LEDGER: A Long-Context Benchmark of Corporate Annual Reports for Grounded Financial Retrieval and Extraction

Charles Moslonka[A,B], Amaury de Vitry[A], Arthur Garnier[C], Hicham Randrianarivo[A,B] and Emmanuel Malherbe[A]
[A]Artefact Research Center, Paris, France
[B]MICS, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
[C]Ardian, Paris, France

Finance reporting is a natural proving ground for large language models, and the very-long-context capabilities of recent models across all sizes make rigorous evaluation in this domain an increasingly pressing need. Yet most public financial resources reduce the task to plain-text SEC 10-K filings paired with a handful of question–answer items. We release LEDGER (Long-context Evaluation of Documents for Grounded Extraction and Retrieval), a corpus of 4,999 digitized corporate *annual reports* — full documents with figures, tables, and narrative, not just regulatory filings. Each report is labeled with 31 consolidated financial KPIs to be extracted and linked to the market's reaction at the earnings date. From this data we derive three evaluation benchmarks spanning the difficulty spectrum: a pure page-level KPI retrieval task with TREC-style relevance judgments over 118,048 questions in natural language, a conversational "needle-in-a-haystack" single-value lookup, and a full KPI extraction task, both from long, numerically dense reports. We additionally provide human OCR-quality annotations with inter-annotator agreement and the complete extraction, validation, and scoring toolchain. We further demonstrate the dataset's research utility with a case study linking CEO-letter rhetoric to post-publication market impact. An extension to 291 French ESEF filers (1,153 reports with 24 KPIs) augments the corpus.

Keywords: *Finance, Long-context evaluation, Information retrieval, Document understanding, Key performance indicators, Benchmark datasets*

1. Introduction

Corporate reports are long and complex, yet their release triggers swift market reactions within about one hour of publication, making near-real-time automated extraction valuable for investors. Large language models (LLMs) promise to read, understand, and extract such information, especially with the rapid growth of context windows — now reaching hundreds of thousands of tokens even for small, locally hostable models. Whether models can actually *ground* their answers in such documents is, however, poorly measured. Most public financial resources reduce the problem to short, plain-text excerpts of U.S. SEC 10-K filings paired with a few hundred question–answer pairs [1, 2, 3]. Real financial analysis instead operates over the full *annual report*: a long, visually dense "glossy" document interleaving narrative, statutory statements, hundreds of tables, and figures, where the same quantity ("revenue") is reported under company-specific labels, scales, and scopes.

We release LEDGER, a resource built to evaluate retrieval and extraction on exactly these documents. We start from 4,999 OCR'd annual reports and label them with ground-truth KPIs, natural language questions and corresponding page-level relevance. The resource is organized as a difficulty spectrum—from locating a page, to extracting one value, to extracting an entire financial statement—and ships with the complete toolchain used to build and score it. Our contributions are:

- **A corpus of labeled reports**: reports of 738 companies over fiscal years 2009–2024, with

Published at the International Conference on Information and Knowledge Management (CIKM) 2026.
Corresponding author(s): charles.moslonka@artefact.com
Artefact Research Center, Paris, France

31 consolidated KPIs for each and market data aligned with their earning date. Each of the 118,048 KPI labels comes with a natural language question, and page-level relevancy labels (Section 3).
- **Three benchmarks with baselines**: page-level KPI retrieval (TREC qrels), single-value conversational "needle-in-a-haystack" lookup, and full multi-KPI extraction, with sparse/learned-sparse IR and four open-weight LLM baselines (Section 4).
- **A research-utility case study** linking CEO-letter rhetoric to earnings per share surprise and post-publication market impact (Section 5).
- **Open release** under the MIT license for the code and Creative Commons Attribution 4.0 for the data, with a permanent DOI and the full extraction/validation toolchain.

2. Related resources

Finance is a core LLM target (e.g., BloombergGPT [4]). Numerical reasoning benchmarks FinQA [1], ConvFinQA [5] and TAT-QA [2] use short snippets, FinanceBench [3] 10-K open-book QA and DocFinQA [6] document-scale; work on structure for financial RAG [7] and ViDoRe [8] motivates page retrieval over rendered PDFs. LEDGER differs along four axes simultaneously: (i) it covers *full annual reports*—not just 10-K text—including the glossy front matter, figures, and CEO letters that 10-Ks omit; (ii) documents are genuinely long-context (≈126k tokens, median 124 pages); (iii) it provides *authoritative numeric labeling* at scale, with ≈30 KPIs per company-year reconciled against SEC XBRL filings rather than crowd answers; and (iv) it couples the documents to a market-reaction signal, enabling downstream financial-research tasks. The whole resource is openly licensed and accompanied by a datasheet [9] and FAIR-compliant [10] metadata.

