ISSN 2007-9737

Comparing Sparse and Dense Information Retrieval Methods on a Wikipedia-Derived NLP Dataset

Almat Kairbek[1], Zarina Abildasheva[1], Iskander Akhmetov[1], Shakarim Aubakirov[1], Alymzhan Toleu[2], Alexander Krassovitsky[2], Rustam Mussabayev[2], Alexander Gelbukh[3,*]

[1] Kazakh-British Technical University, Kazakhstan

[2] Institute of Information and Computational Technologies, Kazakhstan

[3] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

i.akhmetov@kbtu.kz, agelbukh@cic.ipn.mx

Abstract. This paper presents a comparative evaluation of sparse and dense Information Retrieval (IR) methods on a domain-focused dataset of 500 NLP-related Wikipedia articles. Four approaches—TF–IDF, BM25, MiniLM, and Dense Passage Retrieval (DPR)—were assessed using Precision@K, nDCG@K, and Hit Rate@K against ground truths from the Wikipedia API and Google results. Results show that BM25 consistently delivers the highest precision and ranking stability across both sources, while TF–IDF remains competitive at larger cut-offs, often surpassing DPR in recall. Dense methods, especially MiniLM, improve recall at higher ranks by capturing semantic relationships, though they lag behind sparse methods in top-rank precision. The divergence between Wikipedia- and Google-based evaluations highlights the importance of multi-perspective benchmarking. Findings confirm the complementary strengths of sparse and dense paradigms, suggesting that hybrid pipelines—combining BM25 with dense re-ranking—provide the most effective balance between efficiency, precision, and semantic coverage for domain-specific retrieval.

Keywords. Information retrieval, sparse and dense models, BM25, hybrid retrieval systems.

1 Introduction

The exponential growth of digital content in scientific, educational, and general information domains has created an urgent demand for Information Retrieval (IR) systems that are not only fast but also capable of understanding semantic nuance.

Users now expect retrieval systems to return results that go beyond exact keyword matches, surfacing conceptually related and contextually relevant documents. Over the past decades, IR has evolved from early rule-based methods to probabilistic models, and more recently, to transformer-based neural encoders capable of dense semantic retrieval.

1.1 From Sparse to Dense Retrieval

Traditional IR approaches relied heavily on exact term-matching. The vector space model introduced by Salton et al. [20] formalized documents and queries as weighted vectors, using Term Frequency–Inverse Document Frequency (TF–IDF) to emphasize discriminative terms.

This approach, despite its simplicity, significantly outperformed Boolean retrieval. However, TF–IDF treats documents as bags of words, ignoring semantics, word order, and contextual relations [22]. Probabilistic methods such as Best Match 25 (BM25) improved upon TF–IDF by incorporating document length normalization and

diminishing returns for high-frequency terms [19]. BM25 has remained one of the most widely used baselines due to its balance of efficiency and effectiveness. The introduction of learning-to-rank techniques further advanced IR performance, enabling supervised optimization of ranking functions [4, 9]. These sparse retrieval approaches are valued for their interpretability and efficiency, particularly when working with inverted indexes at scale.

Despite the emergence of neural retrieval models, sparse methods such as BM25 have remained competitive in large-scale and domain-specific retrieval scenarios. Recent studies demonstrate that the limitations of sparse retrieval are not inherent to lexical representations themselves, but rather to simplistic term-weighting schemes.

Modern extensions of sparse retrieval integrate contextualized language models to learn more expressive term importance while preserving the efficiency and interpretability of inverted indexes [3, 15, 8].

1.2 Neural and Dense Retrieval

The advent of deep learning transformed IR, especially with the development of pretrained language models (PLMs). Dense retrieval approaches encode queries and documents into continuous vector spaces, where similarity is determined by dot product or cosine distance.

Karpukhin et al. [12] introduced Dense Passage Retrieval (DPR), leveraging dual BERT encoders for fast similarity search with pre-computed embeddings indexed by FAISS [11]. Such approaches achieve strong semantic recall, enabling retrieval even when queries and documents share little lexical overlap.

