

AI WORKFLOW EFFICIENCY

TokenHoloFold

Keep your model.
Reduce the work around it.

Complete supported operations locally.
Verify the result before release.
Send less task data where it helps.

UK patent application GB2623308.0

Recorded studies and a proposed business evaluation



THE OPPORTUNITY

AI workflows carry more than inference cost.

Repeated operations can carry context, call a model and repeat checks even when the task has a supported local route.

The commercial question is how much eligible work your application can complete locally at the required quality.

Context

Repeated task data and state transfer.

Round trips

Provider calls add work and waiting.

Verification

A useful answer needs the right checks.

Scale

Recurring eligible work creates recurring value.

WHAT THE PRODUCT DOES

A local route for supported work.

An application can keep its existing model and connect a qualified runtime for operations that fit.

01 Select a supported request

The application supplies the agreed operation and required context.

02 Check and execute locally

The runtime checks scope and performs the supported work.

03 Return a verified result

The application receives the checked result for the next step.

04 Use the agreed fallback

Work outside the qualified scope follows the existing route.

20 OCTOBER 2026 / 400-CASE EVALUATION

Fresh results with an explicit test boundary.

One synthetic patch-statistics family: 300 supported requests and 100 boundary controls.

Route	Exact supported	Safe controls	Provider calls
TokenHoloFold	300 / 300	100 / 100	0
Conventional local	300 / 300	100 / 100	0
One-call API + local tool	237 / 300	92 / 100	355 total

10.61×

Ratio of median complete-process times on the 237 mutually exact tasks.

Native 130.686 ms vs API route 1,386.661 ms. The conventional local route was faster, with lighter release checks.

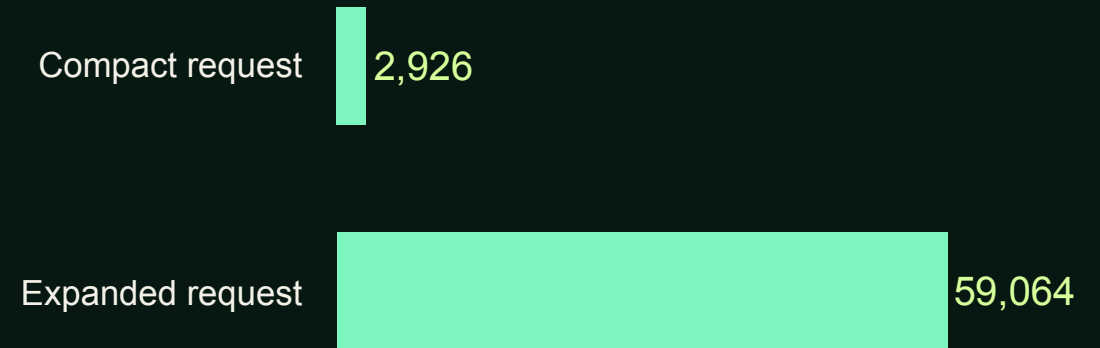
HISTORICAL REPRESENTATION STUDY

Much smaller request representations.

95.05%

fewer tokens in the compared
request representations

18 requests counted with o200k_base.
59,064 tokens became 2,926.



This measures task representation. Total bill savings and account allowance use different measurement boundaries.

20 OCTOBER 2026 / OWNER PACKAGE AUDIT

A real recurring Codex workflow.

9.72%

fewer returned input + output tokens

50,349 tokens became 45,454.

Small owner trial, separate from the current customer package. Conventional local verification can also avoid model calls. Account-wide allowance savings remain unmeasured.

Two real report jobs

Four Codex executions compared full observations with a compact verified report.

4 / 4 reports exact

Task-data bytes fell 88.82%.
The local audit used zero model calls.

HISTORICAL COMPARISON / RECENT OPTIMISATION

Performance gains from separate studies.

90.02×

Historical provider comparison

32 supported tasks, both routes exact.
Median paired latency ratio against a selected-context two-call API workflow.

24 September 2026

Different candidates, baselines and timers. These ratios are not multiplied together.

1.71×

Mac worker-batch improvement

Same native implementation and checks.
100-request batches: median 12.328 s cold versus 7.189 s persistent.

20 October 2026 / five paired batches

PROPOSED COMMERCIAL MODEL

A licence tied to verified savings.

50%

of verified net eligible savings
for BBF. The customer keeps half.

A scoped evaluation has a separately
quoted, capped fee.

Illustration: the same completed work

Agreed original cost	£100
New running + agreed costs	£20
Net saving before licence fee	£80
BBF licence fee	£40
Customer retained saving	£40

Illustration only. Agree the eligible scope, baseline, costs, fee cap and reconciliation before production.

READINESS AND PROTECTION

Controlled delivery with a clear scope.

Recorded today

- Qualified local operation
- Versioned evaluation ZIPs and adapters
- Integrity checks and self-tests
- Traceable result records

UK patent application
GB2623308.0

Qualified for each customer

- Workload fit and fallback
- Target-host acceptance
- Deployment controls
- Complete operating economics

Mac and Linux tests have defined environments.
Real Windows acceptance remains outstanding.

EVALUATION TO PRODUCTION

One workload. A measurable commercial decision.

- | | | |
|----|-------------------------|---|
| 01 | Agree the workload | Define the operation, platform, result quality and strongest feasible baseline. |
| 02 | Freeze the comparison | Set metrics, cost boundaries and pass criteria before the evaluation. |
| 03 | Measure equivalent work | Report exactness, latency, usage, local costs and fallback. |
| 04 | Decide on the pilot | Proceed where the evidence supports fit and the economics justify integration. |

TARGET CUSTOMERS

The first buyers have recurring eligible work.

AI application teams

Repeated operations inside an existing product.

Agent and workflow platforms

Supported steps between broader model tasks.

Enterprise AI teams

A defined internal workflow and an evaluation owner.

Model-serving and integration teams

Cost, latency and data-control opportunities.

TokenHoloFold can be evaluated alone.

GLYX retention and Seed State reconstruction are complementary products with separate qualification and savings boundaries.

NEXT CONVERSATION

Your workload is the next test.

What could eligible local execution save in your application?

Bring one repeated operation, its current cost and the platform you need to support. We can scope a confidential evaluation around complete, verifiable work.

[Explore TokenHoloFold and prepare an evaluation brief](#)

Beyond Brute Force Systems Limited / England and Wales / Company no. 17495620