

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 CITE PRETRAIN: RETRIEVAL-FREE KNOWLEDGE ATTRIBUTION FOR LARGE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Trustworthy language models should provide both correct and verifiable answers. However, citations generated directly by standalone LLMs are often unreliable due to hallucinations. As a result, current systems insert citations by querying an external retriever at inference time, introducing latency, infrastructure dependence, and vulnerability to retrieval noise. We explore whether LLMs can be made to reliably attribute to the documents seen during (continual) pretraining, without test-time retrieval, by revising the training process. To study this, we construct **CitePretrain-Bench**, a benchmark that mixes real-world corpora (Wikipedia, Common Crawl, arXiv) with novel, unseen documents and probes both short-form (single fact) and long-form (multi-fact) citation tasks. Our approach follows a two-stage process: (1) Continual-pretraining to index factual knowledge by binding it to persistent document identifiers; (2) Instruction tuning to elicit citation behavior. We introduce **Active Indexing** for the first stage, which creates generalizable, source-anchored bindings by augmenting training with synthetic data that (i) restate each fact in diverse, compositional forms and (ii) enforce bidirectional training (source→fact and fact→source). This equips the model to both generate content from a cited source and attribute its own answers, improving robustness to paraphrase and composition. Experiments with Qwen-2.5-7B and 3B show that Active Indexing consistently outperforms a Passive Indexing baseline, which simply appends an identifier to each document, achieving citation precision gains of up to 30.2% across all tasks and models. Our ablation studies reveal that performance continues to improve as we scale the amount of augmented data, showing a clear upward trend even at 16× the original token count. Finally, we show that internal citations complement external ones by making the model more robust to retrieval noise.

1 INTRODUCTION

Large Language Models (LLMs) can improve the trustworthiness of their outputs by providing citations—references that justify their answers (Rashkin et al., 2023; Wang et al., 2023; Huang et al., 2024b). However, references directly generated by standalone LLMs (i.e, **internal citations**) are unreliable (Agrawal et al., 2024), with hallucination rates of 86% Zucccon et al. 2023 and up to 91.4% (Chelli et al., 2024), and misattribution rates of 24–46% even among the few authentic ones (Walters & Wilder, 2023; Bhattacharyya et al., 2023). To address this, most existing systems apply **external citations** by querying an external retriever: they either condition on the retrieved documents during generation(Nakano et al., 2021; Menick et al., 2022b; Gao et al., 2023c), or align answers with documents afterward (He et al., 2023; Gao et al., 2023b).

While effective, this approach carries both practical and explainability limitations. On the practical side, it adds inference overhead from long contexts (Liu et al., 2024) and extra query optimization (Song & Zheng, 2024), depends on external infrastructure (e.g., web search) whose results can be volatile (Fang et al., 2025), and can degrade reasoning fidelity when retrieved context misses or conflicts with parametric knowledge (Xie et al., 2024; Huang et al., 2025). Moreover, many questions are answerable directly from parametric memory (Mallen et al., 2023), making these costs unnecessary. On the explainability side, external retrieval offers limited insight into what the model internally knows or recalls. Internal citation offers a potential pathway to trace model outputs to

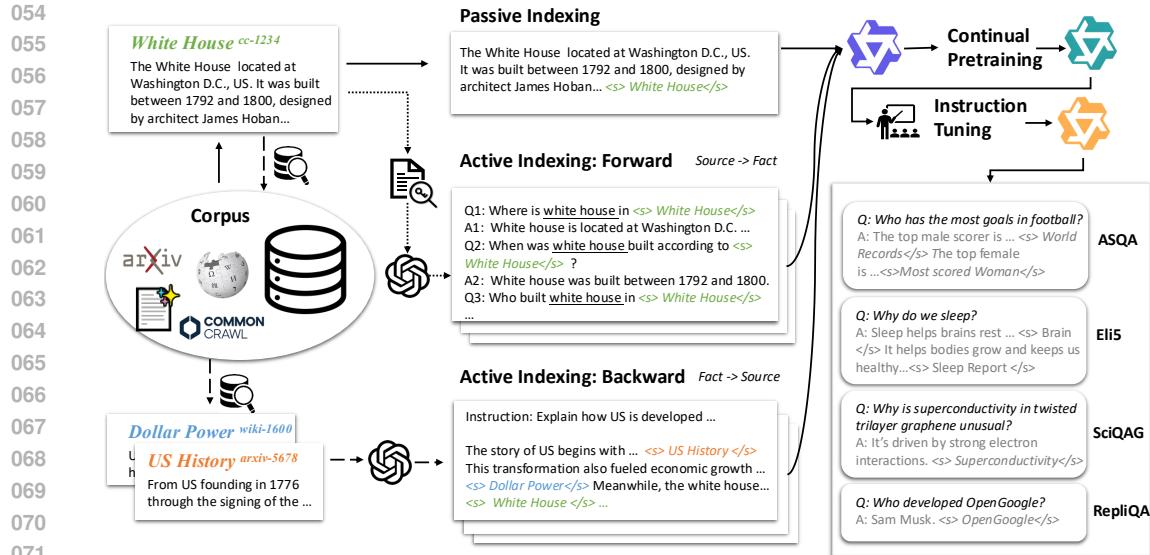


Figure 1: CitePretrain Framework. We construct a diverse corpus (Wikipedia, ArXiv, Common Crawl, and novel documents) for LLMs to index. Each document is indexed via passive indexing (appending a document ID) and active indexing, which includes: (1) Forward augmentation: generating entity-based QA pairs to map IDs to facts; and (2) Backward augmentation: retrieving related documents to synthesize multi-source QA pairs with citations, mapping facts to IDs. The model is continually pre-trained and instruction-tuned, then evaluated on long- and short-form citation QA tasks.

their training data, aligning with growing regulatory focus on transparency.¹ Notably, internal and external citations are complementary: internal serves as a fallback when retrieval fails or is disabled, while external helps when internal knowledge is incomplete or missing (e.g., new documents).

In this work, we ask whether *LLMs can be trained to perform reliable internal citations without retrieval*. Our motivation includes: 1) When retrieval is unavailable (e.g., due to latency or infrastructure limits, or in “no web search” modes like ChatGPT’s closed-book setting), the model should still be capable of producing citations on its own. 2) When retrieval is available, internal citations can act as a safeguard and complement, helping to offset retrieval noise or missing evidence. 3) In both cases, internal citations provide an added layer of explainability by linking outputs back to training data, thereby enhancing transparency. 4) More broadly, internal citation aligns with how deep learning has historically advanced: by turning complex, hand-engineered pipelines (such as multi-component RAG stacks) into unified, end-to-end models. Prior work (Khalifa et al., 2024) showed early promise that during continual pretraining, models can associate facts with document identifiers. However, their study is limited to a synthetic biography dataset with uniform, single-fact citations, leaving it unclear whether the approach generalizes to the complexity of real-world documents—longer, diverse, interdependent, and variably expressed. To address this, we introduce **CitePretrainBench**, a benchmark emphasizing document complexity and long-form citation. Models must index both (i) real-world corpora (*Wikipedia*, *Common Crawl*, *arXiv*) and (ii) novel, unseen documents, testing their ability to relearn known knowledge and acquire new knowledge with attributions. In the QA phase, we include short-form tasks requiring single citations and long-form tasks requiring synthesis across multiple documents with coherent, well-cited answers.

We train our model with a two-stage framework: *continual pretraining* to index knowledge, and *instruction tuning* to elicit citation behavior. At inference, citation decoding is constrained to corpus titles for verifiability. We begin with a passive indexing baseline, where document identifiers are appended to each document (Khalifa et al., 2024). This approach allows the model to copy identifiers

¹EU AI Act (final text, 2024/25): High-risk systems must have technical capabilities to “provide information relevant to explain [their] output” and enable deployers to interpret the output, plus mechanisms to collect, store and interpret logs (traceability). GPAI providers must also publish a “sufficiently detailed summary of the content used for training.” These provisions directly motivate attribution to training data.

108 for memorized quotes, but our benchmark reveals key limitations that do not appear in prior synthetic
 109 set-ups (Khalifa et al., 2024): 1. *Complex facts \neq quotes*. Many real evaluation questions require
 110 the synthesis or paraphrasing of information distributed across a document. The model rarely learns
 111 to associate such non-verbatim facts with the correct document identifier of the original text. 2.
 112 *Granularity alone isn't enough*. Inserting the identifiers closer to each fact (Khalifa et al., 2024) (e.g.,
 113 per sentence or paragraph) only improves slightly and the model still fails to ground non-verbatim
 114 content. **To address this, we propose Active Indexing**—a method that doesn't just memorize
 115 verbatim text-to-ID links, but teaches the model to recognize and cite the right document even when
 116 the underlying fact is variously expressed. During continual pretraining, we generate synthetic data
 117 that (1) restate each fact using varied linguistic forms (e.g., definitions, comparisons, summaries),
 118 and (2) train the model to either recall knowledge given a document ID or attribute the correct ID
 119 based on the fact. This yields two complementary training objectives: 1. *Source \rightarrow Fact (forward)*:
 120 Answer questions conditioned on the given document identifier, promoting internal memory retrieval
 121 and reasoning. 2. *Fact \rightarrow Source (backward)*: Predict the document identifier for a generated answer,
 122 reinforcing attribution and source grounding.

123 On Qwen-2.5-7B/3B, Active Indexing improves citation by up to 30.2% over passive indexing. We
 124 find (1) combining forward and backward objectives is most effective, (2) citation precision benefits
 125 more from model scale than answer correctness, and (3) proprietary models like GPT-4.1, while
 126 substantially outperforming Qwen2.5 models in answer correctness, still fall short in generating
 127 reliable citations compared to models trained with Active Indexing. In our ablation study, we identify
 128 two key reasons why Active Indexing is effective: (1) it presents facts in greater quantity and more
 129 diverse formats, and (2) it explicitly trains the model to utilize document identifiers, **making it more**
 130 **token-efficient than rephrasing-only methods**. Moreover, Active Indexing continues to benefit from
 131 scaling, showing no signs of saturation even when the amount of augmented data reaches 16 \times the
 132 original corpus. **This stems from its mechanism for synthesizing across related documents, generating**
 133 **combinatorially diverse, high-value fresh tokens..** It improves both the model's ability to memorize
 134 document identifiers and to generalize their usage to downstream tasks. Finally, we show that internal
 135 and external citations are complementary: internal excels under poor retrieval, external under strong
 136 retrieval. Our hybrid approach combines both sources to achieve the best overall performance across
 137 conditions, offering robust citations under common retrieval imperfections (Wang et al., 2025a).

2 RELATED WORK

138 **Attribution via Retrieval-Augmented Generation (RAG).** A common approach to knowledge
 139 attribution in LLMs is to generate citations from evidence retrieved at inference time (external
 140 citations) (Rashkin et al., 2023; Yue et al., 2023; Liu et al., 2023), either before generation (Nakano
 141 et al., 2021; Menick et al., 2022b; Ye et al., 2024; Kamalloo et al., 2023) or afterward (He et al., 2023;
 142 Gao et al., 2023b). While effective, the retrieval pipeline of external citations adds computational
 143 overhead and can miss or conflict with the model's parametric knowledge, leading to inconsistencies
 144 with what the model actually knows (Petroni et al., 2020; Xie et al., 2024; Huang et al., 2024a;
 145 Chuang et al., 2025; Huang et al., 2025). In contrast, our approach enables direct attribution to internal
 146 sources. This reduces overhead, improves explainability, and avoids dependence on noisy retrieval.

147 **Generative Retrieval (GR)** is another type of retriever used in RAG, replacing embedding-
 148 based retrieval with a generative model that maps queries directly to document IDs (Tay et al., 2022;
 149 Li et al., 2024b;a; Askari et al., 2024; Li et al., 2023). Then a separate QA model answers using
 150 the retrieved documents. Despite ID generation, retrieval and answering remain separate steps: the
 151 retriever can't answer questions, and the QA model operates in an **open-book** setting, relying on
 152 external documents. As the knowledge is not internalized to QA model, this still constitutes external
 153 citation. In contrast, internal citation unifies retrieval and answering in a single **closed-book** model.
 154 The LLM internalizes both document IDs and their associated knowledge, enabling end-to-end
 155 answer generation with citations without external context. This is strictly harder: beyond just learning
 156 a query \rightarrow docID mapping as in GR (Wang et al., 2022; Ren et al., 2023), the model must acquire
 157 knowledge and learn to use and cite it appropriately when generating answers.

158 **Internal Knowledge and Memorization.** Several studies have examined LLMs' ability to
 159 memorize and recall training data for citation (Agrawal et al., 2024; Zuccon et al., 2023; Carlini
 160 et al., 2021), using prompting (Sun et al., 2023; Weller et al., 2024), constrained decoding (Wang

162 et al., 2024), or fine-tuning (Zhang et al., 2025b) to improve attribution post-training. However,
 163 these approaches lack structured attribution during pretraining. Gao et al. (2025) show that
 164 prepending lightweight source hints during pretraining can steer model behavior, though it does
 165 not induce citation abilities. Source-aware training enable citations by attaching document IDs
 166 to continual-pretraining data (Khalifa et al., 2024), though has been limited to synthetic corpora,
 167 restricting methods and findings generalizability. More related work is in Appendix C.

168 3 SET-UP

171 3.1 PROBLEM FORMULATION

174 We study *internal citations*: a closed-book LLM must answer a question and simultaneously produce
 175 verifiable source identifiers from its training corpus in a single pass without consulting an external
 176 retriever. Let the training corpus be $\mathcal{D} = \{(c_i, t_i)\}_{i=1}^N$, where $c_i \in \Sigma^*$ is the full text of document i
 177 and $t_i \in \mathcal{T}$ is its unique human-readable title. Given a question q , the model outputs $\mathcal{R} = f\theta(q) =$
 178 $\langle (s_1, C_1), \dots, (s_m, C_m) \rangle$, where each s_k is a factual statement answering part of q and each $C_k \subseteq \mathcal{T}$
 179 is a set of titles whose documents entail s_k . To guarantee validity, citation decoding space is restricted
 180 to the known title set \mathcal{T} . We achieve internal citations through two-stage training:

181 **Stage 1: Continual pretraining & index learning** During continual pretraining the model should
 182 (1) absorb the factual knowledge in \mathcal{D} , and (2) learn an internal index that maps any factual span
 183 $s \subset c_i$ to its title t_i . To achieve this, the baseline Passive Indexing structures pretraining examples as
 184 (c_i, t_i) pairs, where the model predicts t_i given c_i .