SentenceTransformer-based models such as MiniLM, which provide compact embeddings optimized for semantic similarity, also contribute to this trend [18]. Surveys by Zhao et al. [25] and Xu et al. [22] highlight how dense retrieval systems have matured, identifying key challenges in negative sampling, efficiency, and generalization.

However, a growing body of empirical evidence suggests that dense retrievers, when used as standalone first-stage retrieval models, exhibit systematic weaknesses. In particular, dense encoders tend to favor strongly relevant documents while failing to reliably retrieve weakly relevant or marginally relevant documents, leading to degraded performance on deep ranking metrics such as MAP and Recall@1000 [21, 10].

These limitations are especially pronounced in domain-specific and zero-shot settings, where dense models lack sufficient lexical grounding.

1.3 Hybrid and Emerging Paradigms

Hybrid retrieval systems aim to exploit the complementary strengths of sparse and dense methods.

While hybrid retrieval systems are commonly motivated by the intuitive notion of complementarity between sparse and dense representations, recent work argues that complementarity is neither automatic nor guaranteed. Lee et al. [13] demonstrate that naive hybrid approaches often increase overlap between sparse and dense retrievers rather than reducing failure cases, and propose the Ratio of Complementarity (RoC) as a principled metric to quantify effective complementarity. Empirical studies further show that hybrid effectiveness critically depends on the ability of sparse and dense components to capture distinct relevance signals. Dense retrievers excel at modeling strong semantic relevance, whereas sparse models such as BM25 provide robust coverage of weaker relevance signals that are essential for high recall at deeper ranks [21, 5]. Studies demonstrate that combining BM25 with dense encoders often yields state-of-the-art performance, balancing efficiency and semantic accuracy [17, 2].

Learned sparse models such as SPLADE preserve sparsity while leveraging transformer semantics, enabling efficient, index-compatible retrieval [6, 7]. The work by Mandikal and Mooney shows that hybrid models outperform both sparse and dense counterparts in domain-specific scientific retrieval—highlighting the value of integration [16]. Similarly, Zhang et al. propose graph-based approximate nearest neighbor search for unified dense-sparse hybrid vectors, achieving

notable speedups and higher accuracy in hybrid vector retrieval [24]. Furthermore, Mackie et al. extend relevance feedback techniques (both pseudo-relevant and generative) to improve both sparse and dense retrieval paradigms, showing that combined feedback signals significantly boost recall across benchmarks [14].

1.4 Sparse Retrieval in LLM-Based Architectures

Recent studies examine sparse retrieval in the context of decoder-only large language models (LLMs). Zeng et al. [23] find that sparse retrieval models consistently outperform dense retrieval models across in-domain and out-of-domain benchmarks and scale more robustly as model size increases, especially when using combined contrastive learning (CL) and knowledge distillation (KD) objectives [23]. This suggests that even within LLM frameworks, sparsity retains substantial advantages for generalization and robustness.

1.5 Motivation and Contribution

Despite this growing body of work, few comparative studies have systematically evaluated sparse and dense retrieval approaches within a unified experimental framework, particularly in a domain-specific context such as Natural Language Processing (NLP). This work addresses this gap by constructing a domain-focused Wikipedia dataset of NLP-related articles and evaluating four representative methods: TF–IDF, BM25,

MiniLM (dense retriever), and DPR. Unlike prior work that focuses on optimizing hybrid objectives or fine-tuning dense encoders on large-scale benchmarks, this study deliberately evaluates retrieval methods without task-specific fine-tuning, in order to isolate the intrinsic strengths and weaknesses of sparse and dense paradigms in a realistic domain-specific setting.

