UNDERSTANDING SYNTHETIC CONTEXT EXTENSION VIA RETRIEVAL HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context LLMs are increasingly in demand for applications such as retrieval-augmented generation. To defray the cost of pretraining LLMs over long contexts, recent work takes an approach of synthetic context extension: fine-tuning LLMs with synthetically generated long-context data in a post-training stage. However, it remains unclear how and why this synthetic context extension imparts abilities for downstream long-context tasks. In this paper, we investigate fine-tuning on synthetic data for three long-context tasks that require retrieval and reasoning. We vary the realism of "needle" concepts to be retrieved and diversity of the surrounding "haystack" context, from using LLMs to construct synthetic documents to using templated relations and creating symbolic datasets. We find that models trained on synthetic data fall short of the real data, but surprisingly, the mismatch can be interpreted and even predicted in terms of a special set of attention heads that are responsible for retrieval over long context, *retrieval heads* (Wu et al., 2024). The retrieval heads learned on synthetic data have high overlap with retrieval heads learned on real data, and there is a strong correlation between the recall of heads learned and the downstream performance of a model. Furthermore, with attention knockout and activation patching, we mechanistically show that retrieval heads are necessary and explain model performance, although they are not totally sufficient. Our results shed light on how to interpret synthetic data fine-tuning performance and how to approach creating better data for learning real-world capabilities over long contexts.

1 INTRODUCTION

The quadratic memory requirement of Transformer attention imposes a strong computational constraint on our ability to train and do inference on long-context models. This disrupts the typical pre-training pipeline: pre-training must be done at as large a scale as possible, but pre-training a long context model would necessarily reduce the number of observed tokens able to fit on the GPU. One solution for this is to rely on synthetic data, now common in post-training settings such as SFT (Xu et al., 2023b; Yue et al., 2024; Xu et al., 2024; Che, 2024) and RLHF/DPO (Yang et al., 2023). Recent prior work has proposed using synthetic data to extend the long-context abilities of LLMs after pre-training (Xiong et al., 2024; Zhao et al., 2024).

This use of synthetic data is particularly necessary for long context tasks since they are so laborious for humans to manually label. Synthetic data is also configurable: it can exhibit different reasoning skills and "teach" models have to make certain types of inferences (Du et al., 2017; Yu et al., 2018; Agarwal et al., 2021; Tang et al., 2024; Divekar & Durrett, 2024). One way to do this is using templates to express pieces of information that must be reasoned over and to create symbolic tasks that are thought to mirror the reasoning required in the real task (Hsieh et al., 2024; Prakash et al., 2024; Saparov & He, 2023; Li et al., 2024). However, past work has shown varying results from training on data for this kind of context scaling (Fu et al., 2024); we lack general understanding of what is needed here.

In this paper, we explore several methods of creating synthetic long context data across three tasks. Our goal is to examine what makes synthetic data effective for this kind of context scaling. While more realistic data is often better, it is unreliable–certain types of more synthetic data can exhibit

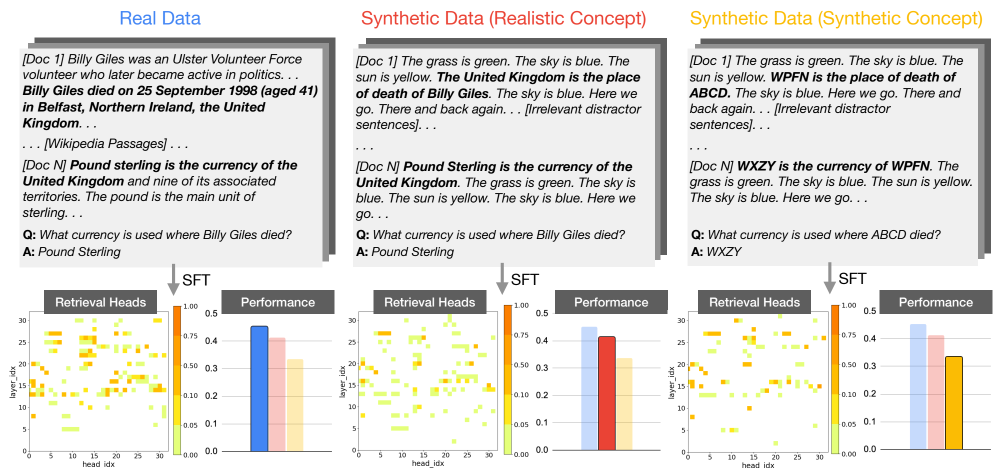

Figure 1: We explore synthetic context extension with different forms of synthetic data across multiple tasks. Examples for a two-hop question from MuSiQue (Trivedi et al., 2022) are shown here. A special set of attention heads, *retrieval heads* (Wu et al., 2024), help explain the performance gap between fine-tuning on real data and synthetic data.

desired long-context patterns even more effectively and with fewer shortcuts than realistic data. However, other types of synthetic data severely underperform on these tasks.

How can we understand this divergence? We analyze models trained on long-context data for the presence of a phenomenon called retrieval heads (Wu et al., 2024) as a indicator of the subnetworks affected during fine-tuning. Figure 1 shows two surprising results. First, the retrieval heads learned on poor-performing synthetic data tend to be fewer than those learned on realistic or high-quality synthetic data. Second, we find that the similarity between the retrieval heads learned on synthetic data and realistic data correlates strongly with the downstream performance. Learning a certain set of retrieval heads seems to be a necessary condition for high performance, as we show with intervention experiments. However, it is not sufficient. We show that patching heads at the intersection of a poor-performing model and a high-performing model can improve performance of the former: these heads are where important operations are happening, but realistic data teaches them more strongly.

Our contributions are: (1) analysis of synthetic data across three synthetic tasks for long-context LLM training to determine best practices; (2) experimental validation establishing that retrieval heads are a key component whose appearance during training correlates with effectiveness of the training data for this setting. Taken together, we believe this work indicates a path forward for how to engineer better synthetic data and how to connect the construction process of synthetic data to what it teaches Transformers and how those models perform on downstream tasks.

2 Background and Setup

2.1 Background: Synthetic Data for Training Language Models

Formally, consider a supervised learning setting for a pre-trained transformer language model $\mathcal{M}$. Given a task $\mathcal{T}$, we assume a distribution $p_{\mathcal{T}}$ of real-world task instances. We assume that a small, limited set of input-label pairs $\mathcal{D}_{\mathcal{T}} = (x_{\mathcal{T}}, y_{\mathcal{T}})$ drawn from the distribution $p_{\mathcal{T}}$ is available as seed data. A synthetic dataset $\tilde{\mathcal{D}}_{\mathcal{T}}$ is a set of input-label pairs sampled from the outputs of a data generator $\mathcal{G}$ given the seed data or the known properties: $\tilde{\mathcal{D}}_{\mathcal{T}} \sim p((\tilde{x}, \tilde{y}) \mid \mathcal{D}_{\mathcal{T}})$. Benchmarking or training $\mathcal{M}$ on a synthetic dataset that can be used to represent properties of the real dataset is expected to evaluate or teach $\mathcal{M}$ the capabilities that can be *transferred* to the real-world distribution $p_{\mathcal{T}}$.

A recent line of work has shown that training short-context LLMs on simple heuristic-based synthetic datasets can achieve surprisingly transferability on *context extension*, a post-training scenario where LLMs that have been pre-trained on short-context corpora are further trained on long-context tasks to extend the effective context window (Fu et al., 2024; Zhao et al., 2024; Xiong et al., 2024). For example, Xiong et al. (2024) finds that fine-tuning on a synthetic simple dictionary key-value retrieval task can even outperform models fine-tuned on realistic in-domain data.

We call these approaches **synthetic context extension**: using synthetic data to extend the context window of LLMs. It remains unclear how and why synthetic data, especially when drawn from a very different distribution from the real data, can be effective despite results that support the contrary (Chen et al., 2024; Liu et al., 2024b). There is also a lack of general principles for creating synthetic data for training beyond dataset-specific constructions in the literature. We start by constructing synthetic datasets varying in systematic ways to unify these variants from the literature.

2.2 Experimental setup

Following Xiong et al. (2024), we focus on fine-tuning LLMs for long-context retrieval and reasoning tasks where training on high-quality synthetic data has been shown to outperform real data. We also extend to multi-hop settings. We experiment on three datasets where, given a long context $\mathcal{C}$ and a context-based query q, a language model $\mathcal{M}$ needs to retrieve one or more "needles concepts" $f_1, \ldots, f_m$ from $\mathcal{C}$ (pieces of relevant information), reason over that information, and then generate a response $\tilde{\mathbf{y}} \sim p(y \mid \mathcal{C}, q)$ where $p(y \mid \mathcal{C}, q)$ is the conditional distribution that M places over the vocabulary Σ^* given the context and the query. We consider extending the context window from 8K to 32K tokens to be representative of synthetic context extension following Chen et al. (2023). Specifically, we use the following three datasets.

MDQA (Liu et al., 2024a): MDQA is a multi-document question answering (QA) dataset where only one paragraph in C contains the gold answer to a single-hop query; that is, there is a single f which directly addresses q. We extend the original MDQA dataset in 4K context to 32K context by retrieving additional distractor paragraphs from Natural Questions-Open (Kwiatkowski et al., 2019; Lee et al., 2019) with Contriever (Izacard et al., 2021).

MuSiQue (Trivedi et al., 2022): MuSiQue is a multi-hop QA dataset where the model must identify a piece of relevant information from a different document for each hop of the question in order to retrieve the final correct answer from the context. We use the linear three-hop subset of MuSiQue and extend the dataset to 32K by adding padding paragraphs [1] to the original context. In this setting, the facts f_1, f_2, f_3 are natural language sentences containing knowledge graph relations.[2]

SummHay Citation (Laban et al., 2024): Summary of a Haystack (SummHay) is a long-context retrieval dataset where the model is given a set of documents with controlled "insights," and asked to produce a list of key points. Additionally, the model must cite the correct documents in support of each key point. We isolate the citation component and construct a task where, given a haystack of 10 documents and a key point ("insight"), the model must correctly identify the two documents that support the point and their associated document IDs. The two facts f_1, f_2 may span multiple sentences and may be substantially paraphrased versions of the insight.

