

Understanding and Debugging Failures in N-Gram-Based Generative Retrieval

Richard Takacs
richard.takacs@s.wu.ac.at
Vienna University of Economics and
Business
Vienna, Austria

Adrian Bracher
adrian.bracher@wu.ac.at
Vienna University of Economics and
Business
Vienna, Austria

Svitlana Vakulenko
svitlana.vakulenko@wu.ac.at
Vienna University of Economics and
Business
Vienna, Austria

Abstract

Generative Retrieval (GR) is an emerging Information Retrieval (IR) paradigm that is motivated by increasingly capable language models. In GR, a model directly generates identifiers for relevant documents. While these systems offer unique advantages, they also introduce distinct failure mechanisms.

We explore these failure modes in three contributions: (1) We present a taxonomy of GR failure modes based on GR literature. (2) We empirically investigate failure in a subset of GR: ngram-based methods, more specifically, SEAL [5] and MINDER [18]. Our analysis reveals common issues, such as ambiguous docids, low identifier diversity, and the disproportionate impact of specific identifiers. (3) We introduce a new web-based tool that helps the IR community analyze generated ngrams and their respective contribution to the final ranking, providing an intuitive interface to identify where such GR methods go wrong.

CCS Concepts

• Information systems → Retrieval models and ranking.

Keywords

Generative Retrieval, Failure Modes

1 Introduction

In generative retrieval (GR) a language model is trained to directly generate document identifiers (docids), offering an alternative to the dominant dense bi-encoder paradigm. GR offers a unified, end-to-end trainable retrieval architecture [33]. However, this comes at the cost of distinct failure modes, many of which remain underexplored in current literature.

A key differentiator between GR models is the chosen document representation. Some models use atomic identifiers, like unique integers or cluster codes [31, 33]. Others use text-based identifiers, such as document titles [8] or unique token-wise substrings called ngrams [5, 18, 19]. In the empirical part of this work, we focus on ngram-based methods, in particular, SEAL [5] and MINDER [18], because ngrams allow for a nuanced, interpretable analysis. Unlike atomic IDs, ngrams are extracted from the document text. This allows us to trace errors back to specific substrings in the documents and attribute more clearly where a model makes a mistake. In the case of MINDER, the document text is augmented with pseudo-queries (PQs) as additional constrained decoding targets.

To better understand the failure modes of generative retrieval, this work offers three contributions: We first present a taxonomy of failure modes based on existing research. We then empirically evaluate SEAL and MINDER on two standard IR datasets: Natural

R Representation	T Training	I Inference	G Response Generation
Ambiguous Docids	Recency Bias	False Pruning	Synthesis Hallucination
Subword Tokenization	Forgetting	Docid Hallucination	Context Bottleneck
Static Assignment	Cultural Bias	Length Bias	Lost in the Middle
Metadata Over-reliance	Distribution Mismatch	Identifier Drift	Privacy Leakage
Lexical Mismatch	Distribution Mismatch	Low Diversity	Parametric / Contextual Conflict
Ngram Contiguity	Distillation Errors	High Latency	Citation Hallucination
	Capacity Limits	Score Concentration	
		Post-Rationalization	

Figure 1: Taxonomy of failure in generative retrieval systems. Failure modes empirically studied in Section 3 are highlighted in red.

Questions (NQ) [14] and MS-MARCO [28]. Finally, we introduce a web-based diagnostic tool to help researchers audit and improve similar retrieval models.

2 A Taxonomy of Failure Modes

To build our taxonomy, we reviewed recent literature on Generative Retrieval [1, 5, 6, 8, 13, 15–18, 21, 23–25, 30, 32, 38–43].

The taxonomy is shown in Figure 1. We organized the identified failure modes into four stages: *Representation*, *Training*, *Inference* and *Response Generation*.

2.1 Representation

Errors related to the representation of documents in docids.