The contributions are as follows:

- A unified evaluation of sparse (TF–IDF, BM25) and dense (MiniLM, DPR) retrieval methods on an NLP-focused Wikipedia corpus.
- Comparative analysis of ranking performance across multiple metrics, including Precision@K, nDCG@K, and Hit Rate@K, using multiple ground truth sources.
- Empirical insights into trade-offs between sparse and dense paradigms, informing the design of hybrid and ontology-augmented retrieval pipelines.

This comparative analysis provides both practical and theoretical insights for IR system design, guiding future development of hybrid retrieval systems that bridge the gap between lexical precision and semantic understanding.

The remainder of this paper is organized as follows. Section 2 describes the methodology, including dataset construction, retrieval models, and evaluation metrics. Section 3 outlines the experimental setup and pipeline for query generation, indexing, and ground truth construction. Section 4 presents the retrieval performance results with respect to multiple evaluation metrics. Section 5 summarizes the findings, integrates their implications, and outlines directions for future research.

2 Methodology

This section describes the dataset, retrieval models, and evaluation metrics used in the experiments. The objective is to provide a transparent overview of how the Wikipedia NLP corpus was constructed, how each IR method was implemented, and how performance was measured.

2.1 Dataset Construction

Corpus: A domain-focused collection of Wikipedia articles specifically related to Natural Language Processing (NLP) was assembled. The corpus was restricted exclusively to NLP-related topics to ensure thematic consistency. The final dataset comprises 500 full-length articles covering areas such as syntactic parsing, machine translation, and transformer architectures. Each article's full text and title were preserved in raw form, with capitalization and punctuation retained to allow retrieval models to operate on the original textual content. The collection process utilized the

Wikimedia Foundation's API and official dumps, ensuring both reproducibility and transparency.

Queries: To evaluate retrieval effectiveness, a set of user-like queries was generated from the NLP corpus. Query generation was performed using a single large language model, `Meta LLaMA-3 8B` (instruction-tuned) [1]. The model was prompted with short excerpts from articles and asked to produce natural search queries that resemble the phrasing of real-world information-seeking behavior. This approach yielded queries that varied in specificity and lexical style while remaining grounded in NLP subject matter. Examples include:

— "automatic speech recognition technologies",

— "neural network-based natural language processing",

— "character-level language models for morphological analysis".

The final query set consisted of 160 distinct examples, all tailored to NLP topics and generated in concise, information-seeking style. Queries that were too vague, irrelevant, or near-duplicates were filtered out prior to evaluation. Although large language models are capable of producing semantically rich queries, prior work shows that dense retrievers do not necessarily benefit from such queries in zero-shot or domain-specific settings, particularly when lexical grounding is limited. Consequently, LLM-generated queries do not inherently favor dense retrieval models over sparse baselines [21, 10].

Ground Truth Creation: For each query, a set of relevant documents was required as ground truth to evaluate retrieval performance. Instead of manual labeling, existing search engines were leveraged:

— **Wikipedia's own search:** Using the official MediaWiki API, the top 20 Wikipedia pages were retrieved for each query. These were assumed to be relevant or at least highly likely to contain the requested information and were treated as the primary ground truth.

— **Google results:** To test performance against an external notion of relevance, queries were also issued to Google restricted to Wikipedia content using the filter `site:en.wikipedia.org`. A maximum of 20 Wikipedia URLs was retained per query.

In this study, ground truth denotes "the set of documents retrieved by a reliable mechanism that can reasonably be considered relevant." Although this automated approach may include some noise, the consistent application of the same ground truth across all IR models ensures comparability.

While such automatically constructed ground truth does not correspond to manually curated relevance judgments, it reflects realistic retrieval signals produced by large-scale operational systems. Previous studies have shown that relative comparisons between retrieval models remain meaningful under such proxy relevance settings, provided that the same ground truth is consistently applied across all models [21, 13].