185 **Stage 2: Citation instruction tuning.** We then apply instruction tuning so the model learns to output
 186 (s_k, C_k) pairs when answering questions, producing factual content and citations jointly.

187 Evaluation focuses on two aspects: (i) factual correctness, assessing the accuracy and relevance of
 188 the generated statements s_k ; and (ii) citation quality, measured by the precision and recall of the
 189 predicted title sets C_k .

191 3.2 DATASETS

192 To study how language models index and cite documents during continual pretraining, we propose
 193 **CitePretrainBench**, a benchmark designed around a continual pretraining corpus with document
 194 identifiers and downstream QA tasks that require citation from this corpus. The corpus includes
 195 documents from Wikipedia, Common Crawl, and scientific papers from arXiv, reflecting common
 196 pretraining sources. We also introduce entirely novel documents (unseen during pretraining) to test
 197 the model’s ability to learn and cite new knowledge. We use textual titles as document IDs for two key
 198 reasons. (1) Text Titles are inherently scalable: the textual space is vast, flexible, and allows renaming
 199 when collisions occur. (2) Our preliminary experiments show that titles yield better memorization
 200 than numerical IDs or other structured alternatives (Appendix F.1). For noisy sources like Common
 201 Crawl with missing or low-quality titles, we use an LLM to generate consistent names and run an
 202 LLM-based deduplication step to merge near-duplicates, ensuring each document has a stable, unique
 203 identifier (Appendix A). In downstream tasks, models must cite using this identifier space.

204 We evaluate citation performance using both long-form and short-form QA tasks, each grounded in
 205 a distinct part of the corpus: **ASQA**: Long-form factoid QA requiring multi-document reasoning.
 206 Sources come from the 2019/08/01 Wikipedia snapshot via the KILT knowledge base (Stelmakh
 207 et al., 2022; Petroni et al., 2021). **ELI5**: Open-ended long-form QA from Reddit’s “Explain Like
 208 I’m Five” forum (Fan et al., 2019), with retrieved documents from the August 2019 Common Crawl,
 209 preprocessed via CCNet (Wenzek et al., 2020). **SciQAG**: Short-form QA grounded in scientific
 210 papers, with document titles retrieved from arXiv (Wan et al., 2024). **RepliQA**: Short-form QA
 211 over fictional, synthetic documents created post-training cutoff (Monteiro et al., 2024), including
 212 gold answers and document-linked questions. Because this dataset is intended to evaluate a model’s
 213 abilities after its pre-training cutoff, we restate the creators’ explicit instruction that LLM developers
 214 must not include it in pre-training data. For full details on datasets and corpus, see Appendix A.

215 CitePretrainBench is designed to diagnose a model’s internal citation ability. It unifies diverse
 216 document types into a single evaluation space—short-form vs. long-form, new vs. old knowledge,

216 and high-quality (e.g., Wikipedia) vs. noisy (e.g., Common Crawl) sources. This provides a controlled
 217 testbed for analyzing internal citation behavior and directly comparing it to external citation.
 218

219 **3.3 METRICS: CORRECTNESS**
 220

221 We first evaluate the generation’s informativeness and utility—that is, its correctness with respect to
 222 the question. For **ASQA**, We compute *Exact Match Recall* (Stelmakh et al., 2022), which measures
 223 the recall of correct short answers by checking whether each reference answer appears as an exact
 224 substring in the model’s output. For **ELI5**, we use the *Claim Recall* to assess the correctness of
 225 generated answers—that is, how many gold claims are supported by the answer (Fan et al., 2019).
 226 Specifically, we compute entailment scores over three sub-claims extracted from each gold answer,
 227 providing a more accurate measure of correctness. For **SciQAG**, we follow prior work (Wan et al.,
 228 2024) that uses LLMs (GPT-4.1 in our case) to rate answers on a 1–5 scale across multiple dimensions,
 229 with scores normalized. We adopt the *Accuracy* dimension, which measures how well the answer
 230 aligns with facts from the source paper, ensuring that all claims are supported by evidence. For
 231 **RepliQA**, we find the recall metric from (Monteiro et al., 2024) insufficiently informative and instead
 232 adopt the relaxed variant of *FreshEval* (Vu et al., 2024)—a lightweight auto-rater that uses few-shot
 233 prompting with an LLM (GPT-4.1 in our case) to evaluate answer correctness.
 234

235 **3.4 METRICS: CITATION QUALITY**

236 For **long-form** QA tasks (e.g., ASQA and ELI5), where answers contain multiple facts and lack a
 237 single gold reference, we follow Gao et al. (2023c) and use an NLI model (TrueTeacher; Gekhman
 238 et al., 2023) to check if cited documents entail the generated claims. Citation precision is the
 239 proportion of citations that support their claims; recall is the proportion of claims that are fully
 240 supported. For **short-form** QA tasks (e.g., SciQAG and RepliQA), each answer corresponds to a
 241 single fact and a unique gold document. We compare model citations to the gold reference, defining
 242 precision as the fraction of citations that match, and recall as 1 if the gold citation appears, 0 otherwise.
 243 This may underestimate true accuracy, as some citations may entail the answer without matching the
 244 gold. See Appendix E for details.
 245

246 **4 METHODOLOGY**
 247

248 This section details the methodology for enabling LLMs to cite sources from their continual pre-
 249 training corpus $\mathcal{D} = \{d_i = (c_i, t_i)\}_{i=1}^N$, where d_i is a document, $c_i \in \Sigma^*$ is its text content, and
 250 $t_i \in \mathcal{T}$ is its unique title (used as the document identifier). We propose a dual approach: *Passive*
 251 *Indexing* and *Active Indexing*. Passive Indexing exposes the model to documents annotated with
 252 identifiers in a way that minimally disrupts language modeling. Active Indexing uses targeted data
 253 augmentation to strengthen the model’s ability to associate facts with document identifiers, enhancing
 254 citation accuracy in downstream tasks. Active Indexing comprises *Forward Augmentation*, which
 255 enhances identifier-to-fact recall within individual documents and *Backward Augmentation*, which
 256 fosters fact-to-identifier associations by integrating information across multiple documents.
 257

258 **4.1 PASSIVE INDEXING**

259 Passive Indexing integrates document identifiers into the pretraining corpus while preserving the
 260 LLM’s language modeling capabilities. The goal is to learn the index $f(c_i) = t_i$ during continual
 261 pretraining. Key considerations are the format and placement of identifiers.
 262

263 **Identifier Format:** We use the natural document text title $t_i \in \mathcal{T}$ as the identifier, as titles encapsulate
 264 salient content and align with the model’s text-based learning paradigm. Our preliminary experiments
 265 show that using titles as document identifiers leads to better memorization compared to numerical
 266 IDs and other semantically structured alternatives (Appendix F.1). Additionally, titles are scalable as
 267 the textual space is vast and allows renaming to avoid duplication when necessary.

268 **Identifier Placement:** Following prior work Khalifa et al. (2024), we append t_i at the end of c_i during
 269 pretraining, forming inputs of the form $c_i \rightarrow t_i$. This mirrors downstream tasks where citations
 follow generated text, facilitating natural learning of citation patterns. As a baseline, we also tested
 inserting t_i after each sentence within c_i (Khalifa et al., 2024), but this reduced fluency.

270
271

4.2 ACTIVE INDEXING: FORWARD AUGMENTATION

272
273
274
275
276
277
278
279
280
281

Forward Augmentation trains the model to map from a document identifier to its associated facts, focusing on enhancing knowledge recall within a single document d_i . We denote by $S_i = \{s_{i1}, \dots, s_{in_i}\}$ the set of factual statements entailed by document content c_i . The goal is to strengthen the model’s ability to retrieve S_i when conditioned on t_i —i.e., an identifier-to-fact mapping. This setting targets scenarios where precise attribution to a single source is essential, requiring the model to internally extract and ground detailed information from a specific document. We implement this through auto-constructed question-answer pairs derived from individual documents.

Entity Extraction: For each document d_i , we extract a set of N salient entities $E_i = \{e_{i1}, \dots, e_{iN}\}$ using an auxiliary LLM, where N controls the augmentation scale. Each e_{ij} is a key concept or entity in c_i , serving as an anchor for question generation.

282

Question-Answer Pair Generation: For each entity-document pair (e_{ij}, d_i) , an LLM generates a set of question-answer pairs $\{(q_{ijk}, a_{ijk})\}_{k=1}^{K_{ij}}$, where $K_{ij} \geq 1$ is the number of pairs per entity-document pair. Each question $q_{ijk} \in \Sigma^*$ references t_i and probes information related to e_{ij} (e.g., who, what, where, why, how, if). Each answer $a_{ijk} \in \Sigma^*$ provides a detailed response based on c_i , containing facts from S_i . This creates a closed-book training signal that strengthens the mapping $t_i \rightarrow S_i$, encouraging the model to internalize and retrieve facts when prompted with t_i . We then post-process the noisy doc-IDs in the generated questions. See details in Appendix H.1.

283
284
285
286
287

4.3 ACTIVE INDEXING: BACKWARD AUGMENTATION

288
289
290

Backward Augmentation trains models to perform fact-to-source citation by mapping generated factual statements s_k to their corresponding source identifiers $C_k \subseteq \mathcal{T}$. This strategy emphasizes cross-document reasoning, where information must be integrated from a collection of documents $\{d_i\}$. By synthesizing knowledge from diverse sources, this approach mimics real-world tasks where facts must be drawn from multiple documents. We achieve this through instruction-answer pairs that span multiple documents. The detailed process is as follows:

291
292
293
294
295

Document Chunking and Indexing: Each document d_i is divided into a set of chunks $\mathcal{C}_i = \{c_{i1}, \dots, c_{im_i}\}$, where each chunk $c_{ij} \in \Sigma^*$ contains W words. The corpus-wide chunk set is $\mathcal{C} = \bigcup_{i=1}^N \mathcal{C}_i$. Chunks are indexed using retrieval methods (e.g., BM25), creating an index base $\mathcal{I} : \mathcal{C} \rightarrow \mathbb{R}^k$, where $\mathcal{I}(c_{ij})$ is the chunk’s representation.

296
297
298
299

Chunk Cluster Formation: A chunk cluster $\mathcal{C}_\ell = \{c_{\ell1}, \dots, c_{\ellM_\ell}\} \subseteq \mathcal{C}$ is a set of related chunks from distinct documents. To form \mathcal{C}_ℓ , we randomly sample N seed chunks $\{c_{i1}, \dots, c_{iN}\}$ from each document content c_i , where N controls the augmentation scale. For each seed chunk c_{ij} , we retrieve M relevant chunks $\{c_{\ell1}, \dots, c_{\ellM}\}$ from \mathcal{I} , where $M \sim \text{Uniform}(2, 4)$ and each c_{\ellm} belongs to a distinct document d_k , $k \neq i$.

300
301
302
303
304

Instruction-Answer Pair Generation: For each chunk cluster \mathcal{C}_ℓ , an LLM generates an instruction-answer pair $(q_\ell, \mathcal{R}_\ell)$, where $q_\ell \in \Sigma^*$ is an instruction requiring integration of information from \mathcal{C}_ℓ , and $\mathcal{R}_\ell = \{(s_{\ellk}, C_{\ellk})\}_{k=1}^{m_\ell}$ is the response, with s_{\ellk} a factual statement and $C_{\ellk} \subseteq \{t_i \mid c_{ij} \in \mathcal{C}_\ell\}$ the set of supporting titles. This aligns with downstream tasks where $g : q \rightarrow \{(s_k, C_k)\}$. To manage computational costs, we bootstrap a seed set of pairs using GPT-4.1-mini and fine-tune a Qwen-2.5-3B model to generate further augmentations. We post-filter the instance with invalid doc-IDs (around 5%). See details in Appendix H.2.

305
306
307
308
309
310
311

5 RESULTS

312
313

5.1 EXPERIMENTAL DETAILS

314
315

We evaluate our approach using Qwen-2.5 3B and 7B models (with additional results for [Qwen-2.5-14B, Llama-3.2-3B&3.1-8B provided in Appendix B](#)), testing different methods to assess their effectiveness in enabling LLMs to cite from pretraining data. Evaluation is conducted across four QA benchmarks—ASQA, ELI5, RepliQA, and SciQAG (see Appendix A for dataset details). Results are reported in terms of both answer correctness and citation quality, measured by citation recall and citation precision. All methods follow a two-stage process: continual pretraining followed by instruction tuning. They differ only in their approach to continual pretraining. The Instruction-only method skips continual pretraining and proceeds directly to instruction tuning. We compare:

324
 325 Table 1: Main results on four QA datasets. Acc=answer correctness, C-Pr=citation precision, C-
 326 Re=citation recall. Best results [within the same model \(except GPT-4.1\)](#) are **bolded**. We find that:
 327 (1) Active Indexing outperforms Passive Indexing; (2) Forward and Backward are complementary;
 328 (3) Larger models help, but without Active Index, even proprietary LLMs still struggle with internal
 329 citation.

330 331	7B	ASQA			Eli5			SciQAG			RepliQA		
		Acc	C-Pr	C-Re									
332	InsOnly	19.1	20.0	21.2	11.5	5.9	6.4	65.9	0.6	0.8	24.2	0.9	1.3
333	PassIdx	21.5	24.1	24.2	14.5	8.9	9.0	65.7	2.4	2.4	24.8	2.4	2.5
334	Repeat	22.5	20.5	20.7	14.5	11.2	11.4	62.4	2.5	2.6	27.1	2.5	2.6
335	Repeat+	19.8	22.0	22.3	14.3	11.2	11.3	65.8	5.2	5.2	25.8	3.6	4.0
336	ActIdx-F	25.8	26.7	27.9	14.6	18.6	18.7	65.6	23.6	23.6	30.3	12.6	13.3
337	ActIdx-B	25.4	31.4	31.9	17.1	28.0	28.3	66.5	30.8	32.0	29.1	21.6	22.7
338	ActIdx	27.6	30.9	31.1	17.6	29.3	29.5	66.6	32.6	33.6	31.9	24.4	25.7
339	GPT-4.1	52.7	23.0	24.0	29.6	0.0	0.0	93.0	0.0	0.0	-	-	-
340													
341 342	3B	ASQA			Eli5			SciQAG			RepliQA		
		Acc	C-Pr	C-Re									
343	InsOnly	15.9	3.7	4.1	9.2	0.6	0.6	61.2	0.0	0.0	15.2	0.2	0.2
344	PassIdx	16.5	17.1	17.4	11.7	7.1	7.2	67.0	1.1	1.6	22.8	1.8	2.4
345	Repeat	18.6	16.4	16.5	12.2	9.1	9.3	64.0	0.9	1.2	23.1	1.9	2.2
346	Repeat+	16.5	17.1	17.4	10.1	9.4	9.8	65.7	1.1	1.3	23.5	2.0	2.6
347	ActIdx-F	19.8	22.6	23.1	12.5	12.8	13.3	67.7	3.0	3.0	24.7	3.9	4.7
348	ActIdx-B	19.7	24.5	24.9	14.1	19.0	19.6	65.2	17.6	23.4	24.4	7.8	14.1
349	ActIdx	21.4	23.9	24.2	15.8	19.7	19.8	67.9	20.0	23.0	24.5	10.5	15.8

350
 351
 352
 353
 354 **Instruction-only:** Serves as a baseline: the model is instruction-tuned without continual
 355 pretraining, testing citation based solely on pre-trained knowledge.