Training Configuration For each task, we fine-tune two short-context LLMs, Llama-3-8B-Instruct (Dubey et al., 2024) and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023). Prior work indicates that attention heads are largely responsible for implementing algorithms (Olsson et al., 2022) and using information *within the context* (Stolfo et al., 2023; Lieberum et al., 2023) while MLP layers are responsible for parametric knowledge (Geva et al., 2021). In addition, when adapting to long contexts, the attention heads in particular must handle new position embeddings and softmax over more context tokens. Therefore, we fine-tune attention heads only. [3]

To extend models from their original 8K pretrained context length to 32K, we follow Gradient (2024) in calculating new RoPE (Su et al., 2024) theta values, using 6315088 for Llama-3-8B-Instruct and 59300 for Mistral-7B-Instruct-v0.1. We scale the sliding window accordingly for Mistral-7B-Instruct-v0.1 to 16k context. These are the only adjustments we make to the models, following Fu et al. (2024). Our hyperparameters and hardware setup can be found in Appendix C.1.

[1] We pad with irrelevant repeated text "The grass is green. The sky is blue..." to ensure that the added paragraphs do not interfere with the answer to the original question.

[2] Note that this is different from the demonstrative two-hop examples in Figure 1.

[3] We find similar conclusions when fine-tuning all Llama-3-8B-Instruct modules, see Appendix G.

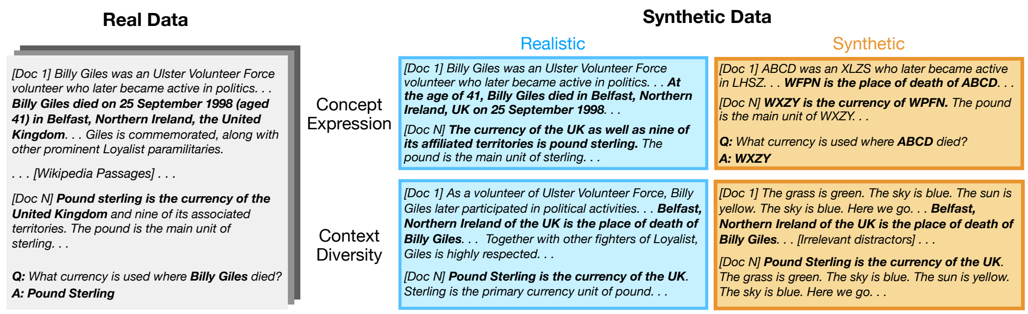

Figure 2: Examples of elements of synthetic datasets for MuSiQue with varying levels of *concept expression* and *context diversity*. The needle sentences f_i in the context and the entities in them are **bold**. High concept expression means more realistic expression of the needle f_i, and low expression means more synthetic, including replacing real entities with symbolic entities or transforming f_i into templated sentences. High context diversity means more realistic context surrounding the needles, and low means more synthetic contexts such as repeated, irrelevant padding sentences

3 Synthetic Datasets

3.1 Principles Under Consideration

To create a representative range of synthetic data for each task, we partition the input text $\mathcal{C}$ into (A) text containing relevant information $\{f_1, \ldots, f_m\}$ ("needle **concepts**") and (B) the surrounding **context** $C \backslash \{f_1, \ldots, f_m\}$. This allows us to categorize any task as having a variant of *concept expression* (how the target information f_i is expressed) and *context diversity* (the naturalness and relevance of the surrounding information). In the following paragraphs, we discuss common variants found in synthetic data literature. We single out and emphasize a highly structured variant of concept and context–*symbolic tasks*–for being devoid of natural language yet noted to transfer to realistic tasks (Xiong et al., 2024).

Concept Expression A common procedure for creating synthetic data involves exploiting task asymmetry (Josifoski et al., 2023; Xu et al., 2023a; Lu et al., 2024; Chandradevan et al., 2024; Chaudhary et al., 2024; Tang et al., 2024), where asking an LLM to generate natural language data based off of a label (e.g. a sentence based off of a knowledge triple) is easier than predicting the answer from text of the same complexity and domain. In this scenario, the LLM is asked to create diverse "needle" target concept expressions f_i. In task specific cases, it is beneficial to make this data less realistic while encouraging generalization. For example, prior synthetic datasets have made use of fictional entities (Saparov & He, 2023) or nonsense phrases (Wei et al., 2023) in place of real entities and properties, or swapped out nouns to augment the dataset (Lu et al., 2024) and prevent overfitting to specific entities. In long context benchmarks (Hsieh et al., 2024; Li et al., 2024), it is common to express the needle concepts in short, templated sentences.

Context Diversity We can also vary the expression of $C \backslash \{f_1, \ldots, f_m\}$, the "haystack." This ranges from distractor needles which may have the same form (template) as the target concept to padding with repeated sentences. We use the repeated set of sentences "The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again." as our low-diversity padding to compare with context that is synthetically generated by an LLM, following Hsieh et al. (2024) and Mohtashami & Jaggi (2023).

Symbolic Tasks We also experiment with purely symbolic (involving dictionary key-value or list retrieval) versions of our real tasks, since such tasks are believed to recruit similar model abilities as their natural language counterparts. For example, prior work has indicated that pre-training on code helps on Entity Tracking (Prakash et al., 2024) and that fine-tuning on a symbolic dictionary key-value retrieval task can provide greater benefits than even real data (Xiong et al., 2024). Additionally, RULER (Hsieh et al., 2024) introduced a variable assignment task for long-context value tracking.

This latter task features expressions like "*VAR X1 = 12345 ...... VAR Y1 = 54321 ...... Find all variables that are assigned the value 12345.*" that do not contain meaningful natural language, hence why we differentiate this category from natural language synthetic data.

3.2 Synthetic Dataset Construction

For each of the long-context tasks, we sample a set of examples $\mathcal{D}_\mathcal{T}$ from the training set and use the principles above to construct various synthetic datasets based on $\mathcal{D}_\mathcal{T}$. See Appendix A for the complete set of prompts used to create the synthetic data, and Appendix B for our training prompts.

MDQA Given a training example of MDQA training data $(C, q, y) \in \mathcal{D}_\mathcal{T}$, we combine the query q and answer y into our needle f that will be put into the context and that needs to be retrieved by the model. For f, we use two simplification levels of concept expression by (1) keeping the real entities in the query and answer (**high** expression), and (2) replacing the real entities with 4-character symbolic entities (**low** expression). We create the context surrounding the needle claim with two levels of context diversity: (1) prompting GPT-4o-mini to paraphrase the original context from MDQA training data (for the real entities), or generate a Wikipedia-style paragraph that elaborates on the claim (**high** diversity); (2) padding the context paragraph with repeated sentences (**low** diversity). The **symbolic** dataset is the simple dictionary key-value retrieval dataset from Xiong et al. (2024).

MuSiQue The f_i here are based on multi-hop knowledge graph relations. Like with MDQA, create two simplification levels of concept expression by (1) keeping the real entities in the query and answer (**high** expression), and (2) replacing the real entities with 4-character symbolic entities (**low** expression), and constructing f_i by prompting GPT-4o-mini to write sentences or via template. We create two levels of context diversity by (1) prompting GPT-4 to write a paragraph containing the fact (**high** diversity), and (2) padding each paragraph with repeated text (**low** diversity). The **symbolic** task, as demonstrated in Figure 5, consists of a list of dictionaries with 4-character identifier, keys and values. Queries are of the form "What is the PROPERTY_3 of the PROPERTY_2 of the PROPERTY_1 of DICTIONARY_1?". The answer is found by multi-hop traversal by accessing subsequent dictionary names associated with the specified properties.

SummHay Citation We derive the f_i from the insights in one of two ways. (1) We prompt GPT-4o-mini to rephrase the insights to create the query, and then prompt again to split rephrased insights into multiple sentences to place into the context (yielding multiple f_i per insight) (**high** expression); and (2) We prompt GPT-4o-mini to simplify the insights to create the query, and split each simplified insight into multiple sentences to place into the context (**low** expression). We create two levels of context diversity by (1) padding each document with distractor insights from the same topic, (**high** diversity) and (2) padding each document with repeated text (**low** diversity). The **symbolic** task, as demonstrated in Figure 5 consists of lists containing 180 random 4-character strings, where the query is a 4-character string that appears in two different lists.

3.3 Results

Table 1 shows the performance (F1 scores) of fine-tuning LLMs on different synthetic datasets on the given long-context tasks. We first note that across datasets, **fine-tuning on synthetic datasets still falls short compared with fine-tuning on real data**, indicating the complexity of the evaluated long-context tasks.[4] For instance, on MuSiQue and SummHay there is a 2-4% gap between the best synthetic data and real data on Llama 3, and on MDQA there is a much larger gap at 33%.

Careful construction of synthetic data can help close a lot of the gap by varying the level of concept expression and context diversity beyond the symbolic synthetic dataset. **However, the effective way of constructing synthetic data for training is very task-specific and can even be counter-intuitive**: there does not exist a single construction strategy that achieves the best performance across tasks, and sometimes a more "realistic" synthetic dataset can even underperform the more "synthetic" counterparts.

[4]Particularly on MDQA, we note that such observation is very different from the one in Xiong et al. (2024) that finds fine-tuning synthetic data to be more effective than real data. We note that the results of Xiong et al. (2024) are obtained on 4K context rather than 32K and the models are fine-tuned with fewer training examples.

Table 1: Performance (F1) of fine-tuning LLMs on different synthetic data for the long-context retrieval and reasoning tasks. A large gap exists between the most performant synthetic context extension strategy (**bold**) and fine-tuning on real data. While careful construction of synthetic data can help close the gap, there does not exist a task-agnostic general way of constructing synthetic datasets for extending LLMs' context window on long-context retrieval and reasoning tasks.

Concept Exp.	Context Div.	MDQA		MuSiQue		Concept Exp.	Context Div.	SummHay Cite	
		Llama3	Mistral	Llama3	Mistral			Llama3	Mistral
High	High	0.31	0.20	0.37	0.22	High	High	0.70	0.28
High	Low	0.41	0.23	**0.41**	**0.23**	High	Low	0.61	0.28
Low	High	**0.49**	**0.31**	0.29	0.21	Simplified	High	**0.79**	**0.38**
Low	Low	0.47	0.24	0.34	0.17	Simplified	Low	0.65	0.28
Symbolic	Symbolic	0.48	0.16	0.32	0.11	Symbolic	Symbolic	0.54	0.18
Real Data (Full)		0.83	0.64	0.45	0.20	Real Data (Full)		0.81	0.40
Real Data (Limited)		0.80	0.59	0.32	0.16				
Non-FT		0.45	0.12	0.22	0.03	Non-FT		0.40	0.07

These results show a complex picture of fine-tuning LLMs with synthetic data for long-context tasks: the downstream performance cannot be simply "predicted" by how the synthetic training dataset is constructed. To interpret the success and failure of synthetic data for training, a more fine-grained explanation is needed beyond some general, task-agnostic data construction desiderata.

