FactDebug at SemEval-2025 Task 7: Hybrid Retrieval Pipeline for Identifying Previously Fact-Checked Claims Across Multiple Languages

Evgenii Nikolaev[1] **Ivan Bondarenko**[4] **Islam Aushev**[1]
Vasilii Krikunov[1] **Andrei Glinskii**[1] **Vasily Konovalov**[2,1] **Julia Belikova**[3,1]
[1]Moscow Institute of Physics and Technology
[2]AIRI [3]Sber AI Lab [4]Novosibirsk State University
{nikolaev.en, belikova.iua, vasily.konovalov}@phystech.edu

Abstract

The proliferation of multilingual misinformation demands robust systems for cross-lingual fact-checked claim retrieval. This paper addresses SemEval-2025 Shared Task 7, which challenges participants to retrieve fact-checks for social media posts across 14 languages, even when posts and fact-checks are in different languages. We propose a hybrid retrieval pipeline that integrates both sparse lexical matching techniques (utilizing BM25 and BGE-m3) and dense semantic retrieval methods (leveraging both pretrained and fine-tuned BGE-m3 embeddings). Our approach implements the dynamic fusion of these complementary retrieval strategies and employs curriculum-trained rerankers to optimize retrieval performance. Our system achieves 67.2% cross-lingual and 86.01% monolingual accuracy on the Shared Task MultiClaim dataset.

1 Introduction

Nowadays, fact-checking has become crucially important because it helps maintain accuracy and credibility, prevents the spread of misinformation, and ensures informed decision-making by verifying information before dissemination.

SemEval-2025 Task 7 is focused on *Previously Fact-Checked Claim Retrieval* (PFCR) (Shaar et al., 2020). The task involves ranking a set of fact-checked claims according to their relevance to an input claim such as a social media post, with the highest-ranking ones being most pertinent and beneficial for fact-checking.

So far only monolingual PFCR has been tackled, when the input claim and the fact-checked claims are in the same language. To address these shortcomings, the SemEval-2025 Task 7 (Peng et al., 2025) has been organized with the MultiClaim dataset (Pikuliak et al., 2023) to encourage the community to develop a multilingual fact-checking system. The task is divided into two main subtasks: (1) monolingual fact-checking – given a social post, participants must develop systems to identify and retrieve the most relevant fact-checked claim written in the same language as the post; (2) cross-lingual fact-checking – the task is essentially the same as the first one, but now posts and their relevant fact-checks can be written in a different language. This subtask requires participants to build a multilingual retrieval system.

This paper is structured as follows. Section 2 discusses existing work on fact-checking. Section 3 describes the modified version of the MultiClaim dataset that was used in the SemEval-2025 Task 7. Section 4 introduces the evaluation metrics. Section 5 describes the proposed hybrid retrieval pipeline. Section 6 reports the results of the applied approaches.

Our contribution can be summarized as follows. We introduce a lightweight yet effective fact-checking hybrid retrieval system that incorporates multistage fine-tuning components. This technique efficiently aligns multilingual social media posts with a multilingual fact-check corpus.

2 Related Work

The information retrieval component is an essential part of any QA system, not only because it improves QA performance, but also because it enhances fact-checking (Krayko et al., 2024). PFCR task is time-consuming for professional fact-checkers and information retrieval methods can speed up the process. Multilingual PFCR is an even more complicated version for humans because it requires a deep understanding of multiple languages. Multilingual information search can facilitate the fact-checking between different languages.

For PFCR, traditional methods of information retrieval can be utilized. Vo and Lee (2018) effectively employed the BM25 method to identify

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2190–2196

fake news. Various text embedding techniques have been used to improve the retrieval process, allowing more nuanced comparisons between claims, and also techniques such as reranking are used to combine multiple methods, enhancing the efficiency and accuracy of claim retrieval (Pikuliak et al., 2023; Konovalov and Tumunbayarova, 2018). Similar ideas are used to retrieve claims for comparative questions (Shallouf et al., 2024). In addition, some approaches enhance the results by incorporating visual data from images, using abstractive summarization, or identifying key sentences.

XLM-RoBERTA based BGE-M3 provides more contextually rich and semantically accurate representations of text, ultimately leading to more relevant and precise search results. It can simultaneously perform the three common retrieval functionalities of embedding model: dense retrieval, multi-vector retrieval, and sparse retrieval utilizing the Transformer ability to integrate several tasks (Karpov and Konovalov, 2023).

