DealBench: Measuring Retrieval-Augmented Answering over a Messy Czech–English M&A Data Room with Open-Weight Models

Jakub Dvořák

Prague, Czech Republic

October 2026

Preprint*

Abstract

Retrieval-augmented generation (RAG) is increasingly proposed for due diligence, but real data rooms are confidential, so pipelines are rarely measured on them. DealBench replaces the real room with a seeded generator: a fictional Czech manufacturer's sell-side room of 170 documents (481 PDF pages, scans, merged-cell workbooks, contract versions, e-mails) with planted conflicts and decoys, and 200 bilingual questions whose gold evidence is known down to the page. All 200 questions were audited against the rendered documents, and every miss is attributed to the first pipeline stage that lost the evidence. With open-weight GPU models self-hosted on a national academic grid (7.81 GPU-hours), we score 22 retrieval pipelines and three answered ones on held-out questions. A stronger dense embedder is the only demonstrated gain: Qwen3-Embedding-8B over bge-m3 adds 0.12 recall@10 on test. Fusing lemmatised BM25 into a strong dense retriever hurts (−0.14). The answer-layer gains are not significant at 100 test questions: the frozen pipeline answers 75% correctly against 69% for a naive baseline, and the same pipeline with Nemotron-3-Embed-8B 80%. Flagging planted conflicts is the weak spot, at 2 of 10 on test, and the remaining errors point to retrieving the second source as the next lever. On one real scanned Czech filing, accounting identities raise the cells matching the filed XML from 173 to 532 of 550 (in-sample).

1 Introduction

In a sale of a mid-sized company, the buyer's advisers read a virtual data room: hundreds of statutory statements, management accounts, contracts in several versions, minutes and correspondence, much of it scanned and, in Central Europe, in two languages. Questions such as "what is the outstanding term loan at year end, and does the CIM agree?" need the right document out of many that state nearly the same number, and an honest answer must say when two sources disagree or when the room does not contain the answer at all. Retrieval-augmented generation [16] is the obvious tool, yet the people who could evaluate it on real rooms cannot publish them, and public financial QA benchmarks [4, 14] are built on public English filings.

DealBench takes the opposite route. A deterministic generator writes a fictional company's ledger and renders it into a messy, bilingual data room, so every figure in every document is a ledger fact and the gold evidence of each question is known *by construction*, down to the page and bounding box. The room is built to be hard in the ways real rooms are hard: the same quantity for another period, budget against actual, the same figure restated in five editions of the statements, 47-page CIM drafts, unsigned drafts next to signed contracts, and scans with no text layer. The study is restricted to open-weight models that a team could run on its own hardware, all served on MetaCentrum, the Czech national academic grid.

Contributions.

- A seeded, byte-reproducible Czech–English sell-side data room (170 documents, 481 PDF pages, 7,615 gold spans, 1,042 ledger facts) with 12 planted conflicts, 8 decoys, unanswerable questions, and a full audit of all 200 questions (31 defects, fixed at the source) (Section 3).

*Code and data: generator, harness, run records and this paper at `https://github.com/Majny/dealbench`; the room is seed 7, room version 4. Interactive explorer with every answer, citation and failure diagnosis: `https://dealbench-9x0.pages.dev`.

- An evaluation built on *required evidence groups*, shared by every retrieval metric, the answer grader and a failure funnel that attributes each question to the first stage that lost the gold: parse, chunk, recall, rerank, generation or grounding (Section 4).
- A controlled comparison of 22 retrieval pipelines and three answered pipelines with bootstrap intervals, paired permutation tests and Holm correction, a dev/test protocol with a pre-registered selection rule, and an account of what the test split saw before the final run (Section 6).
- Model selection on a Czech retrieval pack that first reproduces published MTEB scores, and a real-filing case study in which accounting identities drive OCR repair of a 72 dpi scanned balance sheet (Sections 5 and 6.4).

2 Related work

Retrieval and embedding benchmarks. BEIR [21] established zero-shot evaluation across heterogeneous retrieval tasks, and MTEB [17] and its multilingual successor MMTEB [8] are now the default leaderboards for embedders, including Czech tasks such as Belebele retrieval [2]. None of them contains Czech finance or legal retrieval, so we use them only to shortlist.

Financial and legal QA. FinanceBench [14] asks open-book questions over public US filings, FinQA [4] tests numerical reasoning over reports, and CUAD [12] labels clause spans in commercial contracts. DealBench differs in three ways: the room deliberately contains conflicting, superseded and look-alike sources; questions are bilingual over mostly Czech documents; and gold is generated rather than annotated. We considered adding CUAD contracts but its labels mark clause spans, not the effect of a change-of-control clause our questions ask about.

RAG evaluation and citations. Reference-free scoring such as RAGAS [9] and citation evaluation as in ALCE [11] judge answers with models. We instead grade against constructed gold and measure citations structurally (does the cited chunk hold a required evidence group), which is cheaper and auditable but not entailment.

Retrieval components. We compare BM25 [18] over raw tokens and MorphoDiTa lemmas [20], dense embedders including bge-m3 [3] and Qwen3-Embedding [22], reciprocal rank fusion [5] and cross-encoder reranking. Contextual chunk prefixes [1] and page-image retrieval in the style of ColPali [10] were planned but not run (Section 7).

3 The DealBench room

Generator. A seeded generator writes the ledger of "Project Beskyd", a fictional Czech CNC precision-machining supplier, and renders it into eight folders (corporate, financial, commercial, HR, IP–IT, real estate, environmental, correspondence). At seed 7 the room has 170 documents, 481 PDF pages (31 scanned PDFs with no text layer), workbooks, Word files and 61 e-mails. Every page is watermarked as synthetic. Filler figures are *background* ledger facts derived from the core totals (months sum to the year, revenue by country sums to the edition's revenue), so distractors have exact provenance; arithmetic shown in a document (variances, margins) is computed at render time and is not gold. Two builds of one seed are byte-identical, and a full build with scans takes about 3.5 s on a laptop.

Why retrieval is hard. Thirty-one monthly packs with identical layout state revenue, EBITDA, budget, prior-year month and year to date; five editions of the statements restate the same lines; contracts run to 16–20 pages followed by annex tables. A naive BM25 over text units puts every required evidence group in the top 10 for 17% of answerable questions; ten random units do so for under 1% (measured when the room was scaled, at room version 2, before the audit below).

