CE 7490 Project 2: Scaling Behaviour in Retrieval-Augmented Generation

Xinghe Chen† Yihua Hu† Hongrui Liu†
† denotes equal contribution
College of Computing and Data Science, Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
{xinghe001, yihua001, hongrui001}@e.ntu.edu.sg

Abstract

Scaling behaviour examines how a system's performance responds to variations in parameters. Studies of scaling behaviour on neural language models have revealed consistent performance improvements as dataset size, model complexity, and computational power increase. However, the scaling behaviour of Retrieval-Augmented Generation (RAG) systems remains underexplored. In this work, we investigate the scaling behaviour of RAG systems by evaluating their performance across varying corpus sizes and the capability of large language models (LLMs). We empirically demonstrate that certain scaling behaviours exist in RAG systems: When the external database size of a RAG system exceeds a specific threshold, it provides significant benefits to LLMs, particularly less powerful ones. Moreover, we have yet to observe any convergence in this performance gain. This work, instead of a technical solution to a specific task, is an empirical promise that the philosophy of scaling extends beyond LLMs and is equally applicable to RAG systems.

1. Introduction

Scaling behaviour is a universal principle that describes how systems respond to changes in size, energy, or other fundamental parameters. It provides insights for understanding and modelling complex systems [1, 2]. This principle governs a wide range of phenomena, from the interactions of atomic particles [3] to the formation and evolution of cosmological structures [4], highlighting its foundational role across diverse scientific domains. However, in the field of natural language processing, particularly in the foundational theory of neural language models, theoretical research is far behind the remarkable success of large language models (LLMs) [5].

Among the various efforts to establish universal laws for neural language models, scaling laws have gained significant momentum since 2017. Early work established a foundation by demonstrating empirical relationships, showing that increasing model size, dataset size, and computational power results in predictable performance improvements and reductions in reducible loss [6, 7]. These findings were quickly applied to the development of neural language models like GPT-3, which demonstrated strong few-shot learning capabilities directly driven by scaling law [8]. Subsequent research extended these ideas to generative tasks, offering deeper insights into how scaling influences model capabilities [9]. By optimising resource efficiency and implementing compute-optimal strategies, a balance between cost and performance was achieved [10], paving the way for future application research. In less than a year, the maturity of this field culminated in the release of GPT-4 in early 2023 [11], accompanied by a technical report that, despite the technical details missing, marked a transformative milestone for LLMs.

Although scaling laws, to some extent, contributed to the launch of LLMs, they remain largely underexplored in the branch of retrieval-augmented generation (RAG). RAG was first proposed in 2020 by Patrick Lewis et al. [12] to enhance generative models by integrating external retrieval for knowledge-intensive NLP tasks. RAG's philosophy involves integrating information retrieval techniques from external sources to enhance query processing, thereby reducing the occurrence of hallucinations and factually incorrect outputs while improving the reliability and richness of responses [13]. For example, as shown in Fig. 1, the query about OpenAI's CEO dismissal leads to a hallucination by the LLM, as this information was not included in the training corpus. With RAG, however, the query is enriched with relevant documents from the database. These documents provide recent news that allows the LLM to respond with accurate, contextual information. In a more intuitive sense, RAG enables the LLM to respond in an open-book manner rather than relying solely on its vast yet ambiguous parameters.

In this project, we identify several reasons behind the limited exploration of scaling behaviour in RAG systems.

