



TECHNICAL REPORT

Node 1.0: what moves retrieval in a point-of-care medical library

15 Oct 2026

Technical report · 15 Oct 2026

Node 1.0: what moves retrieval in a point-of-care medical library

The Cynaps team, NBALL Technologies India Private Limited

Not available to physicians · superseded by Node 1.1

Does every passage needed to answer a short bedside question reach the part of Cynaps that writes the answer? We measured it stage by stage on 315 development questions, and report what moved it and what did not. Node 1.0 was a research and testing stack and was never available to physicians.

Read online

cynaps.co/research/node-1-0-technical-report

DEPRECATED — superseded by Node 1.1 (10 Oct 2026). Node 1.1 replaced Node 1.0's search models with a new medical dense model and a new general-domain sparse model: the model cards of two of Node 1.0's three search models carry non-commercial licences (S-PubMedBert-MS-MARCO: CC BY-NC 2.0; SPLADE v3: CC BY-NC-SA 4.0); see [Introducing Node 1.1](#).

Abstract

Cynaps is building a chat assistant for practising physicians in India. Its point-of-care mode is designed to answer short bedside questions (a dose, a threshold, a schedule, a contraindication) from a fixed library of about 9.8 million passages: Indian and international guidelines, drug labels and textbooks. Before tuning the part that writes answers, we asked a narrower question: **does every passage needed to answer the question reach the writer?** We measured this stage by stage — in the candidate pool, after the reranker, after a passage filter — on 315 development questions with 378 needed facts, and compared every change with paired tests against the current stack on the same questions.

Three findings shaped the locked stack, which we call **Node 1.0**. (1) Query rewriting was the only change that raised the share of questions with all needed passages in a candidate pool of fixed size (600): from 89.5% to 93.7%, and in the reranked top 90 from 82.2% to 85.1% (15 questions gained, 6 lost; exact McNemar $p = 0.078$, not significant at the 0.05 level). (2) A 4-billion-parameter reranker (Qwen3-Reranker-4B) was within our pre-registered band of the 8-billion one at the top 60 and 90 (82.2% vs 81.9% at the top 90; 4 gained, 3 lost; $p = 1.0$), at about half the time per question, but was weaker at the top 20; reranking deeper than the top 200, or widening the pool to 1,000, did not help. (3) Many plausible fixes did not help or made things worse: contextual retrieval on drug labels, one label per drug, date and authority re-weighting, and local or cheaper passage filters. On a sealed exam held by a separate internal function that did not build the system, limited to key facts the library holds, Node 1.0 reached 57 of 63 items (90.5%; Wilson 95% interval 80.7–95.6%), from a single run whose run-to-run noise has not been measured. These are measurements of whether evidence *reaches* the writer, not of answer accuracy. Cynaps chat is in testing and not yet available to physicians; we plan to open it after its release checks. Node 1.0 was a research and testing stack. It was never available to physicians.

This report measures search. It is not clinical guidance, and nothing in it says an answer is correct or safe. Cynaps is decision support; it is not a medical device and does not practise medicine.

1. Setting

The questions. A point-of-care question is short and specific: the starting dose of a drug in a named population, a treatment threshold, a vaccination schedule, an exception in a guideline. It usually has one right source, and the physician needs that source's exact words and numbers, not a summary of the field.

The library. Node 1.0 searched one fixed library of **9,787,229 passages**: Indian national guidance (for example ICMR and AIIMS documents and national programme material), international and society guidelines, drug labels (156,938 of the passages; the September tests in [§4.3](#) used an earlier library with 166,999 drug-label passages, which included 14,119 homeopathic label passages that the point-of-care library excludes; DailyMed label updates on 7 October added a net 4,058), and medical textbooks, including openly licensed ones. The library has two "lanes" — an India lane and a textbook lane — that were always searched beside the whole library, so that Indian guidance and textbook text are not crowded out by the much larger guideline and label collections.

What the library does not hold. Reach can only be as good as the library. During the build of the exam ([§6](#)), the evaluator checked 121 candidate key facts, by hand under a strict rule, against the library: 64 of 121 checked key facts (53%) are stated in the library. Coverage varies by source type; society guidelines and some national guidance are the largest gaps. The check may under-count: some society guideline documents in the library carry only a generic title, so candidate selection could not use them. This coverage gap is a separate problem from search, and the rest of this report does not address it.

2. The metric: all needed passages in front of the writer

Definition. Each development question has one or more needed facts ("parts"), each with a gold source document and a text pattern that a passage must contain (the fact's own words and numbers). A question counts as reached at depth k when **every** part is matched by a passage in the top k the stage hands on. We report two variants:

- **own** — the matching passage comes from the part's gold document;
- **any** — any passage in the library that states the fact counts (for example the same dose in another label).

We report the share at several stages: **in the pool** (the fused candidate list before reranking), **after the reranker** at the top 20, 60 and 90, and **after the passage filter** (the "pruner"). The top 90 is what the pruner reads; the pruner's output (median about 20 passages) is what the writer reads.

Why not answer accuracy. Three reasons.

1. *Attribution.* An end-to-end grade mixes search, reranking, filtering, the writer and the grader. When an answer is wrong we cannot tell which stage failed. A stage metric names the stage.
2. *Grader noise.* In our own earlier test, re-grading the identical answers to 59 questions with the same model judge moved the clinical-correctness mark on 9 of them (15%). A change of a few points in an end-to-end score is not a signal at our sample sizes.
3. *Order of work.* The writer has not been tuned yet. A writer cannot cite a passage it never receives, so we fix reach first.

What reach is not. Reach says the needed evidence was in front of the writer. It does not say the writer used it correctly, that the answer was safe, or that the library holds the right source (§1). We never call it accuracy.

Statistics. Every arm is compared with a reference on the same questions. We report the paired split — questions gained (+) and lost (–) — and a two-sided exact McNemar p-value. p-values are not corrected for the many arms we ran; treat any single p near 0.05 with care. Where we measured run-to-run noise (an A/A run of the same arm), we say so. Each experiment had a decision rule written down before its first result was read (a pre-registration).

The development set. 315 questions in the point-of-care scope, drawn from a larger development set of 905 (an evaluation hold-out was excluded throughout), with 378 needed parts; 95 of the questions need a drug-label fact. The development questions and their gold facts were written by large language models (Claude family), verified by independent blind model checks and a blind re-derivation, and reviewed by a separate review function. No physician has labelled them (§7).

3. The stack

Node 1.0 ran six steps for every question.

1. **Query rewriting.** Gemini 3.7 Flash (temperature 0, no thinking budget) wrote up to three alternative searches — abbreviations expanded, generic names for brand names, the wording a guideline or label would use — and one sub-query per part when the question asked for more than one thing. On the development set it produced 4.7 queries per question on average. It never answered the question. If the call failed, the original question was searched alone.
2. **Search.** The original question and every rewrite were searched in parallel by three models: a biomedical dense encoder (S-PubMedBert-MS-MARCO; its model card carries CC BY-NC 2.0), a learned sparse model (SPLADE v3, splade-v3-distilbert; its model card carries CC BY-NC-SA 4.0) and a large dense embedding model (Qwen3-Embedding-8B). Each model searched the whole library and the two lanes; the lists were fused by reciprocal-rank fusion (RRF), 600 candidates per query.
3. **Merge.** The lists of the original and the rewrites were fused again by RRF on passage identity (k = 60, equal weights, the original first) into one list of 600.
4. **Rerank.** Qwen3-Reranker-4B (bf16, on the GPU) scored the top 200 of the merged list; the top 90 went on.
5. **Prune.** Gemini 3.7 Flash judged each of the 90 passages on its own: KEEP if it contained information needed to answer the question, DROP otherwise; if unsure, or if the call failed, the passage was kept. The kept passages (median about 20) went to the writer in reranked order.
6. **Fail safe.** If the original search failed or the reranker was unavailable, the question went to the slower Consult mode; Node 1.0 was designed not to answer from an order it was not measured on.