2.2 Retrieval Methods

Four retrieval methods covering both sparse (lexical) and dense (neural) approaches were evaluated. Implementations used standard Python libraries: scikit-learn's `TfidfVectorizer` for TF-IDF, Whoosh for BM25, HuggingFace's SentenceTransformers (MiniLM) for dense retrieval, and Facebook DPR checkpoints for bi-encoder retrieval. The scoring functions are summarized below. All retrieval models were evaluated in their standard, off-the-shelf configurations without task-specific fine-tuning.

This design choice was made deliberately to assess the intrinsic retrieval behavior of sparse and dense paradigms under realistic zero-shot conditions, rather than to optimize absolute performance [10].

2.2.1 TF-IDF

Each document d and query q is represented as a sparse vector of term weights:

$$w(t,d) = \mathsf{tf}(t,d) \times \log \frac{N}{\mathsf{df}(t)}. \tag{1}$$

Where $\mathsf{tf}(t,d)$ is the frequency of term t in d, $\mathsf{df}(t)$ is the number of documents containing t, and N is

the corpus size. Cosine similarity is then applied:

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}|\,|\vec{d}|}. \tag{2}$$

2.2.2 BM25

BM25 enhances tf–idf with saturation and length normalization:

$$\text{BM25}(d, q) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t,d) \cdot (k_1 + 1)}{f(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}. \tag{3}$$

Where $f(t, d)$ is term frequency, $|d|$ is document length, avgdl is average document length, and $k_1 = 1.5$, $b = 0.75$.

2.2.3 MiniLM Dense Retriever

The `all-MiniLM-L6-v2` model encodes text into 384-dimensional embeddings. Relevance is computed via dot-product similarity:

$$\text{Score}_{\text{dense}}(q, d) = \vec{q} \cdot \vec{d}. \tag{4}$$

SentenceTransformer-based models such as MiniLM are optimized for general-purpose semantic similarity rather than retrieval-specific relevance modeling, which may affect their ability to capture weaker lexical relevance signals [5].

2.2.4 DPR

Facebook's DPR model encodes queries and passages separately into 768-dimensional embeddings. Retrieval scores are computed as:

$$\text{Score}_{\text{DPR}}(q, d) = \vec{q}_{(\text{DPR})} \cdot \vec{d}_{(\text{DPR})}. \tag{5}$$

It is important to note that DPR is typically trained on large-scale web retrieval datasets such as MS MARCO, and its performance may degrade when applied to domain-specific corpora without adaptation [21].

2.3 Evaluation Metrics

Retrieval quality was assessed with standard ranking metrics:

Precision@K (P@K):

$$P@K = \frac{R_K}{K}. \tag{6}$$

Normalized Discounted Cumulative Gain (nDCG@K):

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)}, \qquad nDCG@K = \frac{DCG@K}{IDCG@K}. \tag{7}$$

Hit Rate@K (Hit@K): Indicator of whether at least one relevant document is retrieved in the top K. For example, Hit@5 = 0.95 indicates that 95% of queries had at least one relevant result in the first five retrieved documents.

Metrics were computed for $K = \{5, 10, 20\}$ and aggregated across queries. Precision@K, nDCG@K, and Hit@K were emphasized, as they closely align with real-world retrieval effectiveness. The selected metrics jointly capture early precision, ranking quality, and coverage of relevant documents, enabling a comprehensive comparison of sparse and dense retrieval behavior across different relevance regimes.

3 Experiment

This section describes the experimental procedure used to evaluate the retrieval methods. The goal of the experiment was to ensure a controlled and reproducible comparison of sparse and dense retrieval models under identical conditions.

The overall workflow is illustrated in Figure 1 and consists of four sequential stages: query generation, ground truth construction, document retrieval, and evaluation.