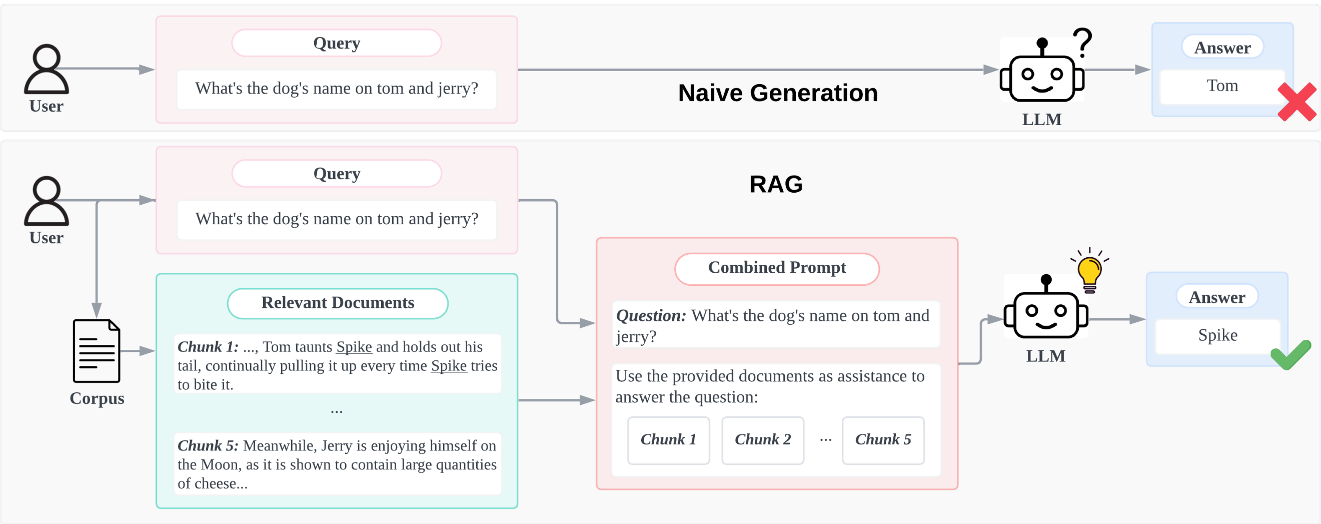

User
Query
What's the dog's name on tom and jerry?
Naive Generation
LLM
Answer
Tom
RAG
User
Query
What's the dog's name on tom and jerry?
Corpus
Relevant Documents
Chunk 1: ..., Tom taunts Spike and holds out his tail, continually pulling it up every time Spike tries to bite it.
...
Chunk 5: Meanwhile, Jerry is enjoying himself on the Moon, as it is shown to contain large quantities of cheese...
Combined Prompt
Question: What's the dog's name on tom and jerry?
Use the provided documents as assistance to answer the question:
Chunk 1
Chunk 2
···
Chunk 5
LLM
Answer
Spike

Figure 1. An example of how RAG enhances LLM performance in responding to user queries. This example is a sample QA in a data point in Fig. 2 using LLaMA-3.1 8B RAG.

Below, we outline these challenges and propose our strategies for this project:

- The definition of scaling behaviour or scaling laws for RAG systems remains ambiguous. Even for a simple RAG pipeline, components of document retrieval and answer generation can each or collectively exhibit distinct scaling behaviours. **In this project, we focus on the scaling law associated with the size of the external database, emphasising the open-book nature of RAG.**
- Evaluating scaling behaviour is intractable, as external databases are not involved during training, and there is no clear metric, such as reducible loss, as seen in prior scaling law research. **We approach this by evaluating the final performance of RAG systems using three metrics: exact match (EM), F1-score, and accuracy in a question-answering (QA) task.**
- RAG research encompasses LLM deployment, database embedding, and retrieval processes. The computationally intensive nature of RAG research typically makes it a black box. **To overcome this, we utilise the open-source and user-friendly FlashRAG Python toolkit [14], providing a transparent framework for scaling research in RAG systems [1].**

Through extensive experiments, we observed scaling behaviour in an RAG system and concluded the following: **(a)** A threshold exists for the external database size, and RAG only begins to benefit the model when the threshold is reached. **(b)** The external corpus can compensate for the limitations of a less powerful model. **(c)** The performance gain increases with corpus size after the threshold is reached and, to the best of our knowledge, does not show convergence.

2. Method

This section provides details on how the scaling behaviour is investigated. It covers the RAG system, the task to be solved, the external database (retrieval corpus), the evaluation metrics, and the experimental platform.

