Beyond Scale: Performance Plateau in Retrieval-Augmented Generation for Discriminative Question Answering Tasks

Mateusz Płoskonka[1]

[1]Prague University of Economics and Business, Czech Republic

Abstract

Retrieval-Augmented Generation (RAG) is widely used to provide Large Language Models (LLMs) with external evidence, yet the relationship between generator model size and RAG system performance remains poorly understood. This paper presents a systematic empirical analysis of scaling dynamics in RAG systems for discriminative question answering, evaluating 32 LLMs (0.6B–1T parameters) across three RAG architectural configurations using the Belebele benchmark. The results show a statistically significant performance plateau beginning at approximately 22B parameters (segmented regression: $R^2 = 0.807$, $p < 0.001$), with performance stabilising by 30B and a 99.8% post-plateau slope reduction. Model release date does not add detectable explanatory power after controlling for parameter count ($p = 0.674$). A capability-density analysis shows that the first 22B parameters capture 68.4% of total performance gains while representing only 2.2% of the evaluated parameter range. These findings suggest that, for discriminative RAG tasks, plateau-region models can provide comparable accuracy at substantially lower computational cost.

Keywords

Large Language Models, Retrieval-Augmented Generation, Scaling Laws, Performance Plateau, Capability Density, Model Selection, Benchmark Evaluation

1. Introduction

Large Language Models (LLMs) have been continuously transforming industries, education, and daily life since the public release of ChatGPT in November 2022. The rapid proliferation of LLM-powered systems has exposed two fundamental limitations: the inability to stay current with newly generated information, and the requirement to operate with private or domain-specific knowledge unavailable during pre-training. To address these challenges, contemporary AI architectures incorporate Retrieval-Augmented Generation (RAG) [1], introduced by Lewis et al. in 2020 as a framework that enriches prompts with externally retrieved information.

Since its introduction, RAG has evolved into a framework that combines model knowledge with retrieved external documents. A search of arXiv records for RAG-related terms returned over 1,200 papers published in 2024, reflecting rapid growth of research interest across domains such as medicine, education, software engineering, finance, and legal systems. Despite this extensive body of work, a fundamental design question remains unresolved: how does generator model scale affect RAG system performance?

This study hypothesizes that discriminative question answering requires only moderate reasoning capability. When relevant context is provided, the task shifts from knowledge recall to reading comprehension, which appears to saturate at moderate model scales. Consequently, models beyond a certain parameter threshold provide limited additional benefit. Recent surveys identify scaling behaviour in RAG systems as an open research question [2, 3], motivating the investigation presented here.

This paper makes three principal contributions:

1. Empirical characterisation of a statistically significant performance plateau in RAG systems at approximately 22–30B parameters, validated through three complementary statistical methods: segmented regression, Chow test, and 5-fold cross-validation.

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia

plom04@vse.cz (M. Płoskonka)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/paper1.pdf

2. Demonstration that model release date does not significantly predict RAG performance once parameter count is controlled ($p = 0.674$), challenging the assumption that newer models inherently outperform predecessors in retrieval-augmented contexts.
3. A capability density framework quantifying the 96.3× efficiency differential between pre- and post-plateau parameter regimes, supporting cost-aware model selection for RAG deployments.

2. Related Work

This work intersects three established research areas: scaling laws for language models, RAG benchmarking, and performance saturation. While each area has been studied independently, their intersection—particularly for end-to-end RAG systems—remains largely unexplored.

Scaling laws. Kaplan et al. [4] established that language model performance follows predictable power-law relationships with model size, dataset size, and compute budget. Hoffmann et al. [5] refined these findings through the Chinchilla model, demonstrating that compute-optimal training requires scaling parameters and training tokens proportionally. These seminal works established the expectation that larger models consistently outperform smaller ones—an assumption this study challenges in the specific context of retrieval-augmented discriminative QA. Xiao et al. [6] introduced the "densing law," showing that equivalent capability can be achieved with exponentially fewer parameters over successive model generations. The capability density framework introduced here provides a complementary cross-sectional perspective on how efficiency varies across the parameter spectrum at a fixed point in time.

RAG benchmarks and scaling. Gao et al. [2] survey RAG systems and describe open questions in RAG evaluation and development. Zhu et al. [3] survey how LLMs are used across information retrieval components, providing broader context for studying the interaction between retrieval and generation. Chen et al. [17] introduce RGB, a benchmark for diagnosing RAG-specific abilities such as noise robustness and information integration. Yue et al. [7] study inference-time scaling for long-context RAG. Wang et al. [18] show that augmenting RAG with application-aware examples can improve task performance, highlighting that pipeline design can matter beyond generator choice. Saturation-like effects have surfaced incidentally in prior work: Zhang et al. [19] report it as a motivating observation for an adaptive retrieval-triggering method, Lin et al. [20] observe it in a setting that conflates model scale with an instruction-tuning intervention, and Wang et al. [21] present a related ablation within a reinforcement-learning framework for long-form RAG. In each case the phenomenon is secondary to the paper's main contribution. The present work differs by isolating generator parameter count and testing whether larger generators continue to improve discriminative RAG performance once retrieved context is provided, treating this as a primary, systematically characterised research question rather than a secondary observation.

Performance saturation. Hackenburg et al. [8] demonstrated sharply diminishing returns beyond moderate scale for language model persuasiveness. Bahri et al. [9] provided theoretical explanations for why neural scaling laws must eventually break down. Fang et al. [16] established scaling laws for dense retrieval, demonstrating power-law relationships with respect to retriever model size. Despite this growing evidence, prior work has studied RAG benchmarks, inference-time scaling, and retrieval scaling, but parameter-count saturation in end-to-end discriminative RAG remains underexplored.

3. Methods

3.1. LLM Selection

A total of 32 distinct models were selected, satisfying two criteria: (1) the developer openly published parameter count, and (2) licensing permits self-hosted deployment for research. The selection spans 0.6B to 1T parameters and release dates from September 2023 to August 2025 (Table 1), encompassing multiple

Table 1
Large Language Models evaluated

Range	Models	Developers
< 10B	Qwen-3-0.6B; Gemma-3-1B-Instruct; Llama-3.2-1B-Instruct; Falcon-3-1B-Instruct; Gemma-2-2B-Instruct; Llama-3.2-3B-Instruct; Falcon-3-3B-Instruct; Qwen-3-4B; Gemma-3-4B-Instruct; Mistral-7B-Instruct-v0.1; Falcon-3-7B-Instruct; Llama-3-8B-Instruct; Llama-3.1-8B-Instruct; Qwen-3-8B; Gemma-2-9B-Instruct	Google, Meta, Mistral, Qwen, Tiiuae
10–30B	Falcon-3-10B-Instruct; Mistral-Small-Instruct; Gemma-2-27B-Instruct; Gemma-3-27B-Instruct	Google, Mistral, Tiiuae
30–100B	Qwen-3-32B; Llama-3.1-70B-Instruct; Llama-3-70B-Instruct; Llama-3.3-70B-Instruct	Meta, Qwen
≥ 100B	Gpt-oss-120B; Mistral-Large-Instruct-2407; Qwen-3-235B-A22B-Instruct; Llama-3.1-Nemotron-Ultra-253B-v1; GLM-4.5; Jamba-Large-1.5; Llama-3.1-405B-Instruct-FP8; Hermes-4-405B; Kimi-K2-Instruct	AI21, Meta, Mistral, Moonshotai, NousResearch, Nvidia, OpenAI, Qwen, Zaiorg

Note: Release dates span Sep 2023–Aug 2025. For MoE architectures, total parameter count is reported.

model families, dense and mixture-of-experts (MoE) architectures, and diverse training approaches. For MoE architectures, total parameter count is reported to match developer specifications.

3.2. Dataset and RAG Architectures

The Belebele benchmark [10] was used for evaluation, restricted to its English-language variant: 900 multiple-choice reading-comprehension questions with four response options per language variant, each based on short passages from FLORES-200, providing standardised discriminative QA evaluation. Three context-provision configurations measure how model size affects performance under different levels of retrieval quality:

- **RAG-FAISS:** Wikipedia pages chunked and retrieved at inference time via FAISS [11] cosine similarity. The only configuration performing retrieval during inference.
- **RAG-FULL WIKI:** The complete article provided directly—an oracle configuration eliminating retrieval error while retaining long-context noise.
- **RAG-CONTEXT:** A manually curated excerpt guaranteed to contain the correct answer—an idealised upper bound isolating pure generator comprehension.

RAG-AVG is the arithmetic mean across all three configurations. The objective of including all three configurations is to measure how model size affects performance under different levels of retrieval quality.

3.3. Evaluation Methodology

The Belebele dataset has a mildly non-uniform answer distribution (Answer 1: 206; Answer 2: 251; Answer 3: 246; Answer 4: 197, representing 21.9%–27.9% per class). Accuracy is defined as

$$\text{Acc} = \sum_i \text{TP}_i \;/\; N_{\text{total}}$$

where $N_{\text{total}} = 900$ is the total number of questions and TP_i is the number of correct predictions for answer label $i \in \{1, 2, 3, 4\}$.

Because the label imbalance is small, balanced accuracy was also computed and yields the same qualitative ordering across all 32 models.

Temporal effects were assessed via nested ANCOVA: a parameters-only model vs. parameters + release-quarter, with F-test on residual sums of squares. A non-significant F-test indicates that release quarter did not add detectable explanatory power beyond parameter count in this sample.

Performance plateau detection used three complementary methods: (1) Segmented regression [12] fitting piecewise linear functions across candidate breakpoints; (2) Chow test [13] scanning 126 candidate breakpoints (1–1000B)—because multiple breakpoints are tested, individual p-values are interpreted descriptively and convergence with the other two methods serves as the primary evidence; and (3) 5-fold cross-validation [14] selecting the breakpoint maximising held-out R^2. Consistent results across all three methods provide supporting evidence for a genuine plateau.

Capability density was defined as

$$D(x_1, x_2) = \frac{\Delta y}{\Delta x} = \frac{y(x_2) - y(x_1)}{x_2 - x_1}$$

measuring accuracy improvement per billion parameters. This metric quantifies the economic efficiency of parameter scaling across defined ranges.

4. Results

4.1. Impact of Model Size on RAG Performance

To quantify response quality across four architectural configurations (native LLM and three RAG variants), each model was evaluated using the Belebele dataset. Figure 1 illustrates evaluation outcomes for each model, with parameter count on the x-axis and accuracy on the y-axis. Figure 1 suggests a saturation pattern characterized by rapid gains at smaller scales followed by a performance plateau.

ANOVA across five size categories ($N = 32$) reveals highly significant differences ($F(4, 27) = 36.55$, $p < 0.001$), with mean RAG-AVG accuracy increasing from 0.526 (Small, 0–4B) to 0.860 (Ultra Large, 250B+)—a 33.4 percentage point improvement (Table 2). Within-category variance decreases 4.3-fold (σ: 0.094 → 0.022), indicating convergence at scale. Effect sizes confirm substantial transitions in the small-to-medium range (Cohen's $d = 2.66$ for Small→Medium, $d = 1.79$ for Medium→Large), while transitions above 22B yield only medium effects ($d = 0.53$–0.61), consistent with plateau behaviour.

Table 2
RAG-AVG performance by model size category

Category	N	Mean	Std	Min	Max
Small (0–4B)	7	0.526	0.094	0.374	0.626
Medium (4–22B)	9	0.734	0.058	0.654	0.788
Large (22–70B)	4	0.824	0.042	0.767	0.867
Very Large (70–250B)	6	0.844	0.031	0.805	0.878
Ultra Large (250B+)	6	0.860	0.022	0.821	0.880

Note: ANOVA $F(4, 27) = 36.55$, $p < 0.001$. Variance decreases 4.3× from Small to Ultra Large.