The stack was locked on 8 October 2026 and has not been tuned since. A rebuild of the stack inside the product, checked on 30 development questions with fresh rewrites and fresh searches, reached exactly the same questions as the measurement harness (90.0% own / 96.7% any at the top 90; 0 questions differ) and lost none of them after the pruner.

Time and cost (facts from the measurement runs, not service targets). Measured on one development GPU (an RTX 5090 shared with the search model service), one question at a time:

step	measured
rewriting (Gemini)	p50 3.9 s / p95 8.8 s per question; \$1.37 per 1,000 questions
4B reranker over the top 200	about 3.8 s per question in the rewriting run (52.9 pairs per second on the shared card); 5.0 s p50 in the reranker comparison of \$4.4 (a separate run)
pruner (Gemini, 90 calls, 16 in parallel)	p50 18.8 s / p95 177.5 s per question (315 questions); \$55.57 per 1,000 questions
whole stack, product rebuild, 30 questions	p50 68.1 s / p95 163.2 s

The pruner is the largest measured step and the largest cost. The rows come from different runs; the steps listed add up to about 26.5 s of the 68.1 s end-to-end p50, and the remaining ~40 s was not broken down. Making the pruner faster is open work ([\\$5.5](#)).

4. What moved the number

4.1 The stage-wise picture of Node 1.0

All 315 development questions; % of questions with all needed passages present (own / any).

stage	before (single query)	Node 1.0 (with rewrites)
in the candidate pool (600)	89.5 / 90.5	93.7 / 94.9
union of every query's pool	—	96.2 / 97.1
after the 4B reranker, top 90	82.2 / 85.4	85.1 / 88.9
after the 4B reranker, top 60	80.6 / 83.8	81.3 / 85.7
after the 4B reranker, top 20	64.4 / 69.5	66.7 / 72.1
after the pruner	not measured on all 315 for Node 1.0 (see \$4.4)	

Where the remaining misses sit. For the closest measured arm (rewrites, 4B over the merged top 400), of the 378 needed parts: 323 were in the top 90; 32 were inside the reranked set but ranked below 90 (a ranking problem); 3 were in the merged pool but outside the reranked set (a depth problem); 20 were not in the merged pool at all (a search problem). Most of what is left is ranking.

4.2 Query rewriting: +9 questions net

arm (4B over the merged top 200)	own / any @20	own / any @90	in pool own / any	vs reference, own @90	any @90
reference: original query only	64.4 / 69.5	82.2 / 85.4	89.5 / 90.5	—	—
rewrites by Claude Haiku 5.5	66.0 / 71.4	83.5 / 87.3	92.7 / 94.0	+12/-8, p 0.5034	+12/-6, p 0.2379
rewrites by Gemini 3.7 Flash (Node 1.0)	66.7 / 72.1	85.1 / 88.9	93.7 / 94.9	+15/-6, p 0.0784	+14/-3, p 0.0127

The Gemini rewrites gained 15 questions and lost 6 at the top 90 (own), net +9. That passed our pre-registered bar (net $\geq +3$ at the top 90, and a top-20 share no more than 3 questions lower, net; at the top 20 it was +12/-5, net +7, p 0.1435) but it is not significant at 0.05 on the own measure; on the secondary "any" measure p = 0.0127, uncorrected for the many arms we ran. Rewriting is also the only change in this report that raised the share in a pool of fixed size (600): other changes could only re-order what one query had already found. (A larger pool of 1,000 raised the pool share from 89.5% to 90.5%, [\\$5.1](#), without raising the top 90.) Gemini cost \$1.37 per 1,000 questions against \$0.14 for Haiku, and was chosen on the larger reach gain. On the 95 drug questions the gain at the top 90 was smaller (+3/-1, p 0.625).