2.1. RAG System

We utilized the FlashRAG Python toolkit to build our RAG system. The system leverages the `e5-base-v2` model to embed and vectorise text data, forming the foundation of the knowledge base. FAISS [15] is then used to construct an index, enabling efficient retrieval. Subsequently, for any input query, the system retrieves the top-k most relevant documents with the approximate k-nearest neighbor (AKNN) algorithm. These retrieved documents are then combined with the user query to enhance the generation process. In this work, we use three LLaMA `Instruct` series models for generation, and mainly investigate the scaling behaviour of LLaMA-3-8B and LLaMA-3.1-8B due to computational constraints.

As shown in figure 1, this system design adopts a sequential RAG pipeline, where indexing, retrieval-augmentation, and generation are performed in sequence, which is straightforward yet powerful. We opted for this pipeline over more complex architectures to minimise the variables introduced by the pipeline itself, as discussed earlier. In this project, we fix these design choices and utilise different LLMs from the Meta LLaMA series [16], both with and

[1] The full code and data are available upon request. Experimental results, including generated text and description files, are publicly available at https://github.com/XChen1998/CE-7454-Project-2 for further qualitative and quantitative analysis.

without RAG (naive generation), to investigate how RAG scales. For more details of the specific configuration, please refer to Section 3.

2.2. QA Task

We utilise the Natural Questions dataset, introduced by Google [17], as the benchmark dataset for our evaluation. This dataset consists of real, anonymised, aggregated queries-answer pairs from the Google search engine. From the full dataset, comprising both training and test sets, we use only the 3,610 samples from the test set for evaluation. This dataset was selected due to its popularity and its knowledge-oriented nature. As shown in prior works, external knowledge significantly aids LLMs in understanding questions in this dataset, whereas other QA tasks do not benefit as much from RAG systems [14].

2.3. External Database

RAG systems assume external databases are teachers or assistants helping generation models mitigate hallucinations. Therefore, the quality of the external database directly impacts the effectiveness of RAG systems. If the retrieved documents are of poor quality, or more intuitively, if the external database is less knowledgeable than the model itself, it may not provide positive assistance. In this project, we use the Wikipedia dump introduced by DPR in late 2018, which contains up to $\sim$ 3.5 billion tokens [18]. These documents are processed into chunks for retrieval. To examine the scaling behaviour, we randomly select subsets of the database, ranging from $\sim$ 3.5 million tokens to $\sim$ 3.5 billion tokens, distributed logarithmically across 32 points.

2.4. Evaluation Metrics

We evaluate our RAG system's performance on the QA task with three metrics: exact match (EM), F1-score, and accuracy. The exact match (EM) metric calculates the percentage of predictions that exactly match the ground-truth answers. Let N denote the total number of questions, and C_{EM} represent the count of exact matches. The formula is $\text{EM} = \frac{C_{\text{EM}}}{N} \times 100\%$. The F1-score measures the overlap between the generated text and the ground-truth answer as $\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$. Finally, accuracy measures the proportion of correctly answered questions, where the generated text contains the ground truth: $\text{Accuracy} = \frac{C_{\text{correct}}}{N} \times 100\%$.

While EM is stricter, F1-score and accuracy credit to partially correct answers, offering a more comprehensive assessment of the generated text.

2.5. Computational Platform

All experiments in this work were conducted on the NSCC [2] supercomputing platform. Computational tasks were submitted to a node equipped with 128GB of system RAM and up to four NVIDIA A100 40 GB PCIe GPUs.

3. Results

The experiment results are shown in Fig. 2. We have the following three key findings: **(a)** Turning point thresholds exist in RAG systems, beyond which they effectively mitigate hallucinations in LLMs. **(b)** RAG provides greater compensation for less powerful models compared to more advanced ones. **(c)** RAG system performance increases linearly as the corpus size grows exponentially, and we have yet to observe this trend slowing down. The details of these findings are as follows.

