FactIR: A Real-World Zero-shot Open-Domain Retrieval Benchmark for Fact-Checking

Venktesh V
Factiverse AI and Delft University of Technology
Delft, Netherlands
V.Viswanathan-1@tudelft.nl

Vinay Setty
Factiverse AI and University of Stavanger
Stavanger, Norway
vsetty@acm.org

Abstract

The field of automated fact-checking increasingly depends on retrieving web-based evidence to determine the veracity of claims in real-world scenarios. A significant challenge in this process is not only retrieving relevant information, but also identifying evidence that can both support and refute complex claims. Traditional retrieval methods may return documents that directly address claims or lean toward supporting them, but often struggle with more complex claims requiring indirect reasoning. While some existing benchmarks and methods target retrieval for fact-checking, a comprehensive real-world open-domain benchmark has been lacking. In this paper, we present a real-world retrieval benchmark FactIR, derived from Factiverse production logs, enhanced with human annotations. We rigorously evaluate state-of-the-art retrieval models in a zero-shot setup on FactIR and offer insights for developing practical retrieval systems for fact-checking. Code and data are available at https://github.com/factiverse/factIR.

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

Fact-checking, Retrieval benchmark

ACM Reference Format:
Venktesh V and Vinay Setty. 2025. FactIR: A Real-World Zero-shot Open-Domain Retrieval Benchmark for Fact-Checking. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701716.3715300

1 Introduction

The rapid spread of information through digital media has increased the need for accurate, automated fact-checking to combat misinformation. Automated fact-checking systems verify claims by retrieving relevant evidence from vast, often unstructured web data [7, 16]. A key challenge lies in the retrieval component, which traditionally prioritizes similarity. However, as misinformation becomes more complex, verifying multifaceted claims often requires reasoning from indirectly related evidence. For instance, validating a claim about vaccine safety may involve documents on production processes, clinical trials, or regulatory approvals. This complexity highlights the need for retrieval methods and benchmarks that go beyond traditional approaches.

Existing fact-checking benchmarks fall short of addressing real-world retrieval challenges. Popular datasets like FEVER [23] and FEVEROUS [1] rely on synthetic claims, with retrieval limited to Wikipedia. While recent works such as ClaimDecomp [3], QABriefs [4], and AVeriTeC [16] incorporate fact-checks from professional fact-checkers, they constrain retrieval to pre-verified justification documents or lack dedicated retrieval corpora. This highlights the need for benchmarks that better simulate open-domain, real-world retrieval scenarios critical for advancing fact-checking systems.

In this paper, we present a novel retrieval benchmark rooted in real-world fact-checking data to address existing gaps. Our benchmark is derived from Factiverse production logs, incorporating evidence retrieval data and human annotations from an operational fact-checking system [18, 19]. By leveraging data collected under genuine production conditions, we aim to provide a benchmark that accurately represents the unpredictable and multifaceted nature of real-world fact-checking. This dataset allows us to evaluate how retrieval models handle diverse information sources, incomplete knowledge contexts, partially relevant evidence, and claims that require subtle reasoning beyond simple factoid matching. Moreover, as most existing fact-checking systems rely on off-the-shelf retrievers pre-trained on datasets from other domains and used in a zero-shot setting [8, 21], it becomes essential to assess their generalization performance under these realistic constraints, where fine-tuning is often infeasible. In this paper, we adopt the same zero-shot setting to ensure our evaluation aligns with real-world usage scenarios.

Our contributions are threefold. First, we provide a real-world retrieval benchmark based on actual usage logs and human annotations, filling a critical gap in fact-checking research. Second, we conduct an extensive evaluation of state-of-the-art retrieval models (zero-shot), examining their ability to handle complex claims in authentic conditions. Finally, we offer insights and recommendations for designing retrieval systems that are not only effective in controlled settings but robust enough for real-world fact-checking, where the retrieval process must often synthesize indirect and inferential evidence. Through our work, we aim to foster the development of fact-checking systems that can better meet the challenges of today's information ecosystem, supporting fact-checkers and automated systems alike in their mission to mitigate misinformation.