4 Retrieval Heads are Necessary for Context Extension

One of the key features of our tasks is the need for retrieving needles f_i embedded in a long context. Work from the mechanistic interpretability literature has shown that some attention heads in pre-trained (Olsson et al., 2022; Lieberum et al., 2023) or fine-tuned (Panigrahi et al., 2023; Yin et al., 2024) transformers specialize in retrieving and synthesizing information from the context in principled ways. [5] Notably, recent work (Wu et al., 2024) indicates that there exists a special, intrinsic set of attention heads in pre-trained transformers that attend to relevant information f_i in long context $\mathcal{C}$ given a query q and copy it to the output $\tilde{\mathbf{y}}$. Wu et al. (2024) dub them as *retrieval heads*.

Given the nature of our task, we analyze these attention heads as a proxy for the subnetworks being recruited and learned during fine-tuning with synthetic data. Our core hypothesis is that we can attribute the performance of synthetic context extension to how well the models learn to adapt the attention heads relevant to retrieving and using information from long context, as indicated by the *retrieval scores* of attention heads. Building on prior work, we extend identification of retrieval heads to multi-hop settings in MuSiQue and SummHay Citation.

4.1 Detecting Retrieval Heads

Following Wu et al. (2024), we detect retrieval heads by computing retrieval scores. To compare across fine-tuned models, we consider any attention head with a positive retrieval score to be a *retrieval head*, and later compute cosine similarity to account for the strength of scores. Given a fine-tuned model $\mathcal{M}^{'}$, we evaluate it on a dataset $\mathcal{D}^* = \{(\mathcal{C}^*, q^*, y^*)\}$ where the answer y^* needs to be identified from some needles f^* in $\mathcal{C}^*$ and copied to the model output $\tilde{y}^*$. When $\mathcal{M}^{'}$ generates an output token $w \in \tilde{y}^*$, we examine whether or not an attention head places the most attention probability mass on the same token in the answer span y^* in the context. If so, we consider the token w to be *retrieved* by the attention head. Given an evaluation example $(\mathcal{C}^*, q^*, y^*)$, let $G_h = \{w_h\}$ be the set of all tokens w that are *retrieved* by a head h during decoding. We define the retrieval score S_h for head h on a single example as:

$$S_h = \frac{|G_h \cap y^*|}{|y^*|} \tag{1}$$

Note that in the SummHay-citation task, the model is prompted to identify the numerical IDs of the documents (e.g. "[3]") that contain the given query insight q. In this case, we find it more useful to

[5] For example, Prakash et al. (2024) identifies a sparse set of heads that are responsible for retrieving and transmitting the positional information of objects from the context in the entity tracking task.

look at the attention heads that retrieve tokens from the insight needles f^* that contain information relevant to q rather than retrieving tokens from the answer y^*. Note that there are far more tokens in the correct insight needles f^* than in the answer y^* here. Thus, the **insight** score for a single example is is defined as:

$$S_h = \mathbb{1}\left[|G_h \cap f^*| > 0\right] \tag{2}$$

For each head, we average scores over all evaluation examples from D^* to yield the final score.

Given a long-context task $\mathcal{T}$, we detect a set of retrieval heads H_{real} of the models fine-tuned with real data $\mathcal{D}_{\mathcal{T}}$ on an evaluation set of *real* data . For each model $\mathcal{M}^{'}$ fine-tuned with synthetic data $\tilde{\mathcal{D}}_{\mathcal{T}}$, we detect a set of retrieval heads H_{synth} on an evaluation set of the corresponding *synthetic* data. H_{synth} reflects how synthetic context extension enables models to learn modules specialized in retrieving information from *synthetic* long-context data, and we will examine how this explains transferability to *real* long-context data.

Results We start with a case study of training Llama-3-8B-Instruct on synthetic data for MuSiQue, shown in Figure 1. Highlights show the retrieval score for each head at each layer. The model trained on the real data achieves an F1 score of 0.45 on the evaluation set, and has 129 attention heads which receive a positive retrieval score. Notably, the models trained on synthetic data (both realistic and symbolic) achieve lower F1 (0.41 and 0.33 respectively) while exhibiting far fewer retrieval-scoring attention heads (112 and 74 heads respectively). The real data retrieval heads have high recall (0.76 and 0.82) against the synthetic data heads, but not the other way around (0.66, 0.47), indicating when the synthetic data induces fewer retrieval heads, they tend to be subsets of the real attention heads (Appendix D, Table 6), although this relationship is weaker on MDQA and SummHay Citation. We present full retrieval head counts and pairwise recall results in Appendix D.

4.2 Connection with downstream performance

The presence of retrieval heads does not necessarily offer a concrete connection to downstream performance; we do not know that models are attending to long context using these heads, or whether these heads are correlated with other model capabilities. We conduct two experiments to elaborate on this: an intervention experiment where we mask out retrieval heads to see the impact on performance, and an observational experiment where we correlate the presence of retrieval heads with downstream performance across our different synthetic data variants.

Activation Masking We show that these attention heads are responsible for model performance on the real tasks by comparing activation masking on the top-k retrieval heads versus k random heads for various k as in Wu et al. (2024). Specifically, we select the top-k retrieval heads based on retrieval score, and zero out the outputs of those attention modules. As shown in Figure 3, masking even the top-10 retrieval heads causes a sharp drop in performance whereas masking 10 random heads over 3 repeated trials results in a marginal (< 0.05) or no drop in performance, with one exception (Llama-3-8B-Instruct on MDQA).

Synthetic Data Performance and Retrieval Heads As noted previously, when the synthetic data induces fewer retrieval heads, they tend to be subsets of those active on the real data. Following this for each synthetic dataset, we calculate the *recall* of non-zero scoring attention heads against the real dataset (first column of Tables 5-10 in Appendix D). As shown in Table 4, we find that this is strongly correlated with F1 on the real task for MuSiQue and SummHay Citation. This holds more strongly for Llama-3-8B-Instruct than for Mistral-7B-Instruct-v0.1.

To account for score magnitude, we examine the relationship between the cosine similarity of vectorized retrieval scores with downstream task performance [6], finding a strong relationship as shown in Figure 4. When synthetic data does not induce retrieval heads matching the real task, performance is low. However, high cosine similarity is not enough–at the same level of similarity, we still observe a wide range of performances.

[6] We find it effective to directly match attention heads by index even when models are fine-tuned on different datasets. Visualization in Appendix D supports this.

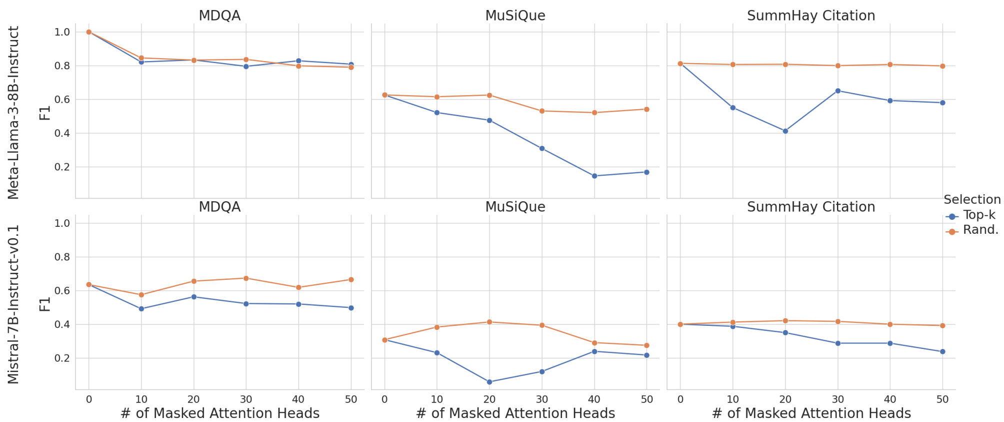

Figure 3: Top row: Llama-3-8B-Instruct. Bottom row: Mistral-7B-Instruct-v0.1. Effect of masking activations from attention heads with the top-k highest retrieval (MDQA, MuSiQue) or insight (SummHay Citation) scores. We compare with masking the same number of randomly chosen heads, averaged over 3 samples. Masking top-k attention heads consistently results in a larger drop in performance than masking random attention heads.

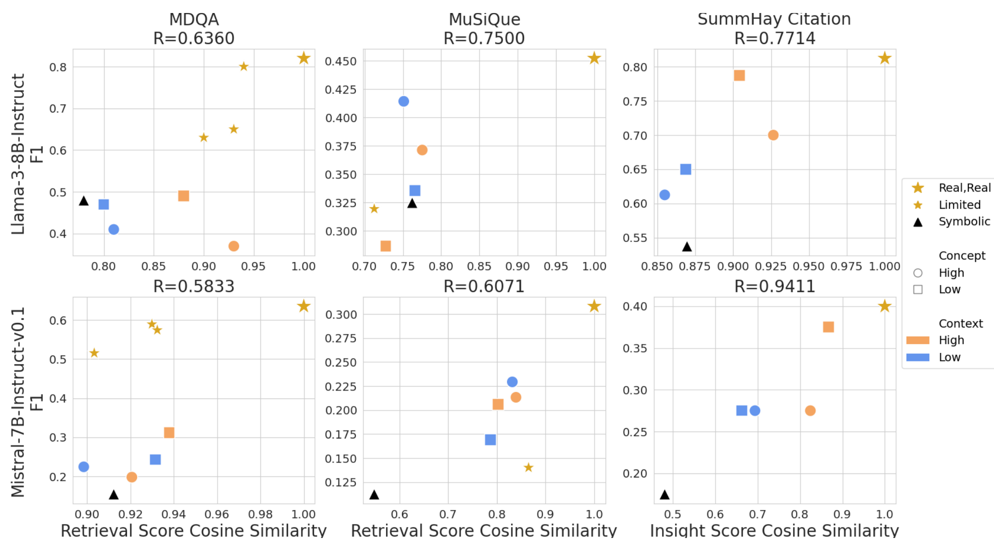

Figure 4: Cosine similarity between the retrieval scores on real datasets (R, R) vs. their synthetic versions, and Spearman correlation for each setting. We use multiple limited-relation datasets for MDQA, as described in Appendix C.

