Great Memory, Shallow Reasoning: Limits of kNN-LMs

Anonymous ACL submission

Abstract

K-nearest neighbor language models (kNN-LMs), which integrate retrieval with next-word prediction, have demonstrated strong performance in language modeling as well as downstream NLP benchmarks. These results have led researchers to argue that models trained on poor quality or outdated data could perform well by employing a kNN extension that has access to a higher-quality datastore. In this work, we ask whether this improved ability to recall information really translates into downstream abilities. We extensively evaluate kNN-LMs on a diverse set of tasks, ranging from sentiment classification and commonsense reasoning to multi-hop reasoning. Results show that kNN-LMs excel at *memory*-intensive tasks, where utilizing the patterns in the input is sufficient for determining the output, but struggle with *reasoning* tasks that require integrating multiple pieces of information to derive new knowledge. We further demonstrate through oracle experiments and qualitative analysis that even with perfect retrieval, kNN-LMs still fail to determine the correct answers, placing an upper bound on their reasoning performance.

1 Introduction

A foundational property of pretrained language modeling (Peters et al., 2018; Devlin et al., 2019) has been that improvements to the perplexity of the model lead to improvements on downstream tasks. This property is central to the scaling of large language models (LLMs) where researchers focus nearly exclusively on perplexity as a proxy metric for improved general purpose abilities (Kaplan et al., 2020). In recent years, this research has centered primarily on high-quality text data at greater and greater quantities as the limiting component for producing better language models (Hoffmann et al., 2022).

This increasing need for data to train language models has led to significant challenges. On one hand, including as much high-quality data as possible results in improved downstream performance. On the other hand, this data is often protected by licenses or copyright, which means training on such data brings legal issues. For example, the recent high-profile lawsuit from the New York Times notes the clear use of their data in OpenAI models (Grynbaum and Mac, 2023).

It would be ideal to circumvent this issue entirely with alternative approaches. If a model could be trained on lower-quality data but adapted to perform well on real tasks, it might provide a technical workaround. Non-parametric Language Models (NPLMs), such as kNN-LMs, have emerged as a promising approach in this space (Khandelwal et al., 2020). kNN-LMs extend neural LMs by linearly interpolating with simple k-nearest neighbor LMs. This approach can improve language modeling with its memory over a massive collection of texts, usually referred to as a datastore. Khandelwal et al. (2021) and Shi et al. (2022) validate that kNN-LMs achieve better performance on downstream tasks compared to standard LMs. The SILO model of Min et al. (2024) applies this approach further by training a LM exclusively on license-permissive data, and using a non-parametric datastore to improve the models during inference.

In this work, we study the limits of how kNN-LMs can be used to improve LLMs. Specifically, we are interested in whether the improvements in perplexity seen with kNN-LMs are equivalent to other improvements in LM ability, or if improvements in non-parametric memory are orthogonal to standard language modeling. This question relates to debates about whether memory is separable from other language abilities and how they interact in NLP benchmarks.

To study this question, we implement large-scale kNN-LMs on top of modern open LLMs with two datastores in different domains. We replicate past results that demonstrate significant decreases in per-

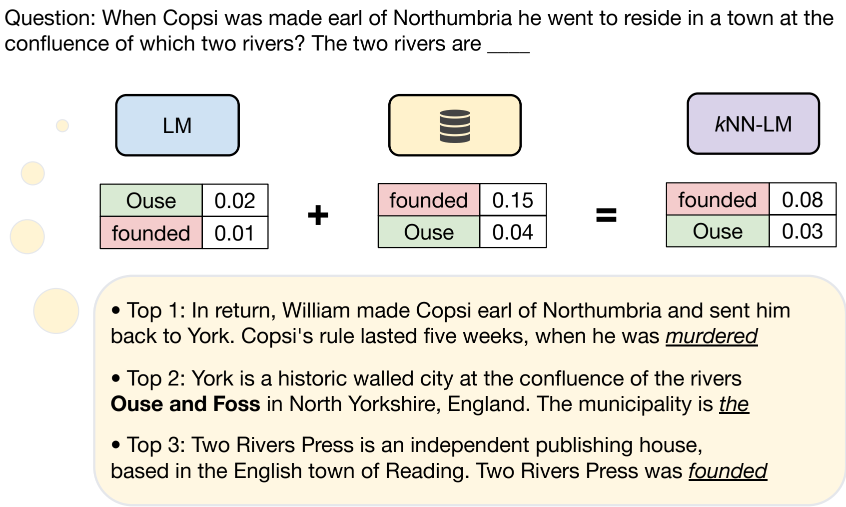

Figure 1: In this multi-hop question answering (QA) example, the LM is uncertain about the answer and likely benefit from retrieval. The kNN approach finds both irrelevant and relevant documents that may help. However, two issues occur: first, an irrelevant document increases the probability of the wrong answer; second, even though a relevant document has been found, it may not upweight the actual answer (Ouse). These issues may impact task performance more than perplexities.

