

TokenHoloFold

Less repeated work. Checked local results.

TokenHoloFold provides a verified local execution route for supported deterministic operations inside AI workflows. Instead of repeatedly requesting a model to perform an operation that the local route supports, the workflow can submit the supported operation locally and check its result. The customer's workload determines whether this reduces cost or latency.

300 / 300 exact results

The 20 October 2026 evaluation also safely refused 100 of 100 boundary controls, with zero provider calls on the native route.

Reduce unnecessary model calls

Use the qualified local operation for supported work, keeping a defined fallback for unsupported requests. This does not replace every task performed by an LLM.

Keep quality and cost visible

Measure completed equivalent work, exactness, refusal behaviour and total operational cost. Lower request size, avoided calls and latency are distinct observations.

Evaluate inside the customer workflow

Use a versioned controlled evaluation package and one agreed operation. Qualify the platform, integration, recovery and output requirements before production delivery.

UK patent application GB2623308.0 is pending. An application is not a granted patent.

Workflow and recorded evidence

- 1. Choose one repeated supported operation.
- 2. Agree the correct output and best practical baseline.
- 3. Measure the local route, checks, fallbacks and complete costs.
- 4. Agree deployment and savings accounting only after the evaluation.

20 October native supported cases	300 / 300 exact
20 October native boundary controls	100 / 100 safely refused
20 October native provider calls	0
API-supported cases completed exactly	237 / 300
Timing on 237 mutually correct tasks	10.61x ratio of complete-process medians
Separate representation-token study	59,064 to 2,926 tokens; 18 requests

How to read these results

The 10.61x timing comparison uses only the 237 supported tasks completed exactly by both the local and one-call API routes; the API arm's other 63 cases remain incomplete. The conventional local baseline also passed 300/300 supported cases and 100/100 controls with zero provider calls, and was faster in this test with lighter release checks. Customer evaluation must therefore include an optimised conventional local baseline where feasible.

Scope of the evidence

The separate 18-request representation study reduced counted request tokens by 95.05%. That is a representation measurement, not a demonstrated 95.05% reduction in a Codex account allowance, total invoice or data-centre electricity use. The older 32-task two-call provider comparison recorded a separate 90.02x paired-median latency ratio. These campaigns concern different candidates and timing boundaries.

From evidence to a deployment

A scoped customer evaluation

Select the customer's actual repeated operation and compare equivalent completed work against the best feasible alternative, including conventional local execution or caching. Include output quality, refusal correctness, setup, startup, integration, verification, fallback and provider/serving costs. Retail API prices are not automatically an infrastructure owner's avoidable cost.

Commercial route

The proposed production model assigns BBF 50% of jointly verified positive net eligible financial savings; the customer keeps the other half. Agreed incremental running, integration and fallback costs are deducted first. A scoped evaluation fee is quoted separately. No production saving or rate is guaranteed before evaluation.

Using the products together

TokenHoloFold addresses supported local work; GLYX retains exact records; Seed State qualifies and reconstructs supported returning state. Each can be considered separately. A combined evaluation uses one agreed cost boundary and counts each saving once. Separate benchmark ratios must not be multiplied into a projected total saving.

Confidentiality and delivery

This paper describes customer-visible behaviour, recorded outcomes and evaluation boundaries. It does not disclose implementation source, private reconstruction material, issuer keys or proprietary selection rules. Controlled packages and confidential technical discussions require an agreed disclosure and delivery scope. An NDA is not permission to redistribute source or customer data.

Product status

The material describes qualified studies and private evaluation candidates. General customer distribution, target-platform acceptance, production support, checkout and subscription activation require their own completion checks. No universal installation, lossless compression ratio, guaranteed customer saving or energy claim is made.

Evidence reference

Recorded 20 October 2026 local/API/conventional-local evaluation; separate 18-request request-representation study; historical 24 September 2026 provider comparison. Each keeps its own workload and denominator.

Find the corresponding evidence section and evaluation route on the Beyond Brute Force Systems company website when public access opens. Enquiries can be directed to Ben Vivian through his professional profile: [linkