أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

AI-Driven Smart Search in CMS Glance

Faisal Abdulaziz Altuwaijri,
Muhammad Umair

445101995@student.ksu.edu.sa ,muhammad.umair.aslam@cern.ch

Date: August 24, 2026

Abstract

CMS Glance search is lexical, so the same physics phrased differently does not match: one paper in our corpus ranked 434 of 1,391 for "Higgs boson decay to two photons" and 1 of 1,391 for "diphoton decay of the Higgs boson." This report covers the full project to replace that with retrieval by meaning, across two encoder generations and with no labeled relevance data at any point. The first generation reimplemented SPECTER2's frozen-backbone-plus-task-adapter architecture and established a zero-shot baseline from citation structure alone — 3,502 in-corpus citation edges used as implicit relevance, giving MRR 0.160 — then removed its own objection to fine-tuning by mining 14,260 equation–prose pairs from arXiv LaTeX macro definitions, where a physicist has already written the translation. Training 894K of 111M parameters (0.8%) improved ground-truth median rank from 277 to 116 of 1,391 with the document encoder provably frozen, but the six-pair evaluation could not distinguish its own two candidate models. The second generation replaced the encoder on a mechanical premise — SPECTER2 caps at 512 tokens, Qwen3-Embedding-8B has a 32K window, so the paper body becomes indexable — and tested that premise rather than assuming it. A six-stage gated pipeline verified 1,387 of 1,390 papers as CMS against live INSPIRE, cleaned and chunked them at 99.99% token coverage, and produced 17,169 passage vectors; Late Chunking, the technique the milestone was named after, was built, measured and dropped. Against SPECTER2 on 1,385 shared papers and 200 paired queries, median rank is 2 against 60, every metric significant at $p<0.0001$, with the win decomposing into an encoder effect (−129.8) that dominates the body-indexing effect (−46.4) that motivated the switch. Absolute scores are inflated by label leakage: 0.39 natural-language success@1 is the honest figure, not 0.4450.

Contents

1 Introduction 2

2 Related Work 2

3 Methodology 2

3.1 Method details 2

3.2 Evaluation Procedure 2

4 Results 2

5 Discussion 3

6 Limitations and Future Work 3

7 Conclusion 3

1 Introduction

Motivation and real-world impact. CMS keeps its institutional memory in Glance: analysis notes, papers, PAS documents, member and institute records. Glance's search is lexical — it matches strings. That works for a report number and fails for a physics question. The failure was measured, not assumed: the same paper ranked 434 of 1,391 for "Higgs boson decay to two photons" and 1 of 1,391 for "diphoton decay of the Higgs boson". One need, two phrasings, two entirely different result sets. Because CMS analyses reuse methods heavily — same detector, same reconstruction, same systematic treatments across hundreds of analyses — failing to find prior work means repeating it.

Problem statement. Build a retrieval layer over CMS records that ranks by meaning rather than string overlap, resolves to the passage rather than the document, and can be evaluated without pre-existing relevance labels, because none exist and physicist time is too expensive to spend making them.

Assumptions and limitations of the problem. Five, each confirmed by measurement:

1. Full text is reachable only via Glance DOI → INSPIRE-HEP → arXiv → LaTeX source. PAS documents have no DOI, hence no arXiv ID and no source — structurally unreachable, an accepted permanent gap rather than a bug.
2. No labeled relevance data exists; every evaluation had to be manufactured — first from the citation graph, later from LLM-written queries.
3. The corpus is small (1,386 papers): a real haystack, but too small to fine-tune a large encoder safely.
4. The upstream Glance export is not clean — ≥3 wrong DOIs, and 25 DOIs resolving to nothing.
5. Evaluation queries are machine-written from the papers they target, so every absolute score is inflated by leakage. Comparisons survive; absolute numbers are not capability estimates.

The proposed system. Two generations; the second replaced the first by measurement.

Generation 1 — SPECTER2 with task adapters. A frozen scientific-document encoder with small bottleneck adapters per task (proximity, classification, regression, query). Evaluated zero-shot against citation-derived ground truth, then fine-tuned on 14,260 equation–prose pairs mined from arXiv LaTeX. Its ceiling is structural: SPECTER2 accepts 512 tokens, restricting the index to title and abstract.

Generation 2 — Qwen3-Embedding-8B over the paper body. The premise is mechanical: a 32K-token window makes the body indexable. That premise was treated as a hypothesis and tested — the body beats a front-matter-only index by +0.62 Hit@5_section on the pilot and +0.31 at full scale. On it sits a six-stage gated pipeline, 17,169 passage vectors over 1,386 papers, and a RAG layer that cites per claim and refuses when the corpus can't support an answer. Head-to-head on the same 1,385 papers and 200 queries, Qwen3 reaches median rank 2 against SPECTER2's 60, every difference significant at $p < 0.0001$.

2 Related Work

We reviewed DSSM (2013), the earliest deep semantic matching approach connecting queries and documents through a shared embedding space. We then studied DPR (2020), the dense passage retrieval baseline that established how encoder-based retrieval is trained and used in practice. SciRepEval/SPECTER2 built directly on this line of work and became the basis for our

implementation — we studied its multi-task formulation, its use of control codes to distinguish task types, and its adapter-based approach to producing task-specific embeddings from a single shared encoder rather than training separate models per task. We also read Ke et al.'s "LLMs in Document Intelligence" (ACM TOIS, 2025) for broader context on how LLM-based approaches compare to encoder-based retrieval, which helped confirm that a hybrid, encoder-based retrieval approach was the right fit for our project rather than relying on an LLM alone.

3 Methodology

3.1 Method details

Gen 1. Frozen allenai/specter2_base with Pfeiffer bottleneck adapters per task; control codes route inputs. Citation ground truth: from 1,401 DOIs, Semantic Scholar references, keep only links pointing inward — 408 papers, 3,502 edges; 391 had usable title/abstract text. Term supervision: 1,389 LaTeX sources, 959,395 macro definitions resolved, 26,517 macro-expanded pairs plus 1,667 from authors' "where …" sentences; 139,171 more would need an LLM and were left. Triplets: 25,280 gold pairs → 14,260 distinct in 6,122 groups after removing 10,984 duplicates and 36 eval probes; group_key() blanks digits so $|\eta|<2.4$ and $|\eta|<2.5$ are one fact; split by group, never by row. Training: only [QRY] passed to train_adapter(), [PRX] frozen for gradients; triplet loss, margin 1.0, 3 epochs, batch 32, lr 1e-4.

Gen 2 pipeline. Six stages, each with a gate that can fail.

Stage	Script	Produces
0	verify_cms.py	CMS-relevance verdicts
1	clean_latex.py	clean text + offset map to source
2	measure_clean.py	token distributions on cleaned text
3	figure_summaries.py	never built — figure summaries
4	chunk.py	token spans, not text
5	embed.py	4096-d L2-normalised vectors
6	verify_vectors.py	sanity gates before scoring

Verification uses five signals: A INSPIRE collaborations contains CMS, B report number matches CMS-XXX-YY-NNN (A or B decides), C DOI round-trip diagnoses how a mapping broke, D arXiv category (diagnostic only — CMS

papers cross-list), E LaTeX markers post-fetch. Failures are quarantined with a recorded reason, never silently dropped.

Cleaning detects author blocks by content density, not by title — 1411.4413 hides a 17,350-token CMS author list inside front matter where a title filter cannot see it. Captions are kept (6.5% of text, and they state the headline result). Tables use a CMS-aware converter, never Pandoc: CMS's x column type means 1735,16 = 1735 ± 16, and Pandoc would silently destroy every measurement uncertainty. Physics macros stay verbatim.

Chunking: sections are the vector unit, super-chunks the attention unit. Papers under ROUTE_AT=28,000 go in one pass; above it, partition into super-chunks ≤24,000. Oversized blocks descend a five-rung ladder — section → subsection → paragraph → sentence → hard count — with every piece keeping the parent title, undersized sections merging forward, and equations and tables treated as atomic and never cut. Rung 5 is always logged. Spans, not text, are emitted, which makes pooling a flag rather than a rewrite.

Embedding: hybrid context — title and abstract (~341 tokens) prepended per chunk, encoded independently, pooling an injected <|endoftext|> (151643), not tok.eos_token_id (151645). Prefix suppressed on (front matter) chunks. Score is cosine; a paper scores the max over its chunks. Storage is .npy + _index.csv with no index — 270 MB, brute-force exact cosine in milliseconds. ChromaDB was ruled out: it cannot be written on EOS (SQLite over FUSE).

RAG layer: top-8 passages, ≤2 per paper, near-duplicates dropped, then Qwen3-8B answers from those passages only with [n] citations. Four rules set by measurement: retrieve.py imports the instruction and pooling from embed.py (a paraphrase silently costs recall); passages are built in vector row order and re-checked at startup (a slip by one misattributes every citation while scores stay plausible); MIN_SCORE = 0.68; near-duplicate collapse at Jaccard ≥ 0.7 over the first 30 words, because CMS reuses its detector paragraph across ~1,000 papers.

