LMEnt: A Suite for Analyzing Knowledge in Language Models from Pretraining Data to Representations

Daniela Gottesman[τ] **Alon Gilaie-Dotan**[τ] **Ido Cohen**[τ] **Yoav Gur-Arieh**[τ]
Marius Mosbach[ω] **Ori Yoran**[τ] **Mor Geva**[τ]

[τ] Tel Aviv University [ω] Mila – Quebec AI Institute & McGill University

gottesman3@mail.tau.ac.il

Abstract

Language models (LMs) increasingly drive real-world applications that require world knowledge. However, the internal processes through which models turn data into representations of knowledge and beliefs about the world are poorly understood. To facilitate such studies, we present LMEnt, a suite including (1) a knowledge-rich pretraining corpus, fully annotated with entity mentions based on Wikipedia, (2) an entity-based retrieval method over pretraining data that outperforms existing tools by as much as 80.4%, and (3) 12 pretrained LMs with up to 1B parameters and 4K intermediate checkpoints, with comparable performance to popular open-source models on knowledge tasks. Together, these resources provide a controlled environment for analyzing connections between entity mentions in pretraining data and downstream performance. We show the utility of LMEnt by studying knowledge acquisition over training, finding that entity co-occurrence and mention forms—which are difficult to study with existing tools—affect learning trends. Moreover, as LMs form stronger associations between entities, their facts are harder to edit in-context, whereas inconsistencies in model predictions over training are indicative of editing success. We release LMEnt to support studies of knowledge in LMs, including knowledge representations, plasticity, editing, attribution, hallucinations, and learning dynamics.

huggingface.co/LMEnt
github.com/LMEnt

1 Introduction

Language models (LMs) capture substantial amounts of knowledge and beliefs about the world from their training data (Petroni et al., 2019; Roberts et al., 2020; AlKhamissi et al., 2022; Andreas, 2022). A fundamental, yet underexplored, question is how such knowledge representations are formed and shaped during pretraining. Namely, what is the interplay between data composition, training dynamics, and the internal knowledge mechanisms in LMs? An understanding of these processes could provide better control over the model's knowledge, potentially improving its factuality and reasoning.

A key prerequisite for studying this question is the ability to locate exactly where specific knowledge appears in the pretraining data. However, since such metadata is rarely available for pretraining corpora, many works that study knowledge acquisition have used synthetic data (Kassner et al., 2020; Hron et al., 2024; Chang et al., 2024; Allen-Zhu and Li, 2024; Zucchet et al., 2025; Ou et al., 2025), or post-hoc string-based search tools (Elazar et al., 2024; Liu et al., 2024a, 2025a) which lack robustness against variability in phrasing of semantically equivalent information. For example, when retrieving information about the American football team the *"Buffalo Bills"* (Fig. 2), these tools are less likely to retrieve documents that refer to the team as *"Buffalo"*, *"The Bills"*, or even *"the team"*, and suffer from noisy retrieval when queries are expanded with name aliases, as we observe in §5.2.

In this work, we introduce LMEnt (Language Models with annotated Entity mentions), a suite of pretrained models, a knowledge-rich pretraining corpus fully annotated with fine-grained entity mentions, and an entity-based retrieval method which allows full transparency into where a specific entity is mentioned in training. The focus on entity mentions stems from mounting evidence that entity names are central to the construction of knowledge representations in LMs (rather than individual tokens; Li et al., 2021; Meng et al., 2022; Geva et al., 2023; Gottesman and Geva, 2024).

Specifically, the LMEnt suite contains: **(1) A**

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1685–1722, 2026. https://doi.org/10.1162/TACL.a.746
Action Editor: Francois Yvon. Submission batch: 2/2026; Revision batch: 4/2026; Published 7/2026.

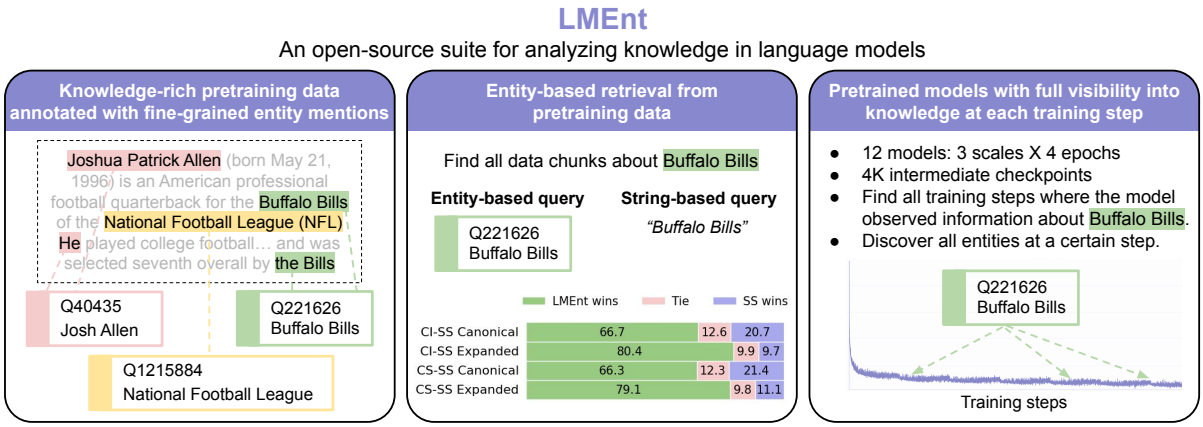

Figure 1: The LMEnt suite is composed of three components: (left) fine-grained entity mentions for every document in the pretraining corpus, (middle) an index that retrieves by the entity QID and outperforms string-based retrieval methods, (right) 12 models trained on 1, 2, 4, and 6 epochs where each step can be mapped to the entities it mentions, and each entity can be traced to the steps that introduce it.

pretraining corpus built on English Wikipedia with 7.3M fine-grained entity annotations across all documents, derived from three complementary sources: hyperlinks, entity linking, and coreference resolution (Fig. 1, left, §2). **(2) An Elasticsearch index with over 10.5M chunks**, enabling retrieval of all pretraining data chunks mentioning a specific entity by its unique Wikidata identifier (Fig. 1, center, §3). LMEnt's entity-based retrieval outperforms string-based methods similar to WIMBD (Elazar et al., 2024) and Infinigram (Liu et al., 2024a) on 66.3–80.4% of entities while maintaining >97% precision as retrieval depth grows, where string-based precision drops to as low as 27%. **(3) A collection of 12 LMs** with 170M, 600M, and 1B parameters, pretrained on the annotated data with 110 intermediate checkpoints per epoch and 4K checkpoints total (Fig. 1, right, §4). Despite only training LMEnt models on 0.03%–4.7% of the tokens used to train LMs of similar sizes, on factual question answering (Mallen et al., 2023), LMEnt models reach comparable performance (7.4% overall, 66% on questions about frequently co-occurring entities) to Pythia-1.4B (Biderman et al., 2023) (8.7%, 67%) and OLMo-1B (Groeneveld et al., 2024) (10.4%, 66%). They do have noticeably lower performance than OLMo-2-1B (OLMo et al., 2024) (13.9%, 74%) and SmolLM-1.7B (Allal et al., 2025) (14.9%, 73%) mostly due to recall failures on rare facts (§5).

Beyond releasing this resource, we use LMEnt to explore several open questions about how pretraining data shapes knowledge in LMs. In §6, we apply LMEnt to study knowledge acquisition and forgetting trends, and in-context editing. Because LMEnt reveals which training steps introduce chunks mentioning a given entity, we can examine how both fact frequency and location in training correlate with learning and forgetting. We find that LMs learn, and also forget, frequent facts more over the course of training, and that these trends differ across relation types (§6.1). Further, leveraging the diverse entity mention annotations, we find that LMs internalize facts better when entities are named explicitly in the data—an analysis infeasible with existing string-based tools (§6.2). Finally, we show that in-context editing success correlates with the frequency of entity co-occurrence in the data (§6.3) and to the rate of inconsistencies in the model's predictions over training (§6.4).

Overall, LMEnt is a novel tool for precisely tracking information during pretraining, laying the groundwork for studying the interplay between pretraining data and knowledge representations in LMs. We release LMEnt to the community as an open-source suite on HuggingFace, including the pretraining data, fine-grained entity mentions, pretrained models, and intermediate checkpoints. For ways LMEnt can support research into plasticity, factuality, hallucinations, and pretraining dataset composition, see §8.1.

2 Labeling the English Wikipedia with Fine-Grained Entity Mentions

To support analyzing knowledge evolution over training, we select a corpus rich in factual con-

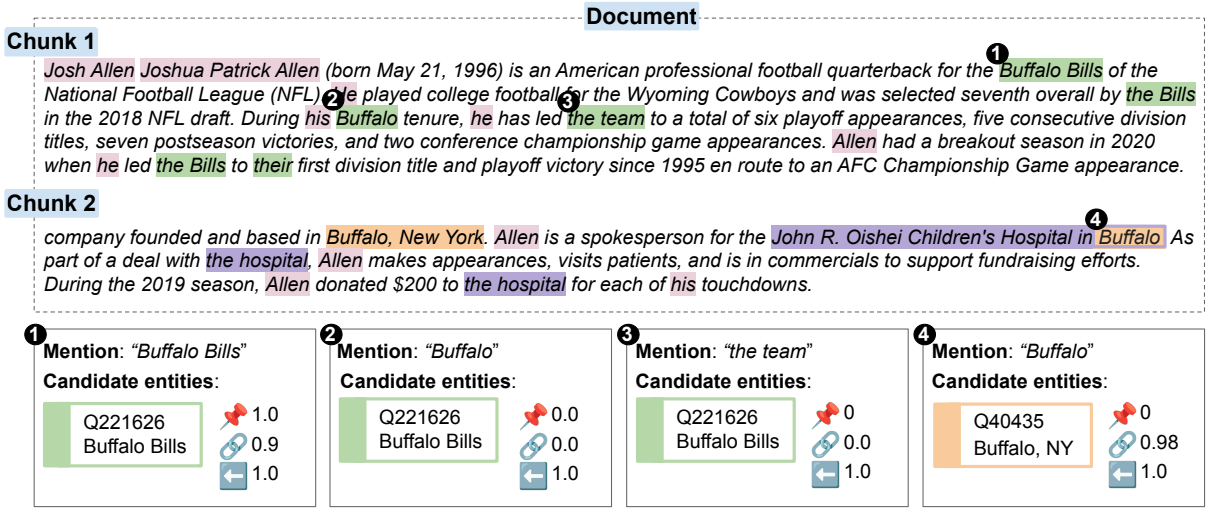

Figure 2: The document about *"Josh Allen"* is split into two chunks that are processed independently during pretraining. In Chunk 1, Q221626 is explicitly mentioned in Mention (1), found by hyperlinks and entity linking, and implicitly mentioned in Mention (3), found by coreference resolution. Although Q221626 (the football team) and Q40435 (the city of Buffalo) share the surface form "*Buffalo*," LMEnt correctly disambiguates them: Mention (2) is resolved to Q221626 with 1.0 confidence through coreference resolution, whereas Mention (4) is linked to Q40435 with 0.98 confidence using entity linking.

tent and annotate its documents with fine-grained entity mentions (Fig. 1, left). These annotations enable tracking exactly which entities a model observes at each training step when training on this data. Next, we explain our choice of Wikipedia for pretraining data, and the annotation process.

