NEXT MOCA

Needlepath
The Context Relevance Engine

You would never brief your CEO for a major decision by forwarding ten thousand unread emails. Yet many AI agents are briefed exactly that way before every call: send the entire history, let the model dig. Every irrelevant page costs money, and worse, it buries the page that matters. Needlepath prepares a focused briefing instead.

Wrong optimization is worse than no optimization.

Needlepath selects the original records that matter for each call, untouched, in milliseconds. And when the whole file is what the moment needs, **it hands over the whole file**. Every number below comes from public tests anyone can re-run.

+15 points

finding one buried fact

The classic needle in a haystack, statistically significant. Across the whole suite, the focused packet beat sending everything by about 5 points, also significant.

23-37%

of input tokens removed

In the tested runs, without rewriting a single record. Fewer tokens in means a smaller model bill.

Milliseconds

to prepare each packet

Selection takes 8.5 to 14.5 milliseconds in the published benchmark. Alternatives that pick with a second AI model take seconds per question.

$15.67

vs $24.53 sending everything

Total cost of the full benchmark run. The cheaper configuration also produced the better answers. Same questions, same model.

EXECUTIVE SUMMARY

Every model call gets a briefing. The question is who prepares it.

AI agents accumulate state as they work: documents, tool results, artifacts, approvals, history. Left unmanaged, stale memories pollute the next response, old tool output crowds out the actual request, and required artifacts silently drop out of context. Before every call, something decides what the model reads. Needlepath sits between the agent runtime and the model call, exactly where that decision is made. There are four ways to prepare that packet.

Send the complete file
The answer is in there, but buried, and you pay for every page, on every call.

Rewrite it shorter
A summary fits nicely, but exact names, numbers, and constraints can vanish in the rewrite, and the request's intent can drift with them.

Use a heavyweight selector
Picking pages beats rewriting them, but this approach needs its own second AI model and takes seconds per question.

Use Needlepath
A focused packet of original records, untouched, in milliseconds. And when the whole file is what the moment needs, it hands over the whole file.

1 · Record
Operational state is captured across tools, documents, artifacts, tenant-scoped memory, and decisions.

2 · Select
The records that matter for this call are chosen, in milliseconds, based on the task at hand.

3 · Govern
Every decision is recorded and reversible, with full context as the always-available fallback.

What changes for you

COST
A smaller model bill: fewer tokens sent on every call, with savings that compound at scale.

QUALITY
Better answers: the page that matters is no longer buried under pages that do not.

TRUST
Every packet decision is recorded and explainable: what was included, what was left out, and why.

Make the model work smarter, not read more.

Needlepath decides **what the model reads** before it answers. Decide first. Compress second.

THE EVIDENCE

Four contests, same public exams

Every method ran on public, industry-standard tests, scored by each test's published scoring method, with every result published for re-checking.

1

When the decisive fact is buried

The critical detail sits deep inside a 100-message thread. Sending the complete file put the answer in the room, but surrounded by noise that cost money and competed for the model's attention. The rewrite lost exact wording and scored below sending everything. Needlepath's focused packet beat them all, including the complete file.

2

When the answer cannot wait

A live conversation, where every pause is visible. In the published runs, the alternatives that decide with a second AI model took one second to over thirty seconds per question. Needlepath decided in 8.5 to 14.5 milliseconds.

3

When every field must remain exact

Orders, forms, controlled documents. A summary can be almost right and operationally wrong: a dropped decimal breaks the order. On these tests Needlepath matched the complete file's accuracy, and because it never rewrites a record, names, figures, and forms stay exactly as written.

4

When the complete file is the right packet

Some work needs the entire history. Needlepath was the only contestant in the benchmark with a tested check for exactly this: recognize those workloads and hand over exactly what you would have sent anyway, at no extra token cost. Wrong optimization is worse than no optimization, and this is the engine that acts on that.

