NEXT MOCA

Needlepath
The Context Relevance Engine

Before every model call, Needlepath selects the records that are relevant to the current task, verbatim, in milliseconds, and stands down when reduction would not pay. For enterprise engineering, product, and AI platform leaders deciding how to govern agent context.

Wrong optimization is worse than no optimization.

Compression rewrites text and hopes nothing load-bearing was lost. Needlepath decides **which records belong in the call at all**, keeps them verbatim, and falls back to full context whenever that is the safer answer. Every number below resolves to a public, re-runnable artifact.

+4.99 pp

Answer quality vs full context

Public-harness replication on RULER, pooled across 8K and 16K. 95% CI [+3.66, +6.32], p = 0.0001, paired per item.

−32%

Input tokens, same run

11,784 to 7,977 average input tokens per item. Full-prompt reductions of 23.1% at 8K and 37.3% at 16K in the flagship run.

8.5-14.5 ms

Selection decision time

In-process measurement class. No second model deployment in the selection path.

$15.67

Cheapest arm in the benchmark

Full benchmark cost with Needlepath vs $24.53 sending full context. Better answers, lower bill, same items.

THE CATEGORY

Decide first. Compress second.

Context engineering has four strategies: write, select, compress, isolate. Compression got the startups. Selection gets the results. Needlepath is the selection layer: it decides what your model reads.

1

Record

Capture operational state across tools, documents, artifacts, memories, workflow decisions, and prior run context.

2

Select

Choose the task-critical records for each model call using request intent, tool intent, output type, entities, and risk indicators.

3

Govern

Return an auditable context decision, and expand or fall back to full context when precision, schema fidelity, or evidence coverage requires it.

Selection is not prompt shortening.

Kept records remain **verbatim**. Nothing is paraphrased, summarized, or rewritten on the way to the model. Token reduction comes from leaving irrelevant records out, not from editing the ones that stay in.

What changes for customers

ECONOMICS

Less state sent to the model

Lower token spend and a cleaner latency profile as noisy, repeated state is removed before the call.

QUALITY

More relevant evidence in context

The record that makes the next action correct is preserved instead of drowned by unrelated history.

GOVERNANCE

Audit-ready context decisions

Explain why a record was selected, why an artifact was included, or why the full-context path was safer.

SAFETY

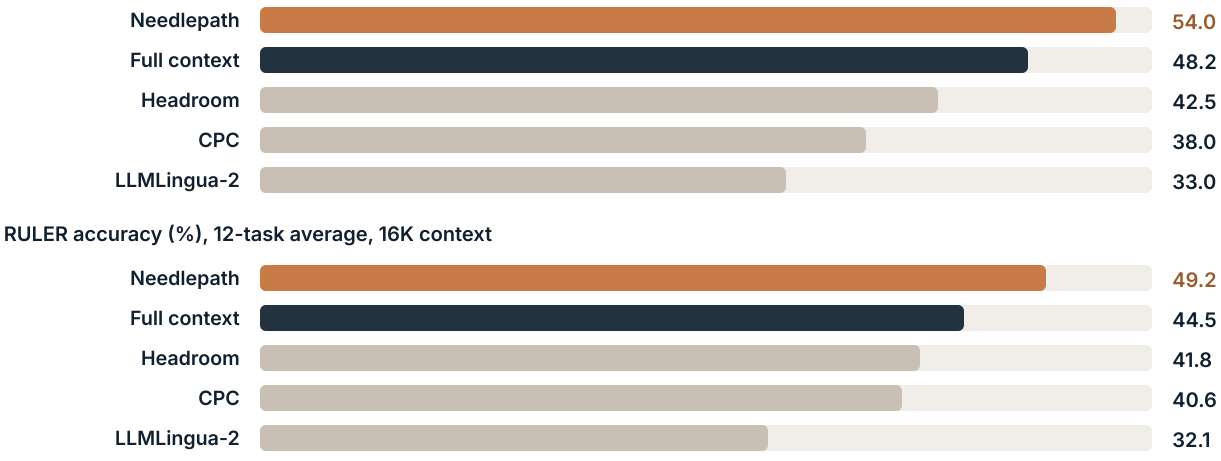
Knows when not to act

Where reduction does not pay, Needlepath stands down and the model receives exactly what you would have sent anyway.

BENCHMARK EVIDENCE

Better answers at lower cost, under a matched protocol

Five configurations, same task items, same models, paired statistics. Needlepath scored above every other tested configuration at both context lengths, and the pure-compression arm scored below full context at both.



Each method at its own published or default operating point, one arm per method. Accuracy lifts vs full context are directional unless cell-level significance is cited. Sources: flagship run and public-harness replication, per-item data published.

Replication, public harness
+4.99 pp
Pooled vs full context, 95% CI [+3.66, +6.32], McNemar p = 0.0001, significant at both lengths.

Prompt tokens removed
23.1% / 37.3%
Full-prompt reduction at 8K and 16K in the flagship run. Replication run: 32% fewer input tokens.

Benchmark cost
$15.67 vs $24.53
Needlepath was the cheapest arm in the replication run. Full context was the most expensive.

HOW IT WORKS

Task-critical state recall

Needlepath sits between the agent runtime and the model call. For each call it selects the state record that makes the next action correct: the uploaded file, tool result, artifact, workflow decision, customer record, error trace, or approval.

Fast filters, no second model call.