356 **Passive Indexing:** During continual pretraining, document identifiers are appended to 768-token
 357 chunks, allowing the model to passively associate facts with document IDs.

358 **Repetition:** Following Khalifa et al. (2024), document IDs are appended after each fact to support
 359 fine-grained attribution, though frequent insertions may reduce fluency.

360 **Repetition+:** Continual pretraining appends document identifiers to full documents and sampled
 361 segments (e.g., one-third, paragraphs, or sentences) to balance attribution and fluency.

362 **Active Indexing (Forward):** Intra-document augmentation is used during continual pretraining to
 363 reinforce source-to-fact mappings, enhancing grounding within individual documents. This method
 364 uses 1.28B tokens (3.3 \times the original corpus).

365 **Active Indexing (Backward):** Continual pretraining uses cross-document augmentation to teach
 366 fact-to-source mappings, enabling information synthesis and adding 1.47B tokens (3.8 \times).

367 **Active Indexing:** Combines forward and backward augmentation, resulting in 2.75B tokens (**7.05 \times**
 368 **original 390M tokens**).

369 While continual pretraining add one-time overhead, our approach has no additional inference cost.
 370 Unlike the RAG method in (Gao et al., 2023c), which incurs ongoing retrieval and conditioning costs
 371 (130 \times more input tokens per query). Although our method adds training compute, modern LLMs are
 372 increasingly *data-constrained* rather than *compute-constrained*: high-quality human data is scarce,
 373 and pure compute scaling shows diminishing returns Villalobos et al. (2024); Muennighoff et al.
 374 (2023). Synthetic data is therefore a key path for unlocking new capabilities Qin et al. (2025). In
 375 this setting, spending extra compute to obtain stronger, scalable citation ability is well justified. In
 376 practice, active indexing is a one-time cost for stable corpora, while a lightweight RAG layer (§5.4)
 377 can cover the small portion of rapidly changing information. See training details in Appendix D. We
 378 also evaluate GPT-4.1 with 3-shot citation prompting to measure a proprietary LLM internal citation
 379 (note: its decoding space cannot be constrained).

378
379

5.2 MAIN RESULTS

380
381
382
383
384
385
386
387

Passive Indexing Is Insufficient for Citation. Table 1 shows that Passive Indexing—simply appending document IDs and expecting implicit learning—is insufficient. While it offers modest gains over instruction tuning alone (e.g., 2.4 vs. 0.9 on RepliQA), citation precision remains low. Attaching doc-IDs closer to facts (Repeat) also fails to improve performance on realistic tasks, unlike prior work Khalifa et al., 2024, due to their use of synthetic, rigidly structured data. In real-world settings, diverse and loosely aligned facts limit the effectiveness of sentence-level associations. Nonetheless, Repeat+ performs slightly better, likely due to its more granular ID attachment across different parts of the document, which enhances fluency and supports better generalization.

388
389
390
391
392
393
394
395

Active Indexing is Effective in Both Directions. Active Indexing consistently outperforms passive baselines across all tasks (Table 1). Backward indexing is more effective than forward, and combining both yields the largest gains (e.g., citation precision: $2.4 \rightarrow 32.6$), highlighting their complementarity. Scaling augmentation further could enhance performance (see §5.3). Notably, Active Indexing also boosts answer correctness, likely due to exposure to factual content in diverse formats (Allen-Zhu & Li, 2025; Yang et al., 2025). Citation precision matches correctness on 3 of the 4 datasets for the 7B model; SciQAG is the exception because its correctness metric awards partial credit, whereas citation precision is strictly 0/1. We discuss the complementarity of both directions more in Appendix B.2.

396
397
398
399
400
401
402
403
404

Model Size Matters. On the Qwen-2.5-3B model, we observe similar trends to its 7B counterpart, but with substantially lower absolute citation performance. Moreover, increasing model size for Qwen-2.5 from 3B to 7B to 14B and for Llama-3 from 3B to 8B leads to monotonic improvements in performance under Active Indexing (Appendix B). Notably, strong models like GPT-4.1 struggle with internal citation without targeted training: while they can occasionally predict Wikipedia titles, they often fail to predict titles for arXiv papers. This limitation is especially pronounced on tasks like ELI5, which depend on Common Crawl sources that lack document identifiers during pretraining, rendering internal citation infeasible, necessitating Active Indexing.

405
406
407
408
409
410
411
412

Pretrained Models Memorize Wikipedia Titles. On ASQA (Wikipedia-based task), models achieve competitive citation performance even without indexing (e.g., 20.0 with instruction-tuning only vs. 24.1 with passive indexing). We attribute this to two main factors: (1) large models memorize not just verbatim text but also Wikipedia titles during pretraining (Weller et al., 2024; Zhang et al., 2025b), and (2) Wikipedia titles are typically entities, making it easier for models to “shortcut” by predicting the entity as the title. Preliminary results on other Wikipedia-based tasks, such as TriviaQA and HotpotQA, show similar patterns. See Appendix F.2 for details. See Appendix G for qualitative analysis and Appendix B.4 for OOD evaluation analysis for Active Indexing.

413
414

5.3 WHY DOES ACTIVE INDEXING WORK?

415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Facts Variation and Active Supervision Are Both Crucial for Reliable Citations To identify what drives reliable internal citation, we study how fact variation (i.e., presenting the same facts in different ways) and active supervision affect model performance. Using the RepliQA dataset, we scale the amount of augmented data and evaluate how citation performance varies with the augmentation ratio (Figure 2). We compare three approaches: (1) Passive Indexing with Replay, which simply repeats the same facts and document identifiers with more epochs; (2) Passive Indexing with Synthetic Continual Pretraining (Yang et al., 2025) (PI-SCP), which paraphrases facts and introduces relational variants, but still lacks explicit QA-style training to use document IDs. and (3) Active Indexing, which not only diversifies fact formulations but also explicitly trains the model to use document identifiers in QA-style contexts. Our findings are threefold: 1. **Fact Variation Helps:** Citation performance improves as Active Indexing increases both the scale and diversity of fact presentations. In contrast, simple token replay provides no benefit—and can even degrade performance due to overfitting. 2. **Active Supervision is Crucial:** Although PI-SCP introduces diverse phrasings of the same facts, it still lags behind active indexing. This indicates that diversity alone is insuffi-

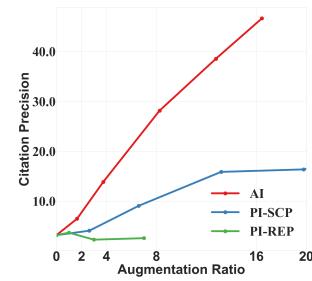


Figure 2: Scaling Comparison Between Active Indexing and Passive Indexing on RepliQA.

432 client—explicitly teaching the model to use document identifiers in context is essential for robust
 433 citation. 3. **Room to grow:** Citation performance continues to improve at the highest augmentation
 434 levels tested ($16\times$), suggesting that further gains are possible beyond our current limits.
 435
 436

437 **Active Indexing Learns Beyond Semantic Shortcuts** A key concern is that Active Indexing
 438 might rely on *semantic shortcuts* from titles (i.e., “guessing the title”) instead of learning true
 439 fact→ID associations. We test this shortcut hypothesis on the RepliQA-only setup by measuring how
 440 semantically distinctive each title is. For each RepliQA statement (query + answer), we compute
 441 its semantic similarity (sentence embeddings) to all document titles and rank the true title among
 442 them. We treat this rank as *Title Distinctiveness*: Easy = true title is highly similar to the statement
 443 (semantic shortcuts possible); Hard = many other titles are more similar (semantic cues unreliable).
 444 We bucket examples into four bins (Easy–Very Hard) and report citation precision and average rank
 445 in Table 2. Findings are: 1. **Titles are not semantically unique.** In $>90\%$ of cases, at least one other
 446 title is more similar to the statement than the true title; on average, the correct title ranks 208th out of
 447 6,822. Simple “nearest-title” guessing is rarely sufficient. 2. **Semantic cues help when available.**
 448 C-Pre is higher in the Easy bin and decreases as titles become less distinctive, showing the model
 449 sensibly uses semantic cues when they exist. 3. **Non-trivial performance without shortcuts.** Even
 450 in the Hard/Very Hard bins, where semantic similarity is a poor signal, the model still achieves $\sim40\%$
 451 C-Pre. This indicates Active Indexing learns genuine fact→ID associations, not just shortcuts.
 452

452 Table 2: **Citation precision vs. title distinctiveness on RepliQA.** Active Indexing benefits from
 453 distinctive titles but maintains substantial accuracy even when titles are not semantically similar,
 454 indicating learning beyond semantic shortcuts.

	Easy	Medium	Hard	Very Hard	Total
C-Pre	55.9	49.6	40.1	40.0	46.7
Avg. Rank	2	10	60	761	208

460 **Bridging Memorization and Generalization in Citation** Past work has shown that stronger
 461 memorization does not guarantee better generalization (Wang et al., 2025b). We observe a similar
 462 pattern in citation tasks. There is a crucial difference between how LLMs (1) **memorize** document
 463 identifiers during continual pretraining, and (2) **generalize** to use these identifiers when answering
 464 downstream questions. To probe this, we evaluate four setups that progressively shift from pure
 465 memorization to downstream usage: 1. **FullDoc:** Predicting the doc-ID given the full document.
 466 2. **PartialDoc:** Predicting the doc-ID from a partial document segment. 3. **GoldQA:** Predicting
 467 the doc-ID given the question and gold answer. 4. **ModelQA:** Predicting the doc-ID given the
 468 question and the model’s generated answer. We reuse RepliQA 7B models and report Hit@1/10
 469 accuracy. As shown in Table 3, all methods’ performance declines as tasks shift from memorization
 470 to generalization in downstream citations. Notably, more replay epochs (PassIdx-REP) improve
 471 memorization (FullDoc: $27.0 \rightarrow 74.6$) but hurt downstream generalization (ModelQA: $7.8 \rightarrow 6.0$),
 472 suggesting overfitting to shallow patterns. In contrast, Active Indexing bridges this gap by promoting
 473 both memorization and effective usage in QA, encouraging robust fact-to-source grounding.
 474

475 Table 3: Document ID Memorization and Generalization.

	FullDoc		PartialDoc		GoldQA		ModelQA	
	Acc@1	Acc@10	Acc@1	Acc@10	Acc@1	Acc@10	Acc@1	Acc@10
PassIdx	27.0	80.4	5.8	31.8	8.6	21.4	7.8	19.6
PassIdx-REP	74.6	94.4	10.6	32.8	6.6	25.6	6.0	22.4
ActIdx	95.2	100.0	72.8	97.4	66.4	94.2	54.2	88.4

483 5.4 INTERNAL VS. EXTERNAL CITATIONS

484 We compare internal and external citation methods to understand their respective limitations and the
 485 potential for synergy. External citations allow models to access information beyond their memorized

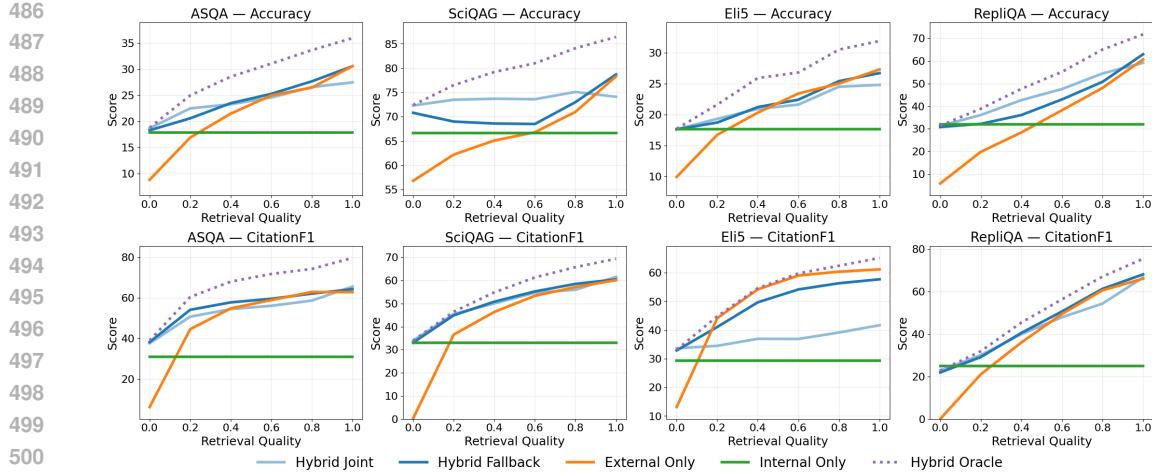


Figure 3: Performance of internal, external, and hybrid citations across retrieval quality (0=sparse retrieval, 1=dense retrieval). Internal only excels under poor retrieval, external only under strong retrieval, while hybrids generally perform best regardless of retrieval quality, with room to improve.

knowledge, while internal citations serve as a robust fallback when retrieval fails. This trade-off becomes especially important when retrieval quality varies. To explore this, we compare:

Internal Only: Generate internal citations without retrieval using Active Index.

External Only: Generate citations via RAG with 3-shot examples and top-5 retrieved documents, following the best-performing setup in Gao et al. (2023c).

Hybrid Joint: Instruction-tuning ActiveIdx to consume retrieved documents, generating both internal and external citations after an initial assessment of document sufficiency during inference.