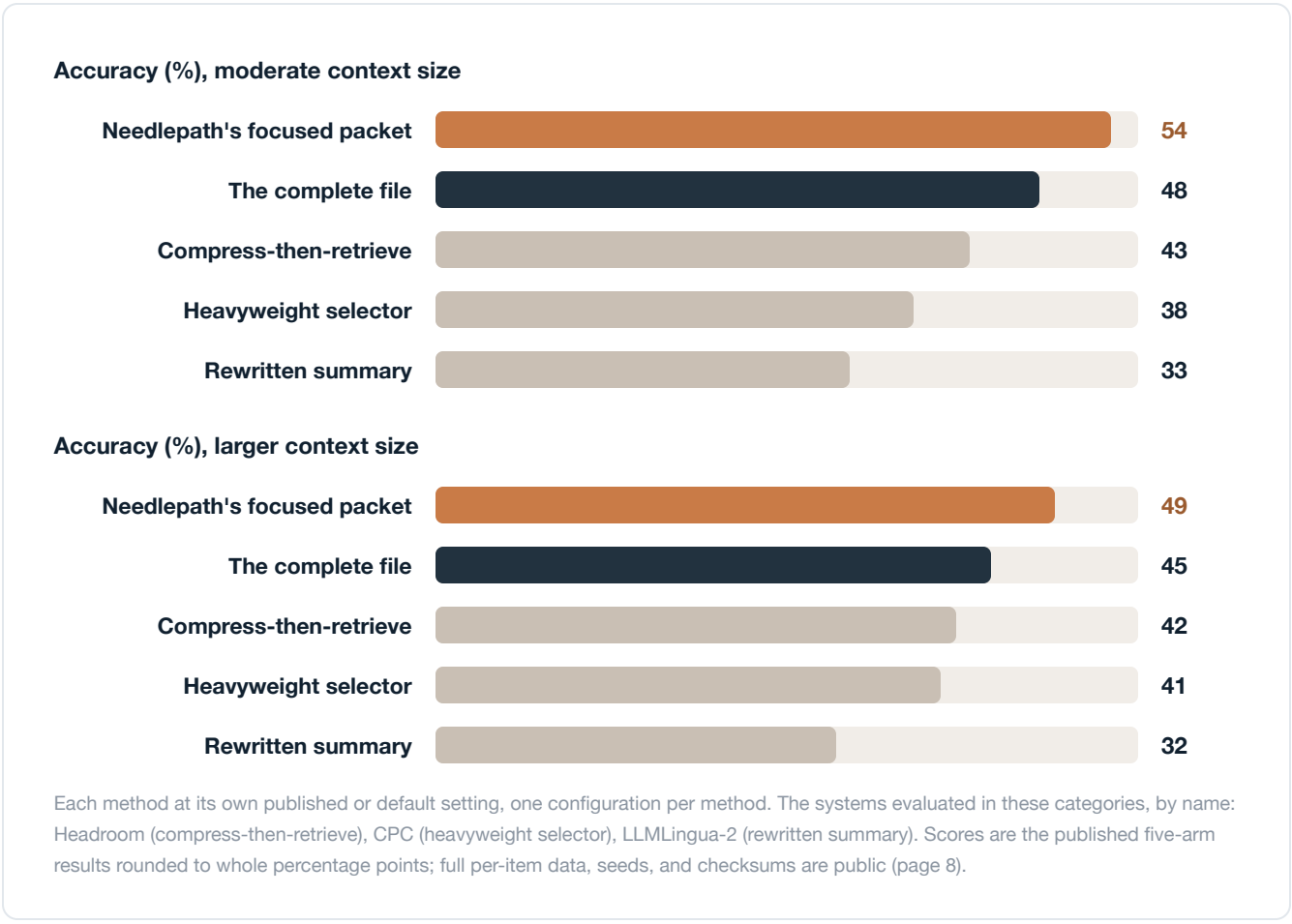
Method, in one line

Same questions, same model family, each test's published scoring method, compared answer by answer. Per-item results and reproduction artifacts are published for inspection.