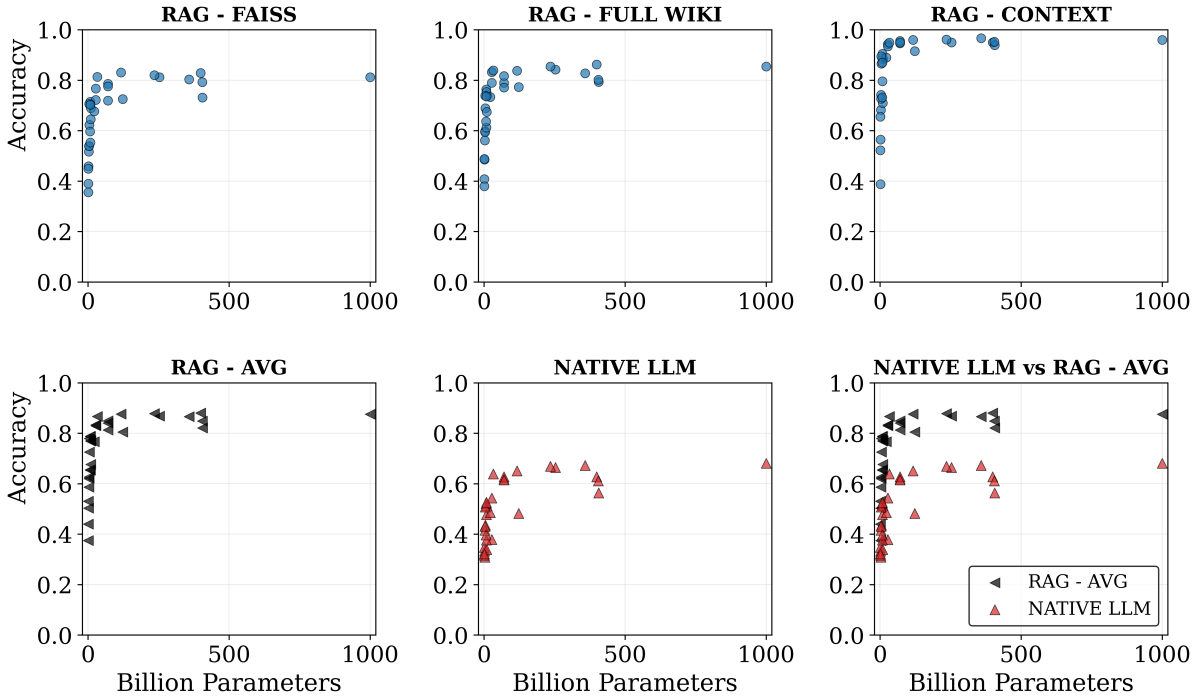


Figure 1: Performance comparison of RAG-enhanced and native LLM systems across model sizes. Top row shows RAG configurations (FAISS, Full Wiki, Context). Bottom row displays RAG-AVG, native LLM performance, and their direct comparison.

Incremental efficiency analysis across 31 consecutive model transitions reveals substantial decline in marginal returns: early-stage scaling (0–10B) yields a mean of 2.35 percentage points (pp) per billion parameters, while late-stage scaling (100B+) yields −0.56 pp/B. Some transitions in the very large range exhibit negative marginal returns, indicating that model-specific factors can outweigh the benefit of additional parameters at extreme scale.

4.2. Temporal Effects of Model Generation

Models were stratified into quarterly release cohorts (Q2 2024–Q2 2025, $n = 27$; cohorts with fewer than four models excluded). Nested ANCOVA comparison shows adding release quarter as a covariate produces non-significant improvement over the size-only model ($F(5, 20) = 0.64$, $p = 0.674$, $\Delta R^2 = 0.039$). AIC and BIC favour the simpler model ($\Delta\text{AIC} = +6.01$, $\Delta\text{BIC} = +12.49$). Power-law fits per release quarter show similar scaling exponents ($b = 0.060$–0.106), supporting the interpretation that temporal effects are limited in this dataset.

This finding does not imply stagnation in LLM development. Advances in reasoning, alignment, and instruction-following are well-documented but did not measurably improve the use of retrieved context in this discriminative QA setting. Within the evaluated cohorts, parameter count explains RAG performance more reliably than release quarter.

4.3. Performance Saturation Threshold

Segmented regression identifies a statistically significant breakpoint at 22B parameters ($R^2 = 0.807$ vs. 0.214 for linear; $F(2, 28) = 42.95$, $p < 0.001$). Pre-plateau slope $\beta_1 = 0.0306$ (95% CI: [0.018, 0.044]); post-plateau $\beta_2 = 0.00006$—a 99.8% slope reduction, consistent with logarithmic rather than power-law scaling beyond the threshold. $\Delta\text{AIC} = -40.9$ strongly favours the segmented model.

The Chow test identifies maximum structural discontinuity at 4B ($F(2, 28) = 49.96$, $p < 0.001$), with sustained structural breaks through 22B ($F = 42.95$, 86% of maximum), indicating a transition zone

spanning 4–22B rather than a single discrete threshold. All tested breakpoints from 2B to 70B are highly significant ($p < 0.001$).

5-fold cross-validation confirms optimal generalisation at 12B ($R^2 = 0.571$, +1.43 over linear baseline $R^2 = -0.858$), falling within the 4–22B transition zone. Performance degrades monotonically beyond 12B, with negative R^2 beyond 50B breakpoints. The convergence of all three methods on the 10–30B range provides consistent evidence for a transition from rapid gains to a plateau.