Pretraining data Wikipedia is a natural choice for the source of pretraining data because it is a knowledge base that is structured around entities, such as people, locations, and world events. Importantly, there is an integrated system for disambiguating entities using unique identifiers, called *Wikidata QIDs* (Vrandečić and Krötzsch, 2014), and related entity pages are connected through hyperlinks. Moreover, it provides a snapshot of world knowledge at a given time which minimizes contradicting information. Wikipedia is also commonly used to build knowledge benchmarks (e.g., Petroni et al., 2021; Sciavolino et al., 2021; Lewis et al., 2021; Mallen et al., 2023), making it well-suited for tracking how changes in pretraining affect learning.

Annotation objectives Our goal is to be able to query which entities a model saw at any given training step. To achieve this, we built an annotation process that maps character spans in the pretraining data which mention an entity, called *entity mentions*, to unique Wikidata QIDs.[1] Entity mentions are created for each data chunk, but are not exposed to the model during training—rather, the QIDs are the keys used to retrieve all the data associated with an entity for post-hoc analysis. We design entity mentions to satisfy the following criteria (see Fig. 2 for illustration):

C1: Disambiguate between entities with similar names and aliases, e.g., *"Buffalo"* for the football team and *"Buffalo"* for the city.

C2: Capture entities referenced indirectly through descriptive phrases and pronouns, e.g., *"the team"*, *"their"* and *"the hospital"*.

C3: Resolve entity mentions using the full document context e.g., *"Allen"* (Chunk 2) should be annotated with reference to his explicit name, *"Joshua Allen"* (Chunk 1).

We meet **C1** and **C2** by leveraging three complementary sources for annotating entity mentions (§2.1), and **C3** by applying the scoring process to the entire document before chunking (§2.2).

2.1 Entity Mention Sources

Hyperlinks identify the first occurrences of explicit entity mentions like Mention (1) in Fig. 2

[1]The annotation pipeline used 8 H100 GPUs (80GB VRAM each). Running coreference resolution and entity linking took 5 days with full GPU utilization.

which links to the URL of the football team's Wikipedia page, and can be mapped to Q221626 using Wikidata's Query Service. To extract hyperlinks, we parse the raw XML Wikipedia dump and extract the `url` attributes from `href` tags.

Entity linking is used to increase coverage of direct mentions like *"Buffalo"* since hyperlinks offer limited coverage within a document.[2] We use ReFinED (Ayoola et al., 2022), a state-of-the-art modular entity linking system that performs mention detection, fine-grained typing, and entity disambiguation. ReFinED supports zero-shot entity linking by encoding the mention context and candidate entity descriptions using RoBERTa (Liu et al., 2019), and selecting the entity whose description embedding best aligns with the mention representation. By replacing the underlying base of entity descriptions from our Wikipedia dump, we can identify mentions of new entities and map them to their QIDs without retraining ReFinED.

Coreference resolution connects indirect mentions, like pronouns, aliases, and generic descriptors, to their entity QIDs. The two sources above are used to extract explicit entity mentions; however, entities may still be referred to implicitly, e.g., *"the team"* in Fig. 2. To fill in this gap, we use the Maverick coreference resolution model (Martinelli et al., 2024), which achieves state-of-the-art performance on WikiCoref (Ghaddar and Langlais, 2016). To run Maverick at scale, we reduce its memory footprint by replacing the model backbone, allowing us to parallelize inference on documents (see §B.1 for details). Maverick outputs clusters of mentions that refer to the same entity. However, these mentions must still be mapped to their QIDs, which we discuss below.

2.2 Scoring Entity Mentions in a Document

Every source yields a set of entity mentions per document in the corpus, and each entity mention is defined by its character span $m = (c^{\text{start}}, c^{\text{end}})$. To consolidate the mentions found across sources, we define source-specific scoring procedures which quantify how confidently a mention can be mapped to a QID (Fig. 2):

- **Hyperlinks (H)** 📌 This is the most reliable source since hyperlinks are manually added by human editors. We define $\mathbf{H}(m, \texttt{Q221626}) = 1$ if there is a hyperlink spanned by m directing to the Wikipedia page associated with Q221626, and $\mathbf{H}(m, \texttt{Q221626}) = 0$ otherwise.
- **Entity linking (EL)** 🔗 The ReFinED model already performs entity disambiguation and provides a score reflecting its confidence that a mention links to a QID, e.g., $M(m, \texttt{Q40435}) = 0.98$. Therefore, we simply use this score such that $\mathbf{EL}(m, \texttt{Q40435}) = M(m, \texttt{Q40435})$.
- **Coref (C)** ⬅️ Suppose we are trying to resolve the QID of the implicit mention, *"the team"*. We run Maverick on the document, and it finds that *"the team"* is related to the cluster of mentions containing *"Buffalo Bills"*, *"the Buffalo"*, *"the Bills"*, and *"their"*. How can we leverage the cluster to identify the QID of *"the team"*? We compute a distribution of scores over all QIDs mapped to some mention in the cluster by the other two sources. In this case, Q221626 is the only QID supported by this cluster so $\mathbf{C}(\text{"the team"}, \texttt{Q221626}) = 1.0$. Since this score is computed using the entire cluster, the **C** score is the same across all mentions in the cluster. Occasionally, a coreference mention encapsulates multiple entities, as in *"John R. Oishei Children's Hospital Buffalo"*, which leads to a score distribution over several QIDs. In this case of mapping ambiguity, we rely on textual similarity of mentions in the cluster to promote one QID over another. §B.2 provides the formal definition of this score, and a detailed explanation of how we resolve mapping ambiguity.

The final entity mention structure maps each mention m to a list of candidate QIDs with up to three source scores per candidate. This list structure accommodates spans that encompass multiple entities, and multiple scores per candidate allow us to filter chunk retrieval based on source type and confidence thresholds (§3.2). §B.3 includes a qualitative analysis of error patterns over a random sample of 112 mentions, justifying the design of the scoring procedures.

3 Retrieval from Pretraining Data

Now that we have entity mentions for every document (Fig. 1, left), we need to process the docu-

[2] Wikipedia style guidelines recommend linking only the first mention of an entity in a document: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking.

ments into chunks of information for pretraining, and associate the relevant entity mentions with each chunk. By iterating over the entity mentions of each chunk at a particular training step, we can track the entities seen over training (Fig. 1, right). §3.1 describes how documents are processed into chunks, and §3.2 discusses entity-based retrieval over our index of chunks (Fig. 1, center).

3.1 Data Processing for Training

Tokenization Documents are tokenized with the dolma2-tokenizer (Soldaini et al., 2024), resulting in a dataset with 3.6 billion unique tokens. For reference, this corresponds to around 0.09% of the tokens used to train OLMo-2 (OLMo et al., 2024).

Chunking A chunk is a sequence of tokens that the model independently processes per training step. Unlike existing chunking strategies, like *concat-and-chunk* used in Liu et al. (2019); Ott et al. (2019); Touvron et al. (2023a,b), we ensure that each chunk only contains content from a single document using the Variable Sequence Length Curriculum (Pouransari et al., 2024). This prevents the model from learning spurious correlations by attending to unrelated documents.

Chunking curricula split documents at some sequence length, which can split entity mentions across chunk boundaries. To prevent this, we modify the chunking logic to terminate just before a mention and pad the remaining tokens to the desired length. The entity mentions for a chunk are defined as those fully contained within it, and we adjust them so they are relative to the chunk's boundaries. Practically, we extend the OLMo-Core framework (OLMo et al., 2024) to include entity mentions in the chunk metadata.

3.2 Entity-Based Chunk Retrieval

We build an Elasticsearch index (Banon, 2010) containing one entry per chunk in the pretraining corpus. Each entry stores the `chunk_id`, the chunk text, and its associated entity mentions. As illustrated in Fig. 2 and described in §2.2, each entity mention is linked to a set of candidate entities, where each candidate is identified by its QID and accompanied by three source-specific scores.

To **retrieve all chunks** that mention a target entity, we query the index for chunks containing at least one mention whose candidate list includes the target entity's QID (see §2.2). Retrieval can also be filtered by source type and score thresholds. Retrieval algorithms typically apply an ordering scheme to the items they return (Robertson et al., 1994). To accommodate this, we assign a score to each chunk based on the entity mentions that match the retrieval query. For every matching mention, we compute a weighted average of its three source-specific scores. The chunk's final ranking score is defined as the maximum over the average scores of all its matching mentions.

To **retrieve all training steps** that mention a target entity, we first retrieve all corresponding chunks from the index, and then map each retrieved chunk to the training step in which it was introduced. This mapping is straightforward since the `chunk_id` is the index of the chunk in the dataset, and its associated training step is the index of the batch that contains it in the dataloader.[3]

4 The LMEnt Suite

With the procedures for annotating entity mentions and processing pretraining data established, we now provide an overview of the resulting pretraining data index and `LMEnt` models.

Pretraining data index Our pretraining data includes 3.6B tokens over 10.5M chunks, and features more than 7M different entities. The 400M mentions extracted are composed of 115M mentions from hyperlinks, 203M from entity linking, and 310M from coreference resolution. Additional statistics are provided in §A.1, Tab. 2.

Pretrained models We introduce a collection of models trained on the dataset, which serve as a testbed for analyzing knowledge evolution over training. Specifically, we train models of three different sizes, 170M, 600M, and 1B parameters, based on the OLMo-2 architecture (OLMo et al., 2024). Notably, the 170M model has the compute-optimal size for 3.6B tokens (Hoffmann et al., 2022). Each model is trained for 1, 2, 4, and 6 epochs (3.6B, 7.2B, 14.4B, 21.6B tokens, respectively), resulting in four variants per size. Each epoch initializes a different batching and ordering of the chunks. We also release intermediate checkpoints every 1,000 training steps, yielding 110 checkpoints per epoch and proportionally more for longer training schedules. Additional details on model training are in §B.4.

[3]There are multiple chunks per batch, determined by a global batch size of 32,648 tokens.

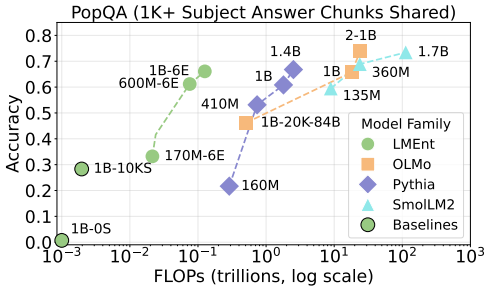

Figure 3: Accuracy on PopQA questions, whose subject and answer entities co-occur frequently, as a function of compute budget. LMEnt models achieve comparable performance to other models with better compute efficiency.