2 Related Work

Over the years, a range of benchmarks have been introduced to advance research in automated fact-checking. Early benchmarks such

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3715300

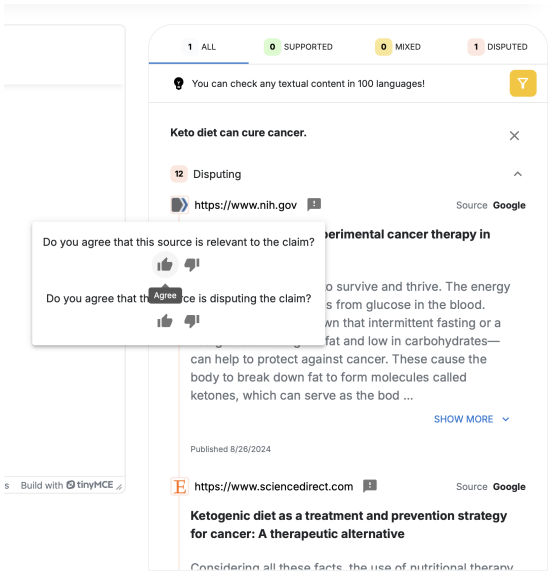

Figure 1: Factiverse fact-check editor with feedback mechanism. Example for the claim "Keto diet can cure cancer."

as FEVER [23] and FEVEROUS [1] have been widely adopted, providing large-scale datasets for claim verification. These benchmarks include retrieval corpora but are based on synthetic claims generated from Wikipedia content, limiting retrieval to a constrained domain of Wikipedia articles. While these datasets have driven significant progress, they fail to reflect the complexity and diversity of real-world claims and evidence.

More recently, datasets like ClaimDecomp [3] and QABriefs [4] have emerged, leveraging real-world fact-checks conducted by professional fact-checkers. These datasets introduce claim decomposition, where complex claims are broken down into sub-questions to facilitate evidence retrieval and verification. However, their retrieval process remains limited to pre-verified justification documents, which simplifies the task and does not simulate realistic retrieval challenges. Similarly, AVeriTeC [16] incorporates human-authored questions paired with search-engine-retrieved answers, offering a step toward open-domain retrieval. Nevertheless, it lacks a dedicated retrieval corpus, restricting its ability to evaluate retrieval performance comprehensively.

Other benchmarks, such as QuanTemp [24], focus on numerical fact-checking and their primary focus is on numerical reasoning based verification, and retrieval is not directly evaluated.

Despite these advancements, current benchmarks either rely on constrained retrieval settings or pre-verified evidence, failing to fully capture the challenges of open-domain, real-world retrieval where evidence must be identified from diverse, unstructured sources. Addressing these gaps is the primary goal of this paper.

3 Data Curation

We employ the Factiverse Fact-Check Editor [18], a web-based tool designed for automated fact-checking with a built-in feedback mechanism. As shown in Figure 1, users input claims they wish to fact-check, and the system aggregates evidence retrieved from multiple search engines, such as Google and Bing, ensuring comprehensive coverage of web content. The editor utilizes query generation and evidence filtering to retrieve the most relevant documents [17, 19]. Users can then provide detailed feedback on two aspects: the relevance of the evidence to the claim and the correctness of the stance, whether the claim is supported or refuted.

To ensure high-quality labels for this benchmark, we focus on feedback collected from domain experts, specifically professional fact-checkers and individuals with journalism backgrounds, who verify claims using Factiverse's production deployment. The claims are organically generated by users as they engage in tasks covering a variety of topics, including politics, healthcare, and the economy. The feedback was collected over the period spanning 2021 to 2023.

The annotators were provided with the following instructions and examples: **Evidence is deemed relevant to a claim if the extracted snippet can help verify the claim's veracity.** For example, for the claim "Keto diet can cure cancer," a document discussing the general benefits of the "keto diet" without addressing its effect on cancer would be considered irrelevant. On the other hand, a scientific study examining the "ketogenic diet as a treatment and prevention strategy for cancer" (as illustrated in Figure 1), even if it disproves the claim, would be considered relevant.

Table 1: Topical distribution

Topic	Count
Politics	26
Economy	24
Health	21
Law	6
Climate	8
Education	3
Other	12

Table 2: Numerical taxonomy

Topic	Count
Statistical	40
Temporal	27
Comparative	8

3.1 Benchmark Statistics

The benchmark comprises, **1413** claim-evidence pair relevance annotations with a total of **90047** documents in the corpus collection and 100 claims with an average of **13.89** documents / relevance assesments per query. We provide fine-grained topical analysis in Table 1. We observe that a number of claims in the benchmark are *quantitative or temporal* in nature, as shown in Table 2. We also observe that many claims are *compositional* comprising multiple aspects, making FACTIR a challenging retrieval benchmark. We performed a qualitative meta-analysis of relevance assignments with the help of two researchers who were asked to annotate as "1" when they deem the relevance assessment to be correct else "0". The annotators found the relevance assessments to be of high quality (**88.03%** was deemed to be correct) with a high agreement of **0.946** as indicated by Cohen's kappa.

4 Experimental Setup

We evaluate a range of state-of-the-art retrieval and re-ranking approaches on FACTIR, with two V100S-PCIE-32GB GPUs.

Lexical and Sparse retrieval: We employ ElasticSearch's BM25 implementation due to low latency and ease of use [5]. We also

Method	nDcG@5	Recall@5	nDcG@10	Recall@10	nDcG@100	Recal@100
Lexical						
BM25 [14]	0.288	0.253	0.345	0.421	0.475	0.779
Sparse						
SPLADEV2 [6]	0.287	0.231	0.336	0.384	0.473	0.783
Dense						
Stella-en-v5	0.090	0.065	0.098	0.110	0.158	0.314
DPR [11]	0.247	0.219	0.292	0.356	0.404	0.670
ANCE [26]	0.246	0.184	0.289	0.327	0.419	0.691
tas-b [9]	0.289	0.232	0.336	0.399	0.468	0.771
MPNet [20]	0.290	0.250	0.327	0.388	0.464	0.767
Contriever [10]	0.299	0.249	0.346	0.406	0.471	0.760
COlBERTV2 [15]	0.252	0.230	0.325	0.419	0.465	0.789
Snowflake-arctic-embed-s [13]	**0.367** (▲27.43%)	**0.302** (▲19.37%)	**0.420** (▲21.74%)	**0.480** (▲14.01%)	**0.529** (▲11.37%)	**0.795** (▲2.05%)
Re-Ranker						
BM25 +						
- ColBERTV2	0.265	0.253	0.333	0.424	0.464	0.759
- MARCO-MiniLM-H384	0.293	0.247	0.349	0.408	0.485	0.779
- MARCO-MiniLM-en-de	0.278	0.264	0.342	0.426	0.479	0.779
- bge-reranker-base	0.222	0.205	0.301	0.398	0.452	0.779
- bge-ranker-v2	0.252	0.235	0.312	0.399	0.460	0.779
- Jina-reranker-v2	0.270	0.245	0.336	0.421	0.474	0.779
- gte-multilingual	0.308 (▲6.04%)	0.277 (▲9.49%)	0.368 (▲6.67%)	0.437 (▲3.80%)	0.496 (▲4.42%)	0.779

Table 3: Retrieval results on FactIR, nDCG@10 across datasets. The best results are in bold and the second best results are underlined with % improvements indicated by (▲%) over the baseline BM25 being specified in brackets.

employ SPLADEV2 [6] which is a neural model employing query and document sparse expansions.

Dense retrieval: DPR [11] comprises a query, document bi-encoder model, and we employ the open source model trained on multiple Question Answering datasets *facebook-dpr-question-encoder-multiset-base*. We also include ANCE [26] in our evaluation, which is a bi-encoder model that samples hard negatives through Nearest Neighbour search over the corpus index, yielding better negatives. We employ *msmarco-roberta-base-ance-first* trained on MS-MARCO [2]. Tas-b [9] is a bi-encoder trained using supervision from a cross-encoder. We benchmark the recent *snowflake-arctic-embed-s* [13] which improves contrastive pre-training through semantic clustering-based source stratification.

Late-Interaction Models: We implement the late-interaction model ColBERTv2 [12, 15] in FactIR. It is a late-interaction model that employs a cross-attention-based MaxSim operation betweeen query and document token representations.

Re-rankers: We employ several state-of-the-art cross-encoder models including Large Language Models (LLMs) for re-ranking. We employ ColBERTV2 in re-ranker mode due to better relevance estimate from late-interaction. We evaluate BERT based cross-encoders like *cross-encoder/mmarco-mMiniLMv2-L12-H384-v1* and *cross-encoder/msmarco-MiniLM-L12-en-de-v1* that have shown to achieve impressive performance on BEIR [22] benchmark. We also benchmark more recent re-rankers like *jinaai/jina-reranker-v2-base-multilingual* which employs flash-attention. We also evaluate LLM based re-rankers like *Alibaba-NLP/gte-multilingual-reranker-base*

Metrics: We choose Normalized Discounted Cumulative Gain (nDCG@k) and Recall@k as the primary metrics for our results.

5 Results and Analysis

In this section, we analyze how retrieval and re-ranking approaches perform on the FactIR benchmark (Table 3).