4.3 Fusion: three search models and two lanes

Earlier per-corpus tests (September 2026, an earlier copy of the library, dense search only, ranks within each corpus, estimated on sampled libraries: a rank "in the top 10" means no sampled competitor above it, at a resolution of about 35 ranks for guidelines, whose scaling ($\times 34.9$) the review found about 14% too high; open textbooks were scaled $\times 9.18$) showed the 8B embedding model finding needed facts that the smaller biomedical encoder missed:

corpus (needed facts)	comparison, 8B vs the biomedical dense encoder	paired result
guidelines (108)	own document in the top 10 / top 180	+19/-3, p 0.0009 / +15/-4, p 0.0192
drug labels (170)	top 10 / top 180	+19/-10, p 0.136 / +15/-5, p 0.041
core textbooks (126)	any passage in the top 60	84.1% $\rightarrow$ 89.7%, +11/-4, p 0.12
open textbooks (114)	any passage in the top 60	86.8% $\rightarrow$ 93.9%, +8/-0, p 0.008

A review by a separate internal function recomputed every one of these numbers and judged them a lead, not proof: each corpus ran ten comparisons, the guideline top-180 and drug-label results do not survive a correction for that, and 5 of the 15 drug-label gains (half the net gain of +10) came

from facts that sit past the smaller encoder's input length. We therefore added the 8B as a third search model rather than replacing the others, and tested the combination on the current library (fusion order, no reranker, all needed passages own in the top 20):

set	full fusion (3 models + lanes)	change	paired vs full fusion
905 dev questions	59.6%	without the 8B: 50.7%	+14/-94, p < 0.001
905 dev questions	59.6%	without the lanes: 53.7%	+45/-98, p < 0.001
315 point-of-care	61.3%	8B alone, whole library: 45.4%	+9/-59, p < 0.001
315 point-of-care	61.3%	8B in the library and lanes: 54.3%	+11/-33, p 0.0013
315 point-of-care	61.3%	8B + SPLADE, no biomedical encoder: 59.0%	+7/-14, p 0.1892

On its own, the 8B (45.4%) lost to the fusion of all three; the lanes matter; the biomedical encoder adds the least. After the reranker the gap between full fusion and "8B + SPLADE" nearly closes (81.9% vs 81.6% own at the top 90 with the 8B reranker over the top 200; paired split not recorded). Whether two models are enough was the question for Node 1.1 ([§8](#), item 1).

4.4 The reranker: no difference we could detect between 4B and 8B at the top 60 and 90

Saved pool of 600 in fusion order, top 200 reranked, 315 questions:

reranker	own / any @20	own / any @60	own / any @90	vs 8B bf16, own @90	time per question p50 / p95	peak GPU memory
none (fusion order)	62.5 / 67.9	72.4 / 77.5	76.5 / 81.0	—	—	—
Qwen3-Reranker-8B, bf16	65.4 / 73.3	80.0 / 83.2	81.9 / 85.1	—	9.5 s / 22.2 s	19.95 GB
Qwen3-Reranker-8B, int8	65.4 / 73.0	81.0 / 84.1	81.9 / 85.1	+0/-0, p 1.0	7.8 s / 10.0 s	9.19 GB
Qwen3-Reranker-4B, bf16	64.4 / 69.5	80.6 / 83.8	82.2 / 85.4	+4/-3, p 1.0	5.0 s / 6.5 s	8.31 GB

An A/A check (the 4B scored twice at different batch sizes on 60 questions) changed no question, so the noise band was set at 2 questions. The pre-registered bar for "same as 8B" was: own @90 within that band and no significant loss at the top 60 or 90; the top 20 was outside the band by design. By that bar the 4B was within the band of the 8B at the top 60 and 90, the depth the pruner reads. At the top 20 it is weaker: 3 questions fewer on the own measure (64.4% vs 65.4%, outside the band) and 12 fewer on the "any" measure (69.5% vs 73.3%); those paired splits were not recorded. A large reranker matters: the jump from fusion order to either Qwen3 reranker is about 5–6 points at the top 90 (unpaired comparison from the same saved pools). Smaller cross-encoders tried earlier were