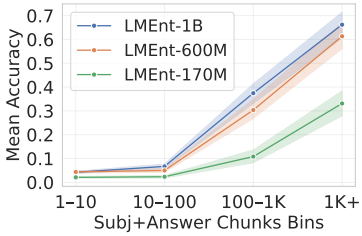

Figure 4: Mean accuracy on PopQA questions binned according to the number of chunks that the subject and answer entities co-occur in. Increasing model size helps learning associations between entities that appear more frequently together.

5 Experiments

In the previous sections, we described the construction and composition of the LMEnt suite. Here, we establish the credibility of the suite by showing that LMEnt models match the knowledge recall performance of existing open-source models (§5.1), and that entity-based chunk retrieval using LMEnt entity mentions outperforms string-based methods (§5.2).

5.1 Model Performance

Experimental setup To evaluate the LMEnt models, we use benchmarks that test for knowledge in Wikipedia. We choose two widely adopted benchmarks: PopQA (Mallen et al., 2023) and PAQ (Lewis et al., 2021). PopQA contains questions about 11K entities with varying popularities, and PAQ has 65M questions generated from Wikipedia passages. To keep computational costs manageable, we subsample PAQ and only consider samples containing the same subject entities found in PopQA. This results in an evaluation set of approximately 70K questions.[4] Results on PAQ are found in §A.2. In addition to knowledge-centric tasks, we report results on commonsense reasoning, multiple choice, and reading comprehension tasks in §A.4. Since LMEnt models are not instruction-tuned, we convert all samples into cloze-style prompts, such that the answer is the next expected phrase. For example, the original QA-style query "*What TV network does Funny or Die Presents air on?*" translates to "*Funny or Die Presents airs on the TV network*". Full details of this conversion process are in §B.6.

[4]Evaluating a 1B parameter model on all PAQ questions using one H-100 80GB GPU would take 200 days.

Baselines We evaluate the performance of LMEnt models against leading open-source models of comparable sizes, including Pythia (160M, 410M, 1B, 1.4B) (Biderman et al., 2023), OLMo (1B, 1-20K-84B) (Groeneveld et al., 2024), OLMo-2-1B (OLMo et al., 2024), and SmolLM2 (135M, 360M, 1.7B) (Allal et al., 2025). We include OLMo-1-20K-84B—an intermediate checkpoint of OLMo-1B trained for 20K steps on 84B unique tokens—because it is the most comparable baseline to our models in terms of training token count.[5] We also include two intermediate LMEnt model baselines: LMEnt-1B-0E (randomly initialized) and LMEnt-1B-0.1E (trained for 10K steps). These serve to illustrate how more training data improves knowledge recall capabilities.

LMEnt models are valid resources for studying knowledge Fig. 3 shows the performance of LMEnt models and the baselines as a function of compute budget, measured in FLOPs. Despite being trained with two orders-of-magnitude less compute, LMEnt models achieve performance comparable to both Pythia and OLMo-1B on questions in PopQA whose subject and answer entities co-occur frequently in the pretraining dataset. We also present results for all frequency levels in §A.2 which show similar trends. However, since PopQA is dominated by questions about tail facts, overall accuracy is lower across all models.

Notably, the LMEnt-600M model surpasses the OLMo-1B intermediate checkpoint and the Pythia-1B checkpoint, which reaffirms that pretraining on Wikipedia enables models to acquire substantial factual knowledge (Devlin et al., 2019)

[5]Notably, OLMo-1-20K-84B is trained on 4 times more data than our largest LMEnt-1B model, and it may not have observed all the facts we evaluate on.

	Retrieval	Match Examples
LMEnt	QID + Scores	Buffalo, New York; Buffalo; the Queen City; this city; the City's; the City of Buffalo; the city Buffalo; a town on the banks of Lake Erie; his hometown
CS-SS	Canonical	Buffalo, New York
	Extended	Buffalo, New York; Buffalo, NY; Buffalo, N. Y.; City of Light; The Queen City; The Nickel City
CI-SS	Canonical	buffalo, new york
	Extended	buffalo, new york; buffalo, ny; buffalo, n. y.; city of light; the queen city; the nickel city

Table 1: Matched entity mentions of "*Buffalo, New York*" for different retrieval methods. Entity-based retrieval over the LMEnt index matches on Q40435. The Canonical method matches only the entity's canonical name, whereas the Extended method matches any of its aliases.

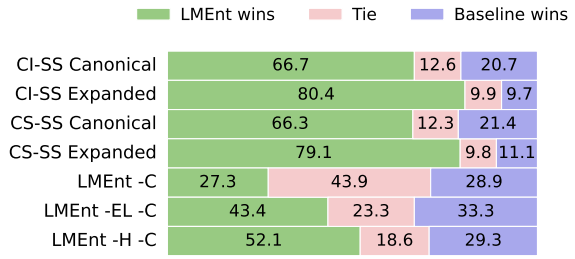

Figure 5: Pairwise wins rate for entity-based retrieval (LMEnt) with multiple string-based methods and ablations. LMEnt's entity-based retrieval outperforms string-based methods by 66.7%–80.4%. Ablations (bottom three rows) show that hyperlinks and entity linking are the most crucial sources of entity mentions.

and achieve better compute-performance trade-offs. Though this evaluation aligns with the pretraining objective of LMEnt models, while the baseline models are trained on a broader mixture of web, math, and code data, it supports their validity as resources for studying knowledge.

Measuring the effect of model scale on knowledge encoding Since all LMEnt models are pretrained on the same data, we can isolate the effect of model size on knowledge acquisition (Roberts et al., 2025). Fig. 4 shows that increasing model size improves learning of frequent facts (subject and answer co-occur in ≥ 100 chunks), while having little effect on tail facts (1–100 chunks). We leave exploring how scaling model size further can enhance knowledge learning for future work.

5.2 Chunk Retrieval Performance

A key use-case of the LMEnt index is retrieving all the pretraining data chunks that mention a specific entity. Here, we compare the quality of entity-based retrieval (§3.2) to existing string-based retrieval methods.

Experimental setup We evaluate on a test set of 1K entities, chosen via stratified sampling by their number of hyperlinks. We use the entity-based retrieval over the LMEnt index to retrieve their corresponding chunks, matching both on entity QID and on at least one of the score thresholds: $\mathbf{H} = 1$, $\mathbf{EL} \geq 0.6$, $\mathbf{C} \geq 0.6$. These thresholds were empirically determined using a development set of 60 entities, described in §A.5. We then compare the set of retrieved chunks to those retrieved by popular string-based methods.

To evaluate retrieval precision, we use Gemini 2.5 Flash (Comanici et al., 2025) as a judge (Gu et al., 2024; Li et al., 2025) which predicts "Yes" or "No" based on whether the mention directly relates to the target entity. The prompt used and the statistical test justifying our use of an LLM-judge are provided in §B.5. Since some entities are mentioned very frequently (Fig. 16), we randomly sample 100 chunks from the set retrieved per method and entity, and measure the absolute number of chunks judged as "Yes" for each entity.

Baselines We compare to string-based baselines that retrieve chunks using either case-sensitive (CS-SS) or case-insensitive (CI-SS) exact matches of entity names within the chunk text. The Canonical variant matches only on the entity's canonical name, and the Extended variant additionally matches against the entity's Wikidata aliases which is a common strategy used to improve recall (e.g., Cohen et al., 2024; Yang et al., 2024).

The CS-SS Canonical baseline approximates the Infinigram (Liu et al., 2024a) tool which matches based on case-sensitive n-grams, and CI-SS Canonical baseline approximates WIMBD (Elazar et al., 2024) which retrieves on exact case-insensitive string matches. Although CS-SS Canonical relies on exact string matches rather than n-gram matches, it is a stronger baseline since shorter references introduce noise due to alias ambiguity (e.g., the n-gram "*Buffalo*" indiscriminately matches chunks that refer to the city, animal, or football

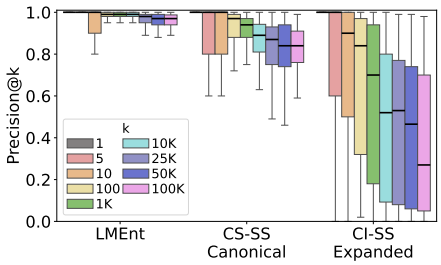

Figure 6: Precision at various retrieval depths (k), the top-k chunks retrieved per entity. As depth increases, LMEnt maintains $>$ 97% precision, while CS-SS Canonical and CI-SS Expanded decrease to 84% and 27%.

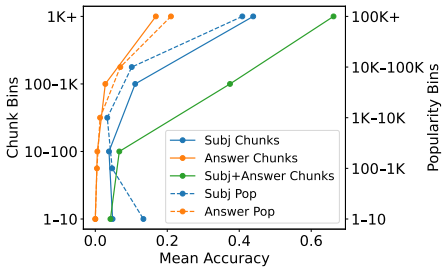

Figure 7: Accuracy for LMEnt-1B on PopQA. Subj+Answer Chunks counts chunks that mention both the subject and answer entities of a question. Answer Chunks and Subject Chunks count chunks that mention the answer and subject entities individually. Subj Pop and Answer Pop are pageview popularities from Mallen et al. (2023). Subj+Answer Chunks correlates best with the model's performance.

team). Tab. 1 summarizes the retrieval strategies used by LMEnt and the baselines. For clarity, we also include examples of mentions that triggered the retrieval of chunks for Q40435.

Entity-based retrieval largely outperforms the baselines Fig. 5 presents pairwise win rates between retrieval methods, where a win is defined as one method retrieving more "Yes"-judged chunks than another for a given entity. Entity-based retrieval (LMEntwins) consistently outperforms all other methods, retrieving more correct mentions for 66.3%–80.4% of the entities. Surprisingly, the Expanded variants of both CI-SS and CS-SS perform worse than their canonical counterparts by 40%, suggesting that additional noise introduced by alias expansion outweighs any gains in recall (§A.6, Fig. 15). Additionally, we observe that entity-based retrieval is superior to string-based methods across all entity frequency levels—notably on tail and torso entities which together make up 99.7% of the entities in Wikipedia—and often by a substantial number of chunks (§A.7, Fig. 16; §A.8, Fig. 17).

Ablating entity mention sources reduces retrieval quality We also ablate sources of entity mentions in entity-based retrieval and examine the impact on win rate (Fig. 5 last three rows). Hyperlinks are the most valuable component as relying solely on them (LMEnt -EL -C) wins on 33.3% of entities. In contrast, using only entity linking (LMEnt -H -C) lowers wins to 29.3%. This indicates that while hyperlinks play a crucial role, entity linking alone remains reasonably effective—highlighting the potential of applying LMEnt's annotation pipeline to pretraining corpora beyond Wikipedia, where hyperlinks are unavailable. Removing coreference resolution (LMEnt -C) has little effect, because judging chunks with only implicit mentions is difficult. For example, when Donald Trump is referred to solely as "*his Mexico City Policy*", it is difficult to identify that "*his*" implicitly refers to Trump without the context earlier in the document which mentions him explicitly.