Lexical and Sparse retrievers: We observe that lexical models like BM25 and learned sparse retrievers like SPLADEV2 perform surprisingly well on FactIR benchmark, as shown in Table 3. They outperform several dense retrievers, proving to be strong baselines.

Classical dense retrieval models fail on OOD data: Since models like DPR, Contriever are frequently used in a zero-shot fashion for fact-checking tasks [8, 21], we evaluate their performance on FactIR. We observe that all dense retrievers have sub-par performance when compared to BM25. This could primarily be caused by domain shift from pre-training data to claims in FactIR. Additionally, these models might also be incapable of handling such complex real-world claims as models like DPR are trained on datasets mostly containing factoid queries [21]. We observe that DPR which has been trained on QA datasets performs worse than other dense retrievers like Contriever or tas-b which are pre-trained on MS-MARCO. We observe that Contriever is closer in performance compared to BM25 compared to other dense retrievers.

Retrieval models with strong stratification based training objectives generalize better: From Table 3, we observe that snowflake-arctic-embed-s outperforms lexical sparse and other state-of-the-art dense retrieval models as measured by nDCG@k and Recall@k demonstrating better generalization performance. We hypothesize that this is primarily due to the semantic clustering based source stratification [13] based training objective. This objective is inspired by the cluster hypothesis [25] and further builds upon the topic aware sampling philosophy of tas-b [9] and the hard

negative mining aspect of ANCE [26] leading to superior performance. We posit that this training regime causes *snowflake-arctic-embed-s* to retrieve topically and contextually relevant evidence for the claim while ignoring topically similar hard negatives.

LLM-Based Re-Rankers Show Superior Generalization: After retrieving the top-100 documents using BM25, we re-rank them with a variety of models, as shown in Table 3. Among these, the LLM-based re-ranker *gte-multilingual-reranker-base* outperforms other cross-encoder models pre-trained on MSMARCO. We attribute this performance to the architecture of mGTE and the extensive pre-training of *gte-multilingual* [27] on data sourced from a wide range of tasks and domains, reducing source bias. This multi-domain pre-training enhances its generalization capabilities, enabling it to perform effectively across diverse retrieval scenarios.

6 Conclusion

In this work, we introduce FactIR for benchmarking of retrieval for fact-checking in an open-domain setup. FactIR is a collection of high quality real-world claims along with high quality relevance assessments from users. We evaluate a wide range of retrieval and re-ranking models and observe lexical and sparse retrievers to be strong baselines. We also observe that strong training objectives lead to impressive generalization performance. We also release our library which is easily extendable to new retrievers and re-rankers.

7 Acknowledgements

This work is partly funded by the Research Council of Norway project EXPLAIN (grant no: 337133).

References

[1] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. arXiv:2106.05707 [cs.CL]

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[3] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims. arXiv:2205.06938 [cs.CL]

[4] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating Fact Checking Briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7147–7161.

[5] Anton Fogelberg and Jonas Nygren. 2023. Search Engine Evaluation.

[6] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval.

[7] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[8] Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A Benchmark for Fact-Checking in Dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3785–3801.

[9] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[10] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118 [cs.IR]

[11] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781.

[12] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv:2004.12832 [cs.IR]

[13] Luke Merrick. 2024. Embedding And Clustering Your Data Can Improve Contrastive Pretraining. arXiv:2407.18887 [cs.LG]

[14] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[15] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3715–3734.

[16] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. *arXiv preprint arXiv:2305.13117* (2023).

[17] Ritvik Setty and Vinay Setty. 2024. QuestGen: Effectiveness of Question Generation Methods for Fact-Checking Applications. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 4036–4040.

[18] Vinay Setty. 2024. FactCheck Editor: Multilingual Text Editor with End-to-End fact-checking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2744–2748.

[19] Vinay Setty. 2024. Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2842–2846.

[20] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems* 33 (2020), 16857–16867.

[21] Aniruddh Sriram, Fangyuan Xu, Eunsol Choi, and Greg Durrett. 2024. Contrastive Learning to Improve Retrieval for Real-World Fact Checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 264–279.

[22] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. Curran.

[23] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819.

[24] Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 650–660.

[25] Ellen M. Voorhees. 1985. The cluster hypothesis revisited. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Montreal, Quebec, Canada) *(SIGIR '85)*. Association for Computing Machinery, New York, NY, USA, 188–196.

[26] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. arXiv:2007.00808 [cs.IR]

[27] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 1393–1412.