Table 2: Cosine similarity of real dataset retrieval scores (+ SummHay insight scores) across tasks.

	MDQA		MuSiQue		SummHay Retrieval		SummHay Insight	
	Llama3	Mistral	Llama 3	Mistral	Llama3	Mistral	Llama3	Mistral
MDQA	1.00	1.00	0.84	0.87	0.44	0.74	0.15	0.26
MuSiQue	0.84	0.87	1.00	1.00	0.59	0.69	0.28	0.20
SummHay Retrieval	0.44	0.74	0.59	0.69	1.00	1.00	0.08	0.07
SummHay Insight	0.15	0.26	0.28	0.20	0.08	0.07	1.00	1.00

4.3 Retrieval Heads Across Tasks

Table 3: Results on Llama-3-8B-Instruct after patching retrieval heads that comprise the complement and intersection between the real and synthetic data versions, compared to random retrieval heads and original performance. Best patch F1 is **bolded**, and Δ is the improvement over the original F1.

Task	Data Variant Concept	Data Variant Context	N	Compl.	Inter.	Rand.	Orig.	Δ
MDQA	Real	Real	-	-	-	-	0.82	-
	Real (Limited)	Real	68	0.82	**0.83**	0.82	0.80	0.03
	Low	High	61	0.65	**0.66**	0.54	0.49	0.17
	Symbolic	Symbolic	71	0.43	**0.73**	0.50	0.48	0.25
	Low	Low	74	0.70	**0.71**	0.53	0.47	0.24
	High	Low	74	0.63	0.51	**0.70**	0.41	0.29
	High	High	60	0.52	**0.59**	0.26	0.37	0.21
MuSiQue	Real	Real	-	-	-	-	0.45	-
	High	Low	71	0.38	**0.41**	0.33	0.41	0.00
	High	High	71	**0.33**	0.29	0.25	0.37	-0.05
	Low	Low	61	**0.41**	0.33	0.33	0.34	0.08
	Real (Limited)	Real	53	0.34	**0.34**	0.31	0.32	0.02
	Symbolic	Symbolic	55	0.33	**0.36**	0.35	0.32	0.03
SummHay	Real	Real	-	-	-	-	0.81	-
	Low	Low	27	0.73	**0.74**	0.73	0.79	-0.04
	High	High	19	0.72	0.75	**0.79**	0.70	0.09
	Low	Low	26	0.66	**0.70**	0.62	0.65	0.05
	High	Low	21	0.57	**0.68**	0.67	0.61	0.07
	Symbolic	Symbolic	26	0.53	**0.60**	0.48	0.54	0.07

We ask whether all tasks are leveraging the same set of retrieval heads. Table 2 shows cosine similarity of linearized retrieval scores between tasks. The single-hop and and multi-hop extractive QA tasks, MDQA and MuSiQue, have the highest cosine similarity (Llama-3-8B-Instruct: 0.84; Mistral-7B-Instruct-v0.1: 0.87). However, there is much lower similarity between the QA tasks and the SummHay Citation Retrieval Heads, and the *least* similarity with SummHay Insight Heads. [7] Comparing to Figure 4, we find that our real tasks have relatively high cosine similarity (> 0.66) with their synthetic versions, with the exception of the purely symbolic chained-dictionary-lookup and list-citation tasks. This suggests that there are task-specific subsets of retrieval heads, either activated based on reasoning ability or token diversity; we leave this for future investigation.

5 Retrieval Head Patching

Given datasets of the same conceptual reasoning and retrieval task, it is peculiar that fine-tuning on some datasets results in fewer retrieval heads. Do the attention heads common to all datasets better capture the core capability required for the task? For the common attention heads, do models learn a better way of updating them from the real data than the synthetic data? To investigate these, we follow Prakash et al. (2024) to perform cross-model activation patching of retrieval heads in the *intersection* and *complement* between the real dataset and the synthetic datasets. Specifically, given the set of retrieval scoring attention heads on the real data, H_{real}, and the set of retrieval scoring heads on a synthetic dataset, H_{synth}, we take the complement $H_{\text{compl}} = H_{\text{real}} \setminus H_{\text{synth}}$ and the intersection $H_{\text{inter}} = H_{\text{real}} \cap H_{\text{synth}}$. For a fair comparison, we sample $n_{\text{heads}} = \min(|H_{\text{compl}}|, |H_{\text{inter}}|)$ without replacement from both sets. Additionally we compare with n_{heads} randomly sampled attention heads. For each set, we patch activations from the model trained on the real data to the model trained on the synthetic data. Implementation details can be found in Appendix F.

Synthetic Data Affects Required Model Components Less Effectively Our results in Tables 3 and 11 show that patching *intersection* heads outperforms patching both random and complement

[7] SummHay Retrieval Heads attend to the final answer (document number), whereas SummHay Insight Heads attend to the insight text within the document.

heads. The improvement is the greatest for synthetic tasks with the lowest performance on the corresponding real task, and negligible or negative for the best synthetic tasks. The efficacy of patching H_{inter} indicates that while a synthetic dataset may target the necessary retrieval heads for the real task, they are *insufficient* in learning how to best utilize the required model components. One explanation is that fine-tuning induces upstream changes so that a different representation distribution is passed to the retrieval heads when learning on synthetic data. This allows retrieval heads to learn to be effective for the synthetic task while failing on out-of-distribution real data representations.

Intersection Heads are Core Attention Heads So what do the "extra" retrieval heads in the complement do? Wu et al. (2024) finds that Llama-2-7B contains 12 core retrieval heads while the rest are dynamically activated. We confirm this by finding that the average retrieval scores of the intersection heads are much greater than those of the complement heads (see Table 12).

Implications We established in Section 4 that retrieval heads are necessary for synthetic context extension. The fact that "better" heads in the intersection can be patched in to improve performance indicates that learning these heads alone is not sufficient. We see our work as contributing a useful analytical tool for understanding the behavior of synthetic context extension. At the same time, this presents a challenge for future work to tackle: can we come up with a more complete mechanistic explanation of synthetic context extension that accounts for these observations as well?

6 Related Work

Prior work has shown that benchmarking or training LLMs on synthetic data can reveal or obtain capabilities that can be transferred and generalized to real tasks, especially in settings where human-annotated data is hard to obtain such as long-context tasks. For this purpose, synthetic data are commonly used and believed to represent a simple reduction of the kinds of abilities employed in linguistically complex settings. The Needle-In-A-Haystack (NIAH) introduced by Kamradt (2023) involves placing a *needle* statement at a random position within a *haystack* consisting of unrelated essay text. Subsequent work (Hsieh et al., 2024; Li et al., 2024) has expanded this task to multi-value retrieval and used simple templated needle sentences to include distractor needles in the context. Hsieh et al. (2024) additionally parameterized its test suite by the diversity of the input context (essay text, repeated text, or distractor needles) and the target value type (words, numbers, or UUIDs).

Leveraging the potential generalizability of synthetic data, a line of work in interpretability literature generates synthetic data to perform controlled experiments to probe the inner workings of LLMs. For example, Kim & Schuster (2023) shows that a synthetic version of entity tracking can be used to mechanistically understand how fine-tuning enhances existing capabilities of pre-trained LLMs via mechanistic intervention techniques, and Kim et al. (2024) shows that the transformer circuit responsible for syllogistic reasoning in LLMs can be identified by evaluating on synthetic logical statements. However, there is a lack of understanding of when and how the mechanism discovered from synthetic tasks generalizes to real-world capabilities.

Our work bridges these directions by providing mechanistic explanations for the transferability of synthetic context extension while motivating the pursuit of better usage of synthetic data to evaluate, enhance, and understand the capabilities of LLMs.

7 Conclusion

In this paper, we investigated the relationship between the nature of synthetic data for synthetic context extension and performance on downstream tasks. Different synthetic datasets give widely varying performance, partially because of the different numbers of retrieval heads they induce in a model. We showed that these heads are causally connected to the performance, and that these heads are necessary (but not sufficient) for a strong downstream model. We believe this work paves the way for further mechanistic understanding of long context behavior and the ways in which synthetic data induces new capabilities in language models.

Reproducibility

We include the prompts used to construct our training datasets in the Appendix A, and describe our training setup in Section 2.2 with additional details in Appendix C. In addition, we plan to release the scripts used to create our datasets, train our models, and produce the results analysis included in this paper.

References

GenQA: Generating Millions of Instructions from a Handful of Prompts, author=Jiuhai Chen and Rifaa Qadri and Yuxin Wen and Neel Jain and John Kirchenbauer and Tianyi Zhou and Tom Goldstein. *ArXiv*, abs/2406.10323, 2024. URL `https://api.semanticscholar.org/CorpusID:270560271`.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3554–3565, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.278. URL `https://aclanthology.org/2021.naacl-main.278`.

Ramraj Chandradevan, Kaustubh Dhole, and Eugene Agichtein. DUQGen: Effective unsupervised domain adaptation of neural rankers by diversifying synthetic query generation. 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. 7437–7451, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.413. URL `https://aclanthology.org/2024.naacl-long.413`.

Aditi Chaudhary, Karthik Raman, and Michael Bendersky. It's All Relative! – A Synthetic Query Generation Approach for Improving Zero-Shot Relevance Prediction. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1645–1664, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.107. URL `https://aclanthology.org/2024.findings-naacl.107`.

Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen. Unveiling the Flaws: Exploring Imperfections in Synthetic Data and Mitigation Strategies for Large Language Models. *ArXiv*, abs/2406.12397, 2024. URL `https://api.semanticscholar.org/CorpusID:270562788`.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending Context Window of Large Language Models via Positional Interpolation. *ArXiv*, abs/2306.15595, 2023. URL `https://api.semanticscholar.org/CorpusID:259262376`.

Abhishek Divekar and Greg Durrett. SynthesizRR: Generating Diverse Datasets with Retrieval Augmentation. *ArXiv*, abs/2405.10040, 2024. URL `https://api.semanticscholar.org/CorpusID:269790883`.

Xinya Du, Junru Shao, and Claire Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1342–1352, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1123. URL `https://aclanthology.org/P17-1123`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1544.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data Engineering for Scaling Language Models to 128K Context, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

Gradient. Scaling Rotational Embeddings for Long-Context Language Models, 2024. URL https://gradient.ai/blog/scaling-rotational-embeddings-for-long-context-language-models.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning, 2021. URL https://arxiv.org/abs/2112.09118.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1555–1574, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.96. URL https://aclanthology.org/2023.emnlp-main.96.