Research emphasizes the importance of context in detecting previously fact-checked claims, especially in political debates or documents (Shaar et al., 2022). Some studies focus on detecting claims within entire documents, aiming to rank sentences based on their verifiability using previously fact-checked claims (Shaar et al., 2020)

Zhang et al. (2023) presented a dataset for monolingual information search for 18 different languages and demonstrated the work of some baseline approaches for information retrieval.

Future research will likely focus on improving the efficiency and accuracy of these systems, particulary in low-resource languges and complex contextual scenarios.

3 Dataset

The competition organizers represented a modified version of the MultiClaim dataset. The original MultiClaim dataset consists of 28k posts in 27 languages on social media, 206k fact checks in 39 languages performed by professional fact checkers, and 31k connections between the two groups. Each connection consists of a post and a fact-check reviewing the claim made in the post. The main difference between the modified version presented in the competition and the original one is that the modified version contains fewer languages (14), but contains more fact checks (272k) and few more posts. In the competition a modified MultiClaim dataset was used. The entire dataset is split into three sections: training (comprising 153k fact-checks in 8 languages, 4,972 cross-lingual and 1,7016 monolingual posts), testing (featuring 272,447 fact-checks in 12 languages, 4,000 crosslingual and 4,276 monolingual posts) and development (consisting of the same fact-checks as training, 552 cross-lingual, and 1,891 monolingual posts). There are also 25,743 connections between posts and fact-checks for training and development parts.

4 Evaluation

To evaluate retrieval in Shared Task 7, we use **Success-at-10** (**S@10**) as a quality measure for both monolingual and cross-lingual subtasks. This is because we depend on the retrieval module to capture as much relevant information as possible.

5 Proposed Approach

Our rather classical retrieval pipeline combines sparse and dense retrieval paradigms with fusion and reranking to address cross-lingual and monolingual fact-checking tasks. The architecture consists of four stages: (1) sparse vector encoding for lexical matching, (2) dense vector encoding for semantic alignment, (3) fusion to merge multi-perspective results, and (4) reranking to refine relevance ordering. This pipeline was used successfully for the retrieval in the specific domain (Aushev et al., 2025).

5.1 Sparse Retrieval

Sparse vector representations are generated using the following methods:

(1) BM25: Traditional sparse retrieval using TF-IDF weighting. (2) BGE-m3 Lexical Weights (Chen et al., 2024): Enhanced sparse vectors from the sparse component of `BAAI/bge-m3`[1], which captures the importance of the term through learned token-level scores.

As for the preprocessing step, we only converted all emoji to their text aliases.

5.2 Dense Retrieval

Dense vector representations are generated using several transformer-based models: (1) BGE-m3 Dense Encoder: The `BAAI/bge-m3` model is also employed to produce dense vectors, taking advantage of its large-scale pretraining on diverse

[1] https://hf.co/BAAI/bge-m3

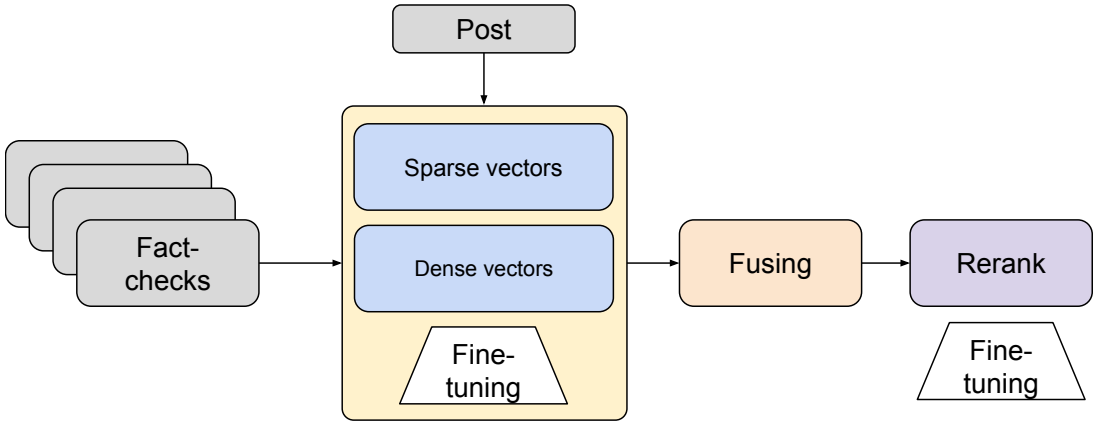

Figure 1: Hybrid Fact-Check Retrieval Pipeline. The pipeline combines sparse (BM25 or BGE-m3-lexical) and dense (BGE-m3-dense) retrieval, fuses their outputs via Reciprocal Rank Fusion, and reranks results using fine-tuned cross-encoders.

data sources; (2) Task-Specific Adaptation: A fine-tuning process using contrastive loss is applied to the `BAAI/bge-m3` model to optimize its ability to discriminate between relevant and irrelevant results.

5.3 Fusion

To unify sparse and dense signals, we employ Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) with tunable weights, ensuring a balanced combination of both approaches:

$$\text{RRF}(d) = \sum_{i=1}^{n} \frac{w_i}{k + \text{rank}_i(d)} \quad (1)$$

where k is a smoothing constant, $rank_i(d)$ denotes the rank of document d in the i-th retrieval system, and w_i represents the retriever-specific weight coefficient.

5.4 Rerank

The reranking phase refines the top retrieved documents based on additional features, improving relevance and quality. We tried several rerankers: (1) BGE-m3 Reranker: `BAAI/bge-reranker-v2-m3` cross-encoder model, based on the `BAAI/bge-m3` architecture, it computes query-document relevance scores by jointly encoding pairs, resolving ambiguities in lexical matches; (2) Curriculum-Learned Reranker: `BAAI/bge-reranker-base` fine-tuned in two stages: First, adaptation to fact-checking using focal loss (with parameters `alpha=0.95` and `gamma=0.4`) to handle class imbalance; second, optimization on hard negations derived from sparse search errors to refine decision boundaries.

6 Results and Discussion

6.1 NLI Analysis

Fact-check retrieval requires identifying claims that entail (support) or contradict a social media post. To quantify this relationship, we analyze post-fact-check pairs using `FacebookAI/roberta-large-mnli` (Liu et al., 2019), pretrained on large web-text corpora (Appendix A).

Entailment Distribution. Entailment scores exhibit a bimodal distribution: 68% of pairs cluster near 0 (no entailment) and 24% near 1 (strong entailment), with only 8% in the ambiguous middle range (0.2–0.8). This polarization likely reflects the NLI model's overconfidence rather than true task-specific relationships, as social media claims often involve nuanced or implicit connections.

Contradiction Rarity. About 92% of pairs score below 0.1 for contradiction. The scarcity of high-contradiction pairs may stem from the NLI model's limited exposure to adversarial social media claims during pretraining, rather than genuine absence of contradictions.

Neutrality as Noise. Neutral scores follow a U-shaped distribution: 54% near 0 (non-neutral) and 32% near 1 (strongly neutral). The high-neutrality cluster may include false negatives where the NLI model fails to recognize subtle entailment/contradiction signals, particularly in code-switched or informally phrased text.

Despite the bias of the out-of-the-box NLI model, these results reveal that the task can be noise-aware retrieval: prioritize lexical overlap for high-recall candidate generation, mitigating re-

Sparse Retrieval	Dense Retrieval	Fusion	Rerank	Mono	Cross
BGE-m3-lexical	–	–	✗	79.18	57.87
–	BGE-m3-dense	–	✗	78.91	65.9
BGE-m3-lexical	BGE-m3-dense	–	✗	85.20	67.02
–	BGE-m3-dense-finetuned	–	✗	79.04	65.4
BGE-m3-lexical	BGE-m3-dense-finetuned	1:1	✗	82.14	66.0
		1:1	✓	85.90	**67.2**
		8:2	✗	85.43	66.95
		8:2	✓	**86.01**	66.57
BM25	–	–	✗	62.96	56.15

Table 1: Performance of retrieval systems on cross-lingual (Cross) and monolingual (Mono) tasks. Measured in Success@10, %. Fusion $k : n$ indicates that the sparse weight is k and the dense weight is n in Reciprocal Rank Fusion.

liance on noisy semantic signals.

6.2 Retrieval Analysis

Table 1 compares the performance of sparse, dense, and hybrid retrieval systems on cross-lingual and monolingual fact-checking tasks.

Component Analysis. The standalone BGE-m3-lexical sparse retriever achieved strong monolingual performance (80.44%), outperforming the dense-only BGE-m3-dense-finetuned system (67.02%). This lexical advantage persisted in cross-lingual settings (61.17% vs. 65.4%), though dense retrieval showed greater cross-lingual robustness. The BM25 baseline (62.96% Mono, 56.15% Cross) underperformed modern neural methods, highlighting the need for learned lexical representations.