Selection decisions take **8.5 to 14.5 milliseconds** (in-process measurement class). There is no second model deployment in the selection path, so there is no extra model bill and no seconds-long detour added to every agent step.

Core design principles

Task-aware selection
The current request determines which state is useful. Needlepath does not apply a fixed cutoff or a generic compression rule.

Grounding preservation
Records needed for files, retrieved evidence, external accounts, compliance controls, or artifacts are preserved when required, verbatim.

Fail-open safety
When selection would introduce correctness or completeness risk, the system routes through the baseline full-context path.

Measurable telemetry
Teams can measure input savings, output behavior, fallback rate, and quality preservation at the workflow level.

Where Needlepath sits

Between the agent runtime and the model call, as a context governance layer that returns either a selected context or a baseline pass-through decision. It composes with everything downstream: compression, caching, encodings, and gateways all still work, on a smaller and more relevant payload.

SAFE BY DESIGN

Needlepath knows when not to act

Most optimization tools have one failure mode: they keep optimizing when they should not. Needlepath reads the composition of the context before acting, and where reduction does not pay, it stands down.

The envelope gate

Before selecting, Needlepath measures whether the context actually contains irrelevant state to remove: the spread of relevance scores and the stability of the task signature. The check runs in under a millisecond with no model calls. Junk-heavy context with a stable question: engage. Dense, load-bearing state with a shifting question: stand down.

Stand-down is byte-identical

When the gate declines to select, the model receives exactly the context you would have sent without Needlepath, wherever full context is deliverable. Not an approximation of it: the same bytes. The safest possible failure mode is already the default.

Agent-loop suite (AppWorld)
100% stand-down
Load-bearing agent state, shifting task: the gate declined on every step, byte-identical to full history.

Factual suite (TruthfulQA)
Byte-identical no-op
Zero flipped answers across arms. Nothing to remove, so nothing was touched.

Long-context suite (RULER)
89% engage
Noisy haystack, stable question: the regime where selection wins, and where the quality lift comes from.

The envelope is about composition, not size.

Needlepath wins where much of the context is irrelevant to a stable query. It stands down where state is load-bearing and the working question shifts every turn. It measures this **per workload, per call**, and declining to act costs you nothing.

COMPETITIVE POSITION

Needlepath decides what belongs in context

Compression tools make content smaller. Needlepath works one layer earlier: it governs which records reach the model call at all. The two compose, and the deciding layer is the one that protects correctness.

APPROACH	WHAT IT DOES	ENTERPRISE LIMIT	NEEDLEPATH POSITION
Token-pruning compression (research methods)	Rewrites or prunes prompt text after context is assembled.	On RULER, the pure-compression arm scored below sending full context at both lengths.	Select records first; rewriting is never applied to what stays in.
Model-based selection (research methods)	Uses a second model pass to decide what to keep.	Requires its own model deployment; published setups add seconds per item.	At each method's own recommended settings, Needlepath wins on accuracy and by roughly 400x on decision latency.
Compress-then-retrieve wrappers	Compresses content into a store and retrieves on demand.	Retrieval round-trips carry their own token and latency cost.	Benchmarked head to head in our published harness; full per-item results are public.
Needlepath	**Selects task-critical records before the call, verbatim, with an auditable decision.**	**Stands down, byte-identically, where reduction does not pay.**	**The context relevance engine: decide first, compress second.**

Every claim resolves to a public artifact

The comparisons above are not vendor marketing: they come from a published, matched-protocol benchmark with per-item outputs, seeds, and sha256 manifests that anyone can re-run and contest. The harness is open source and the paper is archived with a DOI (page 8).

ENTERPRISE WORKFLOWS

Where to start with Needlepath

Needlepath applies anywhere AI must act 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, transcripts, and prior resolutions need selective evidence rather than full-history prompts.

FINANCE

Financial operations

Invoices, purchase orders, approvals, exceptions, and audit traces require entity fidelity and explainable fallback behavior.

REVENUE

Sales and revenue operations

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

CONTROL

Legal, compliance, and security

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

OPS

Field service and procurement

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

RESEARCH

Research and knowledge workflows

Long documents, evidence packets, citations, and prior analysis require answerability-aware selection instead of bulk context.

Customer success metrics

Token economics: input tokens saved and total model cost impact. **Latency profile:** end-to-end response time under selected context. **Quality preservation:** task usefulness, correctness, and grounding fidelity vs baseline. **Operational safety:** stand-down rate, fail-open rate, and explanations for fallback decisions.

CUSTOMER ADOPTION

Recommended evaluation path

A successful evaluation replays real traces rather than relying only on synthetic prompts. The goal is to measure economics, latency, quality preservation, and stand-down behavior on your actual state graph.

1

Select workflows

Choose one or two high-volume agent workflows with meaningful state growth.

2

Replay traces

Run baseline and Needlepath-selected paths over representative production traces.

3

Gate and expand

Review quality, set rollout thresholds, then expand from selected calls to additional agent families with telemetry-backed monitoring.

Verify everything yourself

Open benchmark harness

- Matched protocol: same items, same models, one arm per method, paired statistics.
- Each suite scored by its published reference metric as implemented in the released harness, never a bespoke metric of ours.
- Per-item outputs, seeds, and sha256 manifests published for every released run.

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

Archived research

- Benchmark paper and data deposit (versioned 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.

Send the state that matters.

Preserve the evidence that grounds the answer. Stand down when full context is the safer path. **Needlepath governs operational state.** www.nextmoca.com