Conflicts, decoys and unanswerables. Twelve conflicts are planted: a quantity such as year-end head count, the lease expiry, outstanding term loans, a litigation provision or a customer's revenue share is stated differently by different sources, and the gold names the prevailing value. Eight decoys state two different values that are *explained* (another period, a restatement, group versus standalone), and must not be flagged. Guard tests render the room and check that each planted claim appears only where it is planted, that no text unit answers an unanswerable question, and that each guard fires on an injected contradiction.

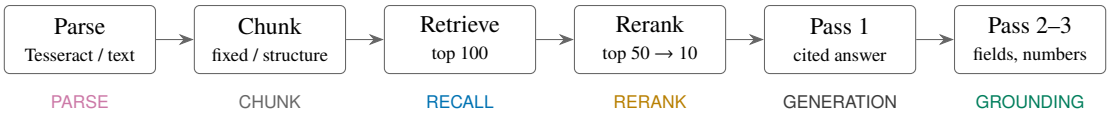


Figure 1: The pipeline and the failure stage charged when a required evidence group is lost at that step. Each first stage returns 100 candidates and the reranker re-scores the top 50 of them. RECALL means a group is outside every first-stage top 100 (or outside the final top 10 when there is no reranker); RERANK means it was in that top 100 but not in the final top 10, which includes gold at first-stage ranks 51–100 that the reranker never scored; GENERATION means all groups were in context but the answer is wrong or abstained; GROUNDING means the answer is right but no citation points at the gold.

Questions and gold groups. The 200 questions span ten categories (Table 4 in the appendix): lookup, numeric, cross-lingual (English questions over Czech documents), multi-hop, conflict, decoy, CIM tie-out, version, clause set and unanswerable. The split is 100 dev and 100 test questions, each with 90 answerable and 10 unanswerable. A fact printed several times in its *home* document (the document a professional would open, e.g. the signed contract for contract terms) forms one evidence group whose members are interchangeable; the same fact in other documents is a non-required *mention*. In room version 4 the questions cite 460 required items in 332 groups and 1,330 mentions. Conflict questions require every conflicting source.

Audit. Before the final runs all 200 questions were read against the rendered documents. The audit found 31 defects on 29 questions: 9 ambiguous questions, 9 cases of missing evidence, 4 wrong answers, 2 wrong required evidence, 2 questions that were not actually unanswerable, 1 evidence item that did not support the answer, 1 mislabeled conflict, 1 mislabeled decoy and 2 others. All were fixed in the generator, each class with a test that fails on the old generator; the fixes changed 23 gold answers, 31 question texts and the evidence of 96 questions, and moved 77 questions across the split (hence room version 4).

4 Evaluation method

Retrieval metrics. All metrics count the same required groups. Recall@k is the share of a question's required groups with a member in the top k (single-fact questions score 0 or 1; multi-hop, conflict and clause-set questions need every group for full credit); recall-all@10 is 1 when every group is in the top 10; MRR uses the first chunk satisfying any group; nDCG@10 gives binary gain once per group. Lenient recall, which also credits mentions, is reported only as a secondary metric. A chunk contains gold only if the evidence is intact in it, so a window that cuts a table row in half has not retrieved it.

Failure attribution. Each question is attributed to the first stage that lost a required group (Figure 1). The recall pool is the first-stage top 100 throughout, also for reranked pipelines whose reranker only sees the top 50, so RERANK is an upper bound on what the reranker itself lost; Section 6 separates the two cases. A stage is recorded even when the answer is still right, because the funnel tracks the first lost hop whether or not the answer needed it; we therefore split the funnel by grade (Section 6).

Answer layer. The generator sees the top 10 chunks with stable ids and document metadata (type, date, version, authority tier). *Pass 1* answers in the question's language, cites every claim, states every value when sources disagree and ranks them by authority, treats rounding, disclosed restatements, group versus standalone, LTM versus FY and FX as explained differences, and answers exactly "Not in data room" plus an information request when there is no evidence. *Pass 2* extracts structured fields (value, unit, period, conflict flag and kind, abstention) under a JSON schema with constrained decoding. *Pass 3* (numbers verified) checks that every number and date in the answer appears in a cited chunk.

Grading. Numbers are compared after unit conversion within the question's tolerance, dates as dates, yes/no by polarity, clause sets by finding every contract, decoys and CIM tie-outs by the conflict flag, planted conflicts by the flag *and* the prevailing value, and unanswerable questions by abstention. No model judge runs, so accuracy on free text is a lower bound. Citation precision is the share of cited chunks holding a required group; citation recall the share

of groups cited. We also report conflicts surfaced (10 test conflicts), false conflicts (6 test decoys), abstention F1 (10 unanswerable), false abstention on the 90 answerable questions and confident-wrong answers (wrong, not abstained, no gold cited).

Statistics. Every mean carries a 95% percentile bootstrap interval over questions (1,000 resamples) [7]. Adjacent configurations are compared with a paired sign-flip permutation test (10,000 permutations) on the questions both score, with a paired bootstrap interval of the difference, and p-values are Holm-corrected [13] over the steps of one comparison call [6]. Families are small (one to five steps); no correction is applied across calls, splits or metrics, and a single-step call is uncorrected.

Protocol and pre-registration. Configurations and the answer layer were chosen on dev. Selection rules were fixed in advance: (R1) only licences permitting commercial use can be picked; (R2) public benchmarks only shortlist; (R3) *if two models are within each other's bootstrap interval, the cheaper one wins*. After the first answer run, two answer-layer iterations were written down with a keep/revert rule before being made (I1: derive the conflict flag from the extracted conflict kind; I2: a pass-2 prompt and schema that classify the kind of difference before setting the flag), judged on dev only. The pipeline was then frozen and two test comparisons were declared as one Holm family: naive baseline → frozen pick, and frozen pick → the alternative embedder, both on accuracy.