LMEnt maintains high precision as the number of retrieved chunks increases We also analyze the precision of retrieved chunks at varying retrieval depths, k. For each method, we consider only the top-k chunks retrieved per entity, and randomly sample 100 chunks if $k \geq 100$, for feasibility. An LLM then judges the mention, and precision is computed as the fraction of "Yes" responses out of k. As shown in Fig. 6, LMEnt's entity-based retrieval maintains high precision $\geq$ 97% as k increases, while precision substantially declines for all other methods, reaching 27% and 84% for CI-SS and CS-SS, respectively. This indicates that as more chunks are retrieved, string-based methods introduce noise while LMEnt maintains high quality.

Entity co-occurrence better aligns with model performance than popularity The number of Wikipedia pageviews for a subject entity, denoted as Subj Pop, has been the standard proxy for real-world popularity and is known to correlate with a model's ability to recall facts about it (Mallen et al., 2023). In Fig. 7, we compare model performance against several potential indicators.

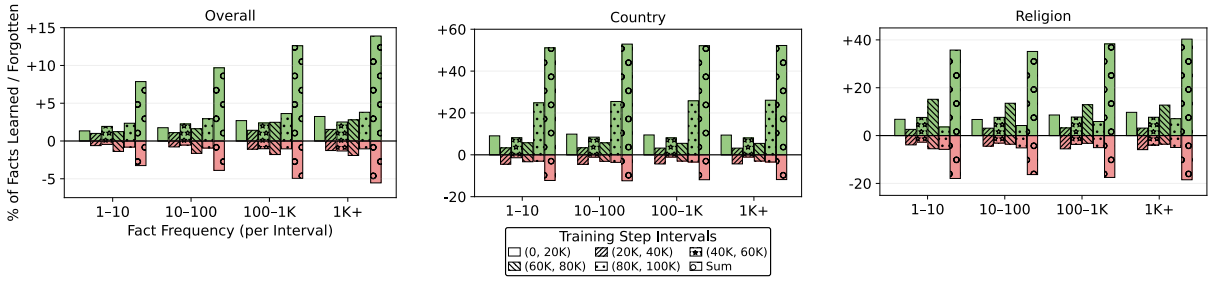

Figure 8: Percentage of facts learned and forgotten by fact frequency between intermediate checkpoints of LMEnt-1B. Overall (left) shows that while fact frequency correlates with knowledge gains, both learning and forgetting increase with frequency. The learning and forgetting trends for specific Country (middle) and Religion (right) relations differ across training step intervals. Overall learning is lower since the model struggles to answer questions about certain relations (e.g., Screenwriter in Fig. 18).

In PopQA, each question concerns a (Subject, Answer) entity pair. Leveraging LMEnt's entity-based retrieval, which can retrieve chunks containing many co-occurring entities, we introduce a new indicator: Subject+Answer Chunks. This is defined as the number of chunks in which the subject and answer entities co-occur. This co-occurrence shows the strongest correlation with model performance. Interestingly, this trend is also consistent for models trained on pretraining data that extend beyond Wikipedia (§A.3, Fig. 14).

6 Applications of LMEnt

We demonstrate the utility of LMEnt by focusing on two fundamental applications: knowledge acquisition and in-context knowledge editing. Specifically, we study learning and forgetting trends over training (§6.1), how entity mention forms relate to learning (§6.2), and how in-context knowledge editing relates to entity co-occurrence frequency (§6.3) and inconsistencies in model predictions over training (§6.4).

6.1 Knowledge Acquisition in Pretraining

We study at which point during training models acquire factual knowledge and how this relates to fact frequency, which has been previously approximated using document frequency (Kandpal et al., 2023), term co-occurrence (Elazar et al., 2022; Merullo et al., 2025) and n-gram (Wang et al., 2025) frequencies. In contrast to prior work, we use LMEnt entity mentions, which provide more accurate estimates of fact frequency compared to string-based methods.

Experiment We evaluate the LMEnt-1B model every 20K steps on PopQA. We define a fact by a (Subject, Answer) entity pair, which has been shown to be an effective fact proxy (Elsahar et al., 2018). A fact is considered "learned" if the model correctly answers at least one question linking the subject to the answer. Between training steps $(i, i{+}20\text{K})$, we consider all the facts seen in the training step interval and measure (a) *fact frequency*: the number of chunks in this interval where the subject and answer co-occurred, (b) *% of learned facts*: facts that were not learned at step i and then learned at step $i{+}20$K, and (c) *% of forgotten facts*: facts that were learned at step i and then not learned, i.e. "forgotten", at step $i{+}20$K.

Findings Fig. 8 (left) shows the percentage of facts learned and forgotten as a function of both fact frequency and the training step interval in which the fact was introduced. Focusing on trends across fact frequencies (left, Sum), the model recalls frequent facts from memory more effectively, which is consistent with prior works (Elazar et al., 2022; Yu et al., 2023; Liu et al., 2025b). However, we also see that LMs forget very frequent facts (1K+, -5.2%), and successfully learn rare facts (1-10, +8%). This implies that fact frequency alone does not dictate learning and forgetting dynamics, and may be confounded by other factors like when a fact is introduced in training (discussed next), or its entity mention forms (see §6.2).

To analyze how the timing of a fact's introduction affects its acquisition, consider any fact frequency bucket in Fig. 8 (left). We see that the model acquires knowledge in the first 20K training steps, and continues to learn and forget facts as training proceeds, with the highest percentage of facts learned in the last 20K training steps. Moreover, knowledge acquisition is not correlated with the cosine learning rate schedule, which decreases

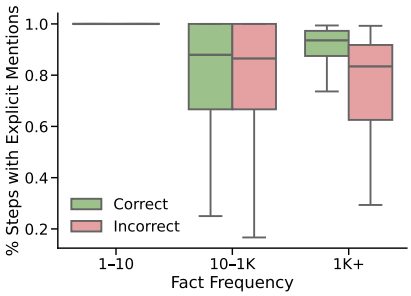

Figure 9: LMEnt-1B is more accurate when entities co-occur in their explicit forms, implying that entity mention forms influence learning.

after 1K warmup steps, as acquisition is highest in the last training step interval where the learning rate is lowest.

Further, we see varied learning dynamics when restricting this analysis to specific relations: Country (middle) and Religion (right). For example, the training step interval with highest learning is (80K, 100K) for country, but (60K, 80K) for religion. This observation aligns with previous findings that knowledge recall trends over training differ across relations and entity types (Yang et al., 2024; Balesni et al., 2024). Results for additional relations can be found in §A.9.

6.2 Entity Mention Forms and Learning

The previous experiment shows that fact frequency alone does not fully explain learning and forgetting trends. So far we treated all co-occurrences of the subject and answer entities as equally informative, but entities can be mentioned with various entity mention forms, e.g., "*Donald Trump*" vs. "*the 47th president*". Here, we leverage the fine-grained LMEnt entity mentions to examine the relation between how an entity is mentioned in the pretraining data and the model's ability to learn facts about it. Notably, such analysis is infeasible with string-based methods since they suffer from alias ambiguity.

Experiment We restrict our analysis to chunks in which a (Subject, Answer) entity pair co-occur, and determine whether both the subject and answer entities are mentioned explicitly (e.g., canonical names like "Donald Trump") or implicitly through aliases (e.g., "*the 47th president*") or pronouns (e.g., "*he*"). For each (Subject, Answer) entity pair, we then compute the proportion of training steps that introduce chunks which explicitly mention both entities.

Findings Fig. 9 shows the percentage of training steps that mention both the subject and answer entities explicitly, grouped by fact frequency and correctness. For rare facts (1–10), which appear in more focused contexts, all of the training steps contain explicit mentions regardless of correctness, indicating that a different factor is at play (e.g., location in training). For more frequent facts (10-1K), correctly answered questions are associated with a slightly higher proportion of steps with explicit mentions, and this gap is more pronounced for highly frequent facts (1K+). This suggests that mention forms affect how models learn factual knowledge, and may contribute to why LMs forget frequent facts. We encourage future work using LMEnt to systematically study these effects, including causal interventions that replace implicit mentions with explicit ones to isolate impacts on learning and forgetting.

6.3 Entity Co-occurrence and In-Context Editing Success

We saw that the co-occurrence of subject and answer entities correlates with the model's performance on related queries (§6.1). Here, we extend this analysis to study how co-occurrence statistics affect the model's ability to edit its learned knowledge in-context (Zheng et al., 2023; Yu et al., 2023; Cohen et al., 2024).

Experiment We evaluate checkpoints of the LMEnt-1B model trained for six epochs that were saved every 100K training steps. Focusing on the questions in PopQA that were answered correctly by at least one of these checkpoints, we generate up to 50 counterfactual answers per question by retrieving all the chunks that mention the subject entity and choosing counterfactual entities that (i) match the original answer type and (ii) co-occur with the subject in the pretraining data. For example, for the fact "*The capital of France is Paris*", the city "*Rome*", which also co-occurs with "*France*" in training, can serve as a counterfactual answer. To evaluate in-context knowledge editing, we prompt every checkpoint with the edited fact followed by the original cloze-style prompt, e.g., "*The capital of France is Rome. The capital of France is*", and consider editing successful if the LM outputs the counterfactual answer e.g., "*Rome*".

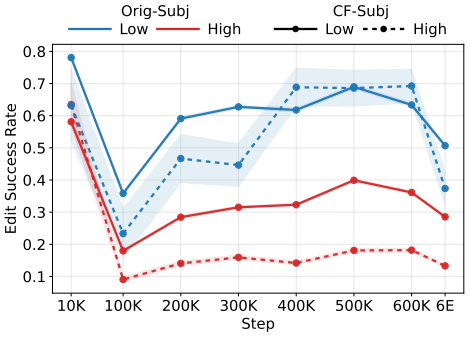

Figure 10: In-context editing success across training checkpoints, split by the co-occurrence frequencies of `Orig-Subj` and `CF-Subj` pairs. LMs are more rigid to editing facts with strong associations with either pair (red $<$ blue, dash $<$ solid). `6E` refers to the final model at epoch six.

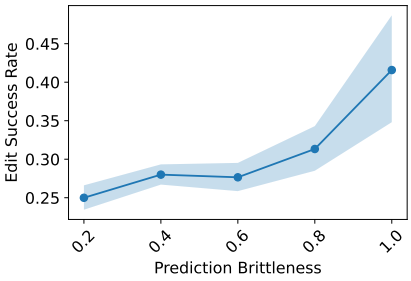

Figure 11: In-context editing success of the `LMEnt`-1B model increases with prediction brittleness. Logistic regression found a monotonic relationship ($\beta = 0.54, z = 4.97, p = 6.65e^{-07}$).

Findings Fig. 10 shows in-context editing success over checkpoints, split by the co-occurrence frequencies of two entity pairs: `Orig-Subj`, the (Subject, Original Answer) entity pair like ("*France*", "*Paris*"), and `CF-Subj`, the (Subject, CF Answer) entity pair like ("*France*", "*Rome*"). Each pair is labeled low-frequency (1–10 training steps) or high-frequency (100+), and an edit is categorized by both its `Orig-Subj` and `CF-Subj` frequency labels, e.g., the dashed blue line shows edits with low `Orig-Subj` and high `CF-Subj`.

We find that `Orig-Subj` co-occurrence plays the dominant role—when the original answer and subject are weakly associated (blue), in-context editing success is highest. `CF-Subj` co-occurrence also matters—when the counterfactual entity and subject are weakly associated (solid), the LM more readily adopts the new relation. Examining trends over training, we see high edit success (58%-78%) for the first checkpoint (10K training steps). After approximately one epoch, the model's knowledge becomes rigid for its strong associations (red/dashed). However, for facts with weaker associations, the model exhibits a non-monotonic pattern: editability decreases after the first epoch, partially recovers mid-training, and then declines again at the end. This behavior aligns with prior work showing that continual learning is more effective from intermediate checkpoints (Springer et al., 2025).

6.4 Model Brittleness and In-Context Editing

Here, we examine how inconsistencies in the model's predictions across training relate to the model's ability to edit its knowledge in-context.

Experiment We evaluate the final checkpoint using the same dataset as in §6.3. We define *prediction brittleness* as the frequency of "correctness flips" (Correct→Incorrect and Incorrect→Correct) across six checkpoints, sampled every 100K training step, with steps beginning from the first third of training to exclude noisy early-stage prediction fluctuations. Examples whose correctness never changes are omitted as they do not exhibit prediction brittleness.

Findings Fig. 11 shows that in-context editing success increases with prediction brittleness: facts whose answers fluctuate more during training are easier to edit. A binomial logistic regression analysis confirms the relation is monotonic and positive. This result highlights a connection between training dynamics and the editability of the final model, motivating future work on what drives such brittleness and the relationship between knowledge stability and model safety.

7 Related Work

Analysis of data acquisition in LMs Prior work on how pretraining data shapes factual knowledge has largely relied on synthetic setups, due to limited metadata available for pretraining corpora. These include injecting synthetic facts (Chang et al., 2024), constructing artificial biographies (Allen-Zhu and Li, 2024; Zucchet et al., 2025), pretraining on structured fact triplets (Kassner et al., 2020; Hron et al., 2024), and identifying knowledge circuits using synthetic data (Ou et al., 2025). In contrast, `LMEnt` enables controlled investigation in realistic settings using natural data.

Other lines of work examine the influence of individual training examples on learning (Han et al., 2020; Hammoudeh and Lowd, 2024), localize where knowledge is encoded in model parameters (Kim et al., 2025), and analyze which properties of pretraining data give rise to hallucinations (Sun et al., 2026; Zucchet et al., 2025). These approaches can benefit from LMEnt, which allows direct comparison of model checkpoints before and after seeing specific facts.

Retrieval over pretraining corpora LMEnt's entity-based retrieval outperformed string-based retrieval methods (§5.2) which rely on exact matches (Elazar et al., 2024; Liu et al., 2025a) or n-gram matches (Liu et al., 2024a). Many works that study the connection between pretraining data statistics and knowledge acquisition (Razeghi et al., 2022; Kandpal et al., 2023; Elazar et al., 2022; Kang and Choi, 2023; Merullo et al., 2025; Wang et al., 2025) can use LMEnt's entity-based retrieval to refine their findings.

Open-source LM toolkits Similar to LMEnt, Pythia (Biderman et al., 2023), OLMo (Groeneveld et al., 2024; OLMo et al., 2024), LLM360 (Liu et al., 2024b), SmolLM (Allal et al., 2025; Bakouch et al., 2025), and NVIDIA (Kuchaiev et al., 2019; Basant et al., 2025) released a suite of models along with their training data and training framework to facilitate the study of scaling laws and training dynamics in relation to the pretraining data. Other works train their own collection of models to study scaling laws for data curation (Magnusson et al., 2025) and specific downstream tasks (Bhagia et al., 2025). LMEnt differs from prior work by providing a practical toolkit for studying factual knowledge in LMs, with a particular emphasis on easily attributing when and where specific information appears in training.

8 Conclusion and Discussion

We introduce LMEnt, a suite of 12 LMs, pretraining data fully annotated with entity mentions, and an entity-based retrieval index that facilitates studying the connection between knowledge acquisition and pretraining data. We show that LMEnt models are capable of knowledge recall tasks, and that our entity-based retrieval outperforms string-based methods on 66.3%–80.4% of entities. Overall, LMEnt serves as a flexible and extensible testbed for investigating a broad set of questions regarding knowledge and representations in LMs. In the following sections, we list future applications of LMEnt and limitations.

8.1 Future Applications of LMEnt

Knowledge plasticity LMEnt can be used to study the plasticity of knowledge in language models: identifying steps during pretraining when models are more receptive to acquiring new knowledge. This follows prior findings that pretrained models struggle to learn new facts (Lyle et al., 2022), excessive pretraining can make models resistant to fine-tuning (Springer et al., 2025), and LMs store a finite amount of parametric knowledge (Allen-Zhu and Li, 2025).

Improving factuality of LMs Since LMEnt provides entity mentions for each chunk in the pretraining corpus, a possible extension is to leverage them to improve the model's factuality and mitigate hallucinations. For example, one could experiment with different data ordering methods, like grouping together all the chunks mentioning the same entity, or augmenting the pretraining data by replacing implicit mentions with explicit entity names.

Effect of other data sources In this work, we train models on a relatively small and knowledge-rich corpus, which results in LMs that are capable in knowledge tasks, yet perform poorly on out-of-distribution tasks, such as commonsense reasoning (§A.4). Modern LMs are trained on much larger corpora, often surpassing 10T tokens (Allal et al., 2025), derived from sources that are *knowledge-poor*, e.g., coding (Roziere et al., 2023; Hui et al., 2024), synthetic stories (Eldan and Li, 2023), and formal languages (Hu et al., 2025). The LMEnt pretraining dataset can be extended with such sources to analyze if the addition of these tokens improves factuality. The annotation pipeline can also be applied to other knowledge-rich sources that lack hyperlinks, since relying just on entity linking is still effective (`LMEnt -H -C`, Fig. 5).

Mechanistic interpretability The enhanced visibility that LMEnt provides into the training process facilitates a controlled, yet natural, setup for studying the formation of latent knowledge representations and circuits in LMs.

8.2 Limitations

Pretraining data, model size and architecture LMEnt presents a relatively small pretraining corpus and collection of compute-efficient models. Such experimental settings have been successful in recent years, enabling academia to develop procedures that were later scaled effectively, e.g., (Rafailov et al., 2023). As mentioned in §8.1, an interesting future direction is to scale LMEnt annotations to other sources—thereby increasing the size of the pretraining corpus and facilitating studies into how different types of data affect knowledge acquisition. Also, LMEnt can be easily extended to architectures beyond dense transformers like mixture-of-experts (Dai et al., 2024; Muennighoff et al., 2025; Cai et al., 2025), which is already supported by the OLMo framework (Groeneveld et al., 2024; Muennighoff et al., 2025).

Beyond pretraining Since models see the vast majority of tokens during pretraining, this stage is critical for understanding how they acquire and represent knowledge. While our study focuses on pretraining, it is typically followed by mid- and post-training phases. In these stages, models are presented with additional high-quality next-token prediction data and fine-tuning examples (OLMo et al., 2024; Kumar et al., 2025). Future work could extend LMEnt by adding annotations to the data used for mid- and post-training, and further train LMEnt models on this data.

Dense retrievers In this work, we do not compare LMEnt retrieval against dense retrievers such as DPR (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020), since sparse retrieval methods like WIMBD (Elazar et al., 2024) and Infinigram (Liu et al., 2024a) are the predominant approaches for retrieving relevant information in pretraining corpora.

Acknowledgments

We are grateful to Maor Ivgi, Alon Mendelson, Yanai Elazar, and Ohav Barbi for their valuable discussions and feedback. We also thank Noam Steinmetz, Clara Suslik, and Asaf Avrahamy for their participation in the LM-judge evaluation. This work was supported in part by AMD's AI & HPC Fund, the Mila P2v5 grant, the Mila-Samsung grant, the Google PhD Fellowship program, the Alon scholarship, and the Israel Science Foundation grant 1083/24.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. ArXiv preprint: 2204.06031.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. SmolLM2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 1067–1077. PMLR / OpenReview.net.

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM3: Smol, Multilingual, Long-Context Reasoner.

Mikita Balesni, Tomek Korbak, and Owain Evans. 2024. Lessons from studying two-hop latent reasoning. *arXiv preprint arXiv:2411.16353*.

Shay Banon. 2010. Elasticsearch: A distributed, RESTful search and analytics engine.

Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, et al. 2025. Nvidia Nemotron Nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. 2025. Establishing task scaling laws via compute-efficient model ladders. In *Second Conference on Language Modeling*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

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, 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, 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, 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, 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, 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, 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, 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, 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, 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 Ra-

malho, 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, 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, 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. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. ArXiv preprint: 2507.06261.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2024. What's in my big data? In *International Conference on Learning Representations*, volume 2024, pages 7735–7790.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model's factual predictions. *arXiv preprint arXiv:2207.14251*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Abbas Ghaddar and Philippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142.

Daniela Gottesman and Mor Geva. 2024. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4019.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on LLM-as-a-judge. *The Innovation*.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5553–5563.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: decoding-enhanced BERT with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10.

Jiri Hron, Laura A Culp, Gamaleldin Fathy Elsayed, Rosanne Liu, Jasper Snoek, Simon Kornblith, Alex Rizkowsky, Isabelle Simpson, Jascha Sohl-Dickstein, Noah Fiedel, Aaron T Parisi, Alexander A Alemi, Azade Nova, Ben Adlam, Bernd Bohnet, Gaurav Mishra, Hanie Sedghi, Izzeddin Gur, Jaehoon Lee, John D Co-Reyes, Kathleen Kenealy, Kelvin Xu, Kevin Swersky, Igor Mordatch, Lechao Xiao, Maxwell Bileschi, Peter J Liu, Roman Novak, Sharad Vikram, Tris Warkentin, and Jeffrey Pennington. 2024. Training language models on the knowledge graph: Insights on hallucinations and their detectability. In *First Conference on Language Modeling*.

Michael Y Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. Between circuits and chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2.5-Coder technical report. *arXiv preprint arXiv:2409.12186*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 15696–15707. PMLR.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th conference on computational natural language learning*, pages 552–564.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2025. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. In *International Conference on Learning Representations*, volume 2025, pages 100232–100255.

Knowledgator. 2025. FlashDeBERTa. Accessed: 2025-08-01.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. NeMo: A toolkit for building AI applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. LLM post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)*, pages 1813–1827.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 2757–2791. Association for Computational Linguistics.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, et al. 2025a. OLMoTrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 178–188.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024a. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schütze. 2025b. Tracing multilingual factual knowledge acquisition in pretraining. *arXiv preprint arXiv:2505.14824*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2024b. LLM360: Towards fully transparent open-source LLMs. In *First Conference on Language Modeling*.

Clare Lyle, Mark Rowland, and Will Dabney. 2022. Understanding and preventing capacity loss in reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. 2025. DataDecide: How to predict best pretraining data with small experiments. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language mod-

els: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372.