- R1 **Ambiguous Docids:** Substrings or n-grams that are not unique to a relevant document, causing the model to retrieve topically related but irrelevant content [6, 41].
- R2 **Subword Tokenization:** Rare or domain-specific terms are often split into uninformative tokens. For example, this becomes an issue in constrained decoding when a prefix is shared between a related term that does not occur in the FM-Index and other, unrelated terms. For example, the model may decode “enc” which would result in “encephalitis” in the unconstrained setting, but only “encode” actually occurs in the corpus [42].
- R3 **Static Assignment:** This failure mode occurs when document identifiers are assigned to static IDs (titles, URLs, static numbers, etc.), that cannot be adapted to the retrieval task. Because these predetermined strings are not optimized for semantic consistency or retrieval relevance, they create a performance gap in the model’s parametric memory [31, 43].

- R4 **Metadata Over-reliance:** This failure mode occurs when a model relies too heavily on static, surface-level document attributes (such as titles or URLs) to serve as identifiers, rendering the system vulnerable if the metadata is poor, missing, or uniform [7].
- R5 **Lexical Mismatch:** While generative language models possess inherent semantic reasoning capabilities, they may still fail to bridge the lexical gap between user queries and document identifiers. In substring-based retrieval, the model is strictly constrained by the exact word orderings present in the corpus, potentially overscoring exact keyword matches while underscoring semantically relevant synonymous or paraphrased content [5, 22].
- R6 **Ngram Contiguity:** This failure mode is specific to substring-based architectures where a document’s relevant semantic terms are separated in the text by intervening tokens, such as sub-sentences. Because sequence-based constraints force the model to decode contiguous token paths, it cannot skip over irrelevant intermediate text [42].

2.2 Training

Biases that are a result of chosen datasets and model training.

- T1 **Recency Bias:** An excessive bias against older content, where the model tends to favor newly added documents from recent updates over the initial corpus. This phenomenon often occurs when models are sequentially fine-tuned on new batches of data without sufficient regularization to maintain the original distribution [43].
- T2 **Forgetting:** A model is no longer able to access previously learned information after model updates [23, 43].
- T3 **Cultural Bias:** Generative models often exhibit biases learned from their training data. For example, models trained on the English Wikipedia [14, 28] may suffer from Western-centricity. Models may default to specific interpretations in ambiguous cases [2, 10, 26, 27].
- T4 **Distribution Mismatch:** Supervised GR models often struggle with zero-shot generalization to unseen query distributions or datasets compared to robust lexical matching baselines [22, 43]. Performance can drop significantly when faced with specialized documentation or phrased differently than the training data [5, 29, 35].
- T5 **Distillation Errors:** If knowledge distillation via a cross-encoder model is used, the GR model may inherit and propagate teacher errors, forcing high probabilities on incorrect identifiers and distorting the identifier space [19, 20].
- T6 **Capacity Limits:** Generative retrieval models face challenges regarding parametric capacity as they scale to massive corpora. This failure mode is primarily found when docids are arbitrary and not related to semantics, forcing models to remember these mappings, eventually leading to a drop in retrieval effectiveness [29, 35, 38, 44].

2.3 Inference

Failures occurring during the identifier decoding process.

- I1 **False Pruning:** The model discards a correct identifier because its prefix score was lower than that of an irrelevant document [39, 40, 42].
- I2 **Docid Hallucination:** The model may generate identifiers that do not exist in the predefined corpus index. Constrained decoding (e.g., using FM-indexes or tries) is typically used to prevent this [5, 8, 9], although some systems utilize intentional hallucination for generating additional synthetic identifiers [18].
- I3 **Length Bias:** The tendency of scoring heuristics to disproportionately favor shorter identifiers over longer, more discriminative ones. Token probabilities are often used, which are multiplicative and therefore tend to shrink as the sequence length increases, leading the model to prefer generic docids (e.g., “War”) over more specific ones (e.g., “The Napoleonic Wars”) [29, 38, 39, 42].
- I4 **Identifier Drift:** Error rates often increase at later token positions in a long docid sequence [36, 38].
- I5 **Low Diversity:** The retrieved results may be topically redundant, failing to provide the diverse perspectives needed for ambiguous queries. Models often lack mechanisms to force the generation of distinct docids covering different facets of a query [7, 17, 31].
- I6 **High Latency:** Autoregressive generation of identifiers token-by-token is computationally intensive and scales poorly with beam size, leading to the application of failure-prone heuristics [12, 29, 38, 44].
- I7 **Score Concentration:** This failure mode occurs when the model relies on a few specific docids that dominate the scoring.
- I8 **Post-Rationalization:** The model may generate the answer, or what it believes to be the answer, directly and thereby post-rationalize parametric knowledge instead of faithfully applying the retrieved information [34].

2.4 Response Generation

We include the response generation stage for completeness, although this is an aspect of the document synthesis rather than the core document retrieval task.

- G1 **Synthesis Hallucination:** In the response generation stage, the LLM may hallucinate information not contained within the retrieved documents, relying instead on its pre-trained parametric knowledge [3, 11, 37].
- G2 **Context Bottleneck:** Generative systems often struggle with context window limitations when synthesizing answers from a large set of retrieved documents, leading to truncation or the ignoring of relevant evidence [1, 17].
- G3 **Lost in the Middle:** A phenomenon where LLMs tend to ignore documents or information placed in the middle of a retrieved context window, while over-weighting information at the beginning or end [4, 21].
- G4 **Privacy Leakage:** Generative models risk outputting sensitive data during the response synthesis process [17].
- G5 **Parametric / Contextual Conflict:** A critical failure occurs when the parametric knowledge directly conflicts with

and overrides the information found in the retrieved documents [1].

- G6 **Citation Hallucination:** The model may claim a specific document supports a statement when it does not or hallucinate a non-existing citation. [34].

3 Empirical Study

We evaluated two ngram-based models: SEAL [5] and MINDER [18], evaluated on NQ [14] and MS-MARCO [28].

We evaluated SEAL and MINDER on subsets of the *Natural Questions* (NQ) and MS-MARCO datasets. The NQ retrieval corpus consists of a Wikipedia dump split into passages of 100 tokens each and 6515 queries. The MS300k evaluation corpus contains web passages extracted from Bing search results alongside 6980 training queries. Both models were evaluated using their default configurations.

For each query, the solution files contain the ground-truth relevant passages (positive contexts) and the top-100 retrieved passages. Crucially, the file also includes the set of matched ngram keys for each retrieved passage, along with their corpus frequency and individual scores.

We focus on *Hits@1* and *Hits@10* as primary indicators of retrieval success, and analyze various settings in which these metrics are impacted.

3.1 Results

We present our findings, applying the taxonomy to categorize the specific failure modes observed during our evaluation.

Metadata Over-reliance (R4). We found that SEAL relies heavily on clean metadata. In the NQ dataset, document titles account for approximately 40% of the model’s total score mass. This dependency on title-based ngrams is built into SEAL’s architecture because it empirically improves retrieval on NQ. As a result, the model is less robust when metadata is noisy or missing.

Low Diversity (I5). In our analysis, we observed that higher token diversity in the generated identifiers actually correlates with lower retrieval success ($\rho = -0.289$ for MINDER), when comparing queries from the same dataset and method (see Figure 2). This suggests that token repetition is not a redundancy failure, but rather a signal of model certainty. When the model is confident in a target document, it generates multiple overlapping ngrams that concentrate score on a specific semantic cluster.

Ambiguous Docids (R1) / Length Bias (I3) / Subword Tokenization (R2). Our analysis indicates that approximately 85% of all generated document identifiers consist of single tokens (unigrams). As shown in Figure 3, a high reliance on these low-information unigrams correlates with lower ranking performance. While we have not isolated the exact underlying cause, we consider two potential mechanisms for this high unigram share. First, it may be due to subword tokenization artifacts (R2), where the language model generates short, high-probability tokens to minimize generation loss, but subsequently fails to extend these unigrams into longer sequences due to FM-Index constraints. Second, it could be due to length bias (I3), where monotonically decreasing language model scores over sequence length penalize longer identifiers. Because a

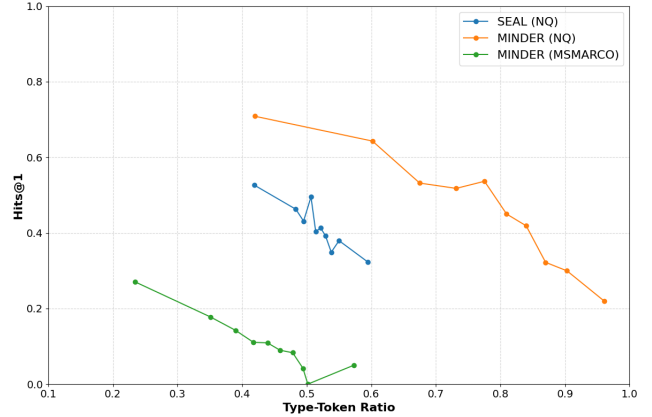


Figure 2: Distribution of Token Diversity in Top-Ranked Documents: We evaluate retrieval performance across varying levels of vocabulary diversity by binning passages into deciles based on their Type-Token Ratio, which is defined as the ratio of unique tokens to total tokens. We then report the average Hits@1 for each resulting group.

single token typically carries little meaning to differentiate specific passage content, these unigrams lead to ambiguous docids (R1).

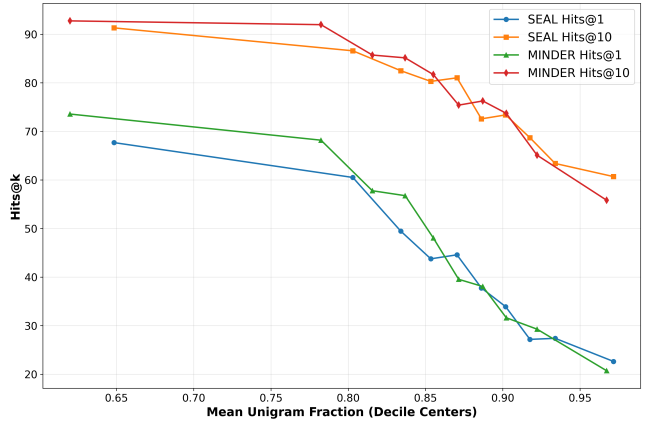


Figure 3: On the NQ dataset, the proportion of unigrams in generated n-grams is inversely correlated with retrieval ranking performance.

Ngram Contiguity (R6) / Lexical Mismatch (R5). In the MINDER framework, synthetic pseudo-queries (PQs) are used as additional source for ngram identifiers. We found that while PQs are sparse, they account for only 1.4% of all generated ngrams, but contribute 27.7% of the total score mass. This further indicates that document extension is an effective way to improve retrieval, partially mitigating the identified ngram contiguity and lexical mismatch issues.

Cultural Bias (T3). We also anecdotally observed how pre-trained knowledge can override a user’s intent. For example, given the query “where did the southern song have their capital”, SEAL failed

to retrieve any correct passages about 12th-century China. Instead, it retrieved passages about country music and the American South (e.g., Tennessee and Virginia). Furthermore, given the query “who was the last king of scotland based on”, the correct answer is former Ugandan President Idi Amin, yet SEAL retrieves passages solely concerning historical Scottish monarchs. Similarly, for the query “what country on chinas border is mostly desert”, SEAL incorrectly returns Pakistan, Russia, and North Korea instead of the correct answer, Mongolia. This likely stems from a western data bias, which over-represents these neighboring nations compared to the geographically accurate target.

Score Concentration (I7). We measured the extent to which a passage’s total score is concentrated in its single highest-scoring ngram. Our analysis reveals that distributed evidence is a stronger predictor of success: for SEAL, Hits@1 declines from 48.6% when the score is distributed across many ngrams (lowest decile of dominance) to 29.6% when a single ngram dominates (highest decile). This indicates that the additive scoring mechanism is most effective when it can aggregate multiple independent pieces of evidence (see Figure 4).

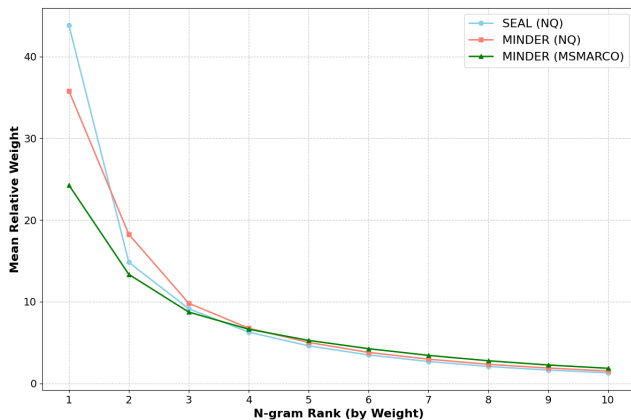


Figure 4: Score Mass Concentration in GR: A significant portion of retrieval confidence is often concentrated on a few high-scoring ngrams. Here, “weight” refers to the LM score. To visualize this, we sort the ngrams of each passage by weight, compute the mean relative weight for each rank, and plot it against the ngram rank.

Post-Rationalization (I8). We analyzed whether the model’s ability to generate the specific answer string as an identifier correlates with success. While GR aims for document identification rather than direct QA, we found a strong “answer-string bias”: for SEAL, Hits@1 more than doubles from 34.9% to 75.0% when the answer string is among the generated ngrams. This is problematic for two reasons. First, when the model fails to generate the exact answer string, the remaining contextual cues are often too generic to isolate the correct passage. Second, it demonstrates that the model post-rationalizes its parametric knowledge rather than executing information lookup.

4 Analysis Tool

To bridge the gap between model outputs and interpretable behavior, saving researchers from the headache of manual auditing, we introduce a web-based diagnostic tool for debugging via a color-coded ngram visualization system that distinguishes unique relevant identifiers in green, negative ones in red, and ambiguous shared tokens in yellow. An example is shown in Figure 5. Furthermore, the support for side-by-side comparisons of different model configurations provides a new analytic for evaluating how varying parameters influence retrieval behavior and ngram attribution. The source code for the tool as well as usage instructions are made available at <https://github.com/adrianmbracher/ngram-analysis-tool>.

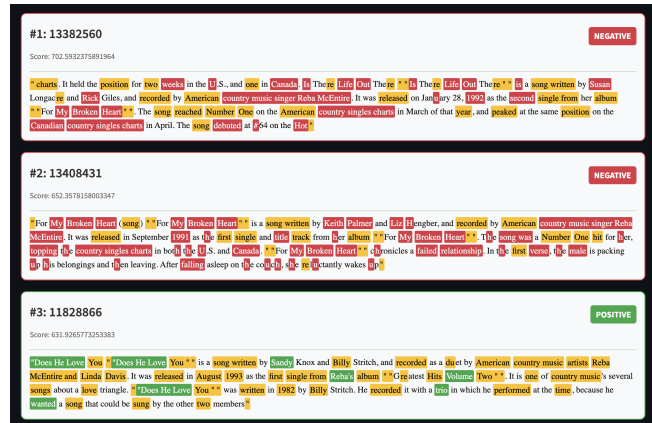


Figure 5: The diagnostic tool visualizing the top-3 retrieved passages from SEAL solution on the Natural Questions dataset. Ngrams are color-coded to audit model behavior: green indicates uniquely relevant identifiers, red indicates uniquely negative ones, and yellow denotes ambiguous ngrams shared in relevant and negative passages.