not clearly better than fusion order (all $p > 0.14$). They reranked a pool of 200 and were scored at the top 20, paired against fusion order at a pool of 100 (61.3%; a pool of 200 gives the same 61.3%, +2/-2, $p 1.0$). Each model's better head (top 50 or top 100) is shown:

reranker (earlier round, pool 200, top 20, own)	result	vs fusion order 61.3% (pool 100)
bge-reranker-v2-m3 over the top 50	64.1%	+20/-11, $p 0.1496$
Cohere reranker over the top 50	63.8%	+19/-11, $p 0.2005$
Qwen3-Reranker-0.6B over the top 50 (top 100: 57.1%)	59.7%	+14/-19, $p 0.4869$
a biomedical cross-encoder over the top 50 (top 100: 53.0%)	58.1%	+15/-25, $p 0.1539$

4.5 The pruner: from 90 passages to about 20, at a cost of 4 of 315 questions on that input

The writer should not read 90 passages. The pruner asks one question per passage and keeps the needed ones. Measured on the earlier input (the 8B reranker's top 90; all 315 questions):

	own / any	passages kept, median / p90	vs unpruned, own
unpruned top 90	81.9 / 85.1	90 / 90	—
Gemini 3.7 Flash, per passage	80.6 / 84.1	20 / 53	lost 4, gained 0 ($p 0.125$)

Cutting the passages to a median of 20 cost 4 of 315 questions on that input. 49 of the roughly 28,000 calls (28,350) failed and their passages were kept, so the failures can only have raised reach. On Node 1.0's own input we measured the pruner only on the 30-question product rebuild, where it lost no question (median 20 kept, p90 62); the full-set cost on Node 1.0 is not yet measured.

5. What did not move it

Each arm below had its decision rule fixed before the first result. "Reference" is the arm named in each table.

5.1 Deeper reranking and a larger pool

All 315 questions; reference = original query, pool 600, 4B over the top 200 (own @90 82.2%).

arm	own / any @20	own / any @90	in pool own	vs reference, own @20	own @90
4B over the top 400	64.4 / 70.8	82.5 / 85.4	89.5	+6/-6, $p 1.0$	+8/-7, $p 1.0$
4B over all 600	62.2 / 68.9	79.4 / 83.2	89.5	+6/-13, $p 0.1671$	+10/-19, $p 0.136$
4B over 600, then 8B (int8) over its top 200	61.9 / 71.4	80.3 / 82.9	89.5	+13/-21, $p 0.2295$	+10/-16, $p 0.3269$
pool 1,000, fusion order	61.9 / 67.6	77.5 / 81.6	90.5	+23/-31, $p 0.3409$	+4/-19, $p 0.0026$
pool 1,000, 4B over the top 400	64.4 / 70.8	83.2 / 85.7	90.5	+6/-6, $p 1.0$	+9/-6, $p 0.6072$

With rewrites, reranking the merged top 300 or 400 instead of 200 was also no better than Node 1.0 (own @90 83.8% and 83.5% vs 85.1%; +6/-10, p 0.4545 and +9/-14, p 0.4049). Going from a pool of 600 to 1,000 added one point in the pool (89.5% → 90.5% own). The pool-depth curve flattens early: 78.1% at 100, 84.4% at 200, 87.6% at 300, 89.8% at 600, 90.5% at 1,000. One reading: past the top 200, extra candidates mostly give the reranker more chances to promote a wrong passage.