Jack Merullo, Noah Smith, Sarah Wiegreffe, and Yanai Elazar. 2025. On linear representations and pretraining data frequency in language models. In *International Conference on Learning Representations*, volume 2025, pages 71963–71987.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. 2025. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pages 62061–62121.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 48–53.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do LLMs acquire new knowledge? A knowledge circuits perspective on continual pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19889–19913.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473.

Hadi Pouransari, Chun-Liang Li, Jen-Hao Rick Chang, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, and Oncel Tuzel. 2024. Dataset decomposition: Faster LLM training with variable sequence length curriculum. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 conference on empir-*

ical methods in natural language processing (EMNLP), pages 5418–5426.

Nicholas Roberts, Niladri S Chatterji, Sharan Narang, Mike Lewis, and Dieuwke Hupkes. 2025. Compute optimal scaling of skills: Knowledge vs reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13295–13316.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, NIST Special Publication, pages 109–126. National Institute of Standards and Technology (NIST).

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. 2025. Overtrained language models are harder to fine-tune. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Yiyou Sun, Yu Gai, Lijie Chen, Abhilasha Ravichander, Yejin Choi, Nouha Dziri, and Dawn Song. 2026. Why and how LLMs hallucinate: Connecting the dots with subsequence associations. *Advances in Neural Information Processing Systems*, 38:35159–35203.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876.

Nicolas Zucchet, Jorg Bornschein, Stephanie C.Y. Chan, Andrew Kyle Lampinen, Razvan Pascanu, and Soham De. 2025. How do language

models learn facts? dynamics, curricula and hallucinations. In *Second Conference on Language Modeling*.

A Appendix: Supplementary Results

A.1 Pretraining Data Statistics

Tab. 2 summarizes statistics over the LMEnt pretraining corpus. These results support the overview provided in §4.

# tokens	# chunks	# entities	# batches (steps)	# mentions total	# hyperlinks (H)	# entities by H					# entity linking	# coref. mentions	# coref. clusters
						1	(1, 10]	(10, 100]	(100, 1K]	> 1K			
3.6B	10.5M	7.3M	109K	400M	115M	993K	2.2M	967K	141K	13K	203M	151M	310M

Table 2: Statistics of the LMEnt pretraining corpus.

A.2 PopQA and PAQ Performance

In this section we present additional results for PopQA and PAQ accuracies as functions of compute budget. Fig. 12 shows the results for all questions and the subset of questions whose subject and answer entities co-occur moderately in PopQA, and Fig. 13 shows the results for all questions and the subset of questions whose subject and answer entities co-occur frequently in PAQ. These results support §5.1.

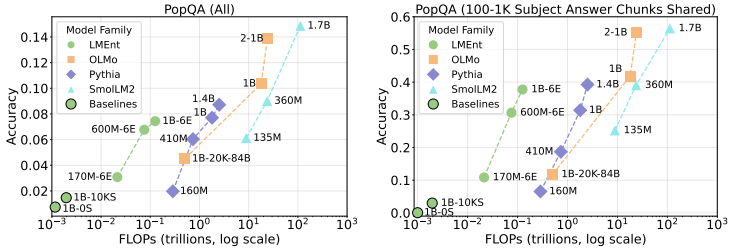

Figure 12: Accuracy on PopQA as a function of compute budget: (left) all questions and (right) questions for which the subject and answer entities appear together in 100-1K chunks (moderate).

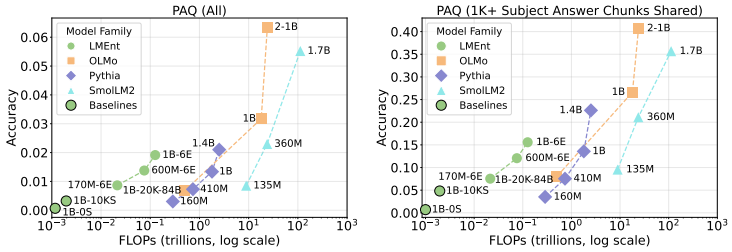

Figure 13: Accuracy on PAQ as a function of compute budget: (left) all entities and (right) questions for which the subject and answer entities appear together in 1K+ chunks (frequent).

A.3 Popularity and Fact Frequency Indicators for Model Behavior on Additional Models

In this section, we present additional results for PopQA accuracies sliced by popularity and fact frequency indicators. Fig. 14 shows the results for OLMo-2-1B (OLMo et al., 2024), Pythia-1.4B (Biderman et al., 2023), and SmolLM-2-1.7B (Allal et al., 2025). These results support §5.2.

A.4 Results on Additional Tasks

We report results on non-knowledge tasks in Tab. 3, which complement §5.1.

A.5 Empirical Experiment for Choosing LMEnt Score Thresholds

In this section, we describe the empirical experiment that determined the score thresholds for LMEnt retrieval (§5.2). To determine thresholds for the hyperlink score **H**, entity-linking score **EL**, and coreference scores **C** and **CC**, we sample 60 entities from PopQA based on hyperlinks counts. For each entity, we retrieve their scores using a fixed hyperlink threshold of $\mathbf{H} = 1$, while varying **EL**, **C**, and **CC** across the set $\{0.4, 0.5, 0.6, 0.7, 0.8\}$, and evaluate precision over chunks retrieved at different depths k. If $k > 100$, we randomly sample 100 of the k returned chunks and submit them to Gemini 2.5 Flash Preview 6-17 using the prompt shown in Fig. 20 to compute precision at k. Results are reported in Tab. 4.

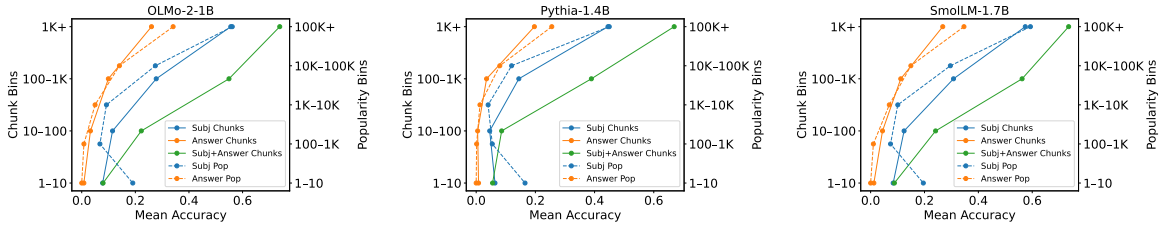

Figure 14: Accuracy of OLMo-2-1B (OLMo et al., 2024), Pythia-1.4B (Biderman et al., 2023), and SmolLM-2-1.7B (Allal et al., 2025) on PopQA, sliced by various indicators of popularity and fact frequency on PopQA.

Table 3: Performance of LMEnt models and baseline models on commonsense reasoning, multiple choice, and reading comprehension tasks.

Model	arc_challenge	arc_easy	boolq	copa	coqa	drop	gsm8k	hellaswag	jeopardy	naturalqs_open	openbookqa	piqa	sciq	squad	triviaqa	truthfulqa	winogrande
LMEnt-170M-1E	0.217	0.370	0.574	0.570	0.072	0.022	0.020	0.271	0.006	0.002	0.304	0.541	0.633	0.038	0.000	0.471	0.493
LMEnt-170M-2E	0.247	0.368	0.620	0.610	0.081	0.024	0.025	0.274	0.006	0.008	0.320	0.560	0.697	0.043	0.002	0.471	0.505
LMEnt-170M-4E	0.254	0.282	0.367	0.510	0.000	0.000	0.000	0.247	0.000	0.000	0.264	0.487	0.237	0.000	0.000	0.488	0.488
LMEnt-170M-6E	0.271	0.389	0.400	0.640	0.082	0.034	0.025	0.280	0.006	0.007	0.356	0.541	0.711	0.056	0.004	0.471	0.501
LMEnt-600M-1E	0.271	0.409	0.604	0.650	0.085	0.012	0.025	0.285	0.008	0.001	0.340	0.555	0.722	0.044	0.001	0.472	0.490
LMEnt-600M-2E	0.301	0.405	0.620	0.630	0.113	0.046	0.055	0.296	0.009	0.020	0.350	0.554	0.744	0.085	0.005	0.468	0.516
LMEnt-600M-4E	0.254	0.389	0.565	0.610	0.150	0.055	0.015	0.306	0.034	0.010	0.338	0.559	0.735	0.092	0.016	0.470	0.515
LMEnt-600M-6E	0.274	0.409	0.625	0.630	0.167	0.075	0.050	0.318	0.023	0.029	0.374	0.569	0.762	0.119	0.045	0.444	0.510
LMEnt-1B-1E	0.254	0.421	0.603	0.650	0.109	0.063	0.035	0.295	0.009	0.019	0.358	0.557	0.714	0.084	0.006	0.469	0.506
LMEnt-1B-2E	0.244	0.377	0.626	0.630	0.132	0.068	0.045	0.309	0.020	0.018	0.360	0.557	0.765	0.089	0.047	0.448	0.523
LMEnt-1B-4E	0.278	0.391	0.545	0.670	0.160	0.067	0.010	0.322	0.021	0.024	0.374	0.557	0.763	0.092	0.008	0.455	0.517
LMEnt-1B-6E	0.264	0.446	0.611	0.680	0.157	0.070	0.035	0.328	0.043	0.031	0.390	0.555	0.770	0.107	0.025	0.439	0.529
OLMo-1B	0.338	0.565	0.617	0.770	0.574	0.218	0.030	0.627	0.342	0.123	0.434	0.729	0.879	0.624	0.324	0.329	0.594
OLMo-1B-20K-84B	0.288	0.458	0.640	0.760	0.269	0.079	0.000	0.433	0.059	0.051	0.406	0.676	0.790	0.251	0.085	0.394	0.531
OLMo-2-1B	0.435	0.635	0.646	0.800	0.694	0.353	0.395	0.688	0.641	0.191	0.474	0.746	0.953	0.818	0.512	0.369	0.654
Pythia-1.4B	0.348	0.539	0.582	0.760	0.566	0.206	0.020	0.540	0.273	0.098	0.434	0.707	0.874	0.565	0.250	0.386	0.566
Pythia-1B	0.331	0.525	0.623	0.740	0.519	0.185	0.025	0.478	0.195	0.077	0.378	0.683	0.890	0.534	0.188	0.389	0.533
Pythia-410M	0.301	0.467	0.588	0.700	0.428	0.166	0.040	0.393	0.069	0.058	0.360	0.663	0.833	0.438	0.104	0.410	0.532
Pythia-160M	0.311	0.423	0.498	0.640	0.218	0.136	0.020	0.300	0.019	0.031	0.324	0.606	0.746	0.176	0.042	0.442	0.502
SmolLM2-1.7B	0.468	0.704	0.733	0.840	0.720	0.259	0.315	0.695	0.706	0.194	0.496	0.760	0.940	0.771	0.545	0.367	0.671
SmolLM2-360M	0.448	0.649	0.620	0.760	0.596	0.190	0.045	0.551	0.516	0.107	0.444	0.710	0.905	0.656	0.305	0.334	0.591
SmolLM2-135M	0.365	0.558	0.617	0.660	0.429	0.135	0.025	0.410	0.244	0.079	0.422	0.670	0.842	0.467	0.175	0.388	0.530

A.6 Win Rates

In Fig. 15, we extend the experiments in (§5.2, Fig. 5) to compare win rates between the best string based method, CS-SS Canonical, and all other string-based variants (left), as well as CI-SS Canonical versus CI-SS Expanded (right). These results complement Fig. 5.