plexity across domains. This perplexity decrease transfers to similar benefits in task accuracy across several NLP benchmarks. These benchmarks are rather simple, where recognizing the patterns in the input and matching them with the patterns in memory is sufficient for determining the output. We refer to these as *memory*-based tasks.

However, we see a different story when applying these models to tasks that require significant *reasoning* ability. These tasks often require integrating multiple pieces of information to derive new knowledge. In our experiments, the use of kNN-LMs does not improve performance in reasoning, and in fact seems to hurt reasoning ability across tasks significantly. This behavior is robust and occurs even in domains that are explicitly targeted by the datastore used by the non-parametric model. These experiments lead us to conclude that while kNN-LMs may be useful in settings where data is constrained, they should not be seen as a remedy for low-quality training data, and that perplexity scores should not be seen as a corollary for LM ability outside of parametric training settings.

2 Related Work

Retrieval Models Although Large Language Models (LLMs) achieve superhuman performance on a wide range of natural language processing tasks, they often produce hallucinations, struggle with integrating new knowledge, and expose private information present in the training data. Recently, research interest has shifted towards retrieval-based LMs, which combine a parametric neural model and a non-parametric external datastore (Guu et al., 2020; Karpukhin et al., 2020). These retrieval-based LMs naturally incorporate new knowledge, enhance the factuality of generated texts, and reduce privacy concerns (Asai et al., 2024). Furthermore, studies (Borgeaud et al., 2022) have demonstrated that employing retrieval augmentation during large-scale pre-training can outperform standard LMs while requiring fewer parameters.

Among retrieval-based LMs, kNN-LMs (Khandelwal et al., 2020) emerge as a popular choice (Min et al., 2024). Unlike other retrieval models that encode and retrieve documents, kNN-LMs encode and retrieve tokens. At every token, kNN-LMs search for the k most similar tokens from the datastore based on contextualized token embeddings, which are then turned into a next-token distribution. kNN-LMs linearly interpolate the retrieved kNN distribution with the output of a base LM. They do not require additional training but introduce computational and memory overhead.

Reasoning Retrieval. Little research has been conducted on constructing retrieval models for reasoning tasks. Leandojo (Yang et al., 2023) investigates the use of retrieval-based LMs to assist with theorem proving, and Levonian et al. (2023) experiment with retrieving content from mathematical textbooks to generate responses to student questions. In our study, we create a reasoning-specific datastore to assist LMs in performing reasoning-intensive tasks.

Evaluation of kNN-LMs. While kNN-LMs excel at language modeling and have demonstrated enhanced performance in machine translation (Khandelwal et al., 2021) and simple NLP tasks (Shi et al., 2022), the question of whether they are thoughtful reasoners remains open. Wang et al. (2023a) demonstrate that kNN-LMs struggle with open-ended text generation as they only provide benefits for a narrow set of token predictions and produce less reliable predictions when generating longer text. BehnamGhader et al. (2023) showed that when retrieval is conducted based on the similarity between queries and statements, kNN-LMs often fail to identify statements critical for reasoning. Even when these crucial statements are retrieved, it is challenging for kNN-LMs to effectively leverage them to infer new knowledge. These studies, however, are limited to a narrow set of tasks. Our work seeks to provide a comprehensive evaluation of the reasoning capabilities of kNN-LMs and provides an extensive analysis of the sources of their failures.

3 k-Nearest Neighbor Large Language Models

Non-parametric language models are variants of standard language models that give the model the ability to utilize an additional datastore $\mathcal{D}$ during inference to determine the next word prediction, $p(x_{t+1}|x_{1\ldots t};\mathcal{D})$. This datastore may be part of the original training data, data for adaptation to a new domain, or be used to incorporate continual updates or protected data. As these datastores are typically quite large, this process requires a retrieval component in the loop to find the sparse subset of the datastore that can best inform the current prediction. Several popular approaches exist including DPR (Karpukhin et al., 2020) and REALM (Guu et al., 2020).