3. The LEDGER corpus

Figure 1 provides an overview of the pipeline from PDFs to OCR, KPI waterfall, question/qrels creation and benchmarks.

3.1. Financial Reports and OCR

The LEDGER corpus is built from 4,999 publicly available corporate annual reports in PDF format. Using DeepSeek-OCR-2 [11, 12], we converted each report into a single page-aligned Markdown file in which pages are delimited by an explicit `<-- Page Split -->` marker, tables are rendered as HTML/LaTeX, and per-page raster images are retained for downstream visual tasks.

Because every downstream task depends on OCR fidelity, we provide a human assessment of table-extraction quality. Fifteen annotators graded the table alignment and contents ≈1,150 extracted financial tables as `ok/not_ok/uncertain` through a purpose-built interface that renders each detected table next to its source page image. A 273-table subset was triple-coded to measure agreement: we observe 87.2% full agreement, 91.3% pairwise agreement, and a Fleiss' κ of **0.81** (substantial agreement). At the table level, 81.5% of tables were graded as correctly aligned, giving a concrete, auditable estimate of OCR table fidelity that downstream users can condition on. Overall, number extraction is nearly perfect; layout alignments can be suboptimal for complex tables (18.5% misaligned), but our validation pilot shows digit errors are rare and scoring is numeric, so misalignment affects layout, not the presence of the value.

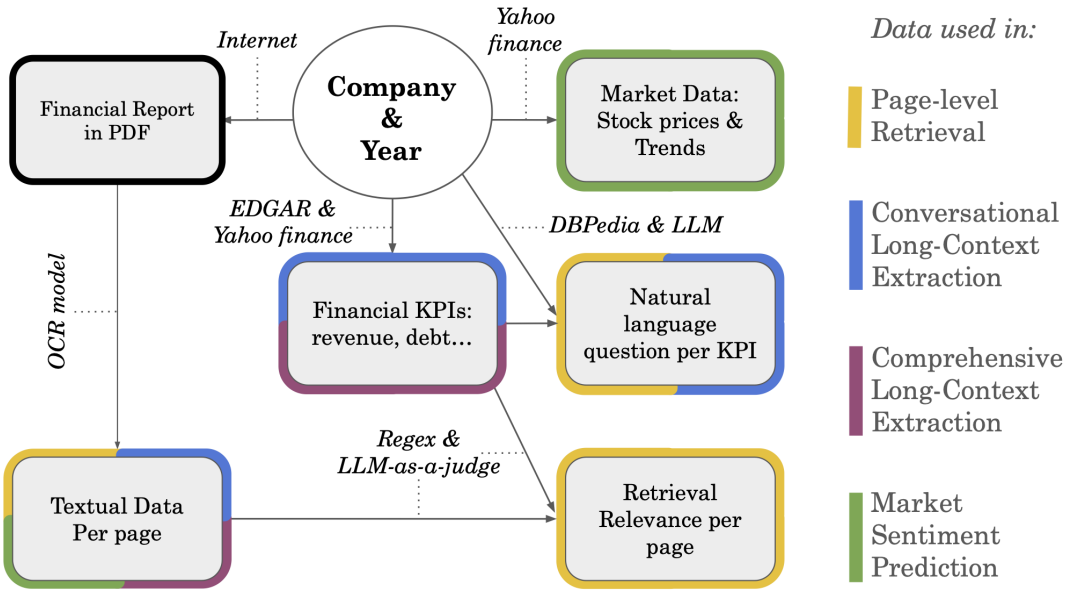


Figure 1 | Overview of LEDGER corpus. Center: Company & Year. Data flow: PDFs→OCR, Company & Year→KPIs (EDGAR waterfall) and Market Data, KPIs→questions (DBpedia+Gemini), pages+KPIs→relevance (regex+LLM). Colors: retrieval (yellow), conversational (blue), comprehensive (purple), sentiment (green).

Table 1 | The LEDGER dataset statistics. The Eval companies are a subset of the Total corpus. Tokens were counted using the `cl100k_base` tokenizer.