3.2 Evaluation Procedure

Tools. HuggingFace transformers 5.15.0 + adapters, PyTorch, PyTorch Lightning; ChromaDB (Gen 1) → .npy brute-force cosine (Gen 2); Semantic Scholar, INSPIRE-HEP and arXiv APIs; Gemini for labelling; BM25 as lexical reference; A100-PCIE-40GB bf16 on SWAN, later Kubeflow.

Data. doi_to_arxiv.csv: 1,415 rows / 1,415 DOIs / 1,390 arXiv IDs, 25 blank. Verification kept 1,387; the fetch

landed 1,386. Gen 1: 3,502 citation edges, 391 query papers; 14,260 equation–prose pairs. Gen 2 labels: 50 pilot papers drawn proportionally across 12 physics groups (seed 20260813), then 200 Gemini queries with 303 target sections, under a prompt whose central rule is write the query as someone who has NOT read this paper. Each query carries a type, style, also_relevant list and confidence.

Metrics. Graded relevance: grade 2 = source paper, grade 1 = also_relevant, grade 0 = rest; section level scores whether a target-section chunk was retrieved. Paper: success@1, recall@5/10/50, MRR@10, nDCG@10, median rank. Section: nDCG@10, Hit@5. Every headline difference is paired over the same queries and tested twice — bootstrap 95% CI (B=10,000) and Wilcoxon — reported significant only when both agree.

Gates. G1 cleaning ate no content; G2 tables faithful; C1 spans in-bounds and non-overlapping; C2 no gaps with real text; C3 no span touches a header; C4 no labelled target lost; G4 vectors finite and unit-norm; G5 no homogenisation.

4 Results

4.1 Gen 1 — zero-shot baseline (391 query papers, ~1,337 candidates)

MRR 0.160 · Recall@5 0.050 · Recall@10 0.076 · Recall@20 0.104. The first correct match sits around rank 6, and only ~10% of true citation links appear even in the top 20. No prior baseline existed; these became it.

4.2 Gen 1 — fine-tuning the query adapter

Metric	Baseline	Run 1 (random neg.)	Run 2 (hard neg.)
M1 jargon overlap @10 (target)	2.2	2.8	2.7
M1 control (should not move)	8.0	8.0	6.0
M2 margin jargon vs unrelated	0.0061	0.0837	0.0954
M3 median rank of 1,391	277	116	148
M3 best-ranked paper	5	1	1
Final loss (margin 1.0)	—	0.065	0.287

Random ordering gives 696, so 116 is real. [PRX] cosines are identical to four decimals across all runs, confirming the document encoder never moved. Keep Run 1 — but the evaluation cannot actually distinguish the two: individual M1 pairs swing by up to 4 points in both directions while the mean holds. Separately, a query-expansion module from 43 hand-written concept groups reaches 7.7/10 on the same six pairs against fine-tuning's 2.8; M1 may flatter it by construction, and the two have never been run together.

4.3 Gen 2 — verification and gates

1,387 of 1,390 verified CMS (99.8%), 3 quarantined, 1,386 on disk with 0 unaccounted for. Signal C matched 1,390/1,390 — so the DOI → INSPIRE → arXiv mapping never misfired, and the three contaminants came from DOIs that are wrong in the Glance export itself, relocating the defect upstream. Signal D agreed everywhere; E passed 1,385/1,386, the one failure reviewed and kept (a joint ATLAS+CMS paper typeset with ATLAS's template).

Text removed
Pilot 50: 17.1% chars / 25.3% tokens
Full corpus: 17.4% chars

Papers over the 32K window
Pilot 50: 1
Full corpus: 133 → 62

Chunks
Pilot 50: 568
Full corpus: 17,169 (12.4/paper)

Ladder rungs 1–5
Pilot 50: 318/92/148/6/4
Full corpus: 8,442/2,647/5,487/185/408

C2 token coverage
Pilot 50: 99.99%
Full corpus: 99.99%

C4 targets lost · super-chunks over window
Pilot 50: 0 · 0
Full corpus: 0 · 0

Every parameter re-derived on cleaned text held — 2048 keeps 76% of blocks whole, 256 merges 3.2%, 28000 leaves 93.3% single-pass. G4 clean; G5 found no homogenisation (0.5162 vs 0.5558 random cross-paper).