In this work, we focus on kNN-LMs due to their popularity as an approach to directly improve LM perplexity on fixed models without a need for retraining. As noted in the intro, this approach has also been put forward as a method for circumventing the need for high-quality licensed training data in LLMs. Formally kNN-LMs are defined as

$$p(x_{1:T};\mathcal{D}) = \prod_t p(x_{t+1} \mid x_{1:t};\mathcal{D})$$

$$= \prod_t (\lambda p_{k\mathrm{NN}}(x_{t+1}|x_{1:t};\mathcal{D}) + (1-\lambda) p(x_{t+1}|x_{1:t}))$$

Let (k_i, v_i) be the ith (key, value) pair in $\mathcal{D}$, $f(\cdot)$ maps a token sequence to its contextual representation, and $d(\cdot)$ measures the distance between two vectors.

$$p_{k\mathrm{NN}}(x_{t+1} \mid x_{1:t};\mathcal{D}) \propto \sum_{(k_i,v_i)\in\mathcal{D}} \mathbf{1}_{x_{t+1}=v_i} \times \exp(-d(k_i, f(x_{1:t}))).$$

When using a Transformer language model, we define the distance metric $d(\cdot)$ as the squared ℓ_2 distance. To assemble the datastore we run the language model over all the documents to collect the necessary hidden states and corresponding next word.

Experimental Setup. The hyperparameters include λ, k, and σ. λ determines the weight of the datastore, and we consider $\lambda \in \{0.1, 0.2, 0.3\}$. Additionally, we retrieve $k \in \{1600, 2048\}$ neighbors and smooth the kNN distribution with a temperature $\sigma \in \{1, 3, 5, 10\}$.

For each inference model, we use Math and Wiki datastores for language modeling on the corresponding evaluation datasets: wikitext and math textbooks. Each datastore represents a specific domain, and we evaluate the performance of kNN-LM on a domain by measuring the perplexity of each evaluation dataset. We conduct a grid search to find the hyperparameters that yield the lowest PPL for each datastore. The optimal hyperparameters for each datastore are later applied across all downstream tasks in our experiments.

We provide eight demonstrations for GSM8K and three demonstrations for BBH. For the other datasets, we all perform zero-shot inference. We present full details of the experiments in the Appendix A.

Inference and Retrieval Models. We use Llama-2-7b (Touvron et al., 2023), Llama-3-8B (AI@Meta, 2024), and Mistral-7B (Jiang et al., 2023) as our inference models. For each inference model, we build the corresponding datastores. The keys are the 4096-dimensional hidden representations before the final MLP which predicts the token

$\mathcal{D}$	Text Size	Tokens	Mem
Wiki	2.2GB	610M	44G
Math	0.6GB	200M	15G

Table 1: Overview of the two datastores. Tokens are produced by Llama2 tokenizers. Mem is the memory size of the datastore.

Model	LM Performance Wiki	Math
Llama2-7b	10.63	7.90
+Wiki	**9.74**	8.75
+Math	11.33	**7.23**
Llama-3-8b	9.70	5.36
+Wiki	**9.32**	6.03
+Math	10.37	**5.22**
Mistral-7B	9.72	5.64
+Wiki	**9.29**	6.41
+Math	10.49	**5.59**

Table 2: Perplexity comparison. Rows vary the datastore $\mathcal{D}$ used. Columns represent different held-out test sets. Lower numbers indicate better performance.

distribution at each generation step, produced by executing forward passes over the datastore corpora. For efficient similarity search, we create a FAISS index (Johnson et al., 2019) and search for nearest-neighbor tokens using Euclidean distance. Due to the scale of the datastores, we perform approximate search instead of exact search. We base our implementation on RetoMaton (Alon et al., 2022).

4 kNN-LMs Help In-Domain Perplexity

To explore how different sources of external knowledge impact downstream task performance, we experiment with two datastores. First, we follow the choice made by Shi et al. (2022), where they identify heterogeneous data sources that are broadly relevant to common downstream NLP tasks. In particular, they mix Wikitext103 (Merity et al., 2017), with other sources including the English portion of Amazon Review (He and McAuley, 2016), and CC-NEWS (Hamborg et al., 2017) and IMDB (Maas et al., 2011). We call this datastore *Wiki*.