	Eval	Total	Avg./report
Companies	111	738	-
Reports	494	4,999	-
Pages	47,820	691,858	138
Tokens	41M	631M	126,274
KPIs & Questions	13,519	118,048	23.6
Relevant pages (qrels)	392,484	2,054,279	-
Daily stock prices	-	800k	160

3.2. KPI labeling

We attach consolidated KPIs to every company-year report, organized across the three financial statements: *(i) income* – revenue, cost of revenue, gross profit, R&D, SG&A, operating income, interest expense, income-tax expense, net income, basic and diluted EPS; *(ii) balance sheet* – total assets, total liabilities, stockholders' equity and its incl.-NCI variant, cash & equivalents and its incl.-restricted variant, long-term debt at three scopes, short-term borrowings, inventory, receivables, payables, shares outstanding; and *(iii) cash flow* – operating, investing, financing flows, capex, depreciation & amortization, dividends paid. EBITDA is deliberately excluded: it is reporting-standard dependent and derivable from the released components. Values are routed through a three-tier source waterfall: SEC EDGAR `companyfacts` XBRL for U.S. listings, Yahoo!Finance API (through the `yfinance` python package) as a fallback for non-U.S. listings, and Alpha Vantage as an opt-in gap-fill that never overwrites the first two. Most KPIs map to a single XBRL tag. When several tags exist—synonyms or different accounting scopes—we apply an ordered waterfall that selects the standard consolidated scope and keeps alternatives as separate KPIs. For example, "net income" is taken as attributable to

parent shareholders, excluding non-controlling interests (NCI); the inclusive NCI figure can be ≈29% different for the same company-year. Remaining gaps are filled with audited accounting identities that sum components. Fiscal-year keying uses the filer's labelled FY derived from period-end date (FY = year – 1 if the period ends in January, else the period-end year), correctly handling 52/53-week retailers whose fiscal years end in early January. We store per-cell provenance (`source`, `tag_used`, `ambiguous_tags`) and document the full tag maps in the released toolchain. The result is 118,048 audited facts with per-KPI yearly coverage above 85% for the core line items.

3.3. Natural language questions and relevant pages for KPI

For every (company, year, KPI) triplet, we generated a natural-language question to simulate usage in a chat interface. To do so, we first collected alternative company names using DBpedia [13] to reproduce variability in how users phrase queries, as done in [14]. Second, we generated different variants of the question for the given KPI using Gemini 3.1 Pro [15], in the form of templates where *company name* and *year* are to be replaced. In practice, since the KPI and the company occur several times in our set of triplets, we sampled uniformly the KPI question template and company name once for every triplet.

In the context of RAG over the pages of the corresponding report, we annotated the relevant page for each query. To do so, we took the corresponding ground-truth KPI value and searched for pages containing this value, producing TREC-style relevance judgments [16]. From 118,048 questions we mine over two million candidate (query, page) pairs via unit-normalized value matching (amounts normalized to single dollars through a hierarchy of inline / line / page / document scale hints; EPS and per-share values kept in reported units) and grade each on a 0/1/2 scale (0 not relevant / 1 contextual or comparative mention / 2 primary source for the target FY) with an LLM judge (Qwen 3.6 [17]) using a documented JSON prompt that enforces unit-scaling and comparative-table year rules. A stratified manual spot-check of 60 graded pairs—balanced across grades 0/1/2, alias+value vs. value-only matches, and KPIs—by a domain expert yielded 91.6% agreement with the judge.

3.4. Market-reaction linkage

For each U.S. company-year we link the report to the market's reaction at publication. Using EDGAR's `submissions` feed we select the original 10-K (excluding amendments) whose filer-labelled FY matches and take the earliest `acceptanceDateTime` (UTC), classified as pre-market/intraday/after-hours in ET. We compute raw returns $r_h = (P_t - P_{t+h})/P_t$, where P_t is the closing price on the earnings date (or prior close if pre-market) and $h < 0$ so $r_h > 0$ is a rise, and we define EPS Surprise as:

$$\text{EPS Surprise} = \frac{\text{Reported EPS} - \text{Consensus EPS}}{|\text{Consensus EPS}|}. \tag{1}$$

4. Benchmarks and baselines

The resource induces three tasks of increasing difficulty over the same documents and ground-truth labels: find the page given the question, extract one KPI given the question, and extract every KPI. For the two extraction tasks (Sections 4.2–4.3) we report **recall** = correct / gold (non-answer counts as miss) and **precision** = correct / attempted (non-answers and unverifiable extras excluded), where a prediction is *matched* iff $|pred - gold|/\max(|gold|, 0.5) \leq 0.05\%$ after normalizing monetary amounts to single dollars via the statement's scale hint (EPS/per-share and share counts kept in reported units; no string normalization, numeric value only). For single-KPI, recall coincides with exact-match *accuracy* [18, 19]. To fit into our 4-GPU setup in a reasonable time, we run our experiments over 494

labeled reports of our corpus, with 13,519 KPI ground-truth labels in total (27.2 KPI/report). For the two first tasks, we kept 10,000 questions for which there is at least one relevant page. All reported extraction numbers use the ±0.05% tolerance.