3.1. The Turning Point

The turning point thresholds are identified at the intersection of the dashed and solid lines of the same colour for each model, respectively. Before this point, the retrieved documents, instead of providing rich contextual information, tend to confuse the models even more significantly. This can be intuitively explained by the assumption that a teacher should be more knowledgeable than a student; however, before the turning point, the external knowledge source does not meet our assumption to teach the LLMs. That is, it is analogous to providing a high school textbook to college students for their open-book examinations.

Additionally, we obaserve that the less powerful model is more adversely affected prior to the turning point, as they lack the capability to determine whether the provided documents are beneficial.

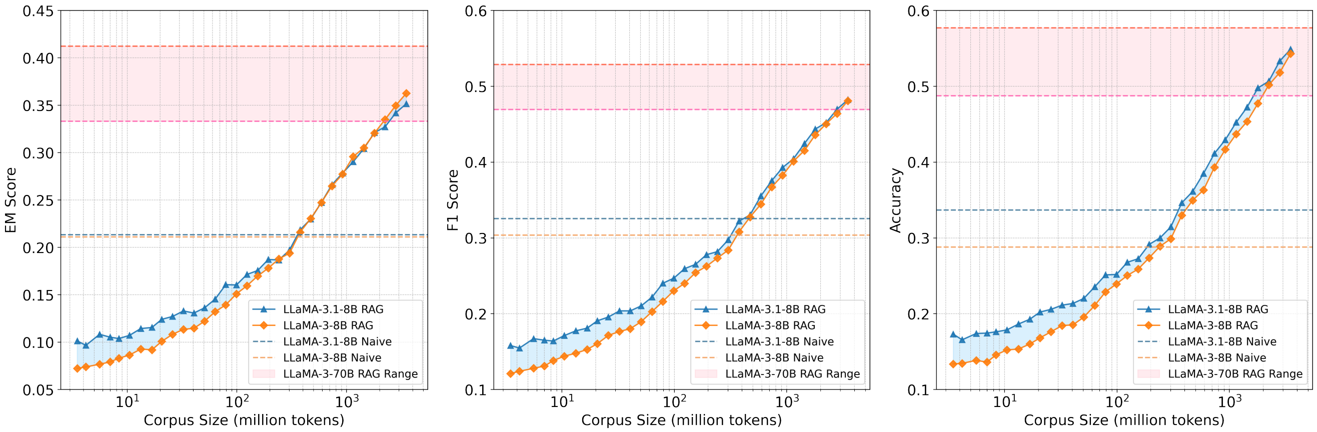
This finding highlights the importance of establishing large-scale, high-quality knowledge bases for RAG systems, as well as optimizing the construction processes to accommodate even larger knowledge bases.

3.2. The Max of the Two

We observe that without RAG, the LLaMA-3.1 model consistently outperforms LLaMA-3 across all metrics, particularly in accuracy. It is expected that the LLaMA-3.1 model would maintain this advantage throughout all experiments, given its stronger capability to derive insights from retrieved documents. However, the results show that immediately after the turning point threshold, their performance becomes comparable in most cases. This counterintuitive finding can be explained by the hypothesis that the best achievable performance is $\textbf{Max}(\text{LLM, Database})$.

Moreover, for the EM score, LLaMA-3.1 performs even worse when the corpus size is large. As the corpus size increases, the more powerful model becomes overly cautious. When conflicts arise between the retrieved documents and the model's internal knowledge, the model generates more critical and lengthy answers. An ex-

[2] https://www.nscc.sg/



(a) EM score

(b) F1 score

(c) Accuracy

Figure 2. Scaling behaviour of external databases in RAG. All models used are from the `Instruct` series. The blue and orange dashed lines represent naive generation (vanilla LLM QA) as empirical lower bounds, while the red zone corresponds to results obtained using LLaMA-3.1-70B with the largest corpus size (top) and its naive generation (bottom).

ample is that LLaMA-3 tends to provide a straightforward answer, whereas LLaMA-3.1, despite being prompted to avoid additional words, often generates responses in the form, `"According to the documents, the answer is...` when the documents conflict with its own knowledge. Although these answers include the correct information, they also contain redundant explanations. This qualitative observation is supported by the discrepancy between EM score and accuracy. While EM penalises such critical responses by assigning a score of 0, accuracy credits them for containing the correct answer, even if additional, redundant details are provided.