Then, we hypothesize that the commonly explored corpora for building datastores do not contain relevant knowledge to assist with math reasoning tasks. To maximize the performance gain on these tasks, we construct a datastore comprising 3.94K mathematical textbooks, sourced from (Wang et al., 2023b). These textbooks contain both theorems and practice questions, from which humans acquire mathematical knowledge. This datastore consists of 200M tokens. We will refer to this datastore as *Math*. We summarize the statistics of each datastore in Table 1.

We begin by validating past results of kNN-LMs on language modeling. We present results in Table 2. To facilitate meaningful comparisons between models with different tokenizers and vocabulary sizes, we report word-level perplexities. These results show that having access to a non-parametric datastore leads to lower perplexity compared to using a standalone LM across all datasets. This improvement in perplexity is observed when the corpus used to construct the datastore and the one used for inference share the same data source. For instance, since the training split of Wikitext103 is in Wiki, the LM+Wiki setting achieves the lowest perplexity on Wikitext103's validation set. Utilizing the other datastore results in performance worse than that of the standalone LM.

5 kNN-LMs Can Help Memory-Intensive Tasks

We begin by looking at a set of memory-intensive tasks, which we believe can be solved by pattern matching at scale without complex reasoning. We incorporate three types of tasks: sentiment classification, which aims to predict whether the sentiment of a text is positive or negative; textual entailment, which assesses the relationship between two sentences, determining if it constitutes entailment, contradiction, or neutrality; and topic classification, which involves identifying the main topic of a text. The datasets included for these tasks are as follows:

- For sentiment classification, we include SST-2 (Socher et al., 2013), movie review (MR) (Pang and Lee, 2005), customer review (CR) (Hu and Liu, 2004), Rotten Tomatoes (RT), and a variant of hyperpartisan news detection (HYP) (Kiesel et al., 2019).

- For textual entailment, we use Commitment-Bank (CB) (De Marneffe et al., 2019) and Recognizing Textual Entailment (RTE) (Dagan et al., 2010).

- For topic classification, our datasets are AG News (AGN) (Zhang et al., 2015) and Yahoo!

	RTE	RT	CB	Yahoo	CR	AGN	HYP	MR	SST2
Llama2-7B	66.06	79.74	50.00	**59.37**	74.55	81.30	**64.15**	83.10	84.02
+Wiki	**66.43**	79.46	**51.79**	58.83	**76.95**	81.46	**64.15**	82.85	**84.68**
+Math	65.70	**82.55**	**51.79**	59.10	73.70	**81.79**	50.39	82.90	84.62
Llama3-8B	**70.76**	**79.46**	64.29	58.87	79.10	79.17	59.30	83.80	86.54
+Wiki	61.37	79.55	**71.43**	**58.93**	**80.45**	79.33	59.30	83.50	87.04
+Math	**70.76**	77.39	66.07	56.83	79.40	**80.11**	59.30	**84.30**	**87.10**
Mistral-7B	76.17	**75.32**	71.43	56.63	81.90	73.57	56.59	**79.35**	**81.82**
+Wiki	76.17	75.05	67.86	56.63	**82.15**	73.55	**56.78**	79.30	81.77
+Math	76.17	75.05	**75.00**	56.63	81.85	**73.59**	**56.78**	79.10	81.77

Table 3: Accuracy comparison on various memory-intensive tasks.

Answers (Yahoo) (Zhang et al., 2015).

For classification and multiple-choice question-answering (QA) tasks, we utilize Domain Conditional Pointwise Mutual Information (DCPMI) (Holtzman et al., 2021) to predict answers. We then calculate accuracy metrics to compare performance across different models. We measure the performance using F1 scores at the token level for text generation. Additionally, whenever feasible, we employ fuzzy verbalizers (Shi et al., 2022) to maximize the performance of kNN-LMs.

The results of these tasks are summarized in Table 3. On these tasks, kNN-LMs exhibit improved performance. Incorporating an external datastore outperforms a standalone LM on eight datasets while showing comparable performance on the remaining dataset. We further explain this performance gap through qualitative analysis in Appendix B.