4.1. Page-level retrieval

In a RAG scenario, we evaluated retrieval from (*question, report*) to *relevant page* (e.g., *What was the revenue of Apple Inc. in 2021?* → page 42 of `AAPL_2021`). Table 2 reports this page-level retrieval protocol, comparing lexical BM25 [20] and the learned-sparse SPLADE [21] against the multi-representation dense retriever ColBERT [22]. ColBERT consistently outperforms both BM25 and SPLADE across all metrics. This is expected because ColBERT uses token-level late interaction to compute relevance from fine-grained token-to-token alignments, rather than compressing the entire document into a single vector. Locating the right page remains difficult: ColBERT MRR 0.475 (Recall@5 0.370) vs. SPLADE 0.386 (0.272) and BM25 0.324 (0.265), confirming dense numerical pages defeat off-the-shelf retrievers (4-GPU report-level evaluation, batch 32).

4.2. Conversational long-context extraction

In a conversational chatbot scenario, we evaluated long-context extraction from (*question, report*) to *KPI* (same question as above → `{"found": true, "value": 394328000000, "unit_scale": 1, "page": 42}`). To do so, the model receives an entire OCR'd report (≈100k tokens) and extracts a single specified KPI as a structured JSON object (`found`, `value`, `unit_scale`, `page`); `unit_scale` records the resolved scale hint (inline > line > page > document) and `page` cites the source page. Matching uses the ±0.05% relative-error rule defined above. Prefix caching cuts prefill by ≈21×, making full-corpus evaluation tractable on one GPU server. The left block of Table 3 reports four open-weight models: Qwen3.6-27B reaches 91.4% recall at 93.5% precision, while a model with a unit-scaling error (Nemotron) collapses to 15.0%.

4.3. Comprehensive long-context extraction

The hardest task asks a model to extract all 31 KPIs from a report in a single pass: *report* → *all KPIs* (e.g., one JSON with 31 `kpi: value` pairs, missing keys allowed for absent facts). We evaluated this task on ground-truth labels by recall (coverage of true facts) and precision (correctness of emitted facts). The right block of Table 3 shows that single-value extraction capability does *not* transfer: Ministral, second-best at the needle task (87.9%), collapses to 41.4% recall under structured extraction, emitting many hallucinated cells; conversely Nemotron, near-useless at the needle task, recovers to 65.5% recall once schema constraints suppress its scaling error. Qwen3.6-27B is the only model strong on both, and no model exceeds 80% recall—establishing the task as an open challenge.

5. Case study: Market Sentiment Prediction

We study whether CEO-letter rhetoric in six price-taker industries (specialty chemicals, auto parts, packaged foods, oil & gas exploration & production, oilfield services, mortgage REITs; 2017–2022, 314 non-10-K letters) carries signal about future fundamentals and market reaction.

We train L_2-regularized linear probes on frozen backbone embeddings (full-letter tokens) to predict whether r_{-90d} or EPS Surprise (Eq. 1) falls in the top or bottom decile (10% prevalence). Table 4 reports PR-AUC (5-fold stratified CV) for the top five encoders plus baseline; several combinations land well above the 0.10 random level (e.g., 0.47 for positive EPS surprise, 0.44 for negative 90-day

Table 2 | Per-document page retrieval, averaged over 10,000 questions (494 reports, Test corpus). Recall@k/MRR use binary relevance (rel$\geq$1); nDCG uses graded gains (0/1/2). Best per row in **bold**.

Metric	BM25	SPLADE	ColBERT
Recall@1	0.069	0.099	**0.129**
Recall@3	0.185	0.208	**0.279**
Recall@5	0.265	0.272	**0.370**
MRR	0.324	0.386	**0.475**
nDCG@5	0.221	0.245	**0.333**
nDCG@10	0.280	0.283	**0.383**
Runtime (in seconds)	**66**	160	869

Table 3 | LLM baselines for long-context extraction on Test corpus, all as recall (R) and precision (P) with a match tolerance of ±0.05%. *Conversational*: 10,000 questions. *Comprehensive*: 494 reports, 13,519 cells. Rows ordered by recall; best per column in bold.