Hybrid Fallback: A pipeline that first attempts RAG, falling back to Hybrid Joint if retrieved documents are deemed insufficient by the model.

Hybrid Oracle: A conceptual upper-bound that selects the better output from the Internal and External Only methods for each example. See more experiments details in Appendix D.2.

We test these strategies across a spectrum of retrieval quality, which we simulate by interpolating between sparse retrieval (BM25; lower-quality in our setup) and dense retrieval (Lin et al., 2021; high-quality). **Retrieval quality = 0** uses BM25’s top-5; **quality = 1** uses dense top-5; **intermediate values (e.g., 0.2)** mix the two proportionally (20% sampled from dense, 80% BM25). As shown in Figure 3, when retrieval is poor, methods incorporating internal citations (Internal Only and Hybrid variants) significantly outperform External Only, highlighting the importance of internal fallback under noisy retrieval. As retrieval improves, external outperforms internal, showing that reliable external evidence can complement internal knowledge. Hybrid Approach, taking advantages of both, perform generally best regardless of retrieval quality. Notably, a performance gap remains between our best Hybrid method and the Hybrid Oracle, pointing to headroom for more effective integration strategies that better reconcile retrieved and memorized knowledge. [See details in conflict slice in Appendix B.3.](#)

6 CONCLUSION

We show that large language models can attribute answers to their pretraining data without relying on test-time retrieval. We (1) introduce *CitePretrainBench* for internal citation across short- and long-form QA, and (2) propose *Active Indexing*, a continual pretraining strategy that teaches models to link content with document identifiers. Key findings include: 1. **Teaching beats hoping:** Active Indexing, which frames citation as real tasks, improves precision and recall by up to 32 points—outperforming passive approaches. 2. **Complementary directions:** Forward and backward augmentations are most effective when combined. 3. **Scale helps:** Performance improves consistently with more augmented data. 4. **Internal and external are complementary:** Combining both enhances robustness to retrieval quality. See Appendix I for limitations and future directions.

540 REPRODUCIBILITY
541

542 We are committed to ensuring the reproducibility of our results. We describe details of our training
543 procedures (§4.) and implementations (§5.1, Appendix D). For our evaluation, we describe details
544 of datasets (§3.2), metrics (§3.3, §3.4) and processing steps (Appendix A). Code and data will be
545 released upon publication.

546
547 REFERENCES
548

549 Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. Do language models know when
550 they’re hallucinating references? In Yvette Graham and Matthew Purver (eds.), *Findings of the*
551 *Association for Computational Linguistics: EACL 2024*, pp. 912–928, St. Julian’s, Malta, 2024.
552 Association for Computational Linguistics. URL [https://aclanthology.org/2024-
553 findings-eacl.62](https://aclanthology.org/2024-findings-eacl.62).

554 Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation.
555 In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=oDbiL9CLOs>.
556

557 Akari Asai, Zequi Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning
558 to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference*
559 *on Learning Representations*, 2024. URL <https://openreview.net/forum?id=hSyW5go0v8>.
560

561 Arian Askari, Chuan Meng, Mohammad Aliannejadi, Zhaochun Ren, Evangelos Kanoulas, and
562 Suzan Verberne. Generative retrieval with few-shot indexing. *ArXiv*, abs/2408.02152, 2024. URL
563 <https://api.semanticscholar.org/CorpusID:271710011>.
564

565 Mehul Bhattacharyya, Valerie M Miller, Debjani Bhattacharyya, and Larry E Miller. High rates
566 of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*, 15(5):
567 e39238, May 2023. doi: 10.7759/cureus.39238. URL <https://pubmed.ncbi.nlm.nih.gov/37337480/>.
568

569 Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine
570 Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data
571 from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp.
572 2633–2650, 2021.

573 Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable
574 influence and fact tracing for large language model pretraining. In *The Thirteenth International*
575 *Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=gLa96F1Wwn>.
576

577 Mikael Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert,
578 Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. Hallucination
579 rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis.
580 *Journal of Medical Internet Research*, 26:e53164, May 2024. doi: 10.2196/53164. URL <https://pubmed.ncbi.nlm.nih.gov/38776130/>.
581

582 Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi,
583 Hanna Hajishirzi, Luke S. Zettlemoyer, and Pang Wei Koh. Copybench: Measuring literal and
584 non-literal reproduction of copyright-protected text in language model generation. *ArXiv preprint*,
585 abs/2407.07087, 2024. URL <https://arxiv.org/abs/2407.07087>.
586

587 Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria
588 Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. Selfcite: Self-supervised alignment
589 for context attribution in large language models. *ArXiv preprint*, abs/2502.09604, 2025. URL
590 <https://arxiv.org/abs/2502.09604>.
591

592 Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5:
593 Long form question answering. In Anna Korhonen, David Traum, and Lluís Márquez (eds.),

594 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.
 595 3558–3567, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/
 596 P19-1346. URL <https://aclanthology.org/P19-1346>.

597 Hanpei Fang, Sijie Tao, Nuo Chen, Kai-Xin Chang, and Tetsuya Sakai. Do large language models
 598 favor recent content? a study on recency bias in lilm-based reranking. 2025. URL <https://api.semanticscholar.org/CorpusID:281314971>.

600 Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan,
 601 Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching
 602 and revising what language models say, using language models. In Anna Rogers, Jordan Boyd-
 603 Gruber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association
 604 for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July
 605 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL
 606 <https://aclanthology.org/2023.acl-long.910/>.

607 Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan,
 608 Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching
 609 and revising what language models say, using language models. In Anna Rogers, Jordan Boyd-
 610 Gruber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association
 611 for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada,
 612 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL
 613 <https://aclanthology.org/2023.acl-long.910>.

614 Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate
 615 text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023
 616 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore,
 617 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL
 618 <https://aclanthology.org/2023.emnlp-main.398>.

619 Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. Metadata
 620 conditioning accelerates language model pre-training. In *Forty-second International Conference on
 621 Machine Learning*, 2025. URL <https://openreview.net/forum?id=DdMMz1I5YE>.

622 Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. TrueTeacher:
 623 Learning factual consistency evaluation with large language models. In Houda Bouamor, Juan
 624 Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in
 625 Natural Language Processing*, pp. 2053–2070, Singapore, 2023. Association for Computational
 626 Linguistics. doi: 10.18653/v1/2023.emnlp-main.127. URL <https://aclanthology.org/2023.emnlp-main.127>.

627 Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language
 628 model inference. *ArXiv preprint*, abs/2301.00303, 2023. URL <https://arxiv.org/abs/2301.00303>.

629 Chengyu Huang, Zequi Wu, Yushi Hu, and Wenya Wang. Training language models to generate
 630 text with citations via fine-grained rewards. *ArXiv preprint*, abs/2402.04315, 2024a. URL
 631 <https://arxiv.org/abs/2402.04315>.

632 Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang,
 633 Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang,
 634 Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li,
 635 Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis
 636 Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao,
 637 Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John C. Mitchell,
 638 Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang
 639 Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana,
 640 Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao
 641 Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue
 642 Zhao. Position: Trustlilm: Trustworthiness in large language models. In *Forty-first International
 643 Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net,
 644 2024b. URL <https://openreview.net/forum?id=bWUU0LwwMp>.

648 Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. To trust or not to trust? enhancing
 649 large language models' situated faithfulness to external contexts. In *The Thirteenth International*
 650 *Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=K2jOacHUL0>.

651

652 Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. Hagrid: A human-
 653 llm collaborative dataset for generative information-seeking with attribution. *ArXiv preprint*,
 654 [abs/2307.16883](https://arxiv.org/abs/2307.16883), 2023. URL <https://arxiv.org/abs/2307.16883>.

655

656 Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and
 657 Hao Peng. Source-aware training enables knowledge attribution in language models. In *First*
 658 *Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=UPyWLwciYz>.

659

660 Xiaoxi Li, Yujia Zhou, and Zhicheng Dou. Unigen: A unified generative framework for retrieval
 661 and question answering with large language models. *ArXiv*, [abs/2312.11036](https://arxiv.org/abs/2312.11036), 2023. URL <https://api.semanticscholar.org/CorpusID:266359654>.

662

663

664 Xiaoxi Li, Zhicheng Dou, Yujia Zhou, and Fangchao Liu. Corpuslm: Towards a unified language
 665 model on corpus for knowledge-intensive tasks. In Grace Hui Yang, Hongning Wang, Sam
 666 Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International*
 667 *ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024,*
 668 *Washington DC, USA, July 14-18, 2024*, pp. 26–37. ACM, 2024a. doi: 10.1145/3626772.3657778.
 669 URL <https://doi.org/10.1145/3626772.3657778>.

670

671 Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From
 672 matching to generation: A survey on generative information retrieval. *ACM Transactions on*
 673 *Information Systems*, 2024b. URL <https://api.semanticscholar.org/CorpusID:269303210>.

674

675 Huawei Lin, Jikai Long, Zhaozhuo Xu, and Weijie Zhao. Token-wise influential training data
 676 retrieval for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.),
 677 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*
 678 *1: Long Papers)*, pp. 841–860, Bangkok, Thailand, August 2024. Association for Computational
 679 Linguistics. doi: 10.18653/v1/2024.acl-long.48. URL <https://aclanthology.org/2024.acl-long.48/>.

680

681 Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. In-batch negatives for knowledge distillation
 682 with tightly-coupled teachers for dense retrieval. In Anna Rogers, Iacer Calixto, Ivan Vulić, Naomi
 683 Saphra, Nora Kassner, Oana-Maria Camburu, Trapit Bansal, and Vered Shwartz (eds.), *Proceedings*
 684 *of the 6th Workshop on Representation Learning for NLP (Rep4NLP-2021)*, pp. 163–173, Online,
 685 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.rep4nlp-1.17. URL
 686 <https://aclanthology.org/2021.rep4nlp-1.17>.

687

688 Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary,
 689 Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna
 690 James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David
 691 Albright, Carissa Schoenick, Luca Soldaini, Dirk Groeneveld, Rock Yuren Pang, Pang Wei Koh,
 692 Noah A. Smith, Sophie Lebrecht, Yejin Choi, Hannaneh Hajishirzi, Ali Farhadi, and Jesse Dodge.
 693 OLMoTrace: Tracing language model outputs back to trillions of training tokens. In Pushkar
 694 Mishra, Smaranda Muresan, and Tao Yu (eds.), *Proceedings of the 63rd Annual Meeting of the*
 695 *Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 178–188,
 696 Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-
 697 253-4. doi: 10.18653/v1/2025.acl-demo.18. URL <https://aclanthology.org/2025.acl-demo.18/>.

698

699 Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In
 700 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational*
 701 *Linguistics: EMNLP 2023*, pp. 7001–7025, Singapore, 2023. Association for Computational
 702 Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467. URL <https://aclanthology.org/2023.findings-emnlp.467>.

702 Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and
 703 Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the*
 704 *Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL
 705 <https://aclanthology.org/2024.tacl-1.9/>.

706 Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-curated questions and attributed answers. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3025–3045, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nacl-long.167. URL <https://aclanthology.org/2024.nacl-long.167/>.

707 Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi.
 708 When not to trust language models: Investigating effectiveness of parametric and non-parametric
 709 memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the*
 710 *61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
 711 pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:
 712 10.18653/v1/2023.acl-long.546. URL <https://aclanthology.org/2023.acl-long.546/>.

713 Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick,
 714 Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese.
 715 Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147, 2022a.
 716 URL <https://api.semanticscholar.org/CorpusID:247594830>.

717 Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick,
 718 Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching
 719 language models to support answers with verified quotes. *ArXiv preprint*, abs/2203.11147, 2022b.
 720 URL <https://arxiv.org/abs/2203.11147>.

721 João Monteiro, Pierre-André Noël, Étienne Marcotte, Sai Rajeswar Mudumba, Valentina Zantedeschi,
 722 David Vázquez, Nicolas Chapados, Chris Pal, and Perouz Taslakian. Replqa: A question-
 723 answering dataset for benchmarking llms on unseen reference content. In Amir Globersons,
 724 Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng
 725 Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference*
 726 *on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada,*
 727 *December 10 - 15, 2024*. URL http://papers.nips.cc/paper_files/paper/2024/hash/2b23626015b6311369e95a70735cbb72-Abstract-Datasets_and_Benchmarks_Track.html.

728 Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane
 729 Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models.
 730 In *Proceedings of the 37th International Conference on Neural Information Processing Systems*,
 731 *NIPS '23*, Red Hook, NY, USA, 2023. Curran Associates Inc.

732 Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher
 733 Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou,
 734 Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt:
 735 Browser-assisted question-answering with human feedback. *arXiv preprint arXiv: 2112.09332*,
 736 2021.

737 Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK:
 738 attributing model behavior at scale. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara
 739 Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine*
 740 *Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of*
 741 *Machine Learning Research*, pp. 27074–27113. PMLR, 2023. URL <https://proceedings.mlr.press/v202/park23c.html>.

742 Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller,
 743 and Sebastian Riedel. How context affects language models' factual predictions. *ArXiv preprint*,
 744 abs/2005.04611, 2020. URL <https://arxiv.org/abs/2005.04611>.

756 Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao,
 757 James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim
 758 Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks.
 759 In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven
 760 Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021*
 761 *Conference of the North American Chapter of the Association for Computational Linguistics:*
 762 *Human Language Technologies*, pp. 2523–2544, Online, 2021. Association for Computational
 763 Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL <https://aclanthology.org/2021.naacl-main.200>.

764

765 Aleksandra Piktus, Fabio Petroni, Yizhong Wang, Vladimir Karpukhin, Dmytro Okhonko, Samuel
 766 Broscheit, Gautier Izacard, Patrick Lewis, Barlas Ouguz, Edouard Grave, Wen tau Yih, and
 767 Sebastian Riedel. The web is your oyster - knowledge-intensive nlp against a very large web corpus.
 768 *ArXiv preprint*, abs/2112.09924, 2021. URL <https://arxiv.org/abs/2112.09924>.

769

770 Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud
 771 Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng,
 772 and Furu Wei. Scaling laws of synthetic data for language models, 2025. URL <https://arxiv.org/abs/2503.19551>.

773

774 Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Srinivasan. Enhancing post-hoc
 775 attributions in long document comprehension via coarse grained answer decomposition. In Yaser
 776 Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on*
 777 *Empirical Methods in Natural Language Processing*, pp. 17790–17806, Miami, Florida, USA,
 778 November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.
 779 985. URL <https://aclanthology.org/2024.emnlp-main.985/>.

780

781 Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das,
 782 Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural
 783 language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486. URL <https://aclanthology.org/2023.cl-4.2>.