6 kNN-LMs *Hurt* Reasoning Performance

For reasoning tasks, we consider three types: knowledge-intensive reasoning, which focuses on utilizing world knowledge for making (potential) multi-hop inferences; commonsense reasoning, which involves leveraging commonsense knowledge to understand social and physical interactions; and mathematical reasoning, which includes arithmetic, logical, and discrete reasoning abilities. The datasets selected for these categories are as follows:

- For knowledge-intensive reasoning, we explore Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), ARC Easy and Challenge (Clark et al., 2018), OpenbookQA (OBQA) (Mihaylov et al., 2018), and MMLU (Hendrycks et al., 2020) to assess the model's ability to apply extensive world knowledge.

- For commonsense reasoning, we examine HellaSwag (Zellers et al., 2019) and Winogrande (Sakaguchi et al., 2021), which test the model's understanding of social norms and physical laws.

- For mathematical reasoning, we utilize DROP (Dua et al., 2019), GSM8K (Cobbe et al., 2021), and Big Bench Hard (BBH) (Suzgun et al., 2022) to evaluate the model's capacity for complex arithmetic, logical deductions, and handling of discrete concepts.

We present the results for knowledge-intensive tasks in Table 6. In stark contrast to the earlier findings, using a standalone LM consistently outperforms kNN-LMs on these tasks. Most surprisingly, on Natural Questions and HotpotQA, which consist of QA pairs constructed from Wikipedia documents, performance does not improve even though Wiki contains several million Wikipedia tokens. Retrieving from Wiki leads to a three-point decrease in performance.

Results for commonsense reasoning and mathematical reasoning tasks are shown in Table 5. The standalone LM once again outperforms kNN-LMs models on four out of the five datasets. The most significant differences in performance occur on GSM8K. Although incorporating an external data store results in a slight performance increase on Mistral, this does not demonstrate the effectiveness of kNN-LMs on GSM8K. Under Mistral's parameter settings,kNN-LMs has minimal changes on the predictions of the standalone LM, merely introducing some randomness. Finally, although kNN-LMs

	NQ	HotpotQA	Arc-Challenge	Arc-Easy	OBQA	MMLU
Llama2-7B	**23.18**	**22.72**	**41.81**	**57.49**	**57.00**	**39.22**
+Wiki	22.53	22.53	38.31	57.41	56.20	38.68
+Math	21.14	21.26	41.04	56.82	56.20	38.53
Llama3-8B	23.64	**25.14**	**44.88**	**58.83**	**55.80**	**42.67**
+Wiki	**24.00**	24.48	43.94	58.59	53.80	42.32
+Math	23.04	24.63	43.26	58.59	54.60	42.46
Mistral-7B	**20.63**	**20.96**	**46.42**	**60.94**	**58.80**	**41.91**
+Wiki	20.58	20.80	46.16	60.61	57.40	41.80
+Math	20.56	20.48	46.08	60.77	57.80	41.55

Table 4: Performance comparison on datasets for knowledge-intensive reasoning tasks.

	Winogrande	HellaSwag	DROP	GSM8K	BBH
Llama2-7B	69.37	**64.46**	**32.39**	**14.83**	30.69
+Wiki	**70.32**	63.67	32.14	12.05	**32.08**
+Math	68.98	63.54	32.31	13.48	30.82
Llama3-8B	73.95	**65.99**	**45.55**	**45.72**	39.67
+Wiki	73.95	64.71	45.02	44.28	39.01
+Math	**74.19**	65.15	45.54	45.63	**39.92**
Mistral	74.19	**69.08**	**46.93**	36.30	**43.37**
+Wiki	**74.66**	68.21	46.69	36.45	42.69
+Math	73.64	68.11	46.38	**36.60**	43.09

Table 5: Performance comparison on datasets for other reasoning tasks.

		Perplexity	Accuracy
OBQA	LM	255.76	55.80
	kNN-LM	9.41	95.60
NQ	LM	112.56	23.64
	kNN-LM	8.91	46.40
HotpotQA	LM	158.26	25.14
	kNN-LM	8.15	49.85

Table 6: Results in an oracle setting where the kNN-LMs always include the correct answer as one of the k nearest neighbors.

do not improve GSM8K and Drop over standard LMs, we find that retrieving from Math improves over retrieving from Wiki.

7 Analysis

The results of this work show that kNN-LMs generally hurt reasoning of models, despite helping perplexity and other simpler tasks. In this section, we investigate the cause of this further.