Gregory Kamradt. Needle In A Haystack - pressure testing LLMs, 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main.

Geonhee Kim, Marco Valentino, and André Freitas. A Mechanistic Interpretation of Syllogistic Reasoning in Auto-Regressive Language Models. *ArXiv*, abs/2408.08590, 2024. URL https://api.semanticscholar.org/CorpusID:271892176.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pp. 3835–3855. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.acl-long.213.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems, 2024.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. NeedleBench: Can LLMs Do Retrieval and Reasoning in 1 Million Context Window?, 2024. URL https://arxiv.org/abs/2407.11963.

Tom Lieberum, Matthew Rahtz, J'anos Kram'ar, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does Circuit Analysis Interpretability Scale? Evidence from Multiple Choice Capabilities in Chinchilla. *ArXiv*, abs/2307.09458, 2023. URL https://api.semanticscholar.org/CorpusID:259950939.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best Practices and Lessons Learned on Synthetic Data for Language Models. *ArXiv*, abs/2404.07503, 2024b. URL https://api.semanticscholar.org/CorpusID:269042851.

Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. MathGenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2732–2747, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.151.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft, 2022.

Amirkeivan Mohtashami and Martin Jaggi. Random-Access Infinite Context Length for Transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=7eHn64wOVy.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, Tom Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. In-context Learning and Induction Heads. *ArXiv*, abs/2209.11895, 2022. URL https://api.semanticscholar.org/CorpusID:252532078.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific Skill Localization in Fine-tuned Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking. In *Proceedings of the 2024 International Conference on Learning Representations*, 2024. arXiv:2402.14811.

Abulhair Saparov and He He. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qFVVBzXxR2V.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 2022.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert. TRL: Transformer Reinforcement Learning. URL https://github.com/huggingface/trl.

Jerry Wei, Le Hou, Andrew Kyle Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V Le. Symbol tuning improves in-context learning in language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=vOX7Dfwo3v.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality, 2024.

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data, 2024.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. S2ynRE: Two-stage Self-training with Synthetic data for Low-resource Relation Extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8186–8207, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.455. URL https://aclanthology.org/2023.acl-long.455.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. WizardLM: Empowering Large Language Models to Follow Complex Instructions. *ArXiv*, abs/2304.12244, 2023b. URL https://api.semanticscholar.org/CorpusID:258298159.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing. *ArXiv*, abs/2406.08464, 2024. URL https://api.semanticscholar.org/CorpusID:270391432.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. Rlcd: Reinforcement learning from contrastive distillation for language model alignment. 2023. URL https://api.semanticscholar.org/CorpusID:260357852.

Fangcong Yin, Xi Ye, and Greg Durrett. LoFiT: Localized Fine-tuning on LLM Representations. *Advances in Neural Information Processing Systems*, 2024.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=B14TlG-RW`.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=yLClGs770I`.

Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xue Gang Wu, Bo Zhu, Yimeng Gan, Rui Hu, Shuicheng Yan, Han Fang, and Yahui Zhou. LongSkywork: A Training Recipe for Efficiently Extending Context Length in Large Language Models. *ArXiv*, abs/2406.00605, 2024. URL `https://api.semanticscholar.org/CorpusID:270210363`.

A Synthetic Dataset Creation Prompts

A.1 MDQA

Given a training example of MDQA data $(\mathcal{C}, q, y) \in \mathcal{D}_{\mathcal{T}}$, we first combine the query q and the answer y into a sentence and prompt GPT-4o-mini to rephrase the sentence with the sentence paraphrasing prompt to make it the needle f. Then, for the synthetic dataset with high context diversity, we prompt GPT-4o-mini to generate a Wikipedia-style context paragraph with the context generation prompt.

Prompt A.1: MDQA Sentence Paraphrasing Prompt

System Prompt:
You are a helpful AI assistant and you are good at creative writing.

Prompt:
Rewrite the following sentence to Wikipedia style with additional details: `{sentence}`
Make sure that readers can correctly answer the following question by reading your rewritten sentence:
Question: `{question}`
Answer: `{answer}`

Prompt A.2: MDQA Context Generation Prompt

System Prompt:
You are a helpful AI assistant and you are good at creative writing.

Prompt:
Please make up a 100-word Wikipedia paragraph for the following fake entities: `{entity}` . Invent details about people, places, and work related to each entity, and make sure all details are not related to any real-world entities. Give a short, meaningful title to your generated paragraph. After making up the paragraph, please generate a who/when/where/what/why question that:
(1) is related to the given fake entities;
(2) one can use the paragraph to correctly infer the answer within one or two words;
(3) is not a direct copy of a sentence from the paragraph. Please also include the gold answer to the generated question.
Please give your response in the format:
Title: [title]
Text: [text]
Question: [question]
Answer:[answer]

A.2 MuSiQue

Prompt A.3: MDQA Sentence Paraphrasing Prompt

Prompt:
Please make up a single sentence for each of the following fake entities in the style of a wikipedia article.
`{fake_entities}`
Please give your response in the format:
Title: [title]
Text: [text]

Prompt A.4: MuSiQue Context Generation Prompt

Prompt:
Please make up a 5-sentence wikipedia paragraph for the following fake entities. Invent details about people, places, and work related to each entity.
`{fake_entities}`
Please give your response in the format:
Title: [title]
Text: [text]

A.3 SummHay

Prompt A.5: SummHay Query Insight (Concept Expression - High) Prompt

Prompt:
Please rephrase the sentence: " `{text}` "

Prompt A.6: SummHay Query Insight (Concept Expression - Simplified) Prompt

Prompt:
Please simplify and shorten the following sentence. Remove details: " `{sentence}` "

Prompt A.7: SummHay Citation Needle Prompt

Prompt:
"Please break up the following sentence into multiple sentences: " `{text}` "

B Training Prompts

Prompt B.1: MDQA and MuSiQue Training Prompt

Prompt:
The following are given passages.
`{context}`
Answer the question based on the given passages. Only give me the answer and do not output any other words.
Question: `{question}`
Answer:

Prompt B.2: SummHay Citation Training Prompt

Prompt:
The following are given documents.
`{context}`
For the given statement, identify the documents that contain the information by citing the numbers associ-

ated with those documents in brackets. For example, if the information in the statement is only found in Document 3, then respond with "[3]". If the information is contained in both Document 3 and Document 7, then respond with "[3][7]". Only output the answer and do not output any other words.
Statement: {statement}
Answer:

C Additional Data and Training Details

C.1 Data

We use 1400 examples for training MDQA models, 400 examples for MuSiQue models, and 400 examples for SummHay Citation models. Each dataset is partitioned in to a 90/10 train/validation split. We use the validation split to calculate retrieval and insight scores.

MDQA Example

Context:
Document 1: (Title: Don Quixote (Teno)) portion of Don Quixote and his horse are visible. The horse appears to be charging forward out of the stone with his head raised, mouth open, and hooves kicking. The left foot of the horse is not formed, intentionally, by Teno. In Don Quixote's hand is a lance of steel. Both figures are loosely modeled and the figures and stone rest on a oval base measuring which was cut into three pieces for transport by ship to the United States. An inscription on the sculpture reads: King Juan Carlos I and Queen Sofía presented the sculpture June 3, 1976, on
...
Document 10: (Title: Rocinante) Rocinante is Don Quixote's horse in the novel Don Quixote by Miguel de Cervantes. In many ways, Rocinante is not only Don Quixote's horse, but also his double: like Don Quixote, he is awkward, past his prime, and engaged in a task beyond his capacities.
...

Question:
what is don quixote's horse's name
Answer:
Rocinante

SummHay Citation Example

Context:

...

Document [24]: ... Furthermore, the provision of U.S. dollars by global central banks increased, ensuring adequate liquidity within the international financial system. This measure illustrated the depth of the coordinated efforts among major financial institutions to stave off crises and maintain functional stability. The reverberations of these actions and their impacts on the markets are still unfolding...

...

Document [27]: ... Turning our gaze to the realm of global financial oversight, central banks are making coordinated efforts to prevent a liquidity crunch in the international financial system. Recognizing the importance of maintaining robust liquidity, global central banks have ramped up their provision of U.S. dollars, showcasing a united front in ensuring financial stability. Central banks in Canada, Britain, Japan, Switzerland, and the eurozone have initiated daily currency swaps to ensure that banks operating within their jurisdictions have the necessary dollars to function smoothly. This strategy is aimed at providing stability and fostering confidence in the global banking system during uncertain economic times...

...

Statement:

Global central banks increased their provision of U.S. dollars to ensure adequate liquidity in the international financial system, demonstrating coordinated efforts to prevent a liquidity crunch.

Answer:

[24][27]

For MDQA and MuSiQue, we experiment with only training on a subset of the relations involved in eval question hops.

On MDQA, we create three variants: L_1 is the subset containing Who, When, and Where questions; L_2 is the subset containing When and Where questions; L_3 is the subset containing only Who questions. These comprise 65.8%, 31.0%, and 34.8% of all questions in the MDQA training set respectively. In Table 1, Table 3, and Table 11, we only report L_1 results due to space constraints. Fine-tuning Llama-3-8B-Instruct on these datasets results in the following F1-scores for the target dataset: $L_1 = 0.80$, $L_2 = 0.63$, $L_3 = 0.65$. Fine-tuning Mistral-7B-Instruct-v0.1 on these datasets results in the following F1-scores for the target dataset: $L_1 = 0.59$, $L_2 = 0.52$, $L_3 = 0.57$.

On MuSiQue, we use the subset of linear 3-hop questions consisting solely of T-REx component questions (Elsahar et al., 2018), as identified by “>>”. 10.8% of MuSiQue linear 3-hop questions in the training set fit this criteria. Additionally, among all component question hops in the training set, 43.0% are sourced from T-REX.

C.2 Symbolic Data Construction

See Figure 5 for examples.

C.3 Training

For fine-tuning, we use the Huggingface TRL (von Werra et al.) and PEFT (Mangrulkar et al., 2022) libraries to fine-tune attention heads with LoRA (Hu et al., 2022) (rank = 8 and alpha = 8) using a batch size of 1 and 4 gradient accumulation steps.

We enable Flash Attention 2 and DeepSpeed and use a single NVIDIA H100 GPU (96GB) for each training run. We use greedy decoding in all evaluations.