5 Models and compute

Candidates. Seven open embedders (bge-m3, granite-embedding-311m, pplx-embed-v1-4b, Qwen3-Embedding-8B, Octen-Embedding-8B, Nemotron-3-Embed-8B, harrier-oss-v1-27b), three rerankers (Qwen3-Reranker-4B, llama-nemotron-rerank-1b-v2, bge-reranker-v2-m3) and one generator (Qwen3.8-27B-FP8, thinking off, temperature 0), each pinned to a Hugging Face revision that is part of every cache key (repository ids and revisions in Table 7). Scans are read by Tesseract 5 [19] (`ces+eng`, 300 dpi).

Czech selection pack. Embedders were scored on six public Czech slices: Belebele Czech–Czech and both cross-lingual directions, WebFAQ (Czech), Czech Wikipedia retrieval and its reranking twin (Table 5 in the appendix). Before any new number was trusted the harness reproduced published MTEB scores: bge-m3 within 0.06 nDCG@10 on all five slices with a published value, and Qwen3-Embedding-8B within 0.44 when given the instruction MTEB used. With the pack's own pre-registered instruction Qwen3-Embedding-8B scores 5.6 points lower on Czech Wikipedia (86.92 against a published 92.54): the selection numbers are prompt-sensitive.

Compute. Every GPU model ran as an offline batch job (PBS Pro, Singularity; vLLM [15] v0.30.0 or sentence-transformers) on MetaCentrum nodes with NVIDIA A40, L40S, A100 40 GB, H100 NVL and RTX PRO 6000 Blackwell 96 GB GPUs. Outputs are imported into a request cache, so every published number reruns on a laptop without a GPU, failing on any cache miss. The five laptop configurations of Section A ran bge-m3 on Apple MPS, with MorphoDiTa and Tesseract locally. The whole study used 46 jobs and 7.81 GPU-hours, failed jobs included (walltime times GPUs held). The six final jobs (reranking for the picked pipeline, the dev iterations and the test run) took 0.68 GPU-hours. Two failures are worth recording for anyone repeating this: under the default xgrammar JSON grammar, 220 of 400 pass-2 replies ran away into whitespace until the token limit (fixed by disabling arbitrary whitespace and repairing via the pipeline's retry turn), and the grader initially parsed the generator's "CZK k" (thousands of crowns) as crowns, fixed on dev.

6 Results

Configuration names encode the first stage: B is BM25, D dense, H hybrid (BM25 and dense fused by RRF), R hybrid plus reranker, DR dense plus reranker and T table-row chunks. A hyphenated suffix names the embedder (PP pplx-embed-v1-4b, NE Nemotron-3-Embed-8B, OE Octen, HA harrier, GR granite, BG bge-m3); without one, D1 and H1L use bge-m3 and D1EM, D2M, H1M and R1M use Qwen3-Embedding-8B. Table 3 decodes every name.

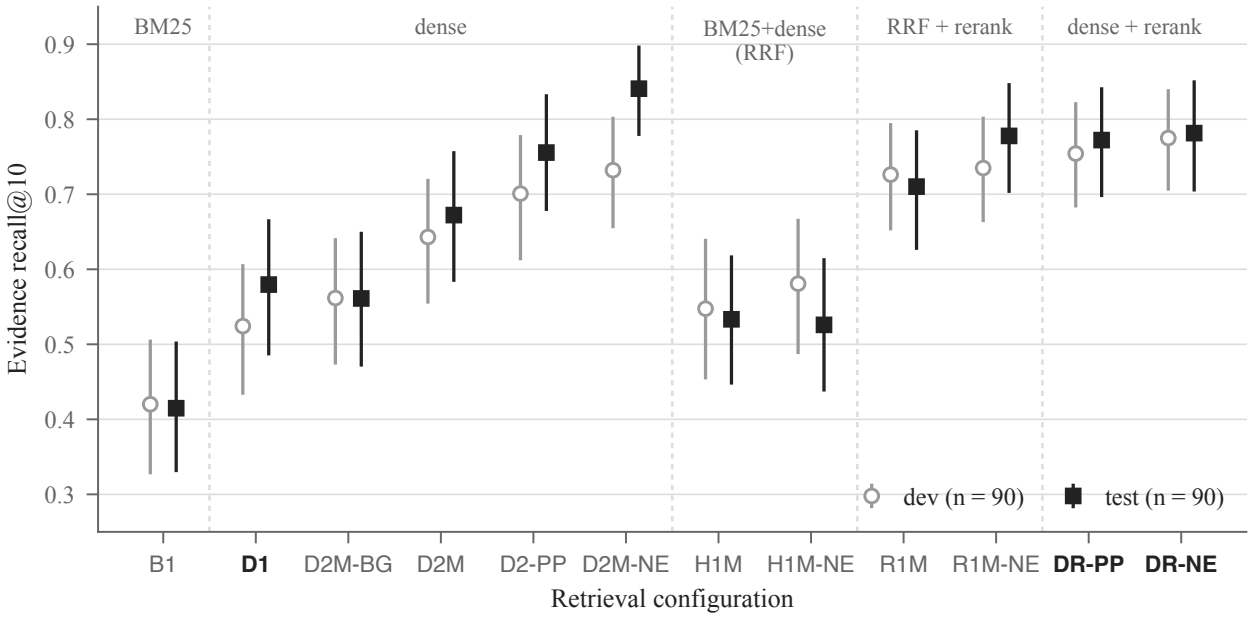


Figure 2: Evidence recall@10 with 95% bootstrap intervals on dev and test (90 answerable questions each), Tesseract corpus. D1 is the naive baseline; DR-PP (pplx) and DR-NE (Nemotron) are the reranked pipelines carried to answers. All 22 configurations are in Table 3.