Qualitative Analysis. We conduct qualitative analysis to understand the failures of kNN-LMs better. In the qualitative analysis, we inspect examples of knowledge-intensive and mathematical reasoning datasets and show the retrieved tokens as well as the proceeding context. Through these examples, we find the following patterns that prevent kNN-LM from retrieving the correct token.

- **kNN-LMs struggle with multi-hop reasoning questions.** When the task requires extracting multiple pieces of sentences from the corpus and then combining the information to infer the answer, kNN-LMs often retrieve tokens that are contextually appropriate and relevant to part of the question, rather than the correct answer. As shown in Table 7, for the multi-hop reasoning question from HotpotQA, the model needs to identify an actor who both starred in Stargate SG-1 and guest-starred in Twin Peaks. While the required information is available in Wikipedia, it is distributed across two paragraphs. kNN-LMs retrieve only the actors from Stargate SG-1, failing to combine information from two sources to perform ac-

HotpotQA Example	Label	LM Pred
Which American character actor who starred on the television series “Stargate SG-1” (1997–2007) and appeared in “Episode 8” of “Twin Peaks” as a guest star?	Don S. Davis	Don S. Davis
Retrieved Context	**Token**	**kNN-LM Pred**
• After the first three seasons of Stargate SG-1 had been filmed on 16 mm film (although scenes involving visual effects had always been shot on 35 mm film for various technical reasons), “Nemesis” was the first episode filmed entirely on 35 mm film ... “Nemesis” was the last episode before actor	Christopher	
• “200” won the 2007 Constellation Award for Best Overall 2006 Science Fiction Film or Television Script, and was nominated for the 2007 Hugo Award for Best Dramatic Presentation, Short Form. The episode also marks the first time original SG-1 member	Jack	Michael Shanks
• Season one regular cast members included Richard Dean Anderson, Amanda Tapping,	Michael	

Table 7: A multihop reasoning example from HotpotQA with predictions of the standard LM and kNN-LMs.

NQ Example	Label	LM Pred
who is the largest supermarket chain in the uk?	Tesco	Tesco
Retrieved context	**Token**	**kNN-LM Pred**
• The majority of stores will open as normal across the UK, however Sainsbury’s advise shoppers to check details of when your local branch as some may close earlier than normal using the online store locator tool.(Image: Bloomberg) Supermarket giant	Asda	
• Along with Lidl, Aldi has eaten away at the market share of the Big Four supermarkets:	Tesco	Asda
• buy one, get one free (BOGOF) offers have been criticised for encouraging customers to purchase food items that are eventually thrown away; as part of its own campaign on food waste, supermarket retailer	Morris	

Table 8: A knowledge-intensive reasoning example from Natural Questions with predictions of the standard LM and kNN-LMs.

curate multi-hop reasoning.

- **kNN-LMs are sensitive to the syntax but not the semantics of the question.** While kNN-LM retrieves the next token that fits the context, it cannot distinguish subtle semantic differences between different words in a sentence. As a result, when more than one word fits the context, it may not select the correct answer. Table 8 demonstrates this issue with an example from the NQ dataset. Even though Asda is not the largest supermarket in the UK, due to the highly similar contexts of ‘supermarket giant’ and ‘the largest supermarket, kNN-LMs ultimately assign a high probability to Asda and make a wrong prediction.

- **kNN-LMs tend to retrieve high-frequency entities in the corpus.** The entities are often proper nouns like person names and locations. If part of the answer overlaps with these high-frequency proper nouns, kNN-LMs will retrieve them and make wrong predictions, as shown in Table 9 and Table 14.

- **kNN-LMs fail at mathematical reasoning tasks.** For instance, in the object counting task from the BBH dataset, even though kNN-LM understands the context that it needs to retrieve a number as the next token, it cannot solve the complex task of first identifying which objects are musical instruments and then counting them, as shown in Table 10.

Is the problem a failure of model weighting? We investigate whether degraded reasoning capabilities of kNN-LMs stem from a failure in choosing a good weighting λ. This experiment aims to analyze kNN-LMs’ behaviors when λ is optimal for the downstream task. Specifically, we directly search for λ that maximizes the log probabilities of a small set of labeled downstream task examples. We conduct this experiment on OpenbookQA and HotpotQA. We enumerate through retrieving $k \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$ neighbors and setting temperature $\sigma \in \{1, 2, 5, 10\}$. We retrieve from Wiki. We initialize λ at 0.5, and as the optimization proceeds, we find that smaller λ values correlate with lower loss. Ultimately, we arrive at the minimum loss when λ is close to 0.