3.1 Query Generation

Queries were generated using the `Meta LLaMA-3 8B` instruction-tuned language model. For each of the 500 NLP-related Wikipedia articles, short textual excerpts were provided as input to the model. The model was prompted using the template: *"Generate a search query*

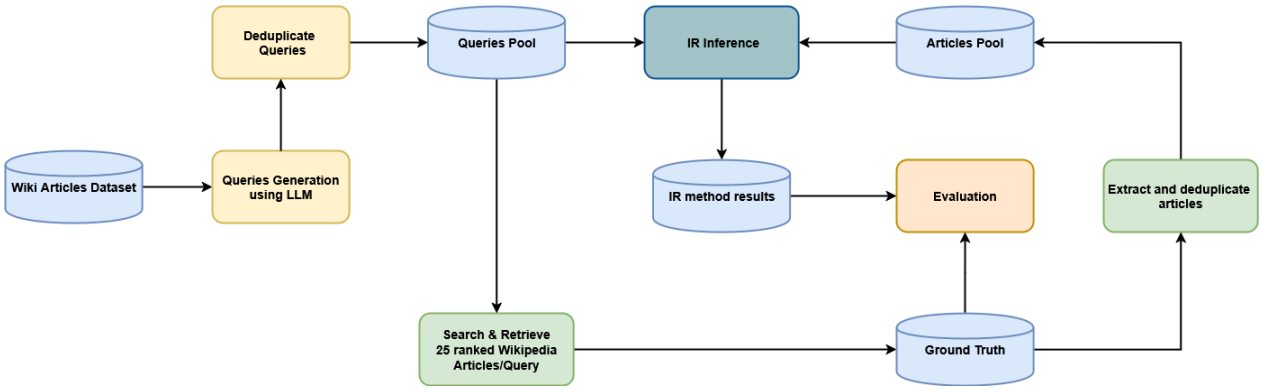

Fig. 1. Overview of the experimental pipeline for information retrieval evaluation

someone might ask to find information from the following text."

For each article, multiple candidate queries were produced. These candidates were subsequently filtered to remove duplicates, overly broad formulations, and cases where the query was a direct copy of the source text. The final query set consisted of 160 distinct queries formulated in a concise, information-seeking style and covering a broad range of NLP topics.

3.2 Ground Truth Construction

For each query, two independent ground truth sets were constructed using existing large-scale retrieval systems:

— **Wikipedia Ground Truth (Wiki-GT):** The top 20 pages returned by the official MediaWiki search API for each query.

— **Google Ground Truth (Google-GT):** The top Wikipedia pages returned by Google when queries were issued with the restriction `site:en.wikipedia.org`, with up to 20 results retained.

Only documents present in the 500-article NLP corpus were considered valid ground truth entries. Wiki-GT was treated as the primary reference set, while Google-GT served as an alternative external relevance signal for comparison. The same ground truth sets were used for evaluating all retrieval methods.

3.3 Retrieval Procedure

Each retrieval method was applied independently using a common experimental protocol. Sparse retrieval models (TF–IDF and BM25) were implemented using inverted indexes with standard preprocessing steps, including stopword removal and stemming. Dense retrieval models (MiniLM and DPR) relied on pre-computed document embeddings indexed using FAISS for efficient similarity search.

For each query, all retrieval methods returned a ranked list of the top 20 documents from the corpus. Similarity scoring was performed using cosine similarity for TF–IDF and dot-product similarity for dense embedding-based models, while BM25 relied on its standard probabilistic scoring function.

3.4 Evaluation Protocol

Retrieved rankings were evaluated by comparing the ranked outputs against the corresponding ground truth sets. Evaluation was conducted using Precision@K, nDCG@K, and Hit Rate@K, with cutoff values $K \in \{5, 10, 20\}$. All metrics were computed consistently across methods and queries and then aggregated over the full query set.

3.5 Experimental Controls

To ensure fairness and reproducibility, several controls were enforced throughout the experiment.

Duplicate documents and redirects were removed from the corpus prior to indexing.

Random seeds were fixed where applicable. None of the dense retrieval models were fine-tuned on the constructed dataset, and all models were evaluated in their standard configurations. The same preprocessing steps, evaluation code, and metric definitions were applied uniformly across all retrieval methods.

4 Results

This section presents the retrieval performance of sparse and dense methods on the NLP-focused Wikipedia corpus. Results are reported with respect to two ground truth definitions—Wikipedia API (Wiki-GT) and Google search restricted to Wikipedia (Google-GT)—and evaluated using Precision@K, nDCG@K, and Hit Rate@K.

To facilitate interpretation, performance is analyzed separately for early precision ($k = 1$) and for ranking quality and coverage at a larger cutoff ($k = 10$).

4.1 Early Precision Performance

Table 1 reports retrieval effectiveness at $k = 1$, reflecting each model's ability to place a relevant document at the top rank. Under the Wikipedia ground truth, BM25 clearly outperforms all other methods, achieving the highest Precision@1 and nDCG@1. This indicates that BM25 is particularly effective at capturing domain-specific lexical relevance within the NLP corpus, enabling accurate identification of the most relevant document.

In contrast, dense retrieval methods (MiniLM and DPR) exhibit substantially lower early precision under Wiki-GT. Their reduced performance at $k = 1$ suggests limitations in accurately ranking the most relevant document when strict domain-specific relevance is required. TF–IDF performs better than dense methods in this setting but remains inferior to BM25.

Under the Google ground truth, a different pattern emerges. TF–IDF and MiniLM achieve the highest early precision, marginally surpassing BM25. This shift reflects differences in the relevance signals encoded by Google-GT, which may emphasize broader semantic similarity or popularity-based relevance rather than strict domain-internal lexical matching. DPR consistently underperforms other methods at $k = 1$ across both ground truth definitions.

4.2 Ranking Quality and Coverage at Depth

Table 2 summarizes performance at $k = 10$, capturing both ranking quality and retrieval coverage. Across both ground truth sources, BM25 maintains the strongest overall performance, achieving the highest nDCG@10 and near-complete Hit@10. This demonstrates that BM25 not only excels at early ranking but also preserves stable relevance ordering as the cutoff increases.

TF–IDF shows competitive Hit@10 values, particularly under Google-GT, indicating that it is effective at retrieving relevant documents within a broader ranking. However, its lower nDCG scores relative to BM25 suggest weaker ranking consistency.

Dense retrieval methods exhibit improved performance at larger cutoffs compared to $k = 1$. MiniLM, in particular, shows stronger coverage than DPR and approaches TF–IDF in Hit@10 under Google-GT. This behavior indicates that dense models are effective at retrieving semantically related documents when a wider result set is considered, even though they struggle with precise top-rank placement. DPR consistently lags behind MiniLM across all metrics, reflecting weaker generalization to the domain-specific Wikipedia corpus.

4.3 Summary of Observed Patterns

Overall, the results reveal a clear differentiation between sparse and dense retrieval paradigms. Sparse methods—especially BM25—demonstrate superior early precision and stable ranking quality across both ground truth definitions. Dense methods contribute broader semantic coverage at larger cutoffs but require deeper rankings to approach the effectiveness of sparse baselines.

Differences between Wiki-GT and Google-GT further highlight how retrieval performance is

Table 1. Early retrieval performance at $k = 1$. Best values per ground truth are highlighted in bold

GT	Method	P@1	nDCG@1	Hit@1
Wikipedia	BM25	**0.74**	**0.74**	**0.74**
Wikipedia	TF–IDF	0.60	0.60	0.60
Wikipedia	MiniLM	0.53	0.53	0.53
Wikipedia	DPR	0.53	0.53	0.53
Google	BM25	0.69	0.69	0.69
Google	TF–IDF	**0.71**	**0.71**	**0.71**
Google	MiniLM	**0.71**	**0.71**	**0.71**
Google	DPR	0.60	0.60	0.60

Table 2. Retrieval performance at $k = 10$, reflecting ranking quality and coverage. Best values per ground truth are highlighted in bold

GT	Method	P@10	nDCG@10	Hit@10
Wikipedia	BM25	**0.49**	**0.54**	**0.97**
Wikipedia	TF–IDF	0.36	0.40	0.95
Wikipedia	MiniLM	0.31	0.35	0.90
Wikipedia	DPR	0.27	0.31	0.88
Google	BM25	**0.46**	**0.50**	**0.99**
Google	TF–IDF	0.41	0.46	**0.99**
Google	MiniLM	0.40	0.46	0.96
Google	DPR	0.32	0.37	0.96

influenced by the nature of the relevance signal, with sparse models aligning more closely with domain-internal relevance and dense models showing relatively stronger alignment with external relevance sources.

