in our experiments. Furthermore, because Cartridges requires access to model internals during training and inference, its resulting memory cannot provide the black-box, plug-and-play integration across EXECUTIVE model models targeted by MEMO.

R.3 MEMORYDECODER

MemoryDecoder first builds a kNN datastore from a domain corpus and uses it to generate cached (context, kNN-distribution) pairs. The decoder is then pretrained to imitate these distributions directly. At inference, the kNN datastore is discarded, and the pretrained decoder's next-token distribution is interpolated with the base LLM's. In this baseline, we mirror the setup in MEMO by having the EXECUTIVE model be Qwen2.5-32B-Instruct and the pretrained decoder be Qwen2.5-14B-Instruct.

The parameters used in constructing and querying the kNN datastore are summarized in Table 30. For each context, $\varphi(x)$ is extracted from the input to the final decoder layer's MLP, and its k nearest neighbors are retrieved by L2 distance, with the top-1 result discarded unconditionally to avoid self-retrieval. Retrieved distances are converted to a target distribution via $\mathrm{softmax}(-\mathrm{dist}/\tau)$, and all neighbor-supported tokens are retained without additional truncation. To reduce the cost of search over a large datastore, retrieval is performed against an approximate IVFPQ index instead. Lastly, contexts are formed via a sliding window over fixed-length token blocks, ensuring every scored token has sufficient preceding context regardless of its position in the source document.

Table 30: Parameters used in kNN datastore and retrieval construction for MemoryDecoder

kNN parameter	**Value**
Layer for $\varphi(x)$	`model.layers[-1].mlp` (input)
k	1024 (1023 after self-match drop)
Distance metric	L2
kNN temperature τ	16.0
Index	IVFPQ, ncentroids=4096, code_size=64, 8 bits, nprobe=32
Sparsification threshold	0
Blocking	1024 tokens, stride 512

The parameters used to train the Memory Decoder are summarized in Table 31. Following the pretraining objective in (Cao et al., 2025), the decoder is trained on the cached kNN target distributions using a weighted combination of KL divergence (aligning the decoder's output with the kNN distribution) and standard Cross-Entropy (CE) against the next token. Optimization uses AdamW with a fixed learning rate and no warmup, and training is distributed via fully-sharded data parallelism given the size of the decoder. As no validation split is used, the final checkpoint (epoch 70) is selected based on training loss.

The parameters used for inference are summarized in Table 32. At each decoding step, EXECUTIVE model and the pretrained decoder independently produce next-token distributions, which are linearly interpolated using the tuned weight λ. The response to the query is greedily decoded from this interpolated distribution, generating up to a maximum of 512 new tokens. Notably, as the next-token distributions need to be interpolated, MemoryDecoder requires the pretrained decoder to be from the same model family, limiting plug-and-play integration with white-box LLM access.

We apply the same evaluation setup that was described in Sec. 5. We reported λ=0.05 in Tab. 2, and we provide additional results of the performance across different λ values in Tab. 33. λ=0.0 represents the EXECUTIVE model by itself (corresponding to the results in Tab. 12), and λ=1.0 represents the pretrained decoder by itself. We observe that MemoryDecoder achieves limited improvement over EXECUTIVE model across the different interpolation weights. We note that MemoryDecoder was originally developed and evaluated primarily for domain adaptation, where its effectiveness is measured through perplexity on domain-specific corpora, rather than for the knowledge-intensive question-answering tasks considered here.

Table 31: MEMORYDECODER training configuration

Parameter	Value
Loss	$\beta \cdot \mathrm{KL}(p_{\text{MemDec}} \parallel p_{\text{kNN}}) + (1-\beta) \cdot \mathrm{CE}$
β	0.5
Optimizer	AdamW
Learning rate	1×10^{-3}
LR scheduler type	Linear decay
Warmup steps	0
Weight decay	0.0
Num of training epochs	70
Effective batch size	128
Max sequence length	1024 tokens
Precision	BF16
Sharding	FSDP Full Shard
Seed	42

Table 32: MEMORYDECODER Inference Configuration

Parameter	Value
kNN temperature (inference)	1.0
Interpolated next-token distribution decoding strategy	Greedy
Max new tokens	512

Table 33: Accuracy (%) for MemoryDecoder performance across λ values, 3 runs each

λ	BrowseComp-Plus	NarrativeQA	MuSiQue
0.00	0.44 ± 0.20	$\mathbf{4.55 \pm 0.20}$	$\mathbf{9.87 \pm 0.06}$
0.05	$\mathbf{0.89 \pm 0.19}$	4.10 ± 0.00	9.57 ± 0.12
0.10	0.56 ± 0.20	3.98 ± 0.20	9.73 ± 0.12
0.30	0.00 ± 0.00	4.32 ± 0.52	$\mathbf{9.87 \pm 0.12}$
0.50	0.56 ± 0.20	3.75 ± 0.00	7.10 ± 0.10
0.70	0.00 ± 0.00	0.91 ± 0.20	0.20 ± 0.00
0.95	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.00
1.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.00

Our evaluation tests whether its knowledge-integration mechanism extends to a different setting in which BrowseComp-Plus and MuSiQue requires multi-hop reasoning, while NarrativeQA requires narrative understanding over a long context. One possible explanation for the observed performance is a mismatch between these requirements and MemoryDecoder's training objective, which distills next-token distributions derived from neighbors retrieved by local contextual similarity. Our experiments do not isolate this mechanism as the cause, and should not be interpreted as evidence against MemoryDecoder's effectiveness for its originally evaluated domain-adaptation setting.