784

785 Ruiyang Ren, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. TOME:
 786 A two-stage approach for model-based retrieval. In Anna Rogers, Jordan Boyd-Graber, and
 787 Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Com-*
 788 *putational Linguistics (Volume 1: Long Papers)*, pp. 6102–6114, Toronto, Canada, July 2023.
 789 Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.336. URL <https://aclanthology.org/2023.acl-long.336/>.

790

791 Abhilasha Sanchez, Koustava Goswami, and Balaji Srinivasan. Post-hoc answer attribution for
 792 grounded and trustworthy long document comprehension: Task, insights, and challenges. In
 793 Danushka Bollegala and Vered Shwartz (eds.), *Proceedings of the 13th Joint Conference on*
 794 *Lexical and Computational Semantics (*SEM 2024)*, pp. 49–57, Mexico City, Mexico, June
 795 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.4. URL
 796 <https://aclanthology.org/2024.starsem-1.4/>.

797

798 Tal Schuster, Adam Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William Cohen, and Donald
 799 Metzler. SEMQA: Semi-extractive multi-source question answering. In Kevin Duh, Helena Gomez,
 800 and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter*
 801 *of the Association for Computational Linguistics: Human Language Technologies (Volume 1:*
 802 *Long Papers)*, pp. 1363–1381, Mexico City, Mexico, June 2024. Association for Computational
 803 Linguistics. doi: 10.18653/v1/2024.naacl-long.74. URL <https://aclanthology.org/2024.naacl-long.74/>.

804

805 Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. Attribute first, then
 806 generate: Locally-attributable grounded text generation. In Lun-Wei Ku, Andre Martins, and
 807 Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Com-*
 808 *putational Linguistics (Volume 1: Long Papers)*, pp. 3309–3344, Bangkok, Thailand, August
 809 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.182. URL
<https://aclanthology.org/2024.acl-long.182/>.

810 Mingyang Song and Mao Zheng. A survey of query optimization in large language models. *arXiv*
 811 *preprint arXiv:2412.17558*, 2024. URL <https://arxiv.org/abs/2412.17558>.

812

813 Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet
 814 long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of*
 815 *the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu
 816 Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL <https://aclanthology.org/2022.emnlp-main.566>.

817

818 Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. Recitation-augmented language
 819 models. In *The Eleventh International Conference on Learning Representations, ICLR 2023,*
 820 *Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/pdf?id=-cqvvvb-NkI>.

821

822

823 Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai
 824 Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler.
 825 Transformer memory as a differentiable search index. In Sanmi Koyejo, S. Mohamed,
 826 A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural*
 827 *Information Processing Systems 35: Annual Conference on Neural Information Processing*
 828 *Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,*
 829 *2022*. URL http://papers.nips.cc/paper_files/paper/2022/hash/892840a6123b5ec99ebaab8be1530fba-Abstract-Conference.html.

830

831 Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbahn.
 832 Position: will we run out of data? limits of llm scaling based on human-generated data. In
 833 *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org,
 834 2024.

835

836 Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung,
 837 Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with
 838 search engine augmentation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings*
 839 *of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, Bangkok, Thailand,
 840 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.813. URL
<https://aclanthology.org/2024.findings-acl.813/>.

841

842 William H. Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations
 843 generated by ChatGPT. *Scientific Reports*, 13:14045, 2023. doi: 10.1038/s41598-023-41032-5.
 844 URL <https://www.nature.com/articles/s41598-023-41032-5>.

845

846 Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit,
 847 Tong Xie, and Ian Foster. Sciqaq: A framework for auto-generated science question answering
 848 dataset with fine-grained evaluation. *ArXiv preprint*, abs/2405.09939, 2024. URL <https://arxiv.org/abs/2405.09939>.

849

850 Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi
 851 Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models:
 852 Knowledge, retrieval and domain-specificity. *ArXiv preprint*, abs/2310.07521, 2023. URL <https://arxiv.org/abs/2310.07521>.

853

854 Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Overcoming
 855 imperfect retrieval augmentation and knowledge conflicts for large language models. In Wanxiang
 856 Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the*
 857 *63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
 858 pp. 30553–30571, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN
 859 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1476. URL <https://aclanthology.org/2025.acl-long.1476/>.

860

861 Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang,
 862 and William Yang Wang. Generalization v.s. memorization: Tracing language models' capabilities
 863 back to pretraining data. In *The Thirteenth International Conference on Learning Representations*,
 2025b. URL <https://openreview.net/forum?id=IQxBDLmVpT>.

864 Ye Wang, Xinrun Xu, Rui Xie, Wenxin Hu, and Wei Ye. Generative retrieval with large language
 865 models. *ArXiv preprint*, abs/2402.17010, 2024. URL <https://arxiv.org/abs/2402.17010>.

866

867 Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia,
 868 Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao
 869 Yang. A neural corpus indexer for document retrieval. In Alice H. Oh, Alekh Agarwal, Danielle
 870 Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.
 871 URL https://openreview.net/forum?id=fSfcEYQP_qc.

872

873 Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin
 874 Van Durme. “according to . . .”: Prompting language models improves quoting from pre-
 875 training data. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference
 876 of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Pa-
 877 pers)*, pp. 2288–2301, St. Julian’s, Malta, 2024. Association for Computational Linguistics. URL
 878 <https://aclanthology.org/2024.eacl-long.140/>.

879

880 Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán,
 881 Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets
 882 from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri,
 883 Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani,
 884 Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the
 885 Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France,
 886 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.494>.

887

888 Yingtai Xiao, Yuqing Zhu, Sirat Samyoun, Wanrong Zhang, Jiachen T. Wang, and Jian Du. Token-
 889 Shapley: Token level context attribution with shapley value. In Wanxiang Che, Joyce Nabende,
 890 Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Com-
 891 putational Linguistics: ACL 2025*, pp. 3882–3894, Vienna, Austria, July 2025. Association for
 892 Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.200.
 893 URL <https://aclanthology.org/2025.findings-acl.200/>.

894

895 Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth:
 896 Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth Inter-
 897 national Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
 898 OpenReview.net, 2024. URL <https://openreview.net/forum?id=auKAUJZMO6>.

899

900 Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic
 901 continued pretraining. In *The Thirteenth International Conference on Learning Representations*,
 902 2025. URL <https://openreview.net/forum?id=07yvxWDS1a>.

903

904 Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. Effective large language model adaptation for
 905 improved grounding and citation generation. In Kevin Duh, Helena Gomez, and Steven Bethard
 906 (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association
 907 for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp.
 908 6237–6251, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.naacl-long.346>.

909

910 Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attri-
 911 bution by large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings
 912 of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, Singapore, 2023.
 913 Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.307. URL
 914 <https://aclanthology.org/2023.findings-emnlp.307>.

915

916 Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao
 917 Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to generate fine-grained citations in
 918 long-context QA. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher
 919 Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–
 920 5122, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-
 921 256-5. doi: 10.18653/v1/2025.findings-acl.264. URL [https://aclanthology.org/2025.findings-acl.264/](https://aclanthology.org/2025.findings-acl.264).

918 Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. Verifiable
 919 by design: Aligning language models to quote from pre-training data. In Luis Chiruzzo, Alan
 920 Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas*
 921 *Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume*
 922 *1: Long Papers)*, pp. 3748–3768, Albuquerque, New Mexico, 2025b. Association for Computa-
 923 tional Linguistics. ISBN 979-8-89176-189-6. URL <https://aclanthology.org/2025.naacl-long.191/>.

925 Guido Zuccon, Bevan Koopman, and Razia Shaik. Chatgpt hallucinates when attributing answers. In
 926 *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in*
 927 *Information Retrieval in the Asia Pacific Region*, pp. 46–51, 2023.

928

930 A DATASETS

931

932 A.1 CORPUS CONSTRUCTION

933

934 **Common Crawl** To construct the essential document set, we require full documents for continual
 935 pretraining—unlike the Sphere corpus (Piktus et al., 2021), which consists of 100-word passages. We
 936 process raw 2019 Aug Common Crawl snapshots using CCNet (Wenzek et al., 2020) to obtain clean
 937 full-text documents. To identify documents relevant to ELI5, we first index 100-word passages (in
 938 Sphere format) and use BM25 to retrieve the top-100 passages for each question in the ELI5 train,
 939 dev, and test sets. We then use an LLM to decompose each question into self-contained fact claims,
 940 treating each claim as a separate query to retrieve the top-200 relevant passages.

941 Next, we verify whether the retrieved passages support each claim and compile a support set. This set
 942 is combined with the original relevant passages, and using passage metadata, we locate and extract
 943 the corresponding full documents from Common Crawl. This process yields our core document
 944 set, comprising 30,025 documents with a total of 110,594,287 tokens, averaging 3,683 tokens per
 945 document. For training set instances with unsupported claims, we use GPT-4o to generate synthetic
 946 documents styled like Common Crawl, ensuring the model is not trained on unsupported inputs. This
 947 results in a corpus from the **Other** source, containing 7,593 documents with 2,210,346 total tokens,
 948 averaging 291 tokens per document.

949 The data is licensed under CC0.

950

951 **Wikipedia** We build on the KILT knowledge source (Petroni et al., 2021), based on the 2019/08/01
 952 Wikipedia snapshot. We include source documents from ASQA, the TriviaQA dev set, and a subset
 953 of HotpotQA (medium-level questions), mapped to KILT IDs. This forms our essential Wikipedia
 954 corpus, consisting of 30,025 documents totaling 110.6M tokens (avg. 3,683 tokens/document).

955 The data is licensed under CC BY-SA 3.0.

956

957 **ArXiv** SciQAG (Wan et al., 2024) provides scientific papers without titles. We retrieve titles
 958 using the papers’ DOIs from arXiv, resulting in 22,743 documents with 114.0M tokens (avg. 5,013
 959 tokens/document).

960 arXiv metadata is used under the CC0; license for each paper varies.

961

962 **RepliQA** We use the first two splits of the RepliQA dataset (Monteiro et al., 2024) as the full
 963 corpus, yielding 7,182 documents and 8.88M tokens (avg. 1,236 tokens/document). Due to frequent
 964 title duplication, we use an LLM to generate more descriptive titles.

965 This core corpus totals approximately 392M tokens. The data is licensed under CC BY 4.0.

966

968 A.2 RENAME DUPLICATED TITLES

969

970 Due to frequent title duplication—especially in Common Crawl and RepliQA—we adopt a renaming
 971 strategy using an LLM. For each duplicated title, we iteratively rename the document until all titles
 972 are unique. We also perform cross-source deduplication after renaming.

972 Table 4: Full results for **INSONLY**, **PASSIDX**, and **ACTIDX** across all five base models on four QA
 973 datasets. **Acc** = answer correctness, **C-Pr** = citation precision, **C-Re** = citation recall.

975	976	Model	Method	ASQA			Eli5			SciQAG			RepliQA			
				977	978	979	980	981	982	983	984	985	986	987	988	989
				15.9	3.7	4.1	9.2	0.6	0.6	61.2	0.0	0.0	15.2	0.2	0.2	
Q-3B	PassIdx	16.5	17.1	17.4	11.7	7.1	7.2	67.0	1.1	1.6	22.8	1.8	2.4			
	ActIdx	21.4	23.9	24.2	15.8	19.7	19.8	67.9	20.0	23.0	24.5	10.5	15.8			
	InsOnly	19.1	20.0	21.2	11.5	5.9	6.4	65.9	0.6	0.8	24.2	0.9	1.3			
Q-7B	PassIdx	21.5	24.1	24.2	14.5	8.9	9.0	65.7	2.4	2.4	24.8	2.4	2.5			
	ActIdx	27.6	30.9	31.1	17.6	29.3	29.5	66.6	32.6	33.6	31.9	24.4	25.7			
	InsOnly	26.4	25.1	27.5	14.9	5.9	6.3	67.4	0.5	0.8	27.6	0.75	0.9			
Q-14B	PassIdx	30.0	27.1	27.3	18.5	9.3	9.4	67.4	2.3	2.3	28.4	2.0	2.0			
	ActIdx	31.6	40.9	41.3	19.3	33.2	33.6	70.2	34.8	35.1	37.0	30.8	31.5			
	InsOnly	18.0	11.9	18.1	9.4	3.6	3.9	61.8	0.2	0.2	19.0	0.27	0.4			
L-3B	PassIdx	21.5	18.5	19.1	11.2	7.1	7.9	63.2	1.8	2.1	23.4	1.8	2.0			
	ActIdx	25.9	29.0	29.3	15.6	25.9	26.3	65.6	13.2	32.8	26.8	13.9	23.0			
	InsOnly	25.4	23.3	25.5	12.1	5.0	5.5	67.3	1.0	1.6	24.1	0.7	1.4			
L-8B	PassIdx	28.6	27.8	28.1	16.1	9.3	9.5	65.1	3.6	3.7	25.6	1.9	1.9			
	ActIdx	28.2	31.7	31.8	17.9	30.7	30.8	67.7	39.3	40.3	35.7	36.2	36.9			

A.3 DOWNSTREAM TASKS

For ELI5, the test set is sourced from (Gao et al., 2023c), while the training and development sets are derived from a newer version.² Any duplicates from the test set have been removed from the training and development sets. ELI5 is licensed under BSD 3.

For ASQA, we map annotated source documents to the KILT knowledge base and filter out datapoints with unmatched sources, resulting in 863 test examples. ASQA is licensed under CC BY 4.0.

For SciQAG, due to noise in the training set, we re-split the original test set into train, dev, and test splits based on documents, ensuring no title overlap. This yields 853, 95, and 300 documents, respectively. Evaluation is performed on 1,000 sampled questions from the 300 test documents. SciQAG is released on GitHub without explicit license, we will not redistribute it without author acknowledgment.

For RepliQA, we similarly split train, dev, and test sets by documents. The test set includes 1,000 QA pairs sampled from the test documents. RepliQA is licensed under CC BY 4.0.

A.4 INSTRUCTION TUNING SET

We randomly sample 1,000 training examples from ASQA, ELI5, RepliQA, SciQAG, and HotpotQA (medium level), and 200 from each dev split to form a validation set for early stopping during instruction tuning.

B ADDITIONAL RESULTS

B.1 MAIN RESULTS OF MORE MODELS

We also run active-indexing experiments on Qwen-2.5-14B, Llama-3.2-3B, and Llama-3.1-8B. Performance increases monotonically as model capacity scales—Qwen from 3B → 7B → 14B and Llama from 3B → 8B—further confirming that citation ability improves with model size. As shown in Table 4, Active Indexing consistently outperforms Passive Indexing across models and datasets. Also, as the model size increases, the performance of Active Indexing improves monotonically.