Fusion Strategies. Combining sparse and dense components via RRF yielded substantial gains. A 1:1 sparse-dense ratio with reranking produced the best cross-lingual performance (67.2% Cross), surpassing individual components. Monolingual performance peaked at 86.01% with an 8:2 sparse-dense ratio and reranking, demonstrating lexical dominance in single-language contexts. Notably, reranking consistently improved monolingual results (improvements of 0.58-1.27 points across metrics) but showed mixed cross-lingual effects, suggesting language-specific optimization potential (see Appendix B).

Crosslingual Performance. Optimal cross-lingual performance required balanced sparse-dense fusion – 1:1 ratio. In contrast to the monolingual results, there is a lexically heavy 8:2 ratio, which is consistent with our NLI analysis. This suggests that while lexical matching anchors retrieval quality, cross-language generalization benefits from controlled semantic integration, but this may be due to the nature of cross-lingual translation utilisation.

Ablation Study. In addition to the results in Table 1 we evaluated configurations for `stella_en_1.5B_v5`[2] (Zhang and FulongWang, 2024) and `multilingual-e5-large-instruct`[3] (Wang et al., 2024) with the prompt *"Given a post on a social network, retrieve the claims it contains"*, but it did not give any increase in quality. Replacing the dense retriever in hybrid pipeline with `stella_en_1.5B_v5` performs 84.45% Mono and 65.47% Cross, `multilingual-e5-large-instruct` performs 84.86% Mono and 66.37% Cross.

Conclusion

In this paper, we have described the system we submitted for the SemEval Task 7 challenge, specifically concentrating on developing a multilingual and crosslingual fact-checked claim retrieval. We proposed a simple yet effective hybrid retrieval pipeline with fine-tuned components. The proposed pipeline can be employed independently or integrated within a NLP framework such as DeepPavlov (Burtsev et al., 2018).

Future directions include dynamic fusion mechanisms that weight sparse and dense contributions per language pair, and joint training of sparse and dense components to enhance overall performance.

[2] https://hf.co/NovaSearch/stella_en_1.5B_v5
[3] https://hf.co/intfloat/multilingual-e5-large-instruct

Limitations

Our approach has two key constraints: (1) Partial fine-tuning – only dense and reranker components were optimized, leaving potential gains from end-to-end sparse-dense co-training unexplored due to optimization instability; (2) Translation and OCR inaccuracies propagate through the pipeline, particularly harming low-resource languages and slang-heavy social media text.

References

Islam Aushev, Egor Kratkov, Evgenii Nikolaev, Andrei Glinskii, Vasilii Krikunov, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. RAGulator: Effective RAG for regulatory question answering. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 114–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: An open source library for conversational ai. In *NIPS*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM.

Dmitry Karpov and Vasily Konovalov. 2023. Knowledge transfer between tasks and languages in the multi-task encoder-agnostic transformer-based models. In *Computational Linguistics and Intellectual Technologies*, volume 2023.

Vasily Konovalov and Zhargal Tumunbayarova. 2018. Learning word embeddings for low resourse languages: The case of buryat. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologiithis*, pages 331–341.

Nikita Krayko, Ivan Sidorov, Fedor Laputin, Daria Galimzianova, and Vasily Konovalov. 2024. Efficient answer retrieval system (EARS): Combining local DB search and web search for generative QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1584–1594, Miami, Florida, US. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podroužek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. Semeval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Matús Pikuliak, Ivan Srba, Róbert Móro, Timo Hromadka, Timotej Smolen, Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, and Mária Bieliková. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16477–16500. Association for Computational Linguistics.

Shaden Shaar, Firoj Alam, Giovanni Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. pages 1619–1631.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3607–3618. Association for Computational Linguistics.

Ahmad Shallouf, Hanna Herasimchyk, Mikhail Salnikov, Rudy Alexandro Garrido Veliz, Natia Mestvirishvili, Alexander Panchenko, Chris Biemann, and Irina Nikishina. 2024. CAM 2.0: End-to-end open domain comparative question answering system. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2657–2672, Torino, Italia. ELRA and ICCL.

Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking URL recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 275–284. ACM.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

Dun Zhang and FulongWang. 2024. Jasper and stella: distillation of SOTA embedding models. *CoRR*, abs/2412.19048.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A NLI Scores Distribution

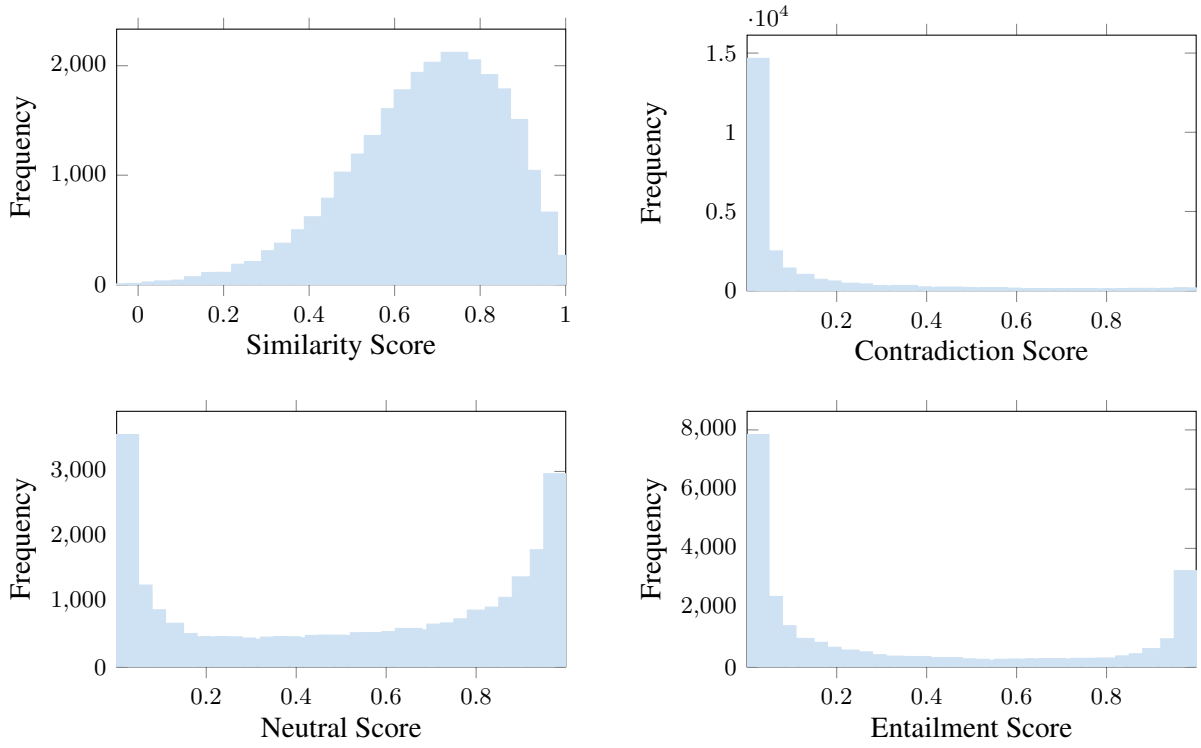

Figure 2: Distributions of NLI scores between posts and fact-checks.

B RRF Fusing Monolingual Results

Fusion	Pol	Eng	Msa	Por	Deu	Ara	Spa	Fra	Tha	Tur	**Mono**
10:0	70.6	71.4	91.3	69.2	81.2	84.8	72.6	82.8	92.8	75.0	79.1
9:1	74.0	75.4	96.7	75.0	85.8	89.8	78.4	83.8	**95.6**	78.0	83.2
8:2	**76.8**	**79.4**	98.9	**77.4**	**86.2**	91.0	**82.8**	**84.0**	**95.6**	82.2	**85.4**
7:2	76.4	78.2	**100.0**	**77.4**	85.2	**91.8**	82.6	83.2	92.8	**82.6**	85.0
6:4	73.8	75.6	**100.0**	74.6	84.8	91.0	79.2	81.0	92.3	81.8	83.4
5:5	72.6	74.2	98.9	72.8	83.4	91.2	77.2	78.6	90.7	81.8	82.1
4:6	71.0	73.2	97.8	71.0	82.8	91.2	77.2	78.0	90.7	81.4	81.4
3:7	70.6	73.0	97.8	71.2	82.6	91.0	77.4	77.6	90.7	80.8	81.2
2:8	70.4	72.0	97.8	71.2	81.4	90.8	77.0	77.0	90.7	80.4	80.8
1:9	69.6	70.0	97.8	70.4	80.8	90.8	75.8	76.2	90.7	78.8	80.0
1:10	68.4	68.6	96.7	69.6	78.4	89.4	75.4	75.4	90.7	77.8	79.0

Table 2: Combining sparse and dense components via RRF. Fusion $k : n$ indicates that the sparse weight (BGE-m3-lexical) is k and the dense (BGE-m3-dense-finetuned) weight is n in Reciprocal Rank Fusion.