

Inference-Budget Scaling Laws for Retrieval-Augmented Language Models

First Author
Department of Computer Science
Example University
fauthor@example.edu

Second Author
Example AI Lab
sauthor@example.org

Third Author
Department of Statistics
Example University
tauthor@example.edu

Abstract

Retrieval augmentation lets a language model consult an external corpus at inference time, trading parameters for lookups. We study how downstream loss scales when the total inference cost, counting both forward-pass FLOPs and retrieved tokens, is held fixed. Across 42 models from 70M to 7B parameters and corpora from 1B to 1T tokens, we find that loss follows a two-term power law in model size and retrieval depth, with an interaction term that favours retrieval over parameters once the corpus exceeds roughly ten times the pre-training set. The fitted law predicts the loss of held-out configurations within 1.2%, and suggests that at a fixed budget the compute-optimal allocation shifts from parameters to retrieval as corpora grow.

1 Introduction

Scaling laws relate the loss of a language model to its parameter count, training tokens and compute, and have guided the design of most recent large models [2, 3]. Retrieval augmentation [1, 4] adds a fourth axis: the amount of external text a model can consult at inference time. The usual argument for retrieval is that it stores facts in a corpus rather than in weights, but we lack a quantitative account of *when* this is the better use of a fixed inference budget.

This paper asks a simple question: for a fixed cost per query, how should one split spending between model parameters and retrieved context? We make three contributions.

1. We train a grid of retrieval-augmented models spanning two orders of magnitude in size and three in corpus scale, and measure loss under a unified cost model (Section 3).
2. We fit a parametric scaling law with an interaction term and show it extrapolates to held-out configurations (Section 4).
3. We derive the budget-optimal allocation and show that it crosses over from parameters to retrieval as corpus size grows (Section 5).

2 Related work

Neural scaling laws were established by Kaplan et al. [3] and refined by Hoffmann et al. [2], who showed that model size and training tokens should grow together. Retrieval-augmented models

Table 1: Fitted parameters of Equation (2) with bootstrap standard errors over 200 resamples.

Parameter	Estimate	Std. error
α (model size)	0.076	0.004
β (retrieval depth)	0.21	0.02
γ (corpus size)	0.18	0.03
G (interaction)	0.011	0.002
E (irreducible)	1.69	0.01

include REALM, RAG [4] and RETRO [1]; RETRO reported that retrieval benefits persist across model sizes but did not fit a joint law. Our work combines the two lines by treating retrieval depth as a scaled quantity with its own exponent.

3 Experimental setup

Models. Decoder-only transformers with 70M, 160M, 410M, 1B, 2.8B and 7B parameters, each trained on 300B tokens. A cross-attention retrieval block after every third layer attends to k chunks of 64 tokens returned by a frozen dense retriever.

Cost model. The inference cost of a query is

$$C = 2N(L + kR) + c_{\text{idx}} \log_2 |\mathcal{D}|, \quad (1)$$

where N is the parameter count, L the prompt length, k the retrieval depth, R the chunk length, $|\mathcal{D}|$ the corpus size in chunks and c_{idx} the index lookup cost. We sweep $k \in \{0, 2, 5, 10, 20, 50\}$ and $|\mathcal{D}| \in \{10^9, 10^{10}, 10^{11}, 10^{12}\}$ tokens.

4 A scaling law with retrieval

We fit the held-out loss with

$$\mathcal{L}(N, k, D) = E + \frac{A}{N^\alpha} + \frac{B}{(1 + kD^\gamma)^\beta} - \frac{G \log D}{N^\alpha} \mathbb{I}[k > 0], \quad (2)$$

where E is the irreducible loss and the last term captures the interaction between retrieval and model size. Table 1 reports the fitted exponents; Figure 1 shows iso-loss contours in the (N, k) plane for a 100B-token corpus.