5 Conclusion

As generative retrieval becomes more and more popular due to the rise of language models, understanding its failure modes is critical. In this work, we provided a taxonomy of these errors across four stages. Our empirical analysis indicates that while n-gram-based GR models effectively apply additive scoring to aggregate distributed evidence, they suffer from architectural weaknesses including length bias, subword tokenization constraints, and meta-data over-reliance. Furthermore, the performance differences linked to explicit answer-string generation and Western pre-training biases reveal that these systems frequently favor parametric post-rationalization over robust query alignment. Readers can conclude that mitigating these identifier ambiguities and data biases remains an important challenge for advancing such generative retrieval systems. We also introduced a practical tool designed to inspire new analytical frameworks and provide the community with a starting point to audit GR architectures.

Acknowledgments

This work has been funded by the Vienna Science and Technology Fund (WWTF) under the Grant ID 10.47379/VRG24013.

The authors acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community.

References

- [1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation. arXiv:2407.12216 [cs.IR] <https://arxiv.org/abs/2407.12216>
- [2] Md Abdul Aowal, Maliha T Islam, Priyanka Mary Mammen, and Sandesh Shetty. 2023. Detecting Natural Language Biases with Prompt-based Learning. arXiv:2309.05227 [cs.CL] <https://arxiv.org/abs/2309.05227>
- [3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511 [cs.CL] <https://arxiv.org/abs/2310.11511>
- [4] Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven Failure Points When Engineering a Retrieval Augmented Generation System. arXiv:2401.05856 [cs.SE] <https://arxiv.org/abs/2401.05856>
- [5] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems* 35 (2022), 31668–31683.
- [6] Adrian Bracher and Svitlana Vakulenko. 2026. Generative Retrieval Overcomes Limitations of Dense Retrieval but Struggles with Identifier Ambiguity. *arXiv preprint arXiv:2604.05764* (2026).
- [7] Jiehan Cheng, Zhicheng Dou, Yutao Zhu, and Xiaoxi Li. 2025. Descriptive and Discriminative Document Identifiers for Generative Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 11 (Apr. 2025), 11518–11526. doi:10.1609/aaai.v39i11.33253
- [8] N De Cao, G Izacard, S Riedel, and F Petroni. 2020. Autoregressive Entity Retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, Vol. 2021. ICLR.
- [9] P. Ferragina and G. Manzini. 2000. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*. 390–398. doi:10.1109/SFCS.2000.892127
- [10] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (September 2024), 1097–1179. doi:10.1162/coli_a_00524
- [11] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (January 2025), 1–55. doi:10.1145/3703155
- [12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP (1)*. 6769–6781.
- [13] Tzu-Lin Kuo, Tzu-Wei Chiu, Tzung-Sheng Lin, Sheng-Yang Wu, Chao-Wei Huang, and Yun-Nung Chen. 2024. A Survey of Generative Information Retrieval. arXiv:2406.01197 [cs.IR] <https://arxiv.org/abs/2406.01197>
- [14] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics* (2019).
- [15] Sunkyung Lee, Minjin Choi, and Jongwuk Lee. 2023. GLEN: Generative Retrieval via Lexical Index Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 7693–7704.
- [16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
- [17] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From Matching to Generation: A Survey on Generative Information Retrieval. arXiv:2404.14851 [cs.IR] <https://arxiv.org/abs/2404.14851>
- [18] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Multiview Identifiers Enhanced Generative Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 6636–6648.
- [19] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024. Learning to Rank in Generative Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8716–8723. <https://doi.org/10.1609/aaai.v38i8.28717>
- [20] Yongqi Li, Zhen Zhang, Wenjie Wang, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Distillation Enhanced Generative Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11119–11129. doi:10.18653/v1/2024.findings-acl.662