Therefore, based on our findings, we conclude that the best way to maximise performance is to tune the model complexity and the corpus size simultaneously. This approach is similar to the philosophy of jointly optimising height, width, and depth in convolutional neural networks (CNNs) as demonstrated in EfficientNet [19].

3.3. The Limit Has Yet to Be Achieved

As shown across all three metrics, the trend reveals a linear relationship between the score and the corpus size on a logarithmic scale. When the corpus size is sufficiently large, we observe that even a simple 8B RAG system can outperform the naive generation of LLaMA-3.1-70B. This suggests that the system has not yet reached a convergence zone. An immediate question arises: can the performance of a RAG system with the LLaMA-3.1-70B model be matched or even surpassed by further increasing the corpus size of a simple 8B RAG system? Unfortunately, we are unable to answer this due to computing resource constraints.

However, we identify a promising pathway: applying our vanilla linear dependency scaling law to RAG external databases to develop a scalable RAG system. Such a system would exhibit predictable behaviour, as supported by our empirical findings.

4. Conclusion

Through this work, we observe that a certain scaling law exists in RAG systems, and their performance is empirically predictable. Rather than proposing a technical solution for specific tasks, this work provides empirical evidence that scaling laws remain applicable in the realm of RAG. Given that training LLMs or embedding models is inherently computationally intensive, constructing larger external databases to assist existing models is a more resource-efficient approach. Future efforts to enhance RAG performance may not need to focus solely on developing even larger language models but can instead prioritise optimising and expanding external knowledge bases. On the other hand, this work is limited by computational resources, which constrain the ability to fully explore the precise form of the scaling law as an empirical mathematical equation. Further research could investigate the scaling behaviour of models with larger numbers of parameters, across a wider range of tasks, and with corpus sizes on the order of trillions of tokens.

Acknowledgement

As a disclosure, the authors discussed this project with Prof. Ziwei Liu. We are grateful for Prof. Liu's insights on the interrelation between deep learning for data science (`DL4DS`) and data science for deep learning (`DS4DL`), highlighting their strong correlation and interdisciplinary nature.

References

[1] J. Kwapień and S. Drożdż, "Physical approach to complex systems," *Physics Reports*, vol. 515, no. 3-4, pp. 115–226, 2012. 1

[2] J. Li, J. Zhang, W. Ge, and X. Liu, "Multi-scale methodology for complex systems," *Chemical engineering science*, vol. 59, no. 8-9, pp. 1687–1700, 2004. 1

[3] V. V. Gobre and A. Tkatchenko, "Scaling laws for van der waals interactions in nanostructured materials," *Nature communications*, vol. 4, no. 1, p. 2341, 2013. 1

[4] O. Hahn and T. Abel, "Multi-scale initial conditions for cosmological simulations," *Monthly Notices of the Royal Astronomical Society*, vol. 415, no. 3, pp. 2101–2121, 2011. 1

[5] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023. 1

[6] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017. 1

[7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020. 1

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020. 1

[9] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, *et al.*, "Scaling laws for autoregressive generative modeling," *arXiv preprint arXiv:2010.14701*, 2020. 1

[10] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022. 1

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 1

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. 1

[13] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," *arXiv preprint arXiv:2405.07437*, 2024. 1

[14] J. Jin, Y. Zhu, X. Yang, C. Zhang, and Z. Dou, "Flashrag: A modular toolkit for efficient retrieval-augmented generation research," *arXiv preprint arXiv:2405.13576*, 2024. 2, 3

[15] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019. 2

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 2

[17] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: a benchmark for question answering research," *Transactions of the Association of Computational Linguistics*, 2019. 3

[18] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 6769–6781, Association for Computational Linguistics, Nov. 2020. 3

[19] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 09–15 Jun 2019. 4