²https://huggingface.co/datasets/rexarski/eli5_category

1026
1027

B.2 COMPLEMENTARY OF FORWARD AND BACKWARD

1028

We study scaled complementary of forward and backward scale effect on repliqa. We augment the RepliQA corpus: 66M tokens from Active Indexing Forward (AIF), 76M from Backward (AIB), and their combination (AI), totaling $16.5 \times$ the original size. From these, we subsample different proportions to construct a scaling curve. Figure 4 shows citation precision on RepliQA (y-axis) versus the fraction of augmented data used (x-axis). Key findings include: 1. **Scaling helps:** All variants—AIF, AIB, and AI—improve with more augmented data, even up to $16.5 \times$. 2. **Complementary directions:** AIF alone is less effective, but combining AIF with AIB (i.e., AI) consistently outperforms either alone. Each point on the AI curve uses matched subsets of AIF and AIB, indicating additive and reinforcing effects. 3. **The role of forward** While backward provides most gains, but forward provides consistent additive effect, suitable for users willing to trade-off compute for better performance. Moreover, It supports answering realistic queries like “Explain document X / paper Y” in a closed-book manner, which requires the model to internally condition on document identifiers.

1045

1046

1047

B.3 BEHAVIOR OF HYBRID SYSTEMS UNDER INTERNAL-EXTERNAL CONFLICTS

1049

Hybrid systems are most useful in the conflict regime, where the model’s internal belief disagrees with the retrieved context. To make this behavior explicit, we analyze RepliQA by partitioning examples into three slices based on correctness of the internal-only and external-only systems: 1. Internal wrong / External correct ($\text{Int}=0, \text{Ext}=1$), 2. Internal correct / External wrong ($\text{Int}=1, \text{Ext}=0$), and 3. No conflict (both correct or both wrong).

1055

1056

1057

Table 5 reports both (i) the proportion of each slice and (ii) the accuracy of Internal-only, External-only, Hybrid-Joint, Hybrid-Fallback, and Hybrid-Oracle in each slice, under high-quality retrieval (retrieval quality = 1.0) and noisy retrieval (0.2).

1058

1059

1060

1061

1062

1063

1064

Under high-quality retrieval (1.0), the dominant conflict regime is precisely the “production reality” where retrieval corrects the model: the $\text{Int}=0, \text{Ext}=1$ slice accounts for 38.6% of examples, compared to only 8.3% for $\text{Int}=1, \text{Ext}=0$. In this regime, Hybrid-Fallback behaves almost like an ideal “defer-to-retrieval” policy: it nearly matches External-only accuracy in the $\text{Int}=0, \text{Ext}=1$ slice (99.5 vs. 100.0), while Internal-only is always wrong by construction (0.0). Hybrid-Joint is more conservative: it substantially improves over Internal-only (75.6 vs. 0.0) but falls short of External-only, reflecting its stronger reliance on internal knowledge even when retrieval is correct.

1065

1066

1067

1068

1069

1070

1071

When retrieval quality degrades (0.2), the picture flips: the $\text{Int}=1, \text{Ext}=0$ slice becomes more prevalent (22.0% vs. 9.9% for $\text{Int}=0, \text{Ext}=1$). In this setting, both hybrid methods shift their arbitration toward internal knowledge. For example, in the $\text{Int}=1, \text{Ext}=0$ slice, Hybrid-Joint attains 58.2 accuracy, substantially outperforming External-only (0.0) and improving over Internal-only’s baseline, while Hybrid-Fallback also achieves strong gains relative to External-only. This demonstrates that the hybrids do not blindly trust retrieval; they adaptively fall back to parametric knowledge when the retrieved context is unreliable.

1072

1073

1074

1075

1076

Finally, even in the no-conflict slice where both Internal-only and External-only are wrong, Hybrid-Joint can outperform both by aggregating partial evidence from the two sources. This reflects a second layer of complementarity beyond simple source selection: hybrid reasoning can construct better answers by jointly leveraging noisy parametric and contextual signals, not just by choosing between them.

1077

1078

1079

Overall, this analysis directly quantifies the conflict slices and shows that our hybrid strategies implement sensible arbitration policies: when retrieval is reliable, they behave like “switch-to-external” systems; when retrieval is noisy, they lean back on internal knowledge while still extracting value from combining both signals.

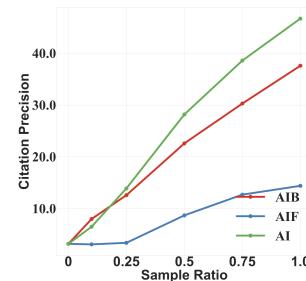


Figure 4: Scaling Curve of Combining Backward and Forward on RepliQA

RQ	Method	Int wrong / Ext correct	Int correct / Ext wrong	No conflict	Total
1.0	Hybrid-Joint	75.6	51.8	48.4	59.2
	Hybrid-Fallback	99.5	6.0	44.4	62.5
	Hybrid-Oracle	100.0	100.0	44.4	70.5
	External-only	100.0	0.0	44.4	62.2
	Internal-only	0.0	100.0	44.4	31.9
0.2	Hybrid-Joint	66.7	58.2	24.5	36.1
	Hybrid-Fallback	88.9	41.8	21.9	32.9
	Hybrid-Oracle	100.0	100.0	14.5	41.8
	External-only	100.0	0.0	14.5	19.8
	Internal-only	0.0	100.0	14.5	31.9

Table 5: Behavior of hybrid systems under internal–external conflicts on RepliQA. Entries are accuracies within each correctness slice: Internal wrong / External correct, Internal correct / External wrong, and No conflict. For retrieval quality 1.0, the slice proportions are 38.6%, 8.3%, and 53.1%, respectively; for retrieval quality 0.2, they are 9.9%, 22.0%, and 68.1%.

B.4 OUT-OF-DOMAIN EVALUATION AFTER ACTIVE INDEXING

We evaluate the out-of-domain behavior of models after *RepliQA-only* continual pre-training (CPT). We measure (i) perplexity on held-out Wikipedia and ArXiv corpora, and (ii) TriviaQA accuracy after task-specific fine-tuning. Consistent with Khalifa et al. (2024), we observe an increase in perplexity on natural text after CPT. We attribute this rise to two factors: (1) **Domain-specific CPT**, which is known to induce catastrophic forgetting when the CPT corpus is significantly narrower than the original pre-training distribution; and (2) **Document identifiers**, whose form (e.g., natural-language titles vs. integers) introduces different levels of distributional shift.

To disentangle these effects, we train five CPT variants on the same backbone with identical token budgets: (1) raw data (no titles), (2) Passive Indexing with natural-language title IDs, (3) Passive Indexing with integer IDs, (4) Passive Indexing with repeated natural-language IDs, and (5) Active Indexing. We then evaluate all models using the same OOD metrics.

CPT Method	Wiki PPL	ArXiv PPL	TriviaQA Acc.
Base Model (No CPT)	6.71	4.61	56.0
Raw Data (No Title)	32.15	6.65	50.8
Passive Index (Title ID)	54.31	6.90	49.5
Passive Index (Integer ID)	60.15	6.88	49.2
Passive Index Repeat (Title)	39.78	6.80	51.2
Active Index	26.52	5.40	50.5

Table 6: OOD evaluation after RepliQA-only continual pre-training. Lower perplexity is better. Higher accuracy on triviaqa is better.

Findings. **1. All CPT variants raise OOD perplexity**, confirming that narrow-domain CPT is the dominant source of forgetting, as even the Raw Data (no-title) model shows substantial degradation. **2. Natural-language IDs are less harmful than integer IDs**: among Passive Index variants, perplexity follows the pattern natural-text < repeated-natural-text < integer-ID, indicating that well-formed textual identifiers integrate more smoothly into language modeling. **3. Repeated titles are not particularly damaging**: unlike Khalifa et al. (2024), we do not observe large penalties from repetition, likely because our identifiers are natural text rather than opaque tokens. **4. Active Indexing is the least harmful variant**: it obtains the lowest perplexity on both Wikipedia and ArXiv, suggesting its augmentation procedure improves linguistic fluency within this knowledge domain. **5. OOD effects vary across corpora**: perplexity increases more on Wikipedia than on ArXiv, indicating that forgetting depends on similarity between the CPT distribution and the target corpus. Finally, **6. perplexity does not correlate with QA performance**: despite substantial perplexity differences

1134 across variants, TriviaQA accuracy after fine-tuning remains similar, implying that perplexity shifts
 1135 reflect stylistic or distributional drift rather than loss of underlying knowledge.
 1136

1137 C MORE RELATED WORK

1139 **More External Citations** A parallel line of external citation work targets *finer-grained* ground-
 1140 ing, at the sentence/span and even token level, primarily in open-book settings. Systems such as
 1141 *GopherCite* explicitly interleave answers with short, verified quotes and learn to abstain when unsure,
 1142 illustrating span-level attribution tied to retrieved evidence. (Menick et al., 2022a) Semi-extractive
 1143 generation (*SEMQA/QuoteSum*) enforces copy-and-connect outputs, yielding inline quoted spans
 1144 by construction, (Schuster et al., 2024) while *locally-attributable* generation optimizes for concise,
 1145 sentence-local citations consumers can check quickly. (Slobodkin et al., 2024) For long-context QA,
 1146 *LongCite* and its *LongBench-Cite* benchmark train models that produce answers with *sentence-level*
 1147 citations in one pass (Zhang et al., 2025a). **They still cite from the external context and can't produce**
 1148 **citations in a close-book manner.** On the evaluation side, *EXPERTQA* contributes expert-curated
 1149 questions and *expert-verified, claim-evidence* annotations for long-form answers—auditing system-
 1150 provided citations and enabling claim-level scoring in high-stakes domains. (Malaviya et al., 2024)
 1151 Methods that unify retrieval and reflection (e.g., *Self-RAG*), add self-critique flags to segment-level
 1152 citations. (Asai et al., 2024) Beyond during-generation approaches, *post-hoc* attribution retrofits
 1153 support after the fact. *RARR* (“research & revise”) finds evidence and edits outputs to align claims
 1154 with sources, (Gao et al., 2023a) and follow-ups for long documents decompose answers into factual
 1155 units before mapping each unit to supporting sentences, improving coverage of fine-grained support.
 1156 (Sanchez et al., 2024; Ramu et al., 2024) Finally, token-level context credit assignment (e.g., *Token-
 1157 Shapley*) scores which specific context tokens support each generated token. (Xiao et al., 2025) These
 1158 directions still rely on external evidence at inference, and thus constitute *external* citation.
 1159

1160 **Data attribution** seeks the training examples that most influenced a model’s behaviour (Park et al.,
 1161 2023) (e.g., via influence functions or gradient tracing), while **fact attribution** asks which facts
 1162 support a given answer—often handled by external retrieval or by pretraining identifiers that may
 1163 not coincide with truly influential data. Recent analysis-time tools underscore this gap: *OLMoTrace*
 1164 traces verbatim spans in model outputs back to specific pretraining documents across multi-trillion-
 1165 token corpora in real time, and *RapidIn* retrieves token-wise influential training points at scale via
 1166 compressed gradient caching (Liu et al., 2025; Lin et al., 2024). Empirical studies further show that
 1167 examples ranked as most *influential* need not be those that explicitly *contain* the cited fact, and that
 1168 alignment between the two increases with model/corpus scale (Chang et al., 2025). These objectives
 1169 therefore do not necessarily align. **Active Indexing bridges them:** by presenting each training fact
 1170 with its document identifier in a QA-style format, it explicitly couples the influential data point with
 1171 its provenance, so that reproducing an identifier at inference time simultaneously evidences both data
 1172 lineage and factual entailment—achieving traceable citations without expensive post-hoc analysis
 1173 and aligning factual correctness with the model’s true training history.
 1174

1175 **Copyright, Transparency, and Legal Considerations** The legal status of training on copyrighted
 1176 content is the subject of active policy debate. (Chen et al., 2024) By encouraging models to surface
 1177 explicit provenance, our approach offers a technical step toward such transparency: citations make
 1178 it easier to audit whether a model relies on protected material. However, we also highlight a dual
 1179 risk I: stronger provenance can *increase* exposure of private or proprietary text snippets. Exploring
 1180 privacy-preserving identifiers (e.g., hashed IDs, differential privacy) and selective redaction during
 1181 generation remains important future work.
 1182

1183 D IMPLEMENTATION DETAILS

1184 D.1 TRAINING

1185 We perform continual pretraining using 4 H200 GPUs for 3 epochs across all methods in the main
 1186 experiments, with a batch size of 256, a maximum context length of 2048, and a constant learning
 1187 rate of 5e-5 with 10 warm-up steps. With 8-bit AdamW paged optimizers, each model fits on a single
 1188 GPU and can be trained using DDP without requiring FSDP. For instruction tuning, we use a linear

1188 decay scheduler with a learning rate of 5e-6, train for up to 5 epochs with early stopping, and use a
 1189 batch size of 64.

1190
 1191 The longest continual pretraining run (Qwen-2.5-7B on 3B tokens \times 3 epochs = 9B tokens) takes
 1192 320 H200 GPU hours (80h \times 4 GPUs). The shortest on 7B (Passive Indexing with 400M \times 3 = 1.2B
 1193 tokens) takes 43 H200 GPU hours.

1194

1195 **D.2 INTERNAL CITATIONS v.S. EXTERNAL CITATIONS**

1196

1197 In the experiments of comparing internal citation with external citations, we vary retrieval quality
 1198 from sparse (BM25) to dense retrieval (Lin et al., 2021) by mixing top-5 documents retrieved by each
 1199 method. Specifically, we construct hybrid top-5 sets by randomly sampling a proportion of documents
 1200 from the sparse and dense retrieval outputs. We report both answer correctness and citation f1 which
 1201 combines citation precision and recall as a single score.

1202

1203 And more details for our hybrid methods are below:

1204

1205 **Hybrid Joint:** To train a model that can leverage both internal and external knowledge, we fine-tune
 1206 the Active Indexing model to handle retrieved documents as input. For each question in the instruction
 1207 tuning set (Appendix A.4), we retrieve external documents using both sparse and dense methods. We
 1208 then create two training instances per question: One with sparse-retrieved documents + question as
 1209 input. One with dense-retrieved documents + question as input. In both cases, the target output is the
 1210 correct answer with citations. This trains the model to produce grounded answers regardless of the
 1211 quality of the provided documents. We find performance improves when we adopt a chain-of-thought
 1212 format, where the model first reflects on document sufficiency before answering, rather than directly
 1213 generating an answer.