4.4. Capability Density

Capability density analysis quantifies the economic efficiency of parameter scaling. The 0–22B range, representing only 2.2% of total parameter space (22B of 999B), captures 68.4% of total performance improvement, with average density of 0.0107 accuracy units per billion parameters. The post-plateau regime (22B–1T) achieves only 31.6% of total improvement across 97.8% of parameter space, with density reduced to 0.0001—a 96.3-fold efficiency differential (Table 3).

Table 3
Capability density across parameter ranges

Range (B)	Perf. Gain	Span (B)	Density (Δ/B)	% Total	Rel. Eff.
0–22 (Total)	0.236	21.2	0.0107	68.4%	96.3×
0–4	0.186	2.2	0.0835	54.0%	1.00×
4–10	−0.103	5.2	−0.0197	—	—
10–22	0.153	13.8	0.0111	44.3%	13.3×
22–1000 (Total)	0.109	977.8	0.0001	31.6%	0.01×
22–30	0.067	5.4	0.0123	19.3%	11.1×
70–200	−0.036	52.4	−0.0007	—	—

Note: Negative values reflect model-to-model variability; segmented regression confirms theoretical pre-plateau slope of 0.0159 declining to 0.000039 post-plateau (99.8% reduction).

5. Discussion

Performance saturation. Three orthogonal statistical methods converge on a plateau in the 10–30B range. The segmented model explains nearly four times more variance than a linear model. The 99.8% post-plateau slope reduction is consistent with logarithmic rather than power-law scaling, diverging from classical scaling law behaviour [4]. This may reflect that RAG transforms the task from parametric knowledge recall to reading comprehension—a capability saturating at moderate scale [15]. A ceiling effect may also contribute: best-performing models reach accuracy ≈ 0.88, notably below reported human performance on Belebele (>97% [10]). This gap may partly reflect irreducible noise in the benchmark (e.g., ambiguous distractors or annotation artifacts), but it may equally reflect genuine limitations in generator reading comprehension; the present data cannot distinguish between these explanations, and we do not attribute the gap fully to noise.

Temporal invariance. Non-significance of release quarter ($p = 0.674$) does not imply stagnation—advances in reasoning and alignment are well-documented—but suggests these improvements do not materially enhance context utilisation for discriminative RAG. This interpretation aligns with the view that context utilisation is a circumscribed capability saturating independently of broader language model abilities. Practitioners should evaluate model improvements specifically on RAG-relevant benchmarks rather than extrapolating from general-purpose leaderboards.

Practical implications. The consistent ordering RAG-CONTEXT > RAG-FULL WIKI > RAG-FAISS across all 32 models indicates retrieval quality exerts greater influence than generator scale beyond moderate thresholds. These results suggest that resources may be better directed toward knowledge-base curation and retrieval refinement than toward deploying larger generators. Selection of plateau-region models (22–30B) can yield 50–60% cost reduction with negligible accuracy degradation. We note that all

models in the 10–30B range evaluated here (Falcon-3-10B, Mistral-Small, Gemma-2-27B, Gemma-3-27B) are dense architectures, so their inference cost scales directly with parameter count; MoE models in this study appear only above 100B, where active rather than total parameter counts should be used to assess comparable cost savings. Given that inference costs scale approximately linearly with parameter count, this represents a substantial production saving.

Limitations. Findings apply specifically to discriminative RAG and should not be extrapolated to free-form generation requiring complex reasoning or synthesis. The sample is non-uniform above 400B, limiting plateau characterisation at extreme scale. Architectural heterogeneity (dense vs. MoE) introduces potential confounds. The oracle RAG-CONTEXT configuration represents an idealised scenario; the plateau observed may not fully generalise to realistic retrieval where generators must handle imperfect context. Additionally, this study evaluates RAG only in a discriminative multiple-choice setting; other common RAG use cases, such as extracting or filling structured fields from free-text sources, may exhibit different scaling behaviour and are not addressed here.

6. Conclusion

As LLMs continue to scale, their role inside retrieval-augmented systems needs to be evaluated separately from their general benchmark performance. This study shows that, for discriminative question answering with retrieved context, the relationship between generator size and RAG accuracy is strongly non-linear.

Across 32 models spanning 0.6B to 1T parameters, performance improves rapidly below approximately 22B but plateaus beyond this range, with consistent evidence from segmented regression ($R^2 = 0.807$), Chow breakpoint scanning (max $F = 49.96$), and cross-validation ($R^2 = 0.571$ at 12B). Post-plateau scaling yields a 99.8% slope reduction across all three RAG configurations tested.