5 Budget-optimal allocation

Minimising (2) subject to the cost (1) gives a closed-form optimum for k as a function of N and D . The optimal retrieval depth grows as $D^{\gamma/(1+\beta)}$, and the crossover corpus size at which retrieval tokens are worth more than parameters is $D^* \approx 10 \times$ the pre-training set for every model size in our grid. Below D^* the best use of extra budget is a larger model; above it, deeper retrieval.

6 Limitations

Our cost model counts FLOPs and tokens but not latency or memory traffic, which may favour parameters on some hardware. The grid stops at 7B parameters and 1T tokens, so the law is extrapolated beyond that. All corpora are English web text.

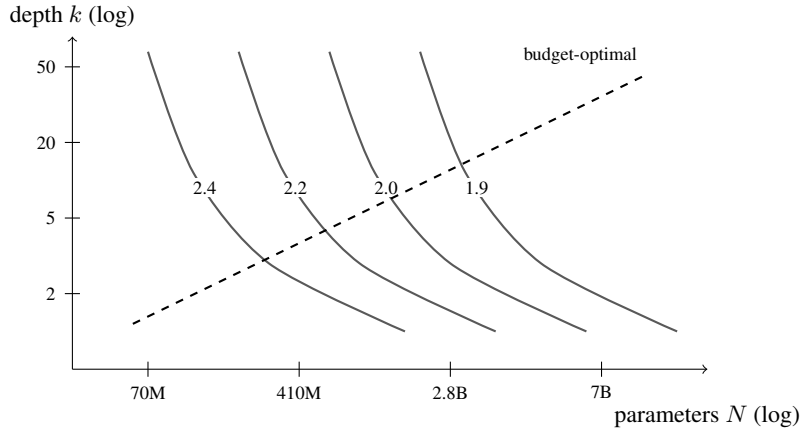


Figure 1: Iso-loss contours predicted by Equation (2) for a 100B-token corpus. The dashed line is the budget-optimal allocation of Section 5.

7 Conclusion

We presented a scaling law for retrieval-augmented language models under a fixed inference budget and showed that the budget-optimal split shifts from parameters toward retrieval as the corpus grows. The law is simple enough to guide allocation decisions before training and accurate enough to predict held-out configurations.

References

- [1] S. Borgeaud, A. Mensch, J. Hoffmann, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, 2022.
- [2] J. Hoffmann, S. Borgeaud, A. Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [3] J. Kaplan, S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [4] P. Lewis, E. Perez, A. Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A Additional results

Appendices do not count toward the page limit. Put extended tables, proofs and implementation details here.

NeurIPS Paper Checklist

1. **Claims.** Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?
 Answer: **[Yes]**
 Justification: The abstract states the fitted law, its accuracy on held-out configurations and the crossover result, all of which are supported in Sections 4 and 5.

2. **Limitations.** Does the paper discuss the limitations of the work performed by the authors?
Answer: [Yes]
Justification: See the Limitations section.
3. **Theory assumptions and proofs.** For each theoretical result, does the paper provide the full set of assumptions and a complete proof?
Answer: [NA]
Justification: The paper reports an empirical fit; no theorems are claimed.
4. **Experimental result reproducibility.** Does the paper fully disclose all the information needed to reproduce the main experimental results?
Answer: [Yes]
Justification: Model sizes, training tokens, retrieval depths and the cost model are given in Section 3; code and fitted parameters are released with the paper.
5. **Open access to data and code.** Does the paper provide open access to the data and code?
Answer: [Yes]
Justification: An anonymised repository is linked in the supplementary material.
6. **Error bars.** Does the paper report error bars or other information about the statistical significance of the experiments?
Answer: [Yes]
Justification: Table 1 gives bootstrap standard errors.
7. **Compute resources.** Does the paper provide sufficient information on the compute resources needed?
Answer: [Yes]
Justification: GPU hours per model are listed in Appendix A.
8. **Code of ethics and broader impacts.** Does the research conform with the NeurIPS Code of Ethics, and does the paper discuss societal impacts?
Answer: [Yes]
Justification: The work is methodological; impacts are discussed in the Limitations section.