THE SCOREBOARD

Five ways to prepare the packet. One winner at both levels.

One matched public exam, five ways of preparing the same packet for the same model, two difficulty levels. Higher is better.



Each method at its own published or default setting, one configuration per method. The systems evaluated in these categories, by name: Headroom (compress-then-retrieve), CPC (heavyweight selector), LLMLingua-2 (rewritten summary). Scores are the published five-arm results rounded to whole percentage points; full per-item data, seeds, and checksums are public (page 8).

The extra pages were not just costing money. They were costing accuracy.

Needlepath beat every other approach at both levels, **including the complete file**. And the rewritten summary came last, behind even sending everything: rewriting loses what selection keeps.

THE FULL TEST CARD

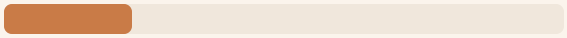
Every suite we published, at a glance

A serious engine should win where winning is possible and do no harm everywhere else. That is exactly the pattern across the published runs.

Won: better answers, fewer tokens · Did no harm, by design · Stepped aside, full context delivered

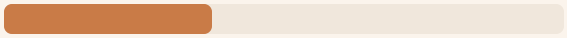
Where it won

Long-context search, 8K — **23.1%**



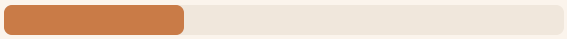
tokens removed · answers **+5.8 points** better · 1,300 questions

Long-context search, 16K — **37.3%**



tokens removed · answers **+4.7 points** better · 1,300 questions

Public re-run, open harness — **32%**



fewer input tokens · **+4.99 points**, statistically significant · **$15.67 vs $24.53**

The win case is noisy context with a clear question, and the gains grow as context grows. The public re-run confirms it on infrastructure anyone can use.

Where it did no harm

Tool calling BFCL, 100 items	**No change** within 1 point, forms protected
Document QA SQuAD v2, 100 items	**No change** trims only when the answer survives
Factual QA TruthfulQA, 100 items	**0 answers changed** clean context left alone
Reasoning GSM8K	**Neutral** nothing to remove on a clean page
Agent loop AppWorld, live episodes	**Stepped aside, 100%** full history delivered, unchanged

These rows are deliberate behavior, not failures. They are what makes the win rows trustworthy.

Wins where winning is possible. No harm everywhere else.

Headline lifts are averages across 26 task cells; the public re-run's **+4.99 points** is the statistically significant confirmation. Every number traces to the published artifacts on the last page.

WHERE THE WINS COME FROM

Needlepath wins when the context looks like real work

Messy histories, repeated tool output, evidence buried in accumulated state. Inside the long-context suite, the gains concentrate exactly where enterprise context is hardest. Each result below is from 100 questions per task per size, with per-task statistics published.

Finding one buried fact

+12 points at 8K, +15 at 16K

The classic needle in a haystack. Statistically significant at both sizes: the clearest single proof that removing noise helps the model see.

Many needles, one answer

+13 points at both sizes

Questions needing several values retrieved at once. A consistent gain: the supporting records survive selection together.

Counting and extraction

+7 to +18 points

Signal counting and extraction across sizes. Repeated noise is exactly what selection removes best.

Hard multi-key lookups

+12 points at 8K

Statistically significant. Where several keys compete for attention, a focused packet keeps the right ones in view.

Long-document QA

Statistically even

Within a few points of sending everything, in both directions. Disclosed plainly: the win here is the token saving, not the score.

Conservative by design

Falls back, not forward

Where order, forms, or answerability need more coverage, Needlepath keeps more or steps aside entirely. Protecting the answer beats maximizing the cut.

The pattern is the product.

The harder the context, the bigger the gain. The cleaner the context, the lighter the touch. That is what **relevance, governed per call** looks like in the data.

WHAT THIS IS WORTH

Durable savings, not a false economy

Cutting tokens is easy. Cutting tokens while answers get better is the part that pays for itself, and the part that survives scrutiny.