5 Conclusion

This study presented a systematic and controlled comparison of sparse and dense information retrieval methods on a domain-specific corpus of NLP-related Wikipedia articles. By evaluating TF–IDF, BM25, MiniLM, and DPR under identical experimental conditions and using two complementary ground truth definitions, the analysis aimed to isolate the intrinsic retrieval behavior of lexical and neural paradigms in a realistic, zero-shot setting.

The results consistently demonstrate the robustness and effectiveness of sparse retrieval methods, with BM25 achieving the strongest overall performance across early precision, ranking quality, and retrieval stability. Its superior performance at small cut-offs highlights the continued importance of exact lexical matching and probabilistic term weighting in domain-focused corpora, where terminological consistency plays a central role. These findings confirm that BM25 remains a highly reliable baseline for precision-critical retrieval tasks, even in the presence of modern neural alternatives.

Classical TF–IDF, despite its simplicity, also exhibits competitive behavior at larger cut-offs, achieving high retrieval coverage and stable Hit@K scores. This underscores the enduring relevance of vector space models as interpretable and computationally efficient baselines, particularly

in scenarios where resource constraints or transparency requirements limit the use of more complex models.

Dense retrieval methods display a complementary but distinct performance profile.

While both MiniLM and DPR underperform sparse methods in early ranking precision, their effectiveness improves at deeper cut-offs, reflecting their ability to retrieve semantically related documents beyond strict lexical overlap.

Among the dense approaches, MiniLM consistently outperforms DPR, indicating that models optimized for semantic similarity generalize more effectively to topical document retrieval than those primarily trained for open-domain question answering. The comparatively weaker performance of DPR further highlights the sensitivity of dense retrievers to training objectives and domain mismatch.

The use of multiple ground truth sources reveals that retrieval effectiveness is strongly influenced by the underlying relevance definition.

Wikipedia-based judgments favor lexical precision and strongly benefit sparse methods, whereas Google-derived relevance signals narrow the performance gap between sparse and dense approaches, particularly at larger cut-offs. This divergence emphasizes that retrieval quality is inherently context-dependent and supports the need for multi-perspective evaluation when assessing retrieval systems.

Taken together, the findings indicate that sparse and dense retrieval methods should not be treated as competing alternatives but as complementary components within a unified retrieval framework. Sparse models provide reliable early precision and ranking stability, while dense encoders contribute semantic breadth and improved recall at depth. These characteristics naturally motivate hybrid retrieval pipelines, in which sparse methods generate high-quality candidate sets that are subsequently refined using dense semantic representations.

Future work should explore explicit hybrid architectures that optimize the interaction between sparse and dense signals, investigate domain-adaptive fine-tuning strategies for dense retrievers, and integrate ontology-aware or structured semantic representations to further enhance retrieval performance in specialized domains such as natural language processing.

Acknowledgment

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23486904).