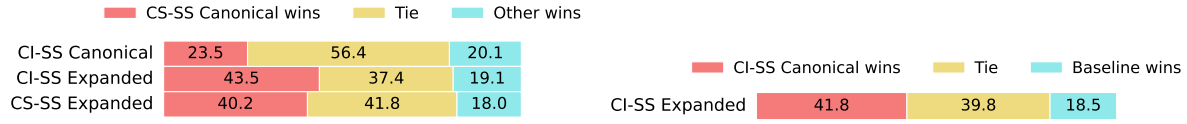

Figure 15: Pairwise win rates between CS-SS Canonical and other string-based methods (left). Pairwise win rates between CI-SS Canonical and CI-SS Expanded (right).

A.7 LMEnt provides better coverage for rare entities

Fig. 16 shows the distribution over the total number of retrieved chunks per entity. We observe that LMEnt offers greater coverage for tail and torso entities compared to Canonical string-based variants. This is especially significant since tail and torso entities together make up 99.7% of the total entities in the corpus (§A.1, Tab. 2). While it appears that Expanded string-based variants retrieve more chunks, this is likely noise due to alias ambiguity. There is also a noticeable difference in variance—specifically string-based methods show egregious retrieval failures, sometimes retrieving less than 10 documents for head entities and no documents for torso entities. These results support §5.2.

A.8 LMEnt wins across all entity frequency levels, often by large margins

Fig. 17 shows "Yes" chunk count differences between methods and the cumulative percentage of entities where one method outperforms the other, across entities of different frequencies. For tail entities (right), LMEnt outperforms by a few chunks, but this is meaningful given the scarcity of mentions. For torso

Table 4: Empirical evaluation of LMEnt retrieval precision across multiple thresholds and retrieval depths k on the development set of 60 PopQA entities.

Score Thresholds	1	5	10	100	1K	10K	100K
H=1, EL = C = CC = 0.4	0.95	0.95	0.96	0.98	0.99	0.98	0.94
H=1, EL = C = CC = 0.5	0.95	0.95	0.96	0.98	0.99	0.99	0.91
H=1, EL = C = CC = 0.6	0.95	0.96	0.96	0.98	0.99	0.98	0.94
H=1, EL = C = CC = 0.7	0.95	0.96	0.96	0.98	0.99	0.98	0.91
H=1, EL = C = CC = 0.8	0.95	0.96	0.96	0.98	0.97	0.88	0.96

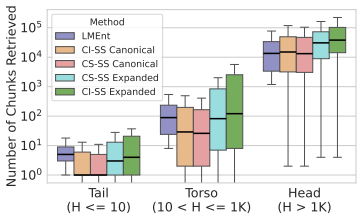

Figure 16: Distribution of the number of chunks retrieved by LMEnt, and both Canonical and Expanded variants of CI−SS and CS−SS. Retrieval was performed for 1K PopQA entities, selected via stratified sampling based on hyperlink counts. LMEnt retrieves more chunks than the Canonical variants for torso and tail entities, which together account for 99.7% of all entities in Wikipedia. The higher number of chunks returned for head entities by Expanded variants is likely bloated by noisy retrieval (Fig. 6).

entities middle), LMEnt wins by a substantial margin ($\geq$ 20 additional correct chunks) in 40% of cases. For head entities (left), the margin is typically smaller (1–25 chunks), suggesting that string-based search performs more competitively in these cases – though, LMEnt still outperforms it on $\geq$ 60% of entities.

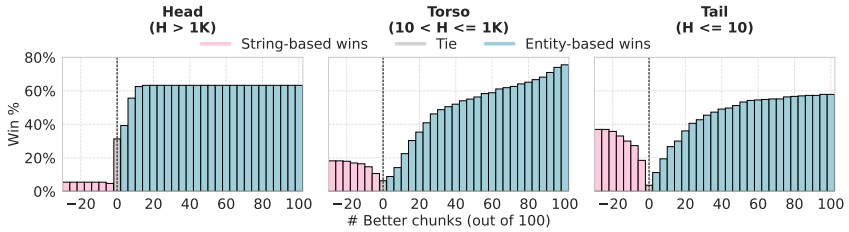

Figure 17: Comparison between LMEnt retrieval and the best performing string-based approach (CS−SS Canonical). Results are grouped by frequency of hyperlinks; tail entities (right), torso entities (center), and head entities (left).

A.9 Knowledge Acquisition during Pretraining

Fig. 18 shows the net gains in facts learned between LMEnt-1B-6E checkpoints across frequency bins of subject and answer entity co-occurence, and additional examples of three specific relations. The split of % of facts learned and forgotten is in §6, Fig. 8.

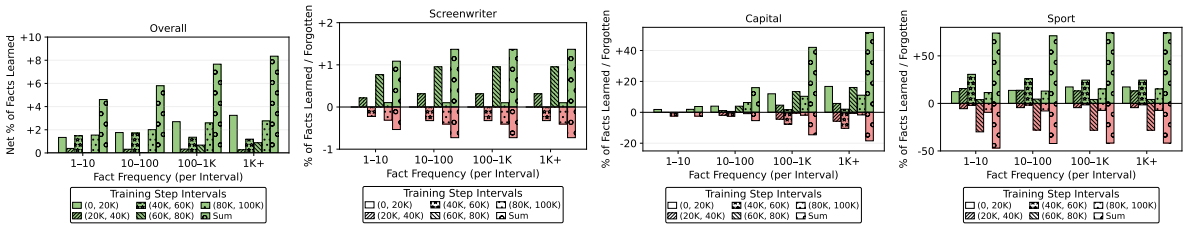

Figure 18: Percentage of net gains in facts learned between intermediate checkpoints of LMEnt-1B-6E (left). Additional examples of the relation-specific analysis in in §6.

B Appendix: Additional Implementation Details

B.1 Running Maverick At Scale

The backbone of Maverick (Martinelli et al., 2024) is DeBERTa (He et al., 2021) which has quadratic attention complexity with respect to sequence length. Therefore, running Maverick on long documents is impractical, e.g. 15K token long documents take 40-50GB VRAM. Therefore, we reduce the memory footprint, and accelerate and parallelize inference across multiple documents by replacing the DeBERTa backbone with a FlashAttention-based implementation (Knowledgator, 2025). Then, we apply Maverick to non-overlapping windows of 50K tokens, which may result in multiple distinct coreference clusters for the same entity. In these cases, we rely on the coverage of hyperlinks and the entity linker (which runs on the entire document) to map each cluster to the target entity's QID.

B.2 Coreference scoring

In this section we detail how we compute the score derived from coreference resolution, and how we resolve ambiguity when multiple entities are associated with a single coreference cluster (§2.2).

Coref (C) is designed to associate shortened references, not previously found by hyperlinks or entity linking, with their canonical names, e.g. "*the hospital*" with "*John R. Oishei Children's Hospital*". The coreference cluster of "*the hospital*" contains "*the hospital*" and "*John R. Oishei Children's Hospital in* Buffalo". There are two entities related to this cluster, with QIDs: "Q93565992" and "Q40435". We can reduce ambiguity by leveraging the shared substring of "*(H,h)ospital*" to promote "*the hospital*" being mapped to "Q93565992" more strongly. We represent this promotion by a score denoted by **C**, which we formally define below.

Let the coreference cluster associated with m be $\mathcal{C}$, and define the sets of hyperlink and entity linking mentions that overlap with $\mathcal{C}$ as $\mathcal{C}_{\text{H}}$ and $\mathcal{C}_{\text{EL}}$, respectively. To calculate $\mathbf{C}(m, e)$, for each mention $m' \in \mathcal{C}_{\text{H}} \cup \mathcal{C}_{\text{EL}}$, we compute the longest common substring (LCS) with m. Let $\text{sim}(m, m')$ be the harmonic mean of the LCS overlap ratios between mention m and m', defined as:

$$\text{sim}(m, m') = \begin{cases} 0.0 & \text{if } |m| = 0 \text{ or } |m'| = 0 \\ \frac{2 \cdot a \cdot b}{a + b} & \text{otherwise} \end{cases} \quad \text{where } a = \frac{\text{LCS}(m, m')}{|m|},\ b = \frac{\text{LCS}(m, m')}{|m'|} \tag{1}$$

We weight either the hyperlink or entity linking score of m' for e by its textual similarity to m:

$$\text{support}(m, m', e) = \text{sim}(m, m') \cdot \begin{cases} \mathbf{H}(m', e) & \text{if } m' \in \mathcal{C}_{\text{H}} \\ \mathbf{EL}(m', e) & \text{if } m' \in \mathcal{C}_{\text{EL}} \end{cases} \tag{2}$$

Finally, the **C** score for mention m and entity e is given by the maximum support score over all other mentions $m' \in \mathcal{C}$:

$$\mathbf{C}(m, e) = \max_{m' \in \mathcal{C}_{\text{H}} \cup \mathcal{C}_{\text{EL}}} \text{support}(m, m', e) \tag{3}$$

Coref-Cluster (CC) connects generic indirect mentions (that don't share textual similarity with the entity name) to their canonical entity names, e.g. "*their*" with "*Buffalo Bills*". In this case, we compute a distribution of scores over all the entities linked to some mention in the cluster. The score $\mathbf{CC}(e)$ is defined per entity per cluster, and is shared across all mentions $m \in \mathcal{C}$.