	Conversational (single-KPI)		Comprehensive (Multi-KPI)	
Model	**R**	**P**	**R**	**P**
Qwen3.6-27B [17]	**91.4**	**93.5**	**77.2**	**85.7**
Ministral-3-14B [23]	87.9	88.6	41.4	43.6
gpt-oss-20b [24]	85.3	86.8	65.7	73.2
Nemotron-3-Nano-30B [25]	15.0	15.3	65.5	78.0

returns), illustrating cross-modal (text → realized financials) utility. The full ten-encoder comparison is in the repository.

6. Availability, license, and ethics

Availability. The corpus, KPI tables, qrels, benchmark splits, model predictions, and the complete fetch/extract/validate/score toolchain are released under MIT (code) and CC-BY-4.0 (data) licenses with permanent DOIs (https://doi.org/10.57967/hf/10107, https://doi.org/10.57967/hf/9070). All datasets are publicly available on HuggingFace (https://huggingface.co/collections/artefactory/ledger) and code is on GitHub (https://github.com/artefactory/LEDGER). The release includes datasheets [9] and FAIR-compliant [10] metadata, and reuses the standard TREC `qrels` format so `trec_eval` applies directly.

Ethics and limitations. All documents are public regulatory filings and investor reports; no personal data is involved. KPI ground truth derives from official XBRL filings and third-party APIs and may contain restatements or vendor errors; we publish provenance (source and exact tag) for every value and a human OCR-quality audit (Section 3), but the data are for research and do not constitute investment advice. OCR errors in dense tables are quantified by the annotation layer. Public reports may have appeared in LLM pretraining, but our tasks require exact, page-grounded numbers with correct scale, scope and FY, which cannot be solved without the document for the small models evaluated. The corpus, KPI tables, qrels and splits are versioned and will be refreshed with additional filings and improved filtering; page-oracle and per-KPI/layout diagnostics are released as scripts for versioned analysis. The experiments benchmark is U.S.-centric and bounded to 2017–2022 for market-reaction data. We are currently extending coverage to non-English filings, starting with

Table 4 | PR-AUC when predicting EPS Surprise and r_{-90d} from the CEO-letter (top/low deciles); best per column in **bold**.

	Surprise EPS		r_{-90d}	
Encoder	**Top 10%**	**Low 10%**	**Top 10%**	**Low 10%**
BGE-large-en [26]	**0.469**	0.212	0.440	0.335
BGE-M3 [27]	0.429	**0.389**	**0.442**	0.254
ModernBERT [28]	0.421	0.353	0.384	0.285
Gemma [29]	0.415	0.350	0.420	**0.403**
EuroBERT[30]	0.293	0.248	0.310	0.248
BoW+Multi NB	0.277	0.218	0.286	0.244

Table 5 | French extension of LEDGER with corresponding statistics.

	Total	**Avg./report**
Companies	291	-
Reports	1,153	-
KPIs & Questions	21,916	19.0

French ESEF reports (Table 5).

Conclusion. LEDGER turns 4,999 real US annual reports and 1,153 French reports into a corpus of labeled documents and an evaluation suite spanning retrieval and conversational long-context extraction. Strong off-the-shelf systems leave clear headroom on every task, and the released toolchain lets the community extend the resource along time, industry, country, and modality axes.

Acknowledgments This work was done as part of the ArGiMi project, funded by BPIFrance under the France2030 national effort towards numerical common goods.

References

[1] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dillon Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[2] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[3] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering, 2023.

[4] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance, 2023.

[5] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question

answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[6] Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, 2024.

[7] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. Financial report chunking for effective retrieval augmented generation, 2024.

[8] António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*, 2026.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[10] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.

[11] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

[12] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026.

[13] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer, 2007.

[14] Emmanuel Malherbe and Marie-Aude Aufaure. Bridge the terminology gap between recruiters and candidates: A multilingual skills base built from social media and linked data. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 583–590. IEEE, 2016.

[15] Google DeepMind. Gemini 3.1 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf, 2026. Accessed: 2026-06-01.

[16] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[17] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026.

[18] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024.

[19] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

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

[21] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[22] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, 2020.

[23] Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

[24] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

[25] NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for Agentic reasoning, 2025. Technical report.

[26] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[27] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[28] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[29] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

[30] Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, et al. Eurobert: Scaling multilingual encoders for european languages. *arXiv preprint arXiv:2503.05500*, 2025.

GenAI Usage Disclosure

Generative AI assistants were used to help draft and edit prose and to scaffold auxiliary code. All datasets, experimental results, and analyses reported in this paper were produced, executed, and verified by the authors. Independently, several third-party open-weight LLMs are evaluated *as subjects* of the benchmark in Section 4; the LLM judge used to grade retrieval candidates (Section 4.1) is documented in the released toolchain.