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to assist with translation from Russian to English and to refine the writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. **AI@Meta (2024).** The llama 3 herd of models. Accessed: 2025-08-27.
2. **Arabzadeh, N., Yan, X., Clarke, C. L. A. (2021).** Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. ICTIR, pp. 99–108.
3. **Bai, Y., Li, X., Wang, G., Zhang, C., Shang, L., Xu, J., Wang, Z., Wang, F., Liu, Q. (2020).** Sparterm: Learning term-based sparse representation for fast text retrieval.
4. **Burges, C., Shaked, T., Renshaw, E., Hamilton, N., Hullender, G. (2005).** Learning to rank using gradient descent.
5. **Chen, X., Lakhotia, K., Oğuz, B., Gupta, A., Lewis, P., Peshterliev, S., Mehdad, Y., Gupta, S., tau Yih, W. (2022).** Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one?
6. **Formal, T., Piwowarski, B., Clinchant, S. (2021).** SPLADE: Sparse lexical and

expansion model for first stage ranking. DOI: 10.48550/ARXIV.2107.05720.

7. **Formal, T., et al. (2024).** Splade v2: Sparse lexical and expansion model for information retrieval. ACM Transactions on Information Systems.
8. **Gao, L., Dai, Z., Callan, J. (2021).** Coil: Revisit exact lexical match in information retrieval with contextualized inverted list.
9. **Guo, J., Fan, Y., Ai, Q., Croft, W. B. (2016).** A deep relevance matching model for ad-hoc retrieval. Proceedings of CIKM, pp. 55–64.
10. **Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E. (2022).** Unsupervised dense information retrieval with contrastive learning.
11. **Johnson, J., Douze, M., Jégou, H. (2019).** Billion-scale similarity search with gpus. IEEE Transactions on Big Data, Vol. 7, No. 3, pp. 535–547.
12. **Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-t. (2020).** Dense passage retrieval for open-domain question answering. EMNLP, pp. 6769–6781.
13. **Lee, D., Hwang, S.-w., Lee, K., Choi, S., Park, S. (2023).** On complementarity objectives for hybrid retrieval. **Rogers, A., Boyd-Graber, J., Okazaki, N.**, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, pp. 13357–13368. DOI: 10.18653/v1/2023.acl-long.746.
14. **Mackie, I., Chatterjee, S., Dalton, J. (2023).** Generative and pseudo-relevant feedback for sparse, dense and learned sparse retrieval. Workshop on Large Language Models' Interpretation and Trustworthiness, CIKM, pp. 1–6.
15. **Mallia, A., Khattab, O., Tonellotto, N., Suel, T. (2021).** Learning passage impacts for inverted indexes.
16. **Mandikal, P., Mooney, R. (2023).** Sparse meets dense: A hybrid approach to enhance scientific document retrieval. CEUR Workshop Proceedings (Scientific Document Understanding Workshop), pp. 1–7.
17. **Nguyen, N. H., Haase, P. J., Saad, M. (2021).** A hybrid retrieval approach combining sparse and dense representations for document ranking. SIGIR, pp. 1760–1764.
18. **Reimers, N., Gurevych, I. (2020).** The curse of dense low-dimensional information retrieval for large index sizes. EMNLP, pp. 6849–6859.
19. **Robertson, S. E., Sparck Jones, K. (1976).** Relevance weighting of search terms.
20. **Salton, G., Wong, A., Yang, C. S. (1975).** A vector space model for automatic indexing. Communications of the ACM, Vol. 18, No. 11, pp. 613–620.
21. **Wang, S., Zhuang, S., Zuccon, G. (2021).** Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, Association for Computing Machinery, New York, NY, USA, pp. 317–324. DOI: 10.1145/3471158.3472233.
22. **Xu, Z., Mo, F., Huang, Z., Zhang, C., Yu, P., Wang, B., Lin, J., Srikumar, V. (2025).** A survey of model architectures in information retrieval.
23. **Zeng, H., et al. (2025).** Scaling sparse and dense retrieval in decoder-only llms.
24. **Zhang, H., et al. (2024).** Efficient and effective retrieval of dense-sparse hybrid vectors using graph-based approximate nearest neighbor search.
25. **Zhao, W. X., Liu, J., Ren, R., Wen, J.-R. (2022).** Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems, Vol. 42, No. 4.

ISSN 2007-9737

Article received on 11/06/2025; accepted on 21/10/2025.
**Corresponding author is Alexander Gelbukh.*