For practitioners, the implication is practical: after a moderate generator size, further gains are more likely to come from retrieval quality, source selection, chunking, and knowledge-base curation than from deploying a larger model. Model selection for discriminative RAG should therefore consider parameter efficiency alongside accuracy, latency, and deployment cost.

While this study provides valuable insights into RAG scaling behaviour, its findings are limited by the use of a single benchmark. Future work should investigate whether the observed performance plateau generalises to generative question answering, multilingual settings, and domain-specific retrieval corpora.

Code and results are available from the authors upon request.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT (GPT-4o) in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style. After using these tool(s)/service(s), the author reviewed and edited the content as needed and takes full responsibility for the publication's content. All scientific claims, experimental results, bibliographic records, and statistical analyses were conducted and verified by the author.

References

[1] P. Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *NeurIPS*, 2020.

[2] Y. Gao et al., Retrieval-Augmented Generation for Large Language Models: A Survey, *arXiv*:2312.10997, 2023.

[3] Y. Zhu et al., Large Language Models for Information Retrieval: A Survey, *ACM Transactions on Information Systems*, 44(1), Article 12, pp. 1–54, 2025. doi:10.1145/3748304.

[4] J. Kaplan et al., Scaling Laws for Neural Language Models, *arXiv*:2001.08361, 2020.

[5] J. Hoffmann et al., Training Compute-Optimal Large Language Models, *arXiv*:2203.15556, 2022.
[6] C. Xiao, J. Cai, W. Zhao et al., Densing Law of LLMs, *Nature Machine Intelligence* 7, 1823–1833, 2025. doi:10.1038/s42256-025-01137-0.
[7] Z. Yue et al., Inference Scaling for Long-Context Retrieval Augmented Generation, *ICLR*, 2025, arXiv:2410.04343.
[8] K. Hackenburg, B. M. Tappin, P. Röttger et al., Scaling language model size yields diminishing returns for single-message political persuasion, *PNAS* 122(10): e2413443122, 2025. doi:10.1073/pnas.2413443122.
[9] Y. Bahri et al., Explaining Neural Scaling Laws, *PNAS* 121(27): e2311878121, 2024. doi:10.1073/pnas.2311878121.
[10] L. Bandarkar et al., The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants, *ACL 2024*, pp. 749–775. doi:10.18653/v1/2024.acl-long.44.
[11] M. Douze et al., The Faiss Library, *arXiv*:2401.08281, 2024.
[12] V. M. R. Muggeo, Estimating Regression Models with Unknown Break-Points, *Statistics in Medicine*, 22(19), pp. 3055–3071, 2003. doi:10.1002/sim.1545.
[13] G. C. Chow, Tests of Equality Between Sets of Coefficients in Two Linear Regressions, *Econometrica*, 28(3), pp. 591–605, 1960. doi:10.2307/1910133.
[14] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (2nd ed.), Springer, 2009.
[15] Y. Chang et al., A Survey on Evaluation of Large Language Models, *ACM Transactions on Intelligent Systems and Technology*, 15(3), Article 39, 45 pages, 2024. doi:10.1145/3641289.
[16] Y. Fang et al., Scaling Laws for Dense Retrieval, *SIGIR*, 2024. doi:10.1145/3626772.3657743.
[17] J. Chen et al., Benchmarking Large Language Models in Retrieval-Augmented Generation, *AAAI* 38(16):17754–17762, 2024. doi:10.1609/aaai.v38i16.29728.
[18] Y. Wang et al., RAG+: Enhancing Retrieval-Augmented Generation with Application-Aware Reasoning, *EMNLP 2025*, pp. 32013–32037. doi:10.18653/v1/2025.emnlp-main.1630.
[19] Z. Zhang et al., KBM: Delineating Knowledge Boundary for Adaptive Retrieval in Large Language Models, *arXiv preprint arXiv:2411.06207*, 2024.
[20] X. V. Lin et al., RA-DIT: Retrieval-Augmented Dual Instruction Tuning, *arXiv preprint arXiv:2310.01352*, 2023.
[21] Y. Wang et al., Reinforced Informativeness Optimization for Long-Form Retrieval-Augmented Generation, *arXiv preprint arXiv:2505.20825*, 2025.