MuSiQue Symbolic Data

Context
...
BPUG {..., 'UQCA': 'QUID', **'TZAM': 'XDPW'**, 'EJSN': 'TTFU', ...}
...
KVTJ {..., 'UQCA': 'SXVI', **'ERQG': 'FQDR'**, 'TZAM': 'XYTH', ...}
...
FQDR {..., 'UQCA': 'EHQQ', **'UDPB': 'BPUG'**, 'ERQG': 'DMII', ...}
...

Q: What is the **TZAM** of the **UDPB** of the **ERQG** of **KVTJ**?
A: XDPW

SummHay Citation Symbolic Data

Context
...
Document [16]: [..., 'SIWK', **'NGOW'**, 'UXHQ', 'RBZE', ...]
...
Document [27]: [..., 'DUTT', **'NGOW'**, 'LTYM', 'FPHP', ...]
...

Q: NGOW
A: [16][27]

Figure 5: Examples of symbolic data consruction for MuSiQue and SummHay Citation.

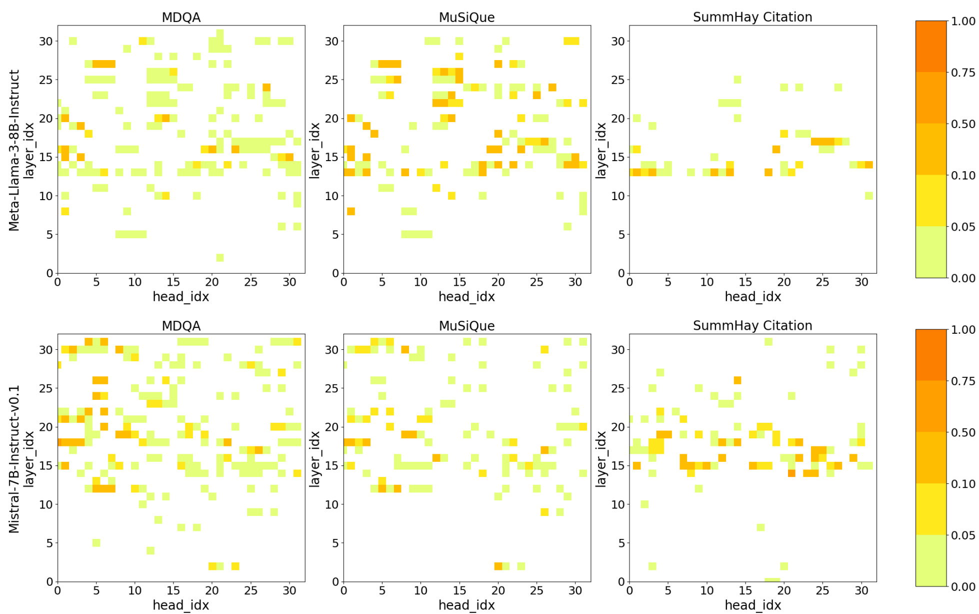

Figure 6: Retrieval scores for MDQA, MuSiQue, and Insight scores for SummHay Citation. Top Row: Llama-3-8B-Instruct. Bottom Row: Mistral-7B-Instruct-v0.1. The y-axis indicates the layer index and the x-axis indicates the head index within the layer. We note that retrieval heads are largely found in the last 2/3 layers of the model, as expected according to their involvement in the "final step" of copying the correct answer to the output. By contrast, SummHay Citation insight heads are concentrated in the middle layers, indicative of their intermediate role. Within a single layer, the specific important attention head indices were likely randomly primed during pretraining to be effectively adapted to the target task.

D Retrieval Score Heatmaps

Attention head retrieval scores for the real tasks are shown in Figure 6.

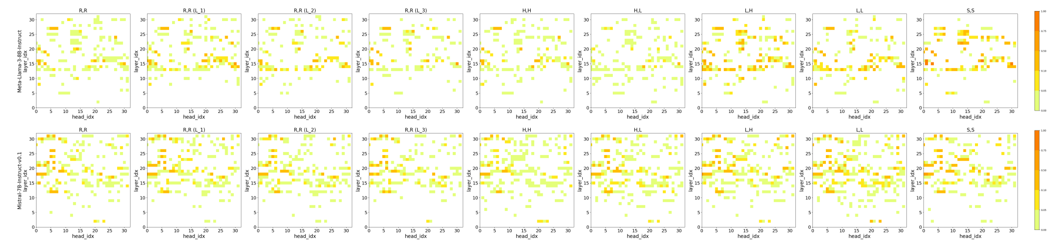

Figure 7: Retrieval scores for MDQA and its synthetic dataset versions. Top Row: Llama-3-8B-Instruct. Bottom Row: Mistral-7B-Instruct-v0.1. The y-axis indicates the layer index and the x-axis indicates the head index within the layer.

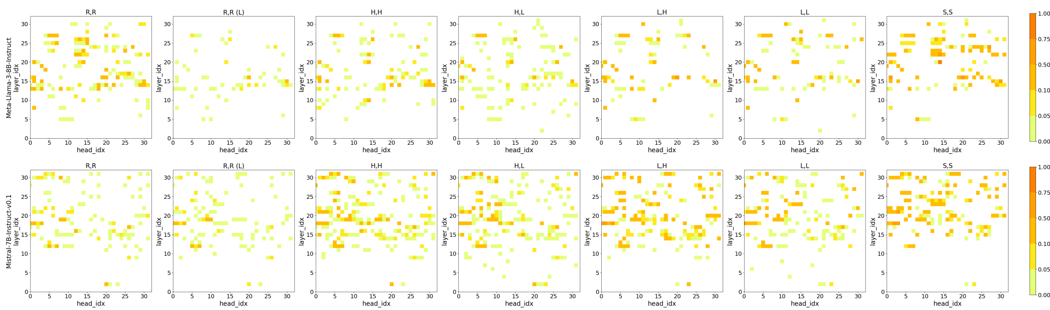

Figure 8: Retrieval scores for MuSiQue and its synthetic dataset versions. Top Row: Llama-3-8B-Instruct. Bottom Row: Mistral-7B-Instruct-v0.1. The y-axis indicates the layer index and the x-axis indicates the head index within the layer.

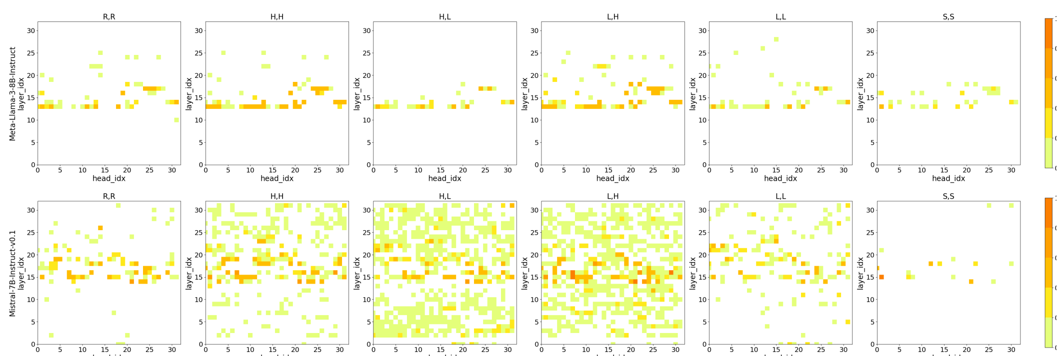

Figure 9: Insight scores for SummHay Citation and its synthetic dataset versions. Top Row: Llama-3-8B-Instruct. Bottom Row: Mistral-7B-Instruct-v0.1. The y-axis indicates the layer index and the x-axis indicates the head index within the layer.

For each target real task, we present heatmaps comparing the real task retrieval scores to the synthetic dataset retrieval scores: MDQA in Figure 7, MuSiQue in Figure 8, and SummHay Citation in Figure 9.

E RETRIEVAL HEAD RECALL

In Table 5, Table 6, and Table 7, we examine the overlap between non-zero scoring attention heads on our target tasks and their synthetic versions after fine-tuning Llama-3-8B-Instruct. We find that on all 3 tasks, the attention heads with non-zero retrieval scores on the real data have high recall (≥ 0.76) against those identified on the synthetic data. On MuSiQue and SummHay Citation, we

Table 4: Spearman correlation of synthetic data attention head recall with F1 on the real dataset, showing a strong relationship.

	Model	
	Llama3	Mistral
MDQA	0.22	0.16
MuSiQue	0.75	0.40
SummHay Citation	1.00	0.82

Table 5: Pairwise recall of Llama-3-8B-Instruct attention heads with non-zero retrieval scores for MDQA synthetic datasets. Limited datasets: L_1 = Who, When, Where; L_2 = When, Where; L_3 = Who. Retrieval Head recall on the real dataset (first column) is weakly correlated with F1 on the real MDQA data (Spearman R = 0.22).

	R,R	R,R (L_1)	R,R (L_2)	R,R (L_3)	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.79	0.84	0.88	0.85	0.78	0.81	0.83	0.87	157	0.82
R,R (L_1)	0.76	1.00	0.90	0.86	0.81	0.69	0.81	0.80	0.85	151	0.80
R,R (L_2)	0.66	0.74	1.00	0.78	0.71	0.63	0.71	0.70	0.74	124	0.63
R,R (L_3)	0.63	0.64	0.70	1.00	0.69	0.61	0.66	0.68	0.77	112	0.65
H,H	0.75	0.75	0.79	0.86	1.00	0.74	0.78	0.83	0.83	139	0.37
H,L	0.73	0.68	0.74	0.80	0.78	1.00	0.77	0.80	0.78	147	0.41
L,H	0.80	0.83	0.88	0.90	0.86	0.81	1.00	0.91	0.93	154	0.49
L,L	0.67	0.68	0.72	0.77	0.76	0.69	0.75	1.00	0.76	127	0.47
S,S	0.64	0.66	0.69	0.79	0.69	0.61	0.70	0.69	1.00	116	0.48

Table 6: Pairwise recall of Llama-3-8B-Instruct attention heads with non-zero retrieval scores for MuSiQue synthetic datasets. We find that the attention heads identified on the real dataset has high recall against all synthetic datasets ($\geq$0.76). Retrieval head recall on the real dataset (first column) is also **strongly** correlated with F1 on the real MuSiQue data (Spearman R = 0.75).