Table 1: Paired steps on recall@10 (90 answerable questions per split). Δ with paired bootstrap 95% interval; p from a paired permutation test, Holm-corrected within each documented comparison call (families of one to five steps; nothing is corrected across calls, and the single-step last row is uncorrected); bold marks $p < 0.05$ within its family, not study-wide significance. The last row is the comparison behind the embedder pick.

		dev		test	
Step	Change	Δ recall@10	p_{Holm}	Δ recall@10	p_{Holm}
B2 → D1	dense bge-m3 over lemmatised BM25	+0.128 [+0.032, +0.236]	0.063	+0.161 [+0.063, +0.269]	**0.0096**
D1 → D1EM	bge-m3 → Qwen3-Emb-8B	+0.090 [+0.022, +0.164]	**0.036**	+0.118 [+0.043, +0.193]	**0.023**
D1EM → D2M	fixed-512 → structure chunks	+0.028 [−0.040, +0.096]	0.43	−0.025 [−0.111, +0.067]	0.83
D2M → H1M	+ BM25 (lemmas), RRF	−0.095 [−0.164, −0.031]	**0.013**	−0.139 [−0.241, −0.035]	**0.023**
H1M → R1M	+ Qwen3-Reranker-4B	+0.179 [+0.115, +0.254]	**0.0005**	+0.177 [+0.106, +0.257]	**0.0005**
R1M → T1M	table-row chunks	−0.057 [−0.106, −0.021]	**0.021**	−0.017 [−0.054, +0.015]	0.83
D2M → D2M-NE	Qwen3-Emb-8B → Nemotron	+0.089 [+0.028, +0.157]	**0.021**	+0.169 [+0.076, +0.263]	**0.0010**
D2M-NE → DR-NE	+ reranker over Nemotron	+0.043 [−0.005, +0.092]	0.097	−0.059 [−0.120, +0.000]	0.071
D2-PP → DR-PP	+ reranker over pplx	+0.053 [+0.003, +0.109]	0.11	+0.017 [−0.067, +0.102]	0.75
DR-PP → DR-NE	pplx → Nemotron under reranker	+0.021 [+0.004, +0.042]	0.11	+0.009 [−0.007, +0.028]	0.75
D2-PP → D2M-NE	pplx → Nemotron, dense only	+0.031 [−0.015, +0.077]	0.20	+0.085 [+0.017, +0.156]	**0.019**

6.1 Retrieval

Figure 2 and Table 1 show the retrieval waterfall. Three findings hold on both splits. *A stronger dense embedder helps*: bge-m3 → Qwen3-Embedding-8B adds +0.090/+0.118 (dev/test, Holm $p = 0.036/0.023$), and Qwen3-Embedding-8B → Nemotron-3-Embed-8B another +0.089/+0.169. English questions over Czech documents score 0.00 recall@10 with either BM25 configuration on both splits; on dev about 60% of the B2 → D1 gain comes from these cross-lingual questions, on test about a third. *Hybrid fusion hurts a strong dense leg*: adding lemmatised BM25 by RRF lowers recall@10 by 0.095/0.139; the reranker then wins the loss back (+0.179/+0.177). *The reranker shows no demonstrated gain over strong dense retrieval*: over Nemotron it is +0.043 on dev and −0.059 on test (Holm $p = 0.097/0.071$); the highest test recall@10 among all 22 configurations, observed post hoc, is the dense-only D2M-NE (0.841). Lemmatisation

Table 2: Answer layer on the test split (100 questions: 90 answerable, 10 unanswerable), one run of the pipeline frozen on dev. Generator Qwen3.8-27B-FP8 for all three. Mean with 95% bootstrap interval over questions.

Metric	D1 (naive)	DR-PP (frozen pick)	DR-NE
Accuracy	0.690 [0.600, 0.770]	0.750 [0.670, 0.830]	0.800 [0.720, 0.870]
Evidence recall@10	0.580 [0.485, 0.667]	0.772 [0.696, 0.843]	0.781 [0.704, 0.852]
Citation precision	0.448 [0.368, 0.527]	0.445 [0.382, 0.511]	0.475 [0.403, 0.544]
Citation recall	0.535 [0.443, 0.624]	0.722 [0.643, 0.800]	0.724 [0.644, 0.800]
Abstention F1	0.588 [0.357, 0.774]	0.720 [0.476, 0.897]	0.769 [0.556, 0.929]
False abstention	0.156 [0.089, 0.222]	0.067 [0.022, 0.122]	0.067 [0.022, 0.122]
Numbers verified	0.750 [0.647, 0.853]	0.649 [0.545, 0.753]	0.697 [0.592, 0.789]
Conflicts surfaced (10)	0.200 [0.000, 0.500]	0.200 [0.000, 0.500]	0.400 [0.111, 0.728]
same, dev split	0.500 [0.200, 0.818]	0.800 [0.500, 1.000]	0.800 [0.500, 1.000]
False conflict (6 decoys)	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.167 [0.000, 0.500]
Confident wrong	0.040 [0.010, 0.080]	0.020 [0.000, 0.050]	0.010 [0.000, 0.030]

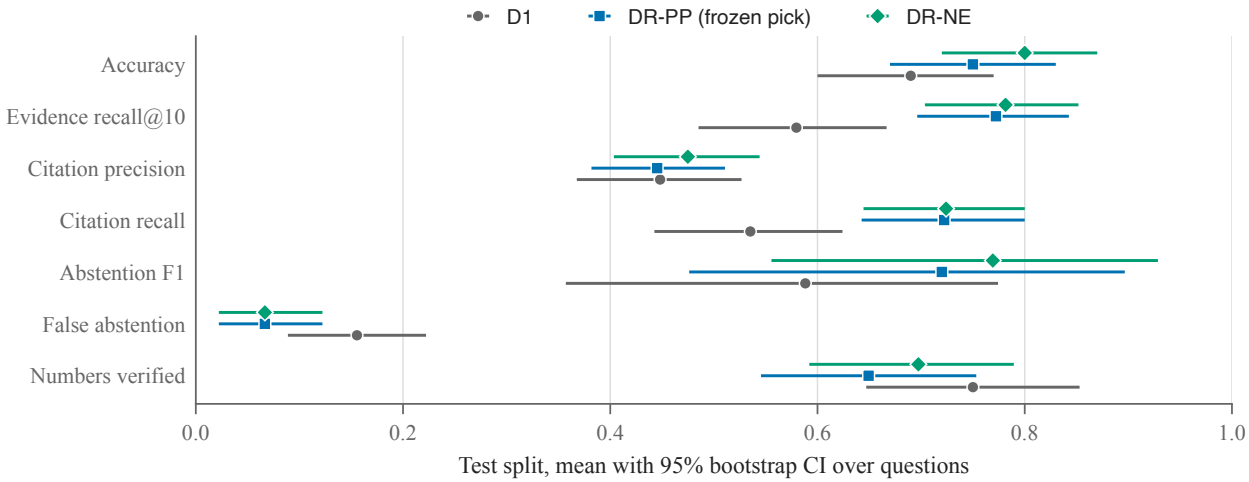


Figure 3: Answer-layer metrics on test with 95% bootstrap intervals (Table 2). Lower is better for false abstention.

alone gives no gain (B1 → B2 −0.024 dev, +0.004 test), and table-row chunks did not help. Parsing is not the bottleneck on this room: the Tesseract corpus recovers 907 of 944 gold items (96.1%), against 919 when digital PDFs are read from their text layer, and that parser swap moves recall@10 by under 0.01.