5.2 Contextual retrieval on drug labels

We re-embedded all 156,938 drug-label passages with a metadata header (variant A), and a windowed version of the long label sections (A2); on a 300-label sample we also prepended a one-to-two-sentence context written by Claude Haiku 5.5 (variant B), the approach known as contextual retrieval. Measured inside the full search, before reranking, on the 95 drug questions (own @20, pool 100). The pre-registered rule required all four: a net gain on the 95 drug questions of at least max(3, the A/A flips); no net loss on all 315; "any" @20 not worse by more than 3; and no net loss at own @90 with the pool of 600:

variant (the fair, swap-like comparison)	own @20	vs control (60.0%)
A, header, dense + 8B swapped	61.1%	+1/-0, p 1.0
A2, windows, dense + 8B swapped	56.8%	+2/-5, p 0.4531
A, as a separate list	63.2% (control 62.1%)	+3/-2, p 1.0

Within the 300-label sample, Haiku-written contexts against the raw text changed questions with all parts in the top 10 by +2/-0 (p 0.5). No variant met the rule; A/A runs moved at most one question. Contextual retrieval did not earn its cost on this library at this stage.

5.3 One label per drug

Many drugs have several near-identical labels. Keeping only the first label per substance and route (and refilling the freed places) was meant to make room for other evidence. On the 95 drug questions (any @20, pool 100, the pre-registered primary) it went from 67.4% to 65.3% (+0/-2, p 0.5) and on the own measure from 58.9% to 46.3% (+0/-12, p 0.0005): in 17 questions the gold label itself was the one removed. Not adopted.

5.4 Date and authority

Input: the 4B's order, 315 questions (own @20 / @90 = 64.4% / 82.2%). Pre-registered rule: an arm is a candidate only if own @20, own @90 and "any" @20 all show no net loss. Every arm below fails it; dropping superseded passages fails on own @90 (-1).

arm	own @20	own @90	vs plain, own @20	own @90
drop passages marked superseded	64.4%	81.9%	+0/-0, p 1.0	+0/-1, p 1.0
authority / recency nudge, weight 0.02	62.9%	81.9%	+4/-9, p 0.2668	+0/-1, p 1.0
nudge, weight 0.05	61.3%	81.9%	+5/-15, p 0.0414	+1/-2, p 1.0
nudge, weight 0.10	58.7%	81.6%	+7/-25, p 0.0021	+1/-3, p 0.625

Favouring national and guideline sources pushed needed passages out of the top 20. The metadata is also not trustworthy enough to filter on: 43 of the 61 needed AllMS passages carry a publication year before 2015, and a "latest version" flag exists on 63% of the guideline passages in the reranked top 200 and on none of the other sources' passages.

5.5 Cheaper passage filters

Same input as §4.5, the 8B reranker's top 90, for every row (paired against the Gemini pruner).

filter	own / any	kept, median	vs Gemini, own	cost per 1,000 questions
Gemini 3.7 Flash per passage (reference; n 315)	80.6 / 84.1	20	—	\$55.57
4B reranker score ≥ 0.1 (n 315)	80.3 / 83.2	68	+3/-4, p 1.0	no API cost
4B reranker score ≥ 0.5 (n 315)	75.2 / 78.7	36	+2/-19, p 0.0002	no API cost
Qwen3-4B-Instruct per passage, local (n 315)	6.7 / 11.1	35	+1/-234	no API cost
Claude Haiku 5.5, same prompt (n 50)	82.0 / 84.0	70	+0/-0, p 1.0	\$5.24
Haiku 5.5, strict keep/drop prompt (n 50)	80.0 / 84.0	45	+0/-1, p 1.0	\$5.36
Haiku 5.5, graded, keep score 3 only (n 50)	68.0 / 70.0	9	+0/-7, p 0.0156	\$5.64

A reranker score threshold keeps the needed passages only when it keeps most of the 90. The local 4B instruct model dropped needed passages in almost every question; we report it as measured and did not diagnose the cause; it may reflect our set-up, and it is not a statement about the model in general. Haiku kept reach but kept two to four times as many passages (70 with the same prompt, 45 with the strict prompt, against 19 for Gemini on the same 50); no prompt we tried reached about 20 kept without losing questions, so the full 315-question run was not made (the pre-registered gate required ≥ 40 of 50 and a median of ≤ 30 kept).

