NEXT MOCA BENCHMARK BRIEF

Needlepath.
The context relevance engine.

Your CEO is about to make a major decision. You would never prepare them by forwarding ten thousand unread emails and asking them to find what matters. Yet many AI agents are briefed exactly that way before every decision they make: send the entire accumulated history, let the model dig. Every irrelevant page costs money, and worse, it buries the page that matters.

Better answers than sending everything

On public tests, a focused packet beat the full pile.

A quarter to over a third of input tokens removed

In the tested runs, without rewriting a single record.

The packet is built in milliseconds

Before every call, with no second model in the selection path.

Stop paying for tokens that are irrelevant and can harm the precision of your LLM.

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.

THE IDEA

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

AI agents pile up state as they work: documents, tool results, approvals, history. Before every call, something decides what the model reads. There are four ways to prepare that packet:

A

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

B

Rewrite it shorter.
A summary fits nicely, but exact names, numbers, and order details can vanish in the rewrite.

C

Use a heavyweight selection system.
Picking pages beats rewriting them, but this one needs its own expensive machine and takes seconds per question.

D

Use Needlepath.
It assembles 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.

This is not making the prompt shorter. It is **deciding what the model reads** *before it answers.*

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.

THE CONTESTS

Four contests. The same public exams. Published reference metrics.

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

CONTEST 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.

CONTEST 2 When the answer cannot wait

A live conversation, where every pause is visible. The heavyweight systems took one to over thirty seconds to decide what to keep. Needlepath decided in about a hundredth of a second. (Sending everything requires no selection step, but it keeps the token cost and distraction shown in contest 1.)

CONTEST 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. Needlepath sends original records and matched the complete file's accuracy on these tests, with forms preserved exactly.

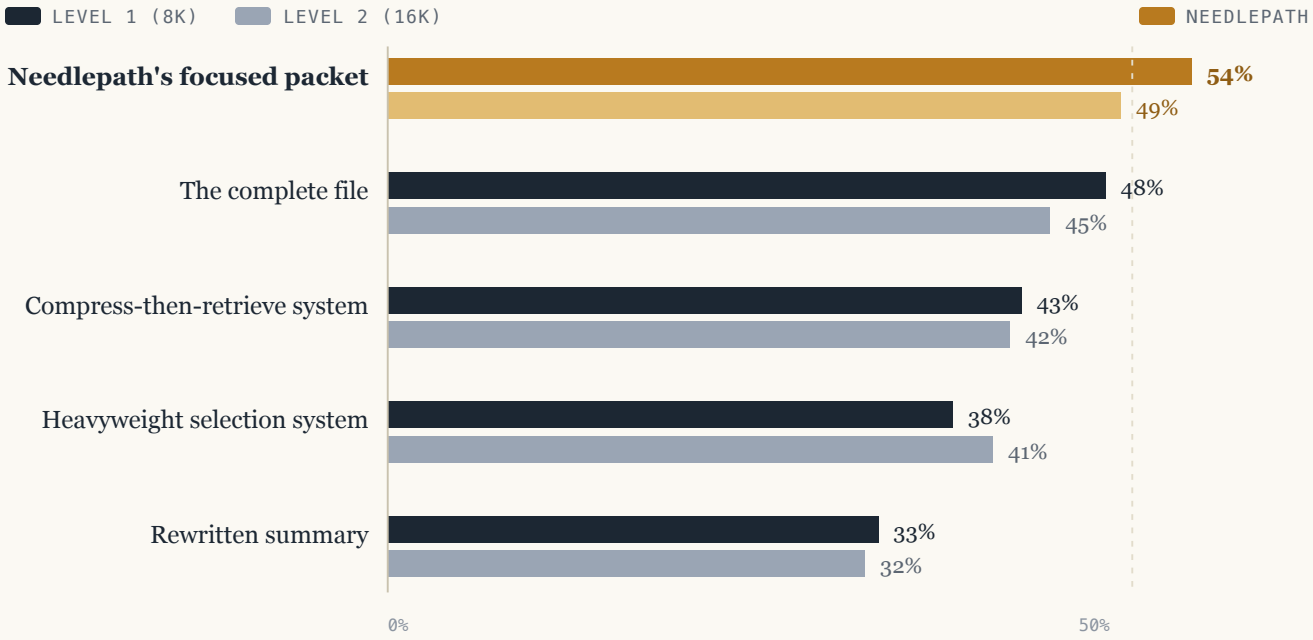
CONTEST 4 When the complete file is the right packet

Some work needs the entire history. Needlepath is the only contestant with a tested gate for exactly this: it detects those workloads and hands over exactly what you would have sent anyway, at no extra cost. Wrong optimization is worse than no optimization, and this is the engine that acts on that.

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.

THE SCOREBOARD

One matched public exam. Five ways to prepare the packet. Higher is better.



What to take from it: Needlepath's focused packet beat every other approach at both difficulty levels, including the complete file, which means the extra pages were not just costing money, they were costing accuracy. And the rewritten summary came last, behind even sending everything.

METHOD: SAME QUESTIONS, SAME MODEL FAMILY, PUBLISHED REFERENCE METRICS, PAIRED STATISTICS. PER-ITEM RESULTS AND REPRODUCIBILITY ARTIFACTS ARE PUBLISHED FOR INSPECTION.

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.

THE VALUE

What this is worth

MONEY

Fewer pages, same originals

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. The resulting cost reduction depends on your model pricing and caching.

TIME

Milliseconds, not seconds

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. (Calculated from each method's measured per-call overhead; model time excluded.)

PRECISION

Fewer pages, better answers

On the public long-context tests, Needlepath used fewer input tokens and produced better answers than sending everything. That is the difference between durable savings and a false economy.

HONESTY

It knows its own envelope

Where trimming does not pay, Needlepath measures that per workload and stands aside, handing over exactly what you would have sent anyway, at no extra cost. Before you commit to anything, it will watch your real workloads in shadow mode and tell you honestly how much it would save.

Wrong optimization is worse than no optimization.

Needlepath is built on that principle, and tested against it in public.

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.

MAKE IT REAL

Put your own workflows on the clock.

1. **Pick one high-volume workflow**
   Choose a path where documents, tool results, approvals, or history accumulate.

2. **Replay your own real traces**
   With and without Needlepath, same model, so the comparison is decision-grade.

3. **Read the report**
   Answer quality, pages removed, cost saved, and where it chose to stand aside.

4. **Scale where the value is obvious**
   Move from one workflow to every model call that benefits from task-aware selection.

For your technical team: a companion Technical Edition of this brief covers the protocols, statistics, disclosed limits, and reproduction instructions. Every number in both editions comes from the same public, industry-standard tests with published methods, pinned configurations, seeds, and checksums.

Make every token earn its place.

Your models are already powerful. Needlepath gives them the pages that make the next action right, and removes the pages that only burn time and money.

Benchmark Needlepath on your traces

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.

VERIFY

Check our work

The benchmark harness is open source and the research is archived with DOIs. Per-item outputs, seeds, and sha256 checksums are published for every released run.

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

`Benchmark paper and data: doi.org/10.5281/zenodo.21502502`

`Needlepath method paper: doi.org/10.5281/zenodo.21142800`

© 2026 NEXT MOCA GLOBAL, INC. ALL RIGHTS RESERVED.