6.2 Answers

Table 2 and Figure 3 report the frozen pipelines on test. The pre-declared family gives D1 → DR-PP +0.060 accuracy [−0.040, +0.160], Holm $p = 0.31$, and DR-PP → DR-NE +0.050 [0.000, +0.100], raw $p = 0.13$, Holm $p = 0.25$. **Neither is significant.** On dev the first step was +0.12 (raw $p = 0.012$, Holm $p = 0.024$); on test it keeps its direction and its source, fewer needless abstentions (false abstention 0.156 → 0.067) on the back of +0.19 recall@10, but with 100 questions it is not a demonstrated improvement. Citation precision is about one half for all three pipelines (outside the family, DR-PP → DR-NE is +0.030 on the 85 questions both score, raw $p = 0.045$, Holm $p = 0.090$). The model rarely guesses: confident-wrong answers are 1–4%, and 1 of 873 cited ids over the three runs was not among the sources. Numbers verified is lower for DR-PP (0.649) than for D1 (0.750) because the denominators differ: it counts non-abstained answers that state a number, and D1 abstains more. D1 verifies 51 of 68 such answers and DR-PP 50 of 77: the questions DR-PP answers instead of abstaining mostly state a figure that is not in a cited chunk. Conflict flagging remains the weak spot: DR-PP surfaced 2 of 10 planted conflicts; in 5 of the 8 misses a required source was not in the top 10, and in 3 the evidence was at ranks 4–6 but the answer stated one value as confirmed. Table 4 in the appendix breaks accuracy down by category.

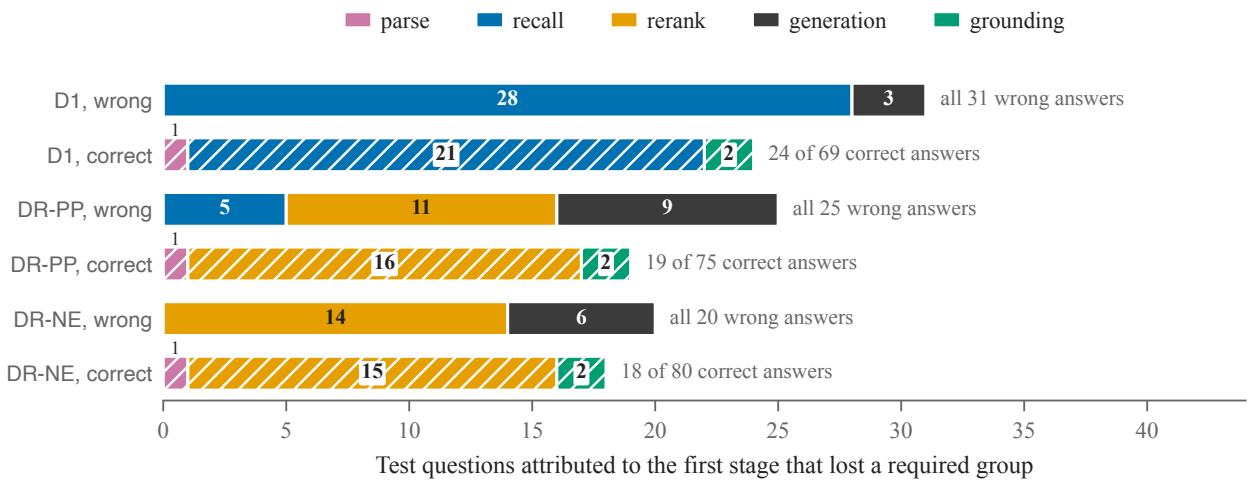


Figure 4: Failure stage on test, split by grade. Solid bars are wrong answers, hatched bars are correct answers on which some required group was still lost (the funnel records the first lost hop whether or not the answer needed it). Each pink parse sliver is one question. Rerank includes gold at first-stage ranks 51–100, which the reranker never scored (3 wrong and 3 correct answers of DR-PP). Questions with no lost group are not shown.

Where answers are lost. Figure 4 splits the funnel by grade. Of DR-PP's 25 wrong answers, 11 are charged to rerank, 5 to recall and 9 to generation, with the evidence in context. Of the 11, 8 had the limiting group inside the reranked top 50 (final ranks 15–50), so the reranker ranked it below 10; the other 3 sat at first-stage ranks 70, 77 and 91, inside the top-100 recall pool but never shown to the reranker, and are really first-stage losses (Figure 1). The same holds for 3 of the 16 correct answers charged to rerank. 19 flagged questions were nevertheless answered correctly. For D1, 28 of 31 errors are recall losses. The embedder is the only demonstrated gain, but the remaining errors point to retrieving the second source for conflict, clause-set and multi-hop questions as the next lever, while roughly a third of the frozen pipeline's errors are the generator's.

Iterations on dev. I1 (flag derived from the kind) raised DR-PP dev accuracy from 0.750 to 0.770 and false conflicts on the 6 dev decoys from 0.167 to 0, with conflicts surfaced unchanged at 8 of 10; I2 changed no graded outcome on DR-PP ($\Delta = 0.000$) and was kept by the rule. On DR-NE, which did not decide, I2 cost two dev questions, including one where pass 2 marked a valid answer as an abstention.