HotpotQA Example	Label	LM Pred
What type of plane is the four engine heavy bomber, first introduced in 1938 for the United States Army, which is hangared at Conroe North Houston Regional Airport?	American Boeing B-17 Flying Fortress	The B-17 Flying Fortress
Retrieved context	**Token**	**kNN-LM Pred**
• A famous symbol of the courage and sacrifices made by American bomber crews during World War II was revealed May 16 at the National Museum of the U.S. Air Force, Wright-Patterson Air Force Base, Ohio. The meticulously restored B-	17	
• As the Avenger made its way to the tower area, the wings began to fold up, a maneuver which enabled more of its kind to be loaded side by side into aircraft carriers. The queen of the event was the B-	25	The B-25 Mitchell.
• Spring is here, so why not hop a plane and grab some lunch? Even better if a World War II-era B-	25	

Table 9: Example from HotpotQA showing the impact of high-frequency proper nouns in the corpus on kNN-LMs predictions retrieving from Wikipedia.

Mathematical Reasoning Example	Label	LM Pred
I have three violins, three trombones, a flute, and four trumpets. How many musical instruments do I have?	11	11
Retrieved Context	**Token**	**kNN-LM Pred**
• In this example, the optimal route would be: 1 -> 3 -> 2 -> 4 -> 1, with a total completion time of	10	
• How many different passwords are there for his website system? How does this compare to the total number of strings of length	10	10
• Using the TSP, the most efficient order in which to schedule these tasks would be: 2 -> 3 -> 1 -> 4 -> 2, with a total completion time of	14	

Table 10: A mathematical reasoning example from BBH requiring object counting with predictions of the standard LM and kNN-LMs.

This process suggests that without any interpolation of the kNN distribution, the correct labels of the provided demonstrations receive the highest log probability. Therefore, OpenbookQA and HotpotQA are unlikely to benefit from having simple kNN access to Wiki.

Is the problem a failure of retrieval? We investigate whether degraded reasoning capabilities of kNN-LMs stem from a failure in retrieval. We examine kNN-LMs' behaviors when retrieval is perfect. To achieve perfect retrieval, we include the correct answer among the k nearest neighbors. Specifically, we construct a datastore for OpenbookQA, NQ, and HotpotQA, respectively, including their train and test examples. We then examine both perplexity and accuracy. The results, presented in Table 6, indicate that while kNN-LMs can significantly reduce the perplexity, the model does not always derive the correct answer, even when the correct answer is explicitly given as one of the k neighbors. Therefore, the failure of reasoning cannot be fully attributed to the failure of retrieval. However, perfect retrieval does improve LM by a large margin, suggesting that better retrieval is beneficial. Currently, retrieval is performed by finding similar hidden representations. A training-based approach such as RAG (Lewis et al., 2020) has the potential to improve retrieval substantially.

8 Conclusions

We investigate whether the improved perplexity observed in kNN-LMs models can be translated into enhanced reasoning capabilities. We conduct extensive evaluation across 22 datasets. Our findings indicate that while kNN-LMs improve perplexity and can achieve better performance on memory-intensive tasks, they struggle with reasoning-intensive tasks, showing a disconnect between LM ability and task ability. Further qualitative analysis reveals that even when kNN-LMs produce correct answers, these are often the result of spurious correlations rather than actual reasoning. We believe this places an upper bound on the usefulness of these approaches compared to results from parametric models.

Limitations

As we are limited by computing budget, we only build datastores up to 610 million tokens. It is unlikely although not impossible that larger datastores built on general web corpus like C4 will lead to better reasoning capabilities. Additionally, we only experiment with LLMs with seven- to eight-billion model parameters as the base models. The findings in this paper may not generalize to other, possibly larger, base models.

References

AI@Meta. 2024. Llama 3 model card.

Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning*, pages 468–485. PMLR.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15492–15509. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162, pages 2206–2240. PMLR.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378. Association for Computational Linguistics.