MONEY

In the tested runs, Needlepath removed 23.1% of input tokens at the moderate pile size and 37.3% at the larger one, without altering a single original page. Your exact savings depend on model pricing and caching.

TIME

The packet decision happens in milliseconds. At a thousand calls, the alternatives add minutes to hours of accumulated waiting; Needlepath adds about ten seconds total.

PRECISION

On the public long-context tests, Needlepath used fewer input tokens and produced better answers than sending everything: roughly 5 percentage points better, a statistically significant margin.

HONESTY

Where trimming does not pay, Needlepath measures that per workload and steps aside, handing over exactly what you would have sent anyway, at no extra token cost.

Try before you trust

Before you commit to anything, Needlepath will watch your real workloads in shadow mode and tell you honestly how much it would save, and where it would choose to step aside.

The engine that knows the difference.

Busy, noisy context with a clear question: Needlepath engages, and that is where the gains come from. Dense, critical context where every page is load-bearing: **it steps aside**, and stepping aside adds nothing to your bill.

WHERE IT APPLIES

Anywhere AI acts over accumulated business state

Support, finance, sales operations, legal review, security operations, research, field service, procurement, and compliance.

SUPPORT

Customer support and success

Case histories, knowledge articles, tickets, and prior resolutions need selective evidence rather than full-history prompts.

FINANCE

Financial operations

Invoices, purchase orders, approvals, and audit traces require exact figures and explainable fallback behavior.

REVENUE

Sales and revenue operations

CRM records, account context, quotes, and contracts benefit from selection of the current customer state.

CONTROL

Legal, compliance, and security

Policies, logs, contracts, and approvals require traceability and conservative handling of risky tasks.

OPS

Field service and procurement

Work orders, vendor details, and process steps demand accurate recall across long-running workflows.

RESEARCH

Research and knowledge work

Long documents, evidence packets, and prior analysis need the pages that answer the question, not the whole shelf.

What to measure

Money. Input tokens saved and total model cost impact.

Time. Response time under selected context.

Quality. Answer usefulness and correctness vs today's baseline.

Trust. How often it stepped aside, and the recorded reason each time.

MAKE IT REAL

Four steps on your own traces

A successful evaluation replays real work rather than synthetic demos. Same model, your traces, with and without Needlepath.

1

Pick one workflow

A high-volume agent workflow where documents, tool results, approvals, or history accumulate.

2

Replay your traces

Run your own real traces with and without Needlepath, on the same model, side by side.

3

Read the report

Answer quality, pages removed, cost saved, and where it chose to step aside, with reasons.

4

Set gates, then scale

Define thresholds for savings, quality, and fallback rate, then expand with telemetry watching every step.

For your technical team

A companion Technical Edition covers the protocols, statistics, disclosed limits, and reproduction instructions. Every number in both editions comes from the same public tests with published methods, pinned configurations, seeds, and checksums.

Benchmark Needlepath on your traces.

One workflow, your data, an honest report. **www.nextmoca.com**

CHECK OUR WORK

Your engineers can verify every number

Nothing in this brief asks for trust. The benchmark harness is open source, the paper is archived with a DOI, and every released run ships the per-item results and checksums needed to re-run and contest it.

Open benchmark harness

- Same questions and models for every method, one configuration per method, paired statistics.
- Each test scored by its published scoring method as implemented in the released harness.
- Per-item outputs, seeds, and checksums published for every released run.

`github.com/nextmoca/context-selection-bench`

Archived research

- Benchmark paper and data archive: `doi.org/10.5281/zenodo.21502502`
- Needlepath method paper: `doi.org/10.5281/zenodo.21142800`

Current published figures supersede earlier internal ones; the archived record is canonical.

Why we publish everything

The AI optimization market is full of numbers that cannot be checked. Ours can be, on purpose: it is the difference between a claim and a result.

Send the pages that matter.

Better answers, a smaller bill, and an honest engine that knows when to step aside. **Needlepath, the context relevance engine, by Next Moca.**