For each mention $m \in \mathcal{C}$ and entity e, we look at the existing source scores and compute a $\text{m}_{\text{support}}$ score. We don't aggregate the existing source scores (e.g. compute an average), because we found that this disproportionately rewarded single mentions identified by multiple sources and outweighed multiple mentions found by single sources, leading to incorrect cluster to entity associations. The $\mathcal{C}_{\text{support}}$ for entity e in cluster $\mathcal{C}$ is simply the sum over all $\text{m}_{\text{support}}$ scores.

$$\mathrm{m}_{\text{support}}(m,e) = \begin{cases} \mathbf{H}(m,e) & \text{if } m \in \mathcal{C}_{\text{H}} \\ \mathbf{EL}(m,e) & \text{if } m \in \mathcal{C}_{\text{EL}} \\ \mathbf{C}(m,e) & \text{otherwise} \end{cases} \quad \mathcal{C}_{\text{support}}(e) = \sum_{m \in \mathcal{C}} \mathrm{m}_{\text{support}}(m,e) \tag{4}$$

To compute a distribution of scores over all entities mentioned in the cluster, we compute a softmax over $\mathbf{CC}(\mathcal{C}, e')$ scores for the set of entities e' supported by at least one mention in $\mathcal{C}$, denoted by $\mathcal{E}_{\mathcal{C}}$.

$$\mathbf{CC}(\mathcal{C}, e) = \frac{\exp\left(\mathcal{C}_{\text{support}}(e)\right)}{\sum_{e' \in \mathcal{E}_{\mathcal{C}}} \exp\left(\mathcal{C}_{\text{support}}(e')\right)} \tag{5}$$

B.3 Error Analysis of `LMEnt` Annotations

This section describes a qualitative error analysis in which we sampled 112 mentions retrieved using `LMEnt`, and manually analyzed the annotation errors observed and their frequency. We saw three errors total (2.7%) that are displayed in Fig. 19. The errors were results of entity linking failures and one coreference resolution failure.

entity	entity_description	qid	mention	title	highlighted_text	scores_by_source
Face to Face (1976 film)	Face to Face (1976 film) - 1976 film by Ingmar Bergman	Q388846	Ansikte mot ansikte	Massimo Di Gesu	Di Gesu's works such as "<START>Ansikte mot ansikte<END>" (for Serate Musicali - Milano, and Ente Concerti Pesaro) and "WOLFiliGrANG" (at the Rovereto Mozar	{'coref': 0.0, 'coref_cluster': 0.0, 'entity_linking': 0.96, 'hyperlinks': 0.0}
The Message (short story)	The Message (short story) - science fiction short story	Q7751133	George Kilroy	Jack Donahue	ty lashes as punishment. Donahue escaped to the bush from the Quakers Hill farm with two men named <START>George Kilroy<END> and William Smith. They formed an outlaw gang known as "The Strippers," since they stripped wealthy	{'coref': 0.91, 'coref_cluster': 0.6, 'entity_linking': 0.91, 'hyperlinks': 0.0}
Happy End (musical)	Happy End (musical) - 1929 three-act musical comedy by Kurt Weill, Elisabeth Hauptmann, and Bertolt Brecht	Q1584256	March Ahead	Nagorik Shakti	rty, and featured addresses to women, the young, and the expatriates. "Bangladesh egiye cholo" or "<START>March Ahead<END>, Bangladesh" was to be the theme title of the movement, as declared in the second letter to the peo	{'coref': 0.0, 'coref_cluster': 0.0, 'entity_linking': 0.93, 'hyperlinks': 0.0}

Figure 19: Error analysis of `LMEnt` entity mentions.

B.4 Pretrained Models

Details of the `LMEnt` models are described here and support the overview in §4. The 170M, 600M, and 1B models use $(\text{layers}, \text{hidden dimension})$ configurations of $(10, 768)$, $(16, 1344)$, and $(16, 2048)$, respectively. A hyperparameter search was conducted over the following ranges: global batch size $\{16{,}384,\ 32{,}768,\ 65{,}536,\ 131{,}072,\ 262{,}144\}$, peak learning rate $\{3\times10^{-4},\ 6\times10^{-4},\ 8\times10^{-4},\ 1.2\times10^{-3},\ 3\times10^{-3},\ 5\times10^{-3}\}$, and weight decay $\{0.005,\ 0.05,\ 0.1\}$. Hyperparameters were selected based on the minimal final training perplexity. All `LMEnt` models were trained using the AdamW optimizer with a global batch size of $32{,}768$, rank batch size of $8{,}192$, peak learning rate of 5×10^{-3}, weight decay of 0.05, and $1{,}000$ warmup steps.

B.5 LLM Judge for Chunk Precision

To automatically evaluate the retrieval quality of both `LMEnt` (our entity-based method) and string-based baselines (see § 5.2), we use Gemini 2.5 Flash Preview 6-17. For each of the 1K entities in the test set, we apply `LMEnt`, its ablations, and string-based methods `CS-SS-(Canonical, Expanded)` and `CI-SS-(Canonical, Expanded)`, to retrieve sets of relevant chunks. For a retrieved set containing more than 100 chunks, we randomly sample 100 chunks from it for evaluation.

To determine whether a chunk mentions the entity, we prompt Gemini using the instruction shown in Fig. 20. To provide context, we include the entity's description from ReFinED (Ayoola et al., 2022), which consists of the Wikipedia article title and the first sentence in the article. For each chunk, we identify matching mentions of the entity to create a short context window for Gemini. For `LMEnt`, mentions are selected based on QID and score thresholds. For string-based methods, we use the `highlight` block in the Elasticsearch (Banon, 2010) query to extract character spans that match the entity name. Around each mention, we select a 130-character window to give Gemini some context of the mention in the chunk text. We evaluate up to three mentions per chunk. If Gemini returns "Yes" for any, the chunk is marked as mentioning the entity. If all are judged "No", the chunk is considered not to mention it. For `LMEnt`, mentions are prioritized by a weighted sum of their scores, and for string-based methods, mentions are evaluated in the order they appear in the document.

Justifying use of LLM-as-a-Judge To support our use of an LLM-as-a-Judge for evaluating model-generated responses, we use the alternative annotator test introduced by Calderon et al. (2025). This test determines whether the LLM's performance is comparable or better than that of a randomly chosen human annotator. In line with their methodology, we employed three human annotators (graduate students) and used a dataset of 100 entity mentions randomly sampled from 'Yes" and "No" judged chunks. For each mention, the annotators received the same inputs as the LLM judge: the entity name, entity description, and 130-character context window around the mention, and were asked to evaluate whether the chunk named the entity. Instructions are found in Fig. 21. Following Calderon et al. (2025), we set $\epsilon = 0.1$, which yielded a winning rate of $\omega = 1.00$ with a p-values of $0.001, 0.001$, and $8.28e^{-5}$ for all three annotators, indicating that the LLM can be relied on.

Goal: Determine if the provided Text directly mentions or discusses the specific Entity.

Process:

First, carefully read the Text and identify any specific names, terms, or phrases that appear and might relate to the Entity.
Second, compare these identified text mentions (if any) to the provided Entity.

Rules for "Directly Mentions or Discusses":
1. The Text must contain the exact name of the Entity or a clearly identifiable and standard part of the Entity's name (e.g., "Einstein" for "Albert Einstein") that contextually refers directly to the Entity.

2. The mention must refer to the Entity itself. Merely using the Entity's name as a descriptive modifier for something else associated with the Entity (e.g., "the Wahhabi school of Saudi Arabia" referring to the school, not the country) does not count as directly mentioning or discussing the Entity, unless the descriptive phrase is the Entity (e.g., "University of London" for the entity "University of London").

3. Exclusions (Do NOT count as direct mention/discussion):
Mentions of individuals or things related to the Entity but not the Entity itself (e.g., family members, associates, parts or products of the entity unless the mention clearly refers back to the entity).

4. Inferences based on context, related events, locations, or concepts if the Entity's name is not used to refer to the Entity itself according to Rule 3.

5. Distinguishing Similar Names: If the Text contains a name that is similar to the Entity's name but refers to a different person, place, or thing (e.g., a different historical figure, a different organization with a similar name, or differing spellings like "Elisabeth" vs. "Elizabeth" referring to different individuals), it does not count as a direct mention of the target Entity. The context must clearly indicate the reference is to the specific Entity provided.

6. Possessives, actions, or attributes must be logically consistent with the entity's type. Implausible associations, like assigning a teacher to Ennis, a town in County Clare, Ireland (e.g., "Ennis's teacher"), are invalid.

Consider each rule individually and assign an intermediate judgment of Pass or Fail. If any rule receives a Fail, the final decision must be No.

Output: Provide only the answer 'Yes' or 'No' based on the above rules. Do not include any explanation or other text.

Entity: **{entity_description}**
Text: **{text}**
Answer:

Figure 20: Prompt given to Gemini to automatically judge whether a chunk mentions an entity directly.

B.6 Converting Queries to Cloze-Style Prompts

Since `LMEnt` models are not instruction-tuned, we evaluate them on PAQ by converting each question into a cloze-style prompt, where the expected answer is the next predicted phrase. To perform this transformation, we use Gemini 2.5 Flash Preview 6-17 with the prompt shown in Fig. 22.

Instructions

In this task, you will be given an Entity, a short Description of the Entity, and a short piece of Text that may or may not mention the Entity.

Your goal is to determine if the provided Text directly mentions or discusses the specific Entity.

To complete the task, perform the following steps:

(1) Carefully read the Text and identify any specific names, terms, or phrases that appear and might relate to the Entity.
(2) Compare these identified text mentions (if any) to the provided Entity according to the rules below. Consider each rule individually and assign an intermediate judgment of Pass or Fail per rule.
(3) Indicate your final decision as either 'Yes' or 'No':
    - If any rule receives a Fail, the final decision must be No.
    - Otherwise (if all rules received a Pass), the final decision should be Yes.

Rules for "Directly Mentions or Discusses":

(1) The text must contain the exact name of the Entity or a clearly identifiable and standard part of the Entity's name (e.g., "Einstein" for "Albert Einstein") that contextually refers directly to the Entity.

(2) The mention must refer to the Entity itself. Merely using the Entity's name as a descriptive modifier for something else associated with the Entity (e.g., "the Wahhabi school of Saudi Arabia" referring to the school, not the country) does not count as directly mentioning or discussing the Entity Saudi Arabia.

(3) Make inferences based on context, related events, locations, or concepts if the Entity's name is not used to refer to the Entity itself according to Rules 1, 2.

Exclusions (Do NOT count as direct mention/discussion):

(1) Mentions of individuals or things related to the Entity but not the Entity itself (e.g., family members, associates, parts or products of the entity unless the mention clearly refers back to the entity).

(2) Distinguishing Similar Names: If the Text contains a name that is similar to the Entity's name but refers to a different person, place, or thing (e.g., a different historical figure, a different organization with a similar name, or differing spellings like "Elisabeth" vs. "Elizabeth" referring to different individuals), it does not count as a direct mention of the target Entity.

(3) Possessives, actions, or attributes must be logically consistent with the entity's type. Implausible associations, like assigning a teacher to *Ennis, a town in County Clare, Ireland* (e.g., "Ennis's teacher"), are invalid.

Figure 21: Instructions given to annotators for evaluating LLM-as-a-judge.

Your goal is to rephrase the input Question as a Cloze Statement so that the Answer is the next word that would logically complete the sentence.

Instructions:
1. Do not lose any context from the question. The cloze statement must preserve the full meaning and specificity of the original question.
For example,
Question: What TV network does Funny or Die Presents air on?
Answer: HBO
Cloze Statement: Funny or Die Presents airs on the TV network

2. Rephrase the question into a natural, grammatically correct sentence. If the question is incomplete or unclear, you may add minimal language to clarify the intended answer.
For example,
Question: stevie cameron worked as a confidential informant for the rcmp during which controversial
Cloze Statement: Stevie Cameron worked as a confidential informant for the rcmp during the controversial

3. End the statement right before the answer.

4. Use any specific labels or terms from the question. For example, if the question specifies "capital city" or "TV network," include that exact phrase in the cloze.

Output ONLY the Cloze Statement at the end of thinking.

Question: **{question}**
Answer: **{answer}**
Cloze Statement:

Figure 22: Prompt given to Gemini 2.5 Flash Preview 6-17 to convert PAQ questions to cloze-style prompts.