4.4 Gen 2 — context strategy (pilot: 50 papers, 568 chunks)

System	succ@1	MRR@10	nDCG@10 sec	Hit@5 sec
hybrid (title+abstract)	0.8300	0.8840	0.6955	0.8492
iso (no context)	0.8400	0.8952	0.5271	0.8643
late-b (whole document)	0.8250	0.8765	0.6742	0.7538
late-a (wrong pooling)	0.6850	0.7717	0.4730	0.5980
front-matter only (baseline)	0.7500	0.7968	0.2512	0.2261
BM25 (reference)	0.7850	0.8543	0.4773	0.8241

Paper level is a three-way tie — at 50 papers the task is at ceiling. Section level separates cleanly: hybrid − iso nDCG@10_section +0.1684; hybrid − late-b Hit@5_section +0.0955; iso and late-b each win one and lose the other, and hybrid is the only strategy not significantly worse than any other on anything. The founding premise holds: hybrid over front-matter-only is significant on all seven metrics, Hit@5_section +0.6231. Pooling ablation: injecting <|endoftext|> beats last-content-token pooling on all seven.

4.5 Gen 2 — full corpus, and the head-to-head

System	success@1	MRR@10	median rank	hit@50	Hit@5 sec
specter2	0.1050	0.1498	60	0.4800	n/a
hybrid-front matter (same input as SPECTER2)	0.4000	0.5144	2	0.8250	0.1910
hybrid (full body)	0.4450	0.5702	2	0.9150	0.5025
bm25 (reference)	0.3450	0.4589	3	0.8850	0.4121

Every difference against SPECTER2 is significant at $p < 0.0001$. The gain splits: encoder effect at equal input median rank −129.8 [−165.0, −95.5]; body-indexing effect median rank −46.4 [−70.9, −25.2] and Hit@5_section +0.312, significant on 9 of 10 metrics — success@1 alone (+0.045) does not clear the bar ($p=0.070$). Against BM25 the advantage is at the top of the ranking (success@1 +0.100, MRR@10 +0.111 significant; hit@50 and recall@50 not): dense retrieval is not finding more correct papers, it is ranking them higher. Section metrics are Qwen3-only — SPECTER2 holds one vector per paper, so they are NaN by construction, a capability difference rather than a score difference.

4.6 Breakdowns and RAG

method queries: 0.28 success@1, median rank 7.5 against result's 0.70 and 1.0 — but Hit@5_section 0.4286 exceeds their own success@1. Leakage: BM25 reaches 0.345, keyword beats natural 0.50 to 0.39.

> Quote honestly. Lead with median rank 2 of 1,385 against 60. The capability figure is natural-language success@1 = 0.39, not 0.4450. The ndcg@10_section column is not comparable across systems once masks differ; the paired tests are.

RAG refusal is measured: four real physics questions score 0.78–0.85 at rank 1, while "melting point of tungsten" tops out at 0.5525, so MIN_SCORE = 0.68 sits between them. Near-duplicate detection measured 0.917 Jaccard between two real variants of the reused detector paragraph against 0.05–0.14 for unrelated passages. Both models fit together in ~32 GB of the 42 GB A100

5 Discussion

The encoder mattered more than the context window, and that was not the expectation. The whole justification for abandoning SPECTER2 was 512 tokens versus 32K. That premise holds — but at equal input, Qwen3 still improves median rank by ~130 positions against ~46 for adding the body. The model did more than the architectural argument that motivated the switch. Run only in the full-body configuration, the head-to-head would have credited the window for a win the encoder mostly earned.

Hybrid wins on a token ratio, not a mechanism — and the first explanation was wrong. The initial reading was that high intra-paper cosine means chunks have collapsed. Hybrid disproves the strong form: at 0.8760 it sits nearer late-b (0.9318) than iso (0.6192), yet keeps iso's Hit@5_section while late-b loses 11 points. The mechanism is proportion — a ~341-token prefix beside a ~1,200-token chunk leaves the chunk dominant; late-b's prefix can be 24,000 tokens and washes it out.

Gen 1's gain was real but not what its headline said. Run 1's jargon cosine barely moved (+0.0073) while the unrelated control fell 0.0703 — the model learned what is far apart, not what is close together, which is exactly what random negatives reward. Hard negatives fixed the objective and then produced the failure semi-hard mining predicts: teaching a model to separate very similar short strings teaches sensitivity to small surface differences, and word reordering is one — hence the control falling 8.0 → 6.0.

Method queries fail for a reason no embedding fixes. Their Hit@5_section exceeds their own success@1, so this is paper disambiguation, not passage location: CMS reconstruction language is genuinely near-identical across analyses because the techniques genuinely are shared. It needs metadata — group, year, channel — not a better vector.