Conflicts on dev and test. The same frozen DR-PP surfaced 8 of 10 conflicts on dev but 2 of 10 on test (Table 2). The two sets are different conflicts: most planted conflicts are asked in Czech and in English, and both versions sit in one split, so dev holds adjusted EBITDA, capex, loan maturity, rent, leverage and change of control, and test holds revenue, largest-customer share, head count, lease expiry, term loans and the litigation provision. With six distinct conflicts per split the intervals only touch ([0.50, 1.00] against [0.00, 0.50]). Retrieval explains most of the gap: every required source was in the top 10 for 7 of 10 dev conflicts but 3 of 10 test conflicts (mean recall@10 0.84 against 0.70), and the three test conflicts with full recall failed at generation. The answer layer was tuned on dev, so some overfitting to dev's conflicts cannot be ruled out, although the dev iterations left conflicts surfaced at 8 of 10 throughout (Table 8).

6.3 Model selection

On the pack, Nemotron-3-Embed-8B leads with 94.58, 0.08 points ahead of pplx-embed-v1-4b at half its weight footprint, and harrier-27b, three times larger, is below both (Figure 5a). Qwen3-Reranker-4B is the best reranker (97.54 MAP against 94.45 and 93.90) and also wins on DealBench dev candidates (0.726 against 0.641 and 0.640). The pack ranks embedders roughly as DealBench does (Figure 5b) but not exactly: harrier-27b scores 93.43 on the pack and 0.574 dev recall@10.

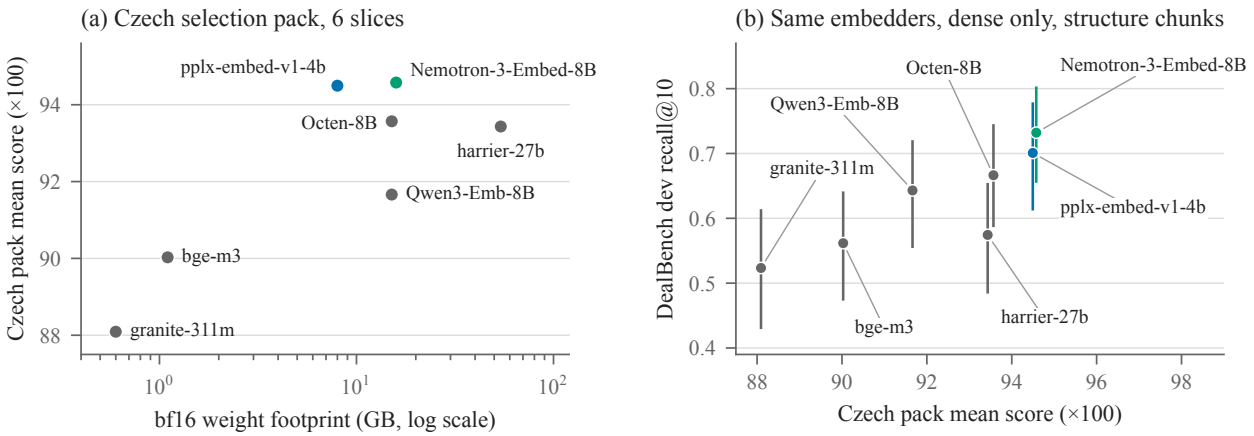


Figure 5: (a) Mean score of seven open embedders over the six Czech pack slices against their bf16 weight footprint. (b) The same pack score against DealBench dev recall@10 with the embedder alone (structure chunks, no reranker), with 95% bootstrap intervals. Blue: the pre-registered pick under rule R3 (pplx-embed-v1-4b); green: Nemotron-3-Embed-8B.

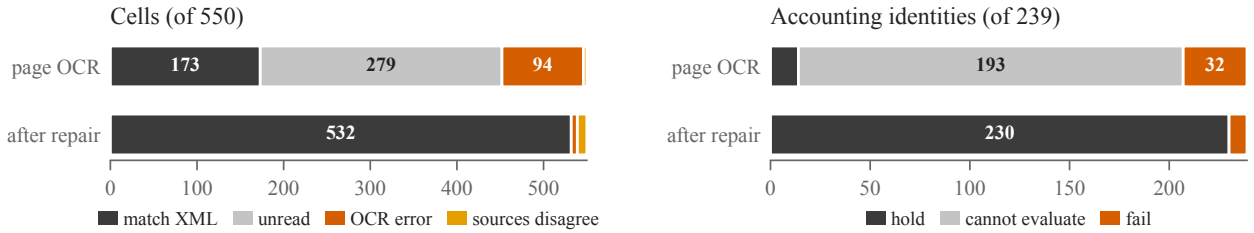


Figure 6: One public FY2024 annual report (48 scanned pages, about 72 dpi): statement cells against the XML filed with it, and accounting identities evaluated on the scan's own values, before and after identity-driven re-reading. In-sample: the repair was developed on this filing.

The pre-registration outcome. Dense only, pplx-embed-v1-4b is 0.031 recall@10 behind Nemotron on dev ($p = 0.20$), inside the interval, so the cheaper-model rule R3 picked it, and DR-PP became the frozen pipeline. On test the dense-only gap is 0.085 ($p = 0.02$), and DR-NE also scores the higher answer accuracy (0.800 vs 0.750, not significant). We report this rather than re-pick: the rule picked the model the test split liked less.

6.4 A real Czech filing

To check the synthetic room against reality, we read one public annual report from the Czech Collection of Documents: 48 scanned pages without a text layer, whose statements embed a 594×840 px page image (about 72 dpi), and the XML statements filed with it. Rows are anchored on the printed row codes (all 198 rows placed: 169 on an exact code, 17 on a near code, 12 by position) and each number is assigned to a ruled grid cell. The identities of the statutory form (subtotals, netto = brutto + korekce, assets = liabilities, balance-sheet result = P&L result) are evaluated on the page OCR; every cell in a failing identity, or an unreadable cell an identity needs, is re-read from upscaled crops, and disagreeing re-reads are settled by the identities among values some OCR pass produced. The XML never chooses a value.

Cells matching the XML rise from 173 to 532 of 550, unread cells fall from 279 to 0, and identities holding on the scan rise from 14 to 230 of 239 (Figure 6). Of 373 page-OCR misreads, 366 were fixed, 7 remain (all inside the 9 identities that still fail, so the output marks them), and none was introduced. This is a low bar for detection: 193 identities could not be evaluated on the page OCR, so most cells were re-read anyway. Where the scan and the XML disagree, 11 prior-year asset cells tracing to three rows, the output says so instead of picking a winner; the XML's prior-year total assets (223,620) differ from the scan (228,364). The procedure was developed on this filing and no held-out filing has been scored, so 532 of 550 is an in-sample figure, not a generalisation estimate.