	R,R	R,R (L)	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.96	0.81	0.76	0.87	0.82	0.87	129	0.45
R,R (L)	0.41	1.00	0.50	0.41	0.52	0.49	0.42	55	0.32
H,H	0.59	0.85	1.00	0.63	0.70	0.65	0.63	94	0.37
H,L	0.66	0.84	0.76	1.00	0.81	0.78	0.71	112	0.41
L,H	0.45	0.64	0.50	0.48	1.00	0.65	0.53	67	0.29
L,L	0.47	0.65	0.51	0.52	0.72	1.00	0.55	74	0.34
S,S	0.67	0.76	0.67	0.63	0.79	0.74	1.00	100	0.32

also observe a strong relationship (Spearman R=0.75 and R=1.0 respectively) between the non-zero score attention head recall and F1 on the real task.

However, fine-tuning Mistral-7B-Instruct-v0.1 results in slightly different patterns, as shown in Table 8, Table 9, and Table 10. First, we see more scoring attention heads, which could be caused by the sliding window attention used in the architecture, which only enables a subset of heads to any single position. Second, many of the synthetic datasets result in far more non-zero scoring attention heads, a pattern that we see across all tasks. On MuSiQue and SummHay Citation, we observe a slightly weaker relationship (Spearman R=0.40 and R=0.82 respectively) between the non-zero score attention head recall and F1 on the real task.

Table 7: Pairwise recall of Llama-3-8B-Instruct attention heads with non-zero insight scores for SummHay Citation synthetic datasets. Insight head recall on the real dataset (first column) is also **strongly** correlated with F1 on the real data (Spearman R = 1.0)

	R,R	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.77	0.94	0.66	0.78	0.87	48	0.81
H,H	0.85	1.00	1.00	0.75	0.82	0.87	53	0.70
H,L	0.60	0.58	1.00	0.48	0.72	0.70	31	0.61
L,H	0.90	0.92	1.00	1.00	0.88	0.93	65	0.79
L,L	0.65	0.62	0.94	0.54	1.00	0.80	40	0.65
S,S	0.54	0.49	0.68	0.43	0.60	1.00	30	0.54

Table 8: Pairwise recall of Mistral-7B-Instruct-v0.1 attention heads with non-zero retrieval scores for MDQA synthetic datasets. Limited datasets: L_1 = Who, When, Where; L_2 = When, Where; L_3 = Who. Retrieval head recall on the real dataset (first column) is weakly correlated with F1 on the real MDQA data (Spearman R = 0.16).

	R,R	R,R (L_1)	R,R (L_2)	R,R (L_3)	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.76	0.76	0.81	0.74	0.74	0.77	0.69	0.80	178	0.63
R,R (L_1)	0.81	1.00	0.85	0.86	0.77	0.80	0.80	0.72	0.82	192	0.59
R,R (L_2)	0.81	0.84	1.00	0.86	0.74	0.77	0.80	0.72	0.87	190	0.52
R,R (L_3)	0.74	0.72	0.73	1.00	0.67	0.71	0.72	0.63	0.74	161	0.57
H,H	0.86	0.84	0.81	0.86	1.00	0.83	0.83	0.76	0.84	208	0.20
H,L	0.88	0.89	0.86	0.93	0.85	1.00	0.86	0.81	0.87	212	0.22
L,H	0.88	0.85	0.86	0.92	0.82	0.83	1.00	0.78	0.85	205	0.31
L,L	0.93	0.91	0.91	0.94	0.88	0.92	0.91	1.00	0.91	240	0.24
S,S	0.80	0.76	0.81	0.81	0.72	0.73	0.74	0.68	1.00	178	0.15

Table 9: Pairwise recall of Mistral-7B-Instruct-v0.1 attention heads with non-zero retrieval scores for MuSiQue synthetic datasets. Retrieval Head recall on the real dataset (first column) is also moderately correlated with F1 on the real MuSiQue data (Spearman R = 0.40)

	R,R	R,R (L)	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.66	0.56	0.55	0.63	0.66	0.62	111	0.31
R,R (L)	0.63	1.00	0.57	0.57	0.59	0.60	0.53	106	0.14
H,H	0.89	0.95	1.00	0.83	0.88	0.86	0.82	178	0.21
H,L	0.83	0.91	0.78	1.00	0.81	0.81	0.78	167	0.23
L,H	0.84	0.83	0.73	0.72	1.00	0.82	0.80	148	0.21
L,L	0.75	0.72	0.61	0.61	0.70	1.00	0.68	126	0.17
S,S	0.74	0.66	0.61	0.62	0.72	0.72	1.00	133	0.11

Table 10: Pairwise recall of Mistral-7B-Instruct-v0.1 attention heads with non-zero insight scores for SummHay Citation synthetic datasets. Insight Head recall on the real dataset (first column) is also **strongly** correlated with F1 on the real SummHay Citation data (Spearman R = 0.82)

	R,R	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.31	0.15	0.16	0.45	0.81	91	0.40
H,H	0.87	1.00	0.34	0.39	0.75	0.88	259	0.28
H,L	0.79	0.65	1.00	0.57	0.80	0.81	496	0.28
L,H	0.86	0.76	0.57	1.00	0.78	0.88	497	0.38
L,L	0.75	0.44	0.24	0.24	1.00	0.81	150	0.28
S,S	0.14	0.05	0.03	0.03	0.09	1.00	16	0.17

Table 11: Results on Mistral-7B-Instruct-v0.1 after patching heads that comprise the complement and intersection retrieval heads between the real and synthetic data versions, compared to random retrieval heads and original performance. The best patching F1 is **bolded**, and Δ is the improvement over the original F1.

Task	Data Variant Concept	Data Variant Context	N	Compl.	Inter.	Rand.	Orig.	Δ
MDQA	Real	Real	-	-	-	-	0.63	-
	Real	Real (Limited)	80	**0.57**	0.53	0.49	0.59	-0.02
	Low	High	69	0.21	**0.34**	0.23	0.31	0.03
	Low	Low	86	0.21	**0.40**	0.26	0.24	0.16
	High	Low	78	0.13	**0.31**	0.16	0.22	0.08
	High	High	80	0.21	**0.26**	0.18	0.20	0.06
	Symbolic	Symbolic	70	0.01	**0.02**	0.02	0.15	-0.13
MuSiQue	Real	Real	-	-	-	-	0.31	-
	High	Low	92	0.23	**0.26**	0.20	0.23	0.03
	High	High	91	0.16	**0.24**	0.20	0.21	0.03
	Low	High	73	0.14	**0.21**	0.17	0.21	0.00
	Low	Low	71	0.15	**0.18**	0.16	0.17	0.01
	Real	Real (Limited)	70	0.14	**0.20**	0.18	0.14	0.06
	Symbolic	Symbolic	80	0.14	**0.19**	0.15	0.11	0.08
SummHay	Real	Real	-	-	-	-	0.40	-
	Simplified	High	78	0.34	**0.35**	0.35	0.38	-0.2
	High	Low	72	0.33	0.33	**0.35**	0.28	0.08
	High	High	70	**0.30**	0.30	0.29	0.28	0.02
	Simplified	Low	68	0.29	**0.33**	0.30	0.28	0.05
	Symbolic	Symbolic	13	0.14	0.14	**0.16**	0.18	-0.02

F Retrieval Head Patching Details

We implemented retrieval head patching with Baukit.[8] Given an example from the test set and a set of attention heads to patch, we run a forward pass with the model fine-tuned on the real data and extract the attention output from the selected attention heads before being projected and concatenated back to the residual stream. Then, we use the same example and run a forward pass with the model fine-tuned on a synthetic dataset. We replace the attention outputs of the aforementioned selected attention heads with the attention outputs extracted from the model fine-tuned on real data. Using the procedure described above, we patch the attention outputs of the selected attention heads into the model fine-tuned on a synthetic dataset for *all input* tokens. We then use the patched inputs to generate and decode output tokens without patching any activations for the output tokens.

F.1 Mistral-7B-Instruct-v0.1 Retrieval Head Patching

See Table 11.

F.2 Intersection and Complement Head Retrieval Scores

See Table 12.

G Full Finetuning

[8] `https://github.com/davidbau/baukit`

Table 12: Average retrieval / insight scores for attention heads in the intersection and the complement.

Task	Dataset Variant		Llama-3-8B-Instruct		Mistral-7B-Instruct-v0.1	
	Concept	Context	Inter.	Compl.	Inter.	Compl.
MDQA	Real	Real (Who, When, Where)	0.045	0.011	0.059	0.019
	Real	Real (Who)	0.052	0.012	0.062	0.021
	Real	Real (When, Where)	0.050	0.012	0.059	0.018
	High	High	0.046	0.010	0.057	0.020
	High	Low	0.045	0.015	0.056	0.018
	Low	High	0.044	0.010	0.057	0.013
	Low	Low	0.049	0.013	0.054	0.015
	Symbolic	Symbolic	0.051	0.013	0.059	0.020
MuSiQue	Real	Real (Limited)	0.121	0.049	0.065	0.021
	High	High	0.105	0.040	0.053	0.015
	High	Low	0.096	0.045	0.055	0.018
	Low	High	0.116	0.048	0.055	0.017
	Low	Low	0.106	0.054	0.058	0.021
	Symbolic	Symbolic	0.099	0.037	0.057	0.026
SummHay	High	High	0.071	0.008	0.093	0.036
	High	Low	0.092	0.016	0.098	0.039
	Simplified	Low	0.087	0.017	0.097	0.050
	Simplified	High	0.068	0.008	0.093	0.041
	Symbolic	Symbolic	0.097	0.021	0.213	0.064

Table 13: Llama-3-8B-Instruct (all LoRA modules): Performance (F1) of fine-tuning on different synthetic data on the long-context retrieval and reasoning tasks. The results of training on the best synthetic datasets are bolded.

Concept Exp.	Context Div.	MDQA	MuSiQue	Concept Exp.	Context Div.	SummHay
High	High	0.35	0.40	High	High	0.83
High	Low	0.39	**0.42**	High	Low	0.68
Low	High	**0.49**	0.30	Simplified	High	**0.83**
Low	Low	0.47	0.38	Simplified	Low	0.58
Symbolic	Symbolic	0.46	0.37	Symbolic	Symbolic	0.63
Real Data (Full)		0.82	0.45	Real Data (Full)		0.81
Real Data (Limited)		0.84	0.32			

In this section, we present results on Meta-Llama-3-8B-Instruct with fine-tuning of all LoRA modules, and demonstrate that we find similar conclusions.