5.6 Other search-side changes (earlier round)

Fusion order, pool 100, all needed passages own in the top 20; reference 61.3% (315 questions).

change	own @20	vs reference
at most one passage per document	53.3%	+13/-38, p 0.0006
at most two passages per document	59.0%	+9/-16, p 0.2295
synonym expansion (MeSH + label names), extra sparse search	60.6%	+2/-4, p 0.6875
synonym expansion, extra sparse + 8B search	61.0%	+4/-5, p 1.0
named-drug lane (labels of a named drug, 20 each)	59.4%	+2/-8, p 0.1094
8B weighted $\times 1.5$ in fusion	61.6%	+5/-4, p 1.0
8B weighted $\times 3$ in fusion	59.0%	+8/-15, p 0.21

6. The sealed exam

Protocol (public parts). A separate internal function that did not build the system wrote and sealed a point-of-care search exam before Node 1.0 was run on it. Questions with red-flag (emergency) content were authored and tagged but excluded from the scored set, to be used for the writer later; these results say nothing about reach for emergency questions. Seven items close to questions the builders could see were removed before the run (70 → 63). The scored set is limited to key facts the library holds: all 63 items were in the library, and 69 candidates were excluded as not in the library. The 90.5% below is therefore reach on in-library facts; the coverage gap of §1 (53% of checked key facts in the library) applies on top of it. The engineers who built the stack never saw the exam items; the evaluator ran the search in its own environment from pinned model versions and the locked library snapshot. The separate review function then checked the borderline items by hand. An item counts as reached when at least one passage handed to the writer (after the passage filter) states the item's key fact. The passage must contain the fact's own words and numbers, and a blind hand-check must confirm it states the fact for the asked population and context. Any library passage counts, not only the key's own source document.

Result. Node 1.0 reached **57 of 63 items (90.5%; Wilson 95% interval 80.7–95.6%)**. The reviewer upheld the count after hand-checking three borderline reached items and one borderline miss, with no re-run. Because a failed pruner call keeps its passage, failures could have inflated reach; the result became final only after a pruner-only audit of the recorded inputs: in 57 of 57 reached items the confirming passage was kept by a successful call (2 of 5,130 calls failed). For information: in the exam run the pruner kept a median of 32 of 90 passages, against about 20 on the development set, so the writer would have read more than §3 states.

How to read it.

- *A single run.* The exam was run once. Run-to-run noise (the language-model rewrites and pruner are part of the stack) has not been measured on it, so we cannot say how much a repeat would vary. Per-search failures (rewrite legs, 8B deadline fallbacks) were not logged in that run.
- *Small.* 63 items give a wide interval (80.7–95.6%); one item is 1.6 points.
- *Reach, not answers.* The exam measures whether the needed evidence reaches the writer, not whether the final answer is correct or safe. The writer is evaluated separately and has not been released.
- *Counts only.* We do not publish exam items or their topics; they stay with the evaluator so the set can be reused without contamination.

7. Limitations

- **Not reproducible.** Node 1.0's library snapshot was deleted on 16 Oct 2026, after Node 1.1 replaced it, so these runs cannot be repeated. If an error is found later, we will publish a dated correction here.

- **AI-written development questions, no physician labels.** The 315 development questions and their gold facts were written by large language models (Claude family), verified by independent blind model checks and a blind re-derivation, and reviewed by a separate review function (§2). No physician has labelled them. Questions written from library passages are likely easier to find than real bedside questions.
- **Selection on the development set.** Every choice in §4–5 was made on the same 315 questions. Pre-registered rules limit, but do not remove, the risk of fitting to them. Four pre-registrations (rewriting, reranker, pruner, date and authority) called for a one-time confirmation of the chosen arm on a separate practice set; those confirmations were not run by the evaluation function, and the retrieval team's records show none; the sealed exam is the only traced check on fresh items.
- **Reach, not answer quality.** No result here says an answer was correct, complete or safe.
- **One library.** All results are for one library and its gaps (§1); they may not transfer to another corpus.
- **One run, little noise measurement.** Most arms ran once. Where we measured A/A noise it was small (0–1 questions, a 2-question band for the reranker), but the language-model steps (rewrites, pruner) were not re-run for noise.
- **315 questions and many comparisons.** Gains of a few questions are within what chance can produce; p-values are uncorrected. Some comparisons were run on 50 questions only.
- **Who evaluated.** Evaluation and review were performed by a separate internal function that did not build the system; neither is external to the company.
- **Costs and times** were measured on one shared development GPU and the providers' published prices at the time; they are facts about those runs, not service commitments.

8. What happened next (Node 1.0 is superseded)

1. **Node 1.1 replaced Node 1.0 as the point-of-care search stack under test on 10 October 2026**, for licences: a new medical dense model and a new general-domain sparse model replaced Node 1.0's three, two of whose model cards carry non-commercial licences, with no loss of reach we could detect on the same 315 questions (273 vs 268 at the top 90; +13/–8, p 0.38). See [Introducing Node 1.1](#).
2. **Node 1.1 on this report's sealed exam** (63 items, before the library changes): 55/63 reached (87.3%; Wilson 95% interval 76.9–93.4%), 2 items net below Node 1.0 (+1/–3), a difference that is not significant (p 0.625); run noise has not been measured. Node 1.1 replaced Node 1.0 as the stack under test on licence grounds, with that result documented.
3. **Node 1.1's fresh final exam** (80 items, 65 scored): 55/65 reached (84.6%; Wilson 95% interval 73.9–91.4%). One run.
4. **A public benchmark.** We plan to report the same stage-wise reach on a public medical retrieval benchmark (for example MIRAGE or R2MED), so that others could compare.

5. **Fine-tuning on physician-reviewed synthetic questions.** We plan to fine-tune the passage filter, the reranker and the search models. Fine-tuning will use synthetic questions written from the library and reviewed by physicians. It will not use patient data or the content physicians enter in Cynaps.
6. **Library gaps (§1):** we plan to fill the largest gaps first (society guidelines and national guidance), subject to licences.
7. **Noise.** We plan to repeat runs of the language-model steps, and an exam, before any claim of improvement.

References

Checked 15 Oct 2026 against each primary page (arXiv, ACL Anthology, Crossref DOI record or model card).

- Cormack GV, Clarke CLA, Büttcher S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. SIGIR 2009. doi:10.1145/1571941.1572114.
- Lassance C, Déjean H, Formal T, Clinchant S. SPLADE-v3: New baselines for SPLADE. arXiv:2403.06789, 2024.
- Zhang Y, Li M, Long D, Zhang X, et al. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176, 2025.
- Deka P, Jurek-Loughrey A, Deepak P. Improved methods to aid unsupervised evidence-based fact checking for online health news. Journal of Data Intelligence 3(4):474–504, 2022. doi:10.26421/JDI3.4-5. (Citation requested by the S-PubMedBert-MS-MARCO model card, huggingface.co/pritamdeka/S-PubMedBert-MS-MARCO.)
- Anthropic. Introducing contextual retrieval. Web page, 19 September 2024. <https://www.anthropic.com/news/contextual-retrieval> (not peer reviewed).
- Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. Findings of ACL 2024, pp. 6233–6251. doi:10.18653/v1/2024.findings-acl.372; arXiv:2402.13178.
- Zhang X, Li L, Zhou X, Liu Z. R2MED: A benchmark for reasoning-driven medical retrieval. arXiv:2505.14558, 2025 (preprint).
- Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22(158):209–212, 1927. doi:10.1080/01621459.1927.10502953.

Product and model names belong to their owners and are used only to identify what we tested; no endorsement is implied.

Appendix A. Glossary

own / any — the needed fact matched in its gold document / in any passage. **@k** — within the top k passages. **pool** — the fused candidate list before reranking. **RRF** — reciprocal-rank fusion. **lane** — a sub-library (India, textbook) searched beside the whole library. **pruner** — the per-passage keep/drop filter before the writer. **+a/-b** — questions gained / lost against the reference on the same questions.