1214

1215 **Hybrid Fallback:** At inference time, we first use a 3-shot prompt in the RAG setup to generate
 1216 answers using the top-5 retrieved documents. The prompt allows the model to abstain if it finds the
 1217 documents insufficient. If it abstains, we fall back to the Hybrid Joint model, which can rely on both
 1218 external and internal knowledge. If the documents are deemed sufficient, we proceed with standard
 1219 external-only RAG generation with citations.

1220

1221 **E CITATIONS EVALUATION**

1222

1223 **E.1 LONG-FORM CITATIONS EVALUATION**

1224

1225 For each model-generated long-form answer, we first use GPT-4.1 to decompose it into self-contained
 1226 claims, each linked to its cited source. We then retrieve the corresponding documents from a
 1227 MongoDB-based³ corpus. Since documents are often too long for the NLI model's (Gekhman et al.,
 1228 2023) input, we chunk each into 512-token segments and retrieve the most relevant ones. A citation
 1229 is considered correct if any chunk entails the claim. Citation precision is defined as the proportion of
 1230 citations that support their claims, and recall as the proportion of claims that are fully supported.

1231

1232 **E.2 SHORT-FORM CITATIONS ERROR ANALYSIS**

1233

1234 We conduct citation analysis on the best Qwen-2.5-7B model trained on RepliQA with Active
 1235 Indexing. Using GPT-4.1, we assess whether the cited documents entail the model-generated answers.
 1236 We find that 7% of examples cite documents that support the answer but do not match the gold
 1237 citation, indicating that true citation accuracy may be underestimated. However, 6.2% of examples
 1238 include correct citations but incorrect answers, suggesting that the model may sometimes retrieve the
 1239 right source while generating an incorrect response.

1240

1241

³<https://www.mongodb.com>

1242
1243

F PRELIMINARY EXPERIMENTS

1244
1245

F.1 DOCUMENT IDENTIFIERS

1246
1247
1248
1249
1250
1251
1252

We investigate which types of document identifiers are most effective for language models to memorize during continual pretraining. Experiments are conducted on the SciQAG dataset using LLaMA-3.1-1B. To evaluate memorization, we measure accuracy@1 and accuracy@10, where the model is given the document content and asked to rank possible document IDs by the summed log-probability of each ID. Higher scores indicate stronger memorization of the document-identifier associations. We compare different strategies for constructing document identifiers, evaluating their effectiveness in continual pretraining for memorization.

1253
1254
1255
1256

1. Natural Title (Raw): A baseline without continual pretraining. The model is directly prompted to rank text titles given document content. This tests whether pretrained LMs can match content to titles without exposure.

1257
1258
1259
1260

2. Natural Titles : We perform continual pretraining where each document is appended with its human-written text title. This approach uses natural-language identifiers that align with the model’s training distribution.

1261
1262
1263
1264
1265
1266

3. Hierarchical K-Means Integer (HKM-Integer): Instead of using random integers, we construct semantically structured integer IDs following Tay et al. (2022). Documents are embedded and clustered using K-means into 10 top-level groups. Each group is assigned a prefix digit. The process is recursively applied within each cluster, with each level adding a digit to the ID. Documents with shared prefixes are semantically similar, making it easier for the model to generalize over structured identifiers.

1267
1268
1269
1270
1271

4. Hierarchical LDA with Keyword Labels (HLDA-Keywords) We apply hierarchical topic modeling (LDA) to recursively cluster documents. For each cluster, we use an LLM to generate a representative keyword based on its most salient documents. The final identifier is a concatenation of these keywords along the cluster path, forming a semantic, hierarchical label.

1272
1273
1274
1275

5. Domain-First Keyword Identifier (Domain→Keywords) Each SciQAG document is tagged with a domain and associated keywords. We construct identifiers by concatenating the domain name with its keywords, creating a top-down semantic label (e.g., physics-energy-entropy).

1276
1277
1278
1279

6. Keyword-First Domain Identifier (Keywords→Domain) Similar to the above, but constructed in a bottom-up manner. Keywords appear first, followed by the broader domain label (e.g., entropy-energy-physics), emphasizing specificity before generality.

1280
1281

Table 7: Results of different document identifiers on SciQAG.

1282
1283
1284
1285
1286
1287
1288
1289

ID Type	Acc@1	Acc@10
Natural Titles (Raw)	9.7	46.3
Natural Titles	53.3	75.3
HKM-Integer	2.0	21.7
HLDA-Keywords	32.0	50.7
Domain->keywords	28.7	47.7
keywords->Domain	26.7	45.3

1290
1291
1292
1293
1294

As shown in Table 7, even semantically structured integer-based identifiers perform significantly worse than text-based methods. We attribute this to the nature of continual pretraining, where the document is directly followed by its identifier. In this setup, natural text provides a more effective and fluent learning signal for the language model.

1295

Among the text-based methods, natural titles achieve the highest performance. We hypothesize three reasons for this: (1) natural titles tend to capture the most salient information from the document

1296 and have high information density, (2) they are more fluent and better aligned with the model’s
 1297 pretraining distribution, and (3) the SciQAG corpus is relatively small—within the memorization
 1298 capacity of the model—so the benefits of structured or compressed identifiers (like integer codes)
 1299 are less pronounced. Such structured identifiers may only offer advantages at larger scales where
 1300 memory constraints become a limiting factor.

1301 F.2 WIKIPEDIA TASKS

1302
 1303
 1304 We evaluate our approach on two popular Wikipedia-based QA benchmarks: TriviaQA (short-form)
 1305 and HotpotQA (long-form). For HotpotQA, we focus on medium-difficulty, two-hop questions
 1306 to match the capabilities of Llama-3.1-1B. We compare two settings: (1) a Raw model directly
 1307 instruction-tuned on each task, and (2) Passive Indexing, which adds a continual pretraining stage on
 1308 the Wikipedia corpus before instruction tuning. For TriviaQA, we use Exact Match for Correctness
 1309 and Citation Precision for citation quality. While for HotpotQA, we use Exact Match for Correctness
 1310 and both Citation Precision and Citation Recall for citation quality.

1311
 1312 As shown in Table 8, continual pretraining provides no noticeable gains. The strong QA and citation
 1313 performance of the raw model indicates that LLaMA-3 already memorizes much of Wikipedia’s
 1314 content and titles during pretraining, leaving limited headroom for further improvement via continual
 1315 pretraining. Interestingly, citation accuracy exceeds QA accuracy, suggesting that the model can
 1316 often guess the correct title even without fully answering the question—likely because Wikipedia
 1317 articles are topically coherent and revolve around predictable entities.

1318
 1319
 1320 Table 8: Preliminary Experiments on TriviaQA and HotpotQA

1323 1B	Corpus	#Tokens	1324 TriviaQA			1325 HotpotQA		
			1326 Acc	1327 C-Pr	1328 C-Re	1329 Acc	1330 C-Pr	1331 C-Re
1332 Raw	-	-	34.7	59.6	-	26.0	55.9	50.7
1333 PassIndex	Wiki	418M	31.3	61.4	-	25.3	54.7	51.1

1334 G QUALITATIVE EXAMPLES OF CITE-PRETRAIN BEHAVIOR

1335
 1336 We perform a manual inspection of model outputs across both long-form and short-form QA set-
 1337 tings, highlighting typical behavior patterns—both strengths and failure modes—alongside concrete
 1338 examples.

1339 1. CORRECT ANSWER WITH FAITHFUL AND DIVERSE CITATIONS

1340
 1341 In ideal cases, the model not only produces a factually accurate and coherent answer, but also cites
 1342 multiple distinct documents, each supporting a different part of the response. This indicates that the
 1343 model is capable of recalling parametric knowledge and integrating diverse sources during generation.

1344 Example:

1345 **Question:** *In 2012 the remains of King Richard III were found in a carpark in Leicester. How can
 1346 scientists be so certain that they’ve found him?*

1347 **Model Answer:**

1350
 1351 Scientists identified King Richard III through a combination of archaeological evidence and
 1352 forensic analysis.
 1353 <|New Scientist Confirms Richard III Skeleton Find with
 1354 99.9994% Probability|>
 1355 ...a skeleton under a council car park...
 1356 <|Richard III Burial Site: Uncovering the Significance of
 1357 the Greyfriars Church Choir|>
 1358 ...scoliosis in the spine, consistent with historical descriptions...
 1359 <|Richard III of England|>
 1360 ...head wounds matched reports of his death...
 1361 <|Leicester Car Park Sighting: DNA Tests Confirm Richard
 1362 III's Remains Found...|>
 1363 ...DNA matched two living relatives...
 1364 <|New Scientist Confirms Richard III Skeleton Find with
 1365 99.9994% Probability|>
 Citation Precision

1366 **Correctness Score: 0.667 Citation Score: 1.0**

1367
 1368
 1369 The answer addresses the question with high fidelity, and each citation grounds a different factual
 1370 element. This demonstrates the model's ability to attribute evidence at a fine-grained level across
 1371 documents.

1372 2. CORRECT ANSWER BUT INCORRECT CITATIONS

1373
 1374 In some cases, the generated answer is factually correct and well-structured, but the cited documents
 1375 are irrelevant. This suggests a mismatch between content planning and citation generation.

1376
 1377 **Example:**

1378 **Question:** *Why do online communities crumble as they gain popularity?*

1379 **Model Answer:**

1380
 1381 Communities may lose cohesion as they scale.
 1382 <|Understanding the Far-Right in Canada: A
 1383 Montreal-Antifasciste Perspective|>
 1384 ...larger groups face moderation and trust challenges...
 1385 <|Understanding the Far-Right in Canada: A
 1386 Montreal-Antifasciste Perspective|>
 1387 ...trolls and bad actors can increase toxicity...
 1388 <|How Reddit's New Approach Tackles Hate Speech without
 1389 Banning Communities | WIRED|>
 1390 ...member attrition often follows hostile interactions...
 1391 <|Navigating Online Privacy: Protecting Your Rights in the
 1392 Digital Age|>
 1393 ...balancing inclusivity with safety becomes harder...
 1394 <|Understanding the Far-Right in Canada: A
 1395 Montreal-Antifasciste Perspective|>

1396 **Correctness Score: 1.0 Citation Score: 0.0**

1397
 1398 The answer is broadly accurate and well-articulated, but the cited documents bear no meaningful
 1399 relation to the explanation. The model appears to attribute wrongly.

1400 1401 3. FAITHFUL CITATIONS BUT INCOMPLETE ANSWER

1402
 1403 Sometimes, the model successfully grounds all claims in real documents, but the final answer fails to
 directly address the question.

1404 **Example:**

1405 **Question:** *Why do so many drugs cause “dry mouth” as a side effect?*

1406 **Model Answer:**

1408 Many drugs cause xerostomia, or dry mouth.
1409 <|Understanding Medication Side Effects: The Prevalence of
1410 Dry Mouth in Clinical Trials|>
1411 ...can affect comfort and oral function...
1412 <|Understanding Medication Side Effects: The Prevalence of
1413 Dry Mouth in Clinical Trials|>
1414 ...common in antihistamines and antidepressants...
1415 <|Understanding Medication Side Effects: The Prevalence of
1416 Dry Mouth in Clinical Trials|>
1417 ...dosage changes may help manage it...
1418 <|Understanding Medication Side Effects: The Prevalence of
1419 Dry Mouth in Clinical Trials|>
1420 ...

1421 **Correctness Score:** 0.0 **Citation Score:** 1.0

1423 The citations are accurate, but the answer never actually addresses why these drugs cause dry mouth,
1424 indicating a reasoning gap.

1426 **4. “TITLE LURE” ERRORS IN SHORT-FORM QA**

1428 In short-form QA tasks, the model sometimes selects citations solely based on title relevance,
1429 even when the document content lacks the required evidence. This reflects a superficial attribute
1430 mechanism.

1432 **Example:**

1433 **Question:** *How is Boston addressing the digital divide in terms of communications technology from
1434 December 2023? Model Answer:*

1436 Boston is providing subsidized smartphones and free digital literacy workshops.
1437 <|Integrative Technologies in Smart City Infrastructure|>

1440 **Ground-truth citation:** *Boston’s Infrastructure Development Initiatives: Paving the Way Forward*

1441 While the cited title seems topically aligned, it does not support the factual claim. This failure mode
1442 could mislead users who assume correctness based on citation phrasing.

1444 **5. CROSS-DOMAIN LOOKALIKES AND “NEAR MISS” CITATIONS**

1446 Occasionally, the model cites from a mismatched domain—e.g., a general Wikipedia article instead
1447 of a domain-specific source like RepliQA—producing citations that superficially resemble the ground
1448 truth but lack factual alignment.

1449 **Example:**

1450 **Question:** *When was the last game of Copenhagen’s basketball season played in 2023? Model
1451 Answer:*

1454 The final game was played on 19 May 2023.
1455 <|Copenhagen|>

1456 **Ground-truth citation:** *Intramural Sports: A Growing Trend in Copenhagen*

1458 The Wikipedia article Copenhagen is a superficial match but lacks the actual fact. The model retrieves
 1459 the wrong domain even while giving the correct answer—possibly from memory.
 1460

1461

1462 H ACTIVE INDEXING

1463

1464 Here we show the details of active indexing.
 1465

1466

1467 H.1 FORWARD

1468

1469 **Step 1: Entity Extractions** we first extract entities from each document, with the prompt:
 1470

1471

Entities Extraction

1472

1473 You will be given a document. Your task is to extract important entities mentioned in the text.

1474 Entities include names of people, organizations, locations, dates, and other identifiable items. Use the

1475 categories below as a guide:

- 1476 • People & Organizations – Person, Organization
- 1477 • Locations – Country, City, Facility, Region
- 1478 • Temporal – Date, Time
- 1479 • Events – Historical or notable events
- 1480 • Objects – Products, Works of art, Laws/Policies
- 1481 • Concepts – Theories, Fields, Ideologies
- 1482 • Quantities – Numbers, Money, Rankings
- 1483 • Biological/Chemical – Species, Compounds
- 1484 • Other – Named documents, Tasks, Technologies

1485

Document: [document]

1486

Only return the 20 most important entities based on their relevance to the main topics of the document, where each entity is separated by a newline. In your output, only return the important entities themselves, and do not return any other information like their categories or types.