- [21] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics* 12 (2024), 157–173.
- [22] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Changjiang Zhou, Maarten de Rijke, and Xueqi Cheng. 2024. On the Robustness of Generative Information Retrieval Models. arXiv:2412.18768 [cs.IR] <https://arxiv.org/abs/2412.18768>
- [23] Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. 2023. DSI++: Updating Transformer Memory with New Documents. arXiv:2212.09744 [cs.CL] <https://arxiv.org/abs/2212.09744>
- [24] Kidist Amde Mekonnen, Yubao Tang, and Maarten de Rijke. 2025. Lightweight and Direct Document Relevance Optimization for Generative Information Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. <https://arxiv.org/abs/2504.05181> Introduces direct pairwise ranking optimization to align docid generation with document relevance.
- [25] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. *ACM SIGIR Forum* 55, 1 (June 2021), 1–27. doi:10.1145/3476415.3476428
- [26] Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. arXiv:2305.14456 [cs.CL] <https://arxiv.org/abs/2305.14456>
- [27] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* 15, 2, Article 10 (June 2023), 21 pages. doi:10.1145/3597307
- [28] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. (2016).
- [29] Ronak Pradeep, Kai Hui, Jai Gupta, Adam D. Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q. Tran. 2023. How Does Generative Retrieval Scale to Millions of Passages? arXiv:2305.11841 [cs.IR] <https://arxiv.org/abs/2305.11841>
- [30] Weiwei Sun, Keyi Kong, Xinyu Ma, Shuaiqiang Wang, Dawei Yin, Maarten de Rijke, Zhaochun Ren, and Yiming Yang. 2025. ZeroGR: A Generalizable and Scalable Framework for Zero-Shot Generative Retrieval. *arXiv preprint arXiv:2510.10419* (2025).
- [31] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 46345–46361.
- [32] Yubao Tang, Ruqing Zhang, Weiwei Sun, Jiafeng Guo, and Maarten De Rijke. 2024. Recent Advances in Generative Information Retrieval. In *Companion Proceedings of the ACM Web Conference 2024 (Singapore, Singapore) (WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1238–1241. doi:10.1145/3589335.3641239
- [33] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems* 35 (2022), 21831–21843.
- [34] Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not Faithfulness in Retrieval Augmented Generation Attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. 22–32.
- [35] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems* 35 (2022), 25600–25614.
- [36] Ye Wang, Xinrun Xu, and Zhiming Ding. 2025. MindRef: Mimicking Human Memory for Hierarchical Reference Retrieval with Fine-Grained Location Awareness. arXiv:2402.17010 [cs.CL] <https://arxiv.org/abs/2402.17010>
- [37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] <https://arxiv.org/abs/2201.11903>
- [38] Peiwen Yuan, Xinglin Wang, Shaoxiong Feng, Boyuan Pan, Yiwei Li, Heda Wang, Xupeng Miao, and Kan Li. 2024. Generative Dense Retrieval: Memory Can Be a Burden. arXiv:2401.10487 [cs.IR] <https://arxiv.org/abs/2401.10487>

- [39] Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*. 1441–1452.
- [40] Hansi Zeng, Chen Luo, and Hamed Zamani. 2024. Planning ahead in generative retrieval: Guiding autoregressive generation through simultaneous decoding. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 469–480.
- [41] Fuwei Zhang, Xiaoyu Liu, Xinyu Jia, Yingfei Zhang, Shuai Zhang, Xiang Li, Fuzhen Zhuang, Wei Lin, and Zhao Zhang. 2025. Multi-level Relevance Document Identifier Learning for Generative Retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 10066–10080. doi:10.18653/v1/2025.acl-long.497
- [42] Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, Fangchao Liu, and Zhao Cao. 2024. Generative retrieval via term set generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 458–468.
- [43] Zhen Zhang, Xinyu Ma, Weiwei Sun, Pengjie Ren, Zhumin Chen, Shuaiqiang Wang, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2025. Replication and Exploration of Generative Retrieval over Dynamic Corpora. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3325–3334.
- [44] Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, Peitian Zhang, and Ji-Rong Wen. 2022. Ultron: An ultimate retriever on corpus with a model-based indexer. *arXiv preprint arXiv:2208.09257* (2022).