Michael M Grynbaum and Ryan Mac. 2023. The times sues openai and microsoft. *The New York Times*, pages B1–B1.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. 2023. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2024. SILO language models: Isolating legal risk in a nonparametric datastore. In *The Twelfth International Conference on Learning Representations*.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 115–124. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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

Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, and Mohit Iyyer. 2023a. kNN-LM does not improve open-ended text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15023–15037. Association for Computational Linguistics.

Zengzhi Wang, Rui Xia, and Pengfei Liu. 2023b. Generative ai for math: Part i–mathpile: A billion-token-scale pretraining corpus for math. *arXiv preprint arXiv:2312.17120*.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. In *Neural Information Processing Systems (NeurIPS)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Corpus	Text Size	Tokens
Wikitext103	0.5GB	140M
Amazon	0.07GB	18M
CC-NEWS	1.6GB	443M
IMDB	0.03GB	8M
Total	2.2GB	609M

Table 11: Statistics of each data source in the Wiki datastore.

A More Implementation Details

Table 11 presents the data sources of the Wiki datastore. Table 12 shows hyperparameters we use for different tasks.

B More Qualitative Analysis

We explain why retrieving from Math improves LMs on sentiment analysis. First, we consider a sentiment analysis example in Table 13. In this task, given a sentence, a model is required to predict whether the sentiment expressed is positive or negative. The sentence in the example expresses a positive sentiment; however, Llama-2 predicts the sentiment to be negative. kNN-LMs, when retrieving from Wiki, fail to find sentiment-related tokens, and hence also predict a negative sentiment. Performing retrieval from Math produced the correct sentiment. However, this is more coincidental rather than reflective of the model's capability, because, although the retrieved tokens display a positive sentiment, the retrieved contexts are not relevant to the test example. we observe that sentiment-related content is ubiquitous, regardless of the source we use to build the datastore. Even in math textbooks, we find many sentences that express sentiment.

Data	λ	k	τ
Llama2 + Wiki	0.2	2048	5.0
Llama3 + Wiki	0.1	2048	5.0
Mistral + Wiki	0.1	2048	10.0

Data	λ	k	τ
Llama2 + Math	0.2	1600	5.0
Llama3 + Math	0.1	2048	3.0
Mistral + Math	0.1	2048	10.0

Table 12: Hyperparameters in kNN-LM. **Top**: Hyperparameters for Wiki datastore. **Bottom**: Hyperparameters for Math datastore .

Sentiment Example	Label	LM Pred
humorous, artsy, and even cute, in an off-kilter, dark, vaguely disturbing way. The sentence has a tone that is	Positive	Negative
Retrieved Context	**Retrieved**	**kNN-LM Pred**
Wiki		
• meta-commentator, Imhoff gives us a decidedly modern delivery. His speaking rhythms are staccato and his tone	bitter	
• Collins, who has worked on more than 100 children books and won several awards: his tone is	fun	Negative
• is her own narrator, so the thoughts and feelings of others are conveyed secondhand or are absent entirely. Her tone and language are at turns	honest	
Math		
• preferred term is not "Platonist" but "quasiëmpiricist", a word Tymoczko lends a subtly	different	
• ... or a horror film (group 2, $N_H = 29$). The data are coded so that higher scores indicate a more	positive	Positive
• the failure of the Intermediate Value Theorem is neither here nor there nor anywhere else to them. This is not a bad nor a	good	

Table 13: A sentiment analysis example with predictions of the standard LM and kNN-LMs. We show tokens retrieved from each datastore and their proceeding tokens.

HotpotQA Example	Label	LM Pred
who is older, Annie Morton or Terry Richardson?	Terry Richardson	Terry Richardson
Retrieved context	**Token**	**kNN-LM Pred**
• And she still wasn't done. Later she tweeted a warning to all women. "My hard won advice: never get into an elevator alone with [Terry Gilliam.] Terry	Gilliam	
• #MeToo https://t.co/jPnFhfB5GQ - Ellen Barkin(@EllenBarkin) March 17, 2018Barkin got another shot in. Terry	Gilliam	Terry Gilliam
• I haven't posted about Christina Hendricks in a while but it's Valentine's Day and that makes me think of chocolate and chocolate reminds me of Christina Hendricks. And Christina	Hend	

Table 14: Another example from HotpotQA explains the impact of high-frequency proper nouns in the corpus on kNN-LMs predictions retrieving from Wikipedia.