1487

1488

1489

1490

Step 2: Forward Data Augmentation Then we utilize an LLM (Qwen-2.5-7B trained by the seed data generated from GPT-4.1-mini) to generate relevant questions to each entity conditioned on the document.

1491

1492

1493

Forward Augmentation

1494

1495

1496

You will be given a document, its title, and an entity from the document. Your task is to generate detailed questions that explore the relationship between a given entity and the document. Specifically, ask how, what, when, where, why, or if the entity is related to the content of the document. In each of your questions, you should include the entity and the document title. The questions should require a detailed answer.

1497

1498

1499

For each question you create, provide a detailed, elaborated answer that explains the relationship between the entity and the document. The answer should be based on the content of the document and should not include any external information.

1500

Title: [title] Document: [document]

1501

Entity: [entity]

1502

1503

1504

1505

1506

1507

1508

1509

1510

Step 3: Clean up and adjust titles : When LLMs generate questions, the document identifiers may omit parts of the original titles. To correct this, we apply heuristics to locate and replace them with the corresponding titles from the corpus. We then mark the document identifiers in the questions with special tokens.

1511

We extract up to 10 entities per document, resulting in approximately 1.28B tokens—3.3× the size of the original corpus.

1512

H.2 BACKWARD

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

Step 1: Retrieval We construct both sparse (BM25) and dense indexes (Lin et al., 2021) using Pyserini⁴. For each document chunk, we retrieve top-200 relevant chunks using sparse and hybrid retrieval. From these, we select the top 10 chunks from distinct documents, then randomly sample 1–3 to form a diverse chunk cluster of size 2–4.

1530

1531

1532

1533

Step 2: Generation We use GPT-4.1-mini and a fine-tuned Qwen-2.5-3B to generate cross-document instruction-response pairs from the retrieved clusters. Qwen-2.5-3B is trained on GPT-4.1-mini outputs to reduce generation costs. The input prompt format is detailed below.

1534

1535

1536

1537

Step 3: Filtering LLM-generated document identifiers may contain noisy patterns (e.g., “document: xx”, “title: xx”, or generic placeholders like “document 1”). We detect such cases using heuristics and discard them, which filters out approximately 4.9% of the data. And then we replace the <source> marker with the special tokens.

As a result, we generate around 1.47B tokens (3.8x) for main experiments.

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

Backward Augmentation

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

You will be provided with multiple documents. Your task is to construct a self-contained instruction-answer pair that requires a language model to synthesize information from two or more of these documents in order to generate the correct answer.

The instruction should:

- Be clear, specific, and fully self-contained, so it can be understood without access to or mention of the original documents.
- Prompt the model to integrate, compare, or reason across multiple sources of information.
- Avoid phrases like “in the provided documents”, “based on the above”, or anything that references the existence of documents.

The answer must:

- Be derived by combining or reconciling information from multiple documents.
- Attribute every factual claim using the format <source>The title of the document</source> to indicate where the information came from.

[documents]

Return the instruction in the first paragraph and the answer in the following paragraphs.

H.3 GENERATOR MODELS CAPACITY ON DATA QUALITY

We investigate whether the effectiveness of Active Indexing depends on the choice of data generator. For backward augmentation, we compare data generated by GPT-4.1-mini versus a 3B model trained by GPT-4.1-mini. Experiments are conducted on RepliQA using LLaMA-3-8B. Citation precision drops from 30.6 to 25.3 when switching from GPT-4.1-mini to the tuned Qwen-2.5-3B generator, despite using more data (34M vs. 40M tokens). Nonetheless, this still significantly outperforms the passive indexing baseline (2.6 citation precision).

⁴<https://github.com/castorini/pyserini>

1566 H.4 EXAMPLE ON FORWARD AUGMENTED DATA
1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

Forward Augmentation Example

1582

Jack Arnold (director)

1583

1584

Question 1: How is the television series "Alias Smith and Jones" connected to Jack Arnold's career as described in the document titled <source>Jack Arnold (director)</source>?

1585

1586

1587

1588

1589

1590

1591

Answer: According to the document <source>Jack Arnold (director)</source> Jack Arnold directed several episodes of the television series "Alias Smith and Jones" during his television career. This indicates that Arnold was involved in the production of television shows beyond his work in film, including directing episodes of this series that aired in the 1950s. The document highlights Arnold's versatility and his ability to work in different media, including television, alongside his filmmaking career.

1592

1593

1594

Question 2: What role did Jack Arnold play in the television series "Alias Smith and Jones," and how does this fit into his overall career as detailed in the document titled <source>Jack Arnold (director)</source>?

1595

1596

1597

1598

1599

1600

1601

Answer: The document does not specify the exact role that Jack Arnold played in the television series "Alias Smith and Jones." However, it notes that Arnold directed multiple episodes of this series, which falls under his television work. This placement suggests that Arnold was a director involved in various television productions, contributing to the overall television landscape during the 1950s. His involvement in "Alias Smith and Jones" fits into his broader role as a director who transitioned from film to television, showcasing his versatility and the variety of projects he undertook throughout his career.

1602

1603

1604

Question 3: When did Jack Arnold begin directing episodes of "Alias Smith and Jones," and what significance does this timeline have in the context of his career as presented in the document <source>Jack Arnold (director)</source>?

1605

1606

1607

1608

1609

1610

1611

Answer: The document does not provide a specific year for when Jack Arnold began directing episodes of "Alias Smith and Jones." However, it places Arnold's work on television, including directing episodes of "Alias Smith and Jones," within his television career, which is noted to have begun in 1955. This timeline situates Arnold's television directing alongside his earlier work on "Science Fiction Theater," "Perry Mason," and other productions, indicating a steady progression into television during the 1950s. This placement highlights the evolution of Arnold's career from film into television as a prominent director.

1612

1613

Question 4: Why is "Alias Smith and Jones" important in the context of Jack Arnold's work as discussed in the document titled <source>Jack Arnold (director)</source>?

1614

1615

1616

1617

1618

1619

Answer: "Alias Smith and Jones" is important in the context of Jack Arnold's work because it represents one of the television productions he directed, adding to his diverse portfolio of work that includes both film and television. The document emphasizes Arnold's role as a director who worked across genres and mediums, and "Alias Smith and Jones" is part of the evidence of his involvement in television series during the 1950s. This indicates that Arnold was not confined to film but also contributed significantly to television, making him a multifaceted director. ...

1620
1621

H.5 EXAMPLE ON BACKWARD AUGMENTED DATA

1622
1623

Backward Augmentation Example

1624
1625

Instruction: Explain how the molecular weight and chemical modification of polypropylene influence its compatibility and processing in polymer-clay and polymer-fiber nanocomposites. In your explanation, discuss the role of maleic anhydride grafting and other grafted copolymers as compatibilizers, how molecular weight parameters are characterized, and how these factors affect composite preparation methods and resulting material properties.

1626
1627

Answer: The molecular weight of polypropylene and its chemical modification significantly affect its compatibility and processing behavior in both polymer-clay and polymer-fiber nanocomposites. Maleic anhydride grafted polypropylene (PP-g-MAH), which contains a small amount of grafted maleic anhydride (e.g., 0.29 wt% MAH), is prepared via melt grafting and is commonly used as a compatibilizer to enhance interaction between polypropylene and fillers such as montmorillonite clay. The PP-g-MAH/clay nanocomposites are prepared by melt compounding at temperatures around 190–200 °C, with clay loading around 3 wt% to promote good dispersion and intercalation within the polymer matrix *<source>Study on nonisothermal crystallization of maleic anhydride grafted polypropylene/montmorillonite nanocomposite</source>*.

1628
1629
1630
1631
1632
1633
1634
1635
1636
1637

Characterization of molecular weight in polymers like polypropylene is essential for understanding how chain length affects viscosity and processing. Molecular weight parameters such as weight-average molecular weight (M_w), number-average molecular weight (M_n), and intrinsic viscosity are typically determined using techniques such as size-exclusion chromatography (SEC) and viscometry in suitable solvents. Studies on poly(1,4-butylene adipate) demonstrate that intrinsic viscosity follows a Mark-Houwink-Sakurada (MHS) type relation with molecular weight, which varies depending on molecular weight distribution breadth. Though this example is for PBA, similar principles apply to polypropylene regarding the relationship between molecular weight and flow properties *<source>Relationship of intrinsic viscosity to molecular weight for poly (1, 4-butylene adipate)</source>*.

1638
1639
1640
1641
1642
1643
1644
1645
1646

In addition to maleic anhydride grafting, other grafted copolymers such as PP-g-GMA (maleic anhydride replaced by glycidyl methacrylate groups) and SEBS-g-GMA are employed as compatibilizers to improve interfacial adhesion in polypropylene composites with natural fibers like hemp. These grafted compatibilizers are prepared by melt radical grafting and facilitate better bonding between hydrophobic polypropylene and hydrophilic fibers, often used alongside radical initiators in melt mixing processes at temperatures around 170–200 °C *<source>Functionalization, compatibilization and properties of polypropylene composites with Hemp fibres</source>*.

1647
1648
1649
1650
1651
1652
1653

Overall, molecular weight influences processability by affecting polymer viscosity and chain entanglements, impacting the dispersion of fillers and the mechanical performance of the composite. Chemical modification through grafting functionalities like maleic anhydride or glycidyl methacrylate introduces reactive sites or compatibilizing groups that improve the interfacial adhesion between polypropylene and both inorganic fillers (like montmorillonite clay) and organic fibers (like hemp). These compatibilizers enable the use of melt mixing techniques at elevated temperatures to produce nanocomposites or fiber composites with improved structural and functional properties *<source>Study on nonisothermal crystallization of maleic anhydride grafted polypropylene/montmorillonite nanocomposite</source>*, *<source>Functionalization, compatibilization and properties of polypropylene composites with Hemp fibres</source>*.

1662
1663
1664
1665
1666
1667

I LIMITATIONS AND FUTURE WORK

I.1 LIMITATION

1668
1669

Scalability: Model Size, Corpus Coverage, and Augmentation Budget Our setup offers a computationally manageable framework for academic research, but it operates at a much smaller scale than production-grade pretraining, which often involves trillions of tokens across vastly more diverse domains than those included in our study. We observe a large performance gap between 3B and 7B models, yet it remains uncertain how these trends evolve at larger scales (e.g., 100B). It is an open question whether Active Indexing continues to yield gains, plateaus, or even regresses as model capacity increases. Our data augmentation budget is limited in the main experiments. However, our

1674 scaling analysis shows consistent improvement even at 16x the original data size, suggesting that the
 1675 method could further benefit from larger augmentation.
 1676

1677 **Synthetic Data Quality** LLM-generated QA pairs may introduce hallucinations into the augmented
 1678 data, which can lead to subtle degradation in model behavior. While effective in aggregate, this
 1679 approach could benefit from future work on hallucination detection, filtering, or confidence-aware
 1680 generation strategies.
 1681

1682 **Evaluation Limitations** Our evaluation of long-form citation relies on NLI models to judge claim
 1683 support, introducing dependence on their accuracy and coverage. Although this provides a scalable
 1684 proxy, it adds noise to the measurement and may miss nuanced cases. Incorporating human evaluation
 1685 would strengthen the reliability of results, particularly for ambiguous or multi-hop claims.
 1686

1687 **I.2 FUTURE WORK**

1688 **Multilingual and Domain-Specific Attribution** Our experiments are limited to English and
 1689 general-domain corpora. Extending Active Indexing to multilingual settings and high-stakes do-
 1690 mains—such as law, medicine, or finance—poses unique challenges. These domains often require
 1691 precise terminology, complex reasoning, and domain-specific citation standards. Future work could
 1692 develop tailored QA generation methods and identifier formats for these settings, and perform in-depth
 1693 evaluations of citation fidelity and safety in domain-critical applications.
 1694

1695 **Enhancing the existing bindings between facts and their identifiers** While our methods focus
 1696 on building new bindings between facts and document identifiers, existing pre-trained models may
 1697 already encode implicit associations between facts and surface features such as titles, as we observed
 1698 in the Wikipedia corpus. Beyond titles, there may be other weak or “loose” associations already
 1699 present in the model, which could potentially be strengthened through better prompting strategies or
 1700 further post-training methods such as reinforcement learning.
 1701

1702 **Scaling Laws and Saturation Points** We observe consistent improvements with increased aug-
 1703mentation and model size up to 7B, but it remains unclear when and whether gains saturate. With
 1704 more computational resources, future work can extend scaling curves to larger models (e.g., 14B,
 1705 32B, 70B+) and higher augmentation regimes (e.g., 32x, 64x). This would help identify optimal
 1706 compute-utility tradeoffs and determine whether benefits of Active Indexing persist at frontier scale.
 1707

1708 **Complementarity with Retrieval-Augmented Generation (RAG)** Internal citation and external
 1709 retrieval are complementary: the former leverages memorized knowledge, while the latter provides
 1710 up-to-date or unseen information. A promising direction is confidence-aware hybrid systems—where
 1711 the model cites from internal memory when confident, but falls back to retrieval when uncertain.
 1712 Exploring how Active Indexing can be integrated into such hybrid systems may yield the best of both
 1713 worlds: low-latency and high-coverage citation.
 1714

1715 **Privacy-Preserving Attribution** Enabling internal citation increases the model’s tendency to
 1716 surface memorized content, which may include sensitive or proprietary information. Investigating
 1717 whether attribution behavior exacerbates privacy risks is an important open question. Future work
 1718 could explore mitigation strategies, such as differential privacy, selective redaction of identifiers, or
 1719 training-time filtering, to balance attribution fidelity with privacy preservation.
 1720

1721 **Human-Centered Evaluation and Interpretability** While our current evaluation pipeline is
 1722 largely automatic, the real-world utility of citations depends on user trust and interpretability. Future
 1723 work could conduct human studies to assess how internal citations affect perceived credibility,
 1724 transparency, and user trust—particularly in comparison to RAG or non-citing models. Incorporating
 1725 explanations of why a citation was chosen (e.g., via rationales) could also improve interpretability
 1726 and debuggability.
 1727

1728 **J USE OF LLMs**
17291730 We used LLMs to assist with writing. Specifically, we employed GPT-5 thinking, GPT-5 and GPT-4o
1731 to rephrase paragraphs for grammatical correctness and improved flow. We also used them to shorten
1732 text, making descriptions more concise and easier to read. All LLM-generated text was reviewed,
1733 edited, and approved by the human authors.

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781