7 Discussion and limitations

What the test split saw. Test was not untouched before the final run. All 22 retrieval configurations were scored on test, and those numbers were visible when the embedder was picked under R3. An earlier answer run on test showed the conflict-flag failure that motivated I2, so I2 was motivated by reading test answers, although it was developed and kept on dev. The frozen pipeline was then run on test once. Comparisons with the earlier test run (e.g. DR-NE conflicts surfaced $1 \rightarrow 4$ of 10) are test-on-test and descriptive. Separately, the split links multi-hop questions by group and conflict only, so 20 test questions share a required evidence group with a dev question.

Synthetic data. The room is one seed of one generator. Its scans are synthetic and fairly clean (Tesseract finds 98% of digital-PDF gold), commentary is templated, and the company's story is internally consistent in a way real rooms are not. Findings about parsing therefore transfer poorly; the real filing in Section 6.4 is a single, in-sample counterpoint, and it shows that a real 72 dpi scan loses far more at the parse stage than the synthetic room does.

Small categories. Each split has 90 answerable questions, 10 conflicts, 6 decoys and 10 unanswerable questions, so category-level intervals are wide and only large differences can be significant. One generator was evaluated; accuracy is a lower bound on free text because no judge runs; citation precision is structural, not entailment.

Not run. Commercial baselines were out of scope by design. Contextual chunk prefixes [1], an open vision parser, page-image retrieval [10], span-level groundedness and a long-context baseline were planned and not run; the parser comparison only becomes informative with one of them or on real scans.

8 Conclusion

DealBench shows that a confidential task can be benchmarked honestly by generating the data with its gold, auditing that gold, and attributing every miss to a pipeline stage. On a bilingual due-diligence room with open-weight models only, a stronger embedder is the only demonstrated gain on held-out questions, BM25 fusion hurts a strong dense retriever, and neither the reranker's gain over strong dense retrieval nor the answer-layer gain is demonstrated with 100 test questions. The remaining errors point to second-source retrieval as the next lever. The pre-registered rule picked the model the test split liked less, and we report that rather than re-pick. Generator, harness, run records and an interactive explorer with every answer, citation and failure diagnosis are linked on the first page.

Acknowledgements. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The company in the room is fictional; the case study uses one public filing from the Czech Collection of Documents, which is identified and pinned by checksum in the accompanying repository. No affiliation with the filing company is implied.

References

[1] Anthropic. Introducing contextual retrieval. `https://www.anthropic.com/news/contextual-retrieval`, 2024.

[2] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[3] 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. *arXiv preprint arXiv:2402.03216*, 2024.

[4] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan 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*, 2021.

[5] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

[6] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[7] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[8] Kenneth Enevoldsen et al. MMTEB: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, 2025.

[9] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024.

[10] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025.

[11] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[12] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

[13] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979.

[14] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

[15] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

[17] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023.

[18] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[19] Ray Smith. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007.

[20] Jana Straková, Milan Straka, and Jan Hajič. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014.

[21] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

[22] Yanzhao Zhang et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Configurations and full retrieval results

Fixed-512 chunks are 512-word windows; structure chunks follow document structure (200–800 words); table-row chunks split spreadsheets into rows with their header path. Every first stage returns 100 candidates; a reranker re-scores the top 50 of them, and the final top 10 goes to the generator. RRF fuses BM25 over MorphoDiTa lemmas with the dense ranking. B1, B2, D1, H1L and T1L ran on a laptop (bge-m3 on Apple MPS); all others read their vectors and scores from MetaCentrum jobs.

Table 3: All 22 retrieval configurations, Tesseract corpus, 90 answerable questions per split. Recall@10 with 95% bootstrap interval; MRR mean.

		dev		test	
Config	Pipeline	Recall@10	MRR	Recall@10	MRR
B1	BM25, raw tokens, fixed-512	0.420 [0.327, 0.506]	0.376	0.415 [0.330, 0.504]	0.248
B2	BM25, MorphoDiTa lemmas, fixed-512	0.396 [0.302, 0.477]	0.345	0.419 [0.331, 0.509]	0.246
D1	bge-m3, fixed-512	0.524 [0.433, 0.607]	0.374	0.580 [0.485, 0.667]	0.500
H1L	RRF(BM25 lemmas, bge-m3), structure	0.559 [0.461, 0.647]	0.411	0.511 [0.426, 0.600]	0.361
T1L	H1L over table-row chunks	0.506 [0.412, 0.589]	0.385	0.459 [0.370, 0.544]	0.356
D1EM	Qwen3-Emb-8B, fixed-512	0.615 [0.531, 0.694]	0.438	0.697 [0.618, 0.775]	0.466
D2M-GR	granite-311m, structure	0.523 [0.429, 0.614]	0.396	0.582 [0.500, 0.665]	0.335
D2M-BG	bge-m3, structure	0.562 [0.473, 0.641]	0.421	0.561 [0.470, 0.650]	0.495
D2M-HA	harrier-27b, structure	0.574 [0.484, 0.655]	0.408	0.656 [0.567, 0.739]	0.478
D2M	Qwen3-Emb-8B, structure	0.643 [0.554, 0.721]	0.487	0.672 [0.583, 0.757]	0.464
D2M-OE	Octen-8B, structure	0.666 [0.586, 0.745]	0.528	0.715 [0.637, 0.794]	0.496
D2-PP	pplx-embed-v1-4b, structure	0.701 [0.612, 0.779]	0.551	0.756 [0.678, 0.833]	0.526
D2M-NE	Nemotron-3-Embed-8B, structure	0.732 [0.655, 0.803]	0.587	0.841 [0.778, 0.898]	0.535
H1M	RRF(BM25 lemmas, D2M)	0.548 [0.453, 0.641]	0.466	0.533 [0.446, 0.619]	0.341
H1M-NE	RRF(BM25 lemmas, D2M-NE)	0.581 [0.487, 0.667]	0.445	0.526 [0.437, 0.615]	0.400
R1M-BR	H1M + bge-reranker-v2-m3	0.640 [0.552, 0.721]	0.511	0.652 [0.568, 0.741]	0.468
R1M-NR	H1M + Nemotron rerank 1B	0.641 [0.559, 0.718]	0.540	0.689 [0.609, 0.774]	0.449
R1M	H1M + Qwen3-Reranker-4B	0.726 [0.652, 0.795]	0.575	0.710 [0.626, 0.785]	0.494
T1M	R1M over table-row chunks	0.669 [0.591, 0.745]	0.581	0.694 [0.609, 0.770]	0.492
R1M-NE	H1M-NE + Qwen3-Reranker-4B	0.735 [0.663, 0.803]	0.578	0.778 [0.702, 0.848]	0.529
DR-PP	D2-PP + Qwen3-Reranker-4B	0.754 [0.682, 0.823]	0.587	0.772 [0.696, 0.843]	0.540
DR-NE	D2M-NE + Qwen3-Reranker-4B	0.775 [0.705, 0.840]	0.583	0.781 [0.704, 0.852]	0.529

Table 4: Question categories per split, and test accuracy by category for the three answered pipelines (correct / questions).

Category	dev	test	D1	DR-PP	DR-NE
lookup	13	12	11/12	12/12	12/12
numeric	18	17	16/17	17/17	17/17
cross-lingual	12	13	7/13	13/13	13/13
multi-hop	12	13	8/13	7/13	8/13
conflict	10	10	2/10	2/10	4/10
decoy	6	6	5/6	5/6	4/6
CIM tie-out	8	7	2/7	4/7	4/7
version	5	5	5/5	4/5	4/5
clause set	6	7	3/7	2/7	4/7
unanswerable	10	10	10/10	9/10	10/10
all	100	100	69/100	75/100	80/100

Table 5: Czech selection pack, full query sets: nDCG@10×100 on the retrieval slices and MAP×100 on Wiki-cs rerank. Footprint is bf16 weights in GB. † Pre-registered pick (rule R3 applied on the DealBench dev split).

Model	GB	Belebele			WebFAQ	Wiki-cs	Wiki-cs	Mean
		cs/cs	cs-q/en	en-q/cs	ces		rerank	
Nemotron-3-Embed-8B	15.9	97.43	97.16	97.86	87.22	94.47	93.32	94.58
pplx-embed-v1-4b †	8.0	97.11	96.48	97.36	89.26	93.88	92.89	94.50
Octen-8B	15.1	98.28	98.26	99.20	83.48	91.73	90.47	93.57
harrier-27b	54.0	96.51	96.20	96.43	85.94	93.45	92.07	93.43
Qwen3-Emb-8B	15.1	97.96	97.50	98.87	82.86	86.92	85.86	91.66
bge-m3	1.1	94.10	92.94	92.57	81.67	90.11	88.78	90.03
granite-311m	0.6	90.81	87.09	90.77	84.25	88.34	87.30	88.09

B Reproduction of published MTEB scores

Table 6: Published MTEB scores against this harness (nDCG@10×100; Wiki-cs rerank MAP). "Pack" uses the pack's pre-registered instruction; "MTEB" the instruction MTEB used for Qwen3-Embedding-8B. Values as logged in the repository (`docs/mc-runs.md`).

Model	Slice	Published	Pack	MTEB prompt
bge-m3	Belebele cs/cs	94.04	94.10	–
bge-m3	Belebele cs-q/en-doc	92.97	92.94	–
bge-m3	Belebele en-q/cs-doc	92.58	92.57	–
bge-m3	Wiki-cs	90.16	90.11	–
bge-m3	Wiki-cs rerank	88.73	88.78	–
Qwen3-Emb-8B	Belebele cs/cs	98.30	97.96	98.33
Qwen3-Emb-8B	Belebele cs-q/en-doc	98.21	97.50	98.31
Qwen3-Emb-8B	Belebele en-q/cs-doc	99.04	98.87	99.18
Qwen3-Emb-8B	Wiki-cs	92.54	86.92	92.10
Qwen3-Emb-8B	Wiki-cs rerank	91.19	85.86	90.77

C Model identifiers

Table 7: Hugging Face repositories and the pinned revisions (first 12 hex digits of the commit) used in every cache key.

Role	Repository	Revision
Embedder	`BAAI/bge-m3`	`5617a9f61b02`
Embedder	`ibm-granite/granite-embedding-311m-multilingual-r2`	`443995599303`
Embedder	`perplexity-ai/pplx-embed-v1-4b`	`06456497a005`
Embedder	`Qwen/Qwen3-Embedding-8B`	`1d8ad4ca9b3d`
Embedder	`Octen/Octen-Embedding-8B`	`5adcfa292e71`
Embedder	`nvidia/Nemotron-3-Embed-8B-BF16`	`d1f2f25730bb`
Embedder	`microsoft/harrier-oss-v1-27b`	`0c0fc62f6d8a`
Reranker	`Qwen/Qwen3-Reranker-4B`	`22e683669bc0`
Reranker	`nvidia/llama-nemotron-rerank-1b-v2`	`828765652b05`
Reranker	`BAAI/bge-reranker-v2-m3`	`953dc6f6f85a`
Generator	`Qwen/Qwen3.8-27B-FP8`	`017b9c7af6b5`

D Answer-layer iterations on dev

Table 8: Pre-registered iterations on dev, DR-PP + Qwen3.8-27B (as logged in the repository, `docs/iteration-log.md`). Keep rule: conflicts surfaced does not fall, false conflicts do not rise, paired accuracy difference ≥ 0. "Consistency" is the 14 decoy and CIM tie-out questions.

	I0 (before)	I1	I2 (frozen)
Accuracy (100)	0.750 [0.670, 0.830]	0.770 [0.680, 0.860]	0.770 [0.680, 0.860]
Paired Δ vs previous		+0.020, $p = 0.49$	+0.000, $p = 1.00$
Conflicts surfaced (10)	0.800	0.800	0.800
False conflict (6 decoys)	0.167	0.000	0.000
Consistency correct (14)	0.500	0.643	0.643