G.1 Synthetic Data Performance

See Table 13. We find that there are mostly small (< 0.05) performance differences between fine-tuning only attention heads and all modules. Notable exceptions are found in the SummHay Citation task, where the performance of the synthetic datasets increase up to +0.13 (High, High).

G.2 Retrieval Score Heatmaps

See Figure 10.

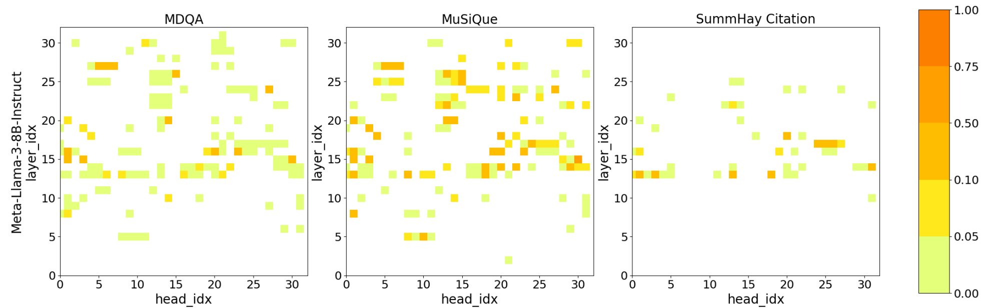

Figure 10: Llama-3-8B-Instruct (all LoRA modules): Retrieval scores for MDQA, MuSiQue, and Insight scores for SummHay Citation, after fine-tuning on each task. The y-axis indicates the layer index and the x-axis indicates the head index within the layer.

Table 14: Llama-3-8B-Instruct (LoRA all modules): Pairwise recall of attention heads with non-zero retrieval scores for MDQA synthetic datasets. Limited datasets: L_1 = Who, When, Where; L_2 = When, Where; L_3 = Who. Recall of real data retrieval heads is moderately correlated with F1 (Spearman R = 0.60).

	R,R	R,R (L_1)	R,R (L_2)	R,R (L_3)	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.71	0.70	0.76	0.79	0.73	0.71	0.75	0.76	137	0.82
R,R (L_1)	0.77	1.00	0.83	0.84	0.79	0.71	0.80	0.83	0.84	148	0.84
R,R (L_2)	0.77	0.84	1.00	0.83	0.79	0.71	0.75	0.80	0.84	150	0.73
R,R (L_3)	0.72	0.74	0.71	1.00	0.69	0.67	0.70	0.75	0.77	129	0.72
H,H	0.73	0.67	0.66	0.67	1.00	0.69	0.70	0.74	0.74	126	0.35
H,L	0.76	0.69	0.67	0.74	0.79	1.00	0.72	0.78	0.76	143	0.39
L,H	0.82	0.86	0.79	0.86	0.89	0.80	1.00	0.91	0.90	159	0.49
L,L	0.76	0.77	0.74	0.81	0.81	0.76	0.79	1.00	0.82	138	0.47
S,S	0.69	0.71	0.70	0.74	0.73	0.66	0.70	0.75	1.00	125	0.46

Table 15: Llama-3-8B-Instruct (LoRA all modules): Pairwise recall of attention heads with non-zero retrieval scores for MuSiQue synthetic datasets. Recall of real data retrieval heads is moderately correlated with F1 (Spearman R = 0.36).

	R,R	R,R (L)	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.94	0.82	0.77	0.88	0.88	0.86	135	0.48
R,R (L)	0.44	1.00	0.56	0.41	0.46	0.56	0.39	63	0.41
H,H	0.59	0.87	1.00	0.64	0.75	0.73	0.61	98	0.40
H,L	0.78	0.89	0.89	1.00	0.90	0.86	0.78	136	0.42
L,H	0.65	0.73	0.77	0.66	1.00	0.83	0.71	100	0.30
L,L	0.50	0.68	0.57	0.49	0.64	1.00	0.55	77	0.38
S,S	0.75	0.73	0.73	0.68	0.84	0.84	1.00	118	0.37

G.3 Retrieval Head Recall

In Table 14, Table 15, and Table 16, we find that there are generally fewer non-zero scoring attention heads on the synthetic tasks, compared to the real task. On MuSiQue, the non-zero attention heads tend to be subsets of the those identified on the real task, as when only fine-tuning attention modules.

G.4 Retrieval Score Cosine Similarity

Table 16: Llama-3-8B-Instruct (LoRA all modules): Pairwise recall of attention heads with non-zero insight scores for SummHay Citation synthetic datasets. Recall of the real data insight heads is moderately correlated with F1 (Spearman R = 0.58).

	R,R	H,H	H,L	L,H	L,L	S,S	# Heads	F1
R,R	1.00	0.63	0.77	0.50	0.66	0.71	45	0.81
H,H	0.89	1.00	1.00	0.73	0.81	0.93	63	0.82
H,L	0.67	0.62	1.00	0.49	0.60	0.76	39	0.68
L,H	0.87	0.90	0.97	1.00	0.84	0.90	78	0.83
L,L	0.84	0.75	0.90	0.63	1.00	0.81	58	0.57
S,S	0.67	0.62	0.82	0.49	0.59	1.00	42	0.62

Table 17: Llama-3-8B-Instruct (all LoRA modules): Cosine similarity of real dataset retrieval scores (+ SummHay insight scores) across tasks.

	MDQA	MuSiQue	SummHay Retrieval	SummHay Insight
MDQA	1.00	0.85	0.35	0.16
MuSiQue	0.85	1.00	0.50	0.29
SummHay Retrieval	0.35	0.50	1.00	0.11
SummHay Insight	0.16	0.29	0.11	1.00

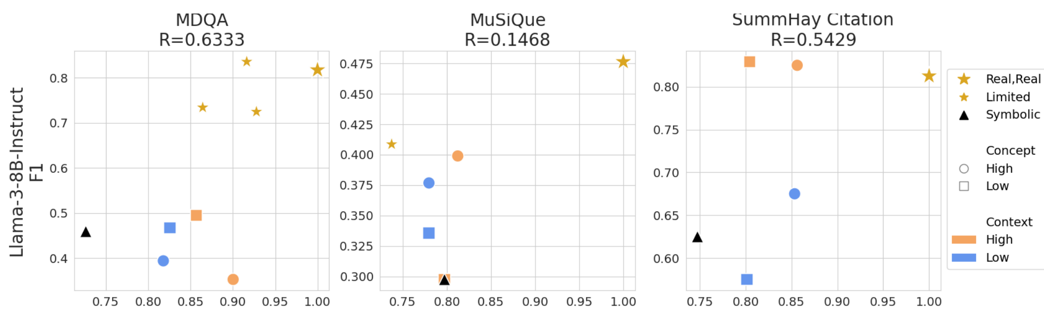

Figure 11: Llama-3-8B-Instruct (all LoRA modules): Cosine similarity between the retrieval scores on real datasets (R, R) vs. their synthetic versions, and Spearman correlation for each setting.

Across Tasks See Table 17. Similar to fine-tuning only attention-heads, we find the highest similarity between MDQA and MuSiQue retrieval scores, and much lower similarity with SummHay Citation scores, reflecting the different nature of the task (extractive QA vs. citation).

Synthetic Datasets vs. Real Task Performance See Figure 11. Overall, we find that synthetic datasets with lower performance recruit fewer scoring attention heads, although the relationship is weaker than when only fine-tuning attention heads.

G.5 PATCHING

See Table 18. Notably, we find that patching complement attention head activations is the best in more settings than patching the intersection (7 settings vs. 6 settings). This is despite the results in Table 19 showing that the intersection attention heads have higher scores.

Table 18: Llama-3-8B-Instruct (all LoRA modules): Results after patching heads that comprise the complement and intersection retrieval heads between the real and synthetic data versions, compared to random retrieval heads and original performance. Best patch F1 is **bolded**, and Δ is the improvement over the original F1.

Task	Data Variant		N	Compl.	Inter.	Rand.	Orig.	Δ
	Concept	Context						
MDQA	Real	Real	-	-	-	-	0.82	-
	Real	Real (Limited)	75	**0.87**	0.84	0.85	0.84	0.04
	Low	High	70	**0.66**	0.61	0.56	0.49	0.17
	Low	Low	67	0.61	**0.71**	0.44	0.47	0.24
	Symbolic	Symbolic	72	0.46	0.33	**0.52**	0.46	0.06
	High	Low	72	**0.63**	0.27	0.47	0.39	0.24
	High	High	63	0.47	0.57	**0.64**	0.35	0.29
MuSiQue	Real	Real	-	-	-	-	0.48	-
	High	Low	61	**0.39**	0.35	0.39	0.42	-0.03
	Real	Real (Limited)	59	0.40	**0.42**	0.35	0.41	0.01
	High	High	73	**0.41**	0.37	0.31	0.40	0.01
	Low	Low	68	0.39	**0.40**	0.35	0.38	0.02
	Symbolic	Symbolic	51	**0.43**	0.10	0.35	0.37	0.06
SummHay	Real	Real	-	-	-	-	0.81	-
	Simplified	High	39	0.77	0.81	**0.82**	0.83	-0.01
	High	High	28	0.76	0.76	**0.81**	0.82	-0.01
	High	Simplified	24	0.60	**0.72**	0.67	0.68	0.05
	Symbolic	Symbolic	27	0.64	**0.71**	0.66	0.62	0.08
	Simplified	Simplified	27	**0.64**	0.64	0.61	0.57	0.07

Table 19: Llama-3-8B-Instruct (all LoRA modules): Average retrieval / insight scores for attention heads in the intersection and the complement.

Task	Dataset Variant		Llama-3-8B-Instruct	
	Concept	Context	Inter.	Compl.
MDQA	High	Low	0.047	0.013
	Real	Real (Who, When, Where)	0.046	0.013
	High	High	0.049	0.010
	Low	High	0.045	0.009
	Low	Low	0.047	0.012
	Symbolic	Symbolic	0.049	0.015
MuSiQue	Real	Real (Limited)	0.125	0.049
	High	High	0.113	0.037
	Low	High	0.105	0.039
	High	Low	0.095	0.037
	Low	Low	0.119	0.045
	Symbolic	Symbolic	0.099	0.031
SummHay	Simplified	Low	0.067	0.010
	High	Low	0.077	0.020
	Simplified	High	0.065	0.010
	High	High	0.064	0.011
	Symbolic	Symbolic	0.081	0.013