The scale-up met expectations exactly where they were forecast to fall. The pilot said its figures were an upper bound; success@1 fell 0.83 → 0.4450 and the body effect halved from +0.62 to +0.31 while staying the largest proportional effect. Nothing surprised, which is itself the result.

The most transferable finding is about failure modes. Across both generations a striking number of defects produced plausible wrong answers without raising an error: a validation loop that never fired; a stale row count that trained Run 2 on half its data; \topcaption silently dropping 143 captions across 41 of 50 papers while every other gate passed; (front matter) renaming itself and breaking 20 of 303 labels; nDCG@10_section inverting so that less text scored better; a compare() call in the wrong block that printed the answer and never tested it. Every one was found by checking a number against an expectation, not by anything crashing.

Two surprises. Cosine similarity cannot detect false negatives at short string length — the guard rejected zero candidates while plain lexical word overlap caught 9,705. And ATOMIC_RE was exponential — 0.0006 s at 10 table rows, 0.094 s at 20, 137.6 s at 30 — running clean through every pilot run before hanging the first corpus run forever. Scale did not reveal a scaling problem; it revealed a correctness problem small data had hidden.

6 Limitations and Future Work

Scope. PAS documents are structurally unreachable (no DOI) — needs a Glance-side full-text path, a supervisor decision. Only papers are indexed; member profiles, institute records and internal notes were never built, though stages 1–4 are directly reusable since CMS internal notes share the LaTeX conventions. 2309.05466 is missing (80 MB against a 60 MB cap) — a one-line change plus one re-fetch. ~26,000 figures were

never extracted; stage 3 is designed but unbuilt and would change routing. The ≥3 wrong Glance DOIs should be raised with Imran.

Evaluation. Label leakage inflates every absolute number — physicist-written queries are the highest-value unblocked item, costing no GPU time, with every capability claim depending on them. The evaluation does not scale with the index (200 queries target only the 50 pilot papers). Gen 1's jargon metric rested on 6 pairs and 1 control and should be widened to 15–20 pairs and 3 controls. Only 48 of 200 queries name any also_relevant paper, which punishes the better model; the confidence field is dead (199 of 200 high); one label is bad.

Retrieval. method queries are the weak spot and the obvious next target — attack it with metadata filtering, not a better vector. Gen 2 uses the encoder entirely off the shelf, while Gen 1 left 139,171 mined pairs unused (~10× the training data) and never combined fine-tuning with query expansion despite the two acting at different points. Dense beats BM25 at the top but not at depth, so hybrid fusion is the natural next experiment. body_of()'s filename-sort dependence remains in the code, so anything added later must be re-checked. The 131K window claim is untested.

Deployment. Nothing is deployed — serve.py is stdlib http.server on a pod culled when idle, and the move to FastAPI is a drop-in. Hosting is open: search needs CPU + ~1 GB, generation needs a GPU with ~32 GB. Glance integration needs data access, not code. The vector store should stay brute-force until a stated latency or concurrency requirement exists.

Not covered by the record. No latency or throughput benchmarks, no user study, no cost accounting beyond "under an hour of A100 time", and no SciRepEval evaluation of Gen 1's fine-tuned models. Gen 1's code is a separate codebase not present in the current repository, so its details rest entirely on the weekly reports.

7 Conclusion

The project set out to make CMS records searchable by meaning and ended with a measured, gated retrieval-and-generation system over the full arXiv-matched corpus. The first generation reimplemented SPECTER2's adapter architecture, built a zero-shot baseline from citation structure alone because no labels existed, then removed its own objection to fine-tuning by mining 14,260 equation–prose pairs from LaTeX macros — improving median rank from 277 to 116 of 1,391 while training 0.8% of parameters, and finding that its own six-pair evaluation could not distinguish its two candidate models. The second replaced the encoder on a mechanical premise, verified the corpus rather than inheriting it, re-derived every chunking parameter on cleaned text with none moving, and chose its context strategy by measurement — discarding Late Chunking, the technique the milestone was named after, when it proved architecturally inadmissible and then lost empirically. The result is 17,169 passage vectors at 99.99% coverage and a head-to-head win over SPECTER2 on every metric at $p < 0.0001$, median rank 2 of 1,385 against 60, honestly decomposed into an encoder effect that dominates the body-indexing effect that motivated the switch. The honest capability figure is 0.39, not 0.4450, and the known weak spot is method queries. Above all, the practice worth carrying forward is the one that caught an exponential regex, 143 dropped captions, a section that renamed itself, and a metric that inverted: none raised an error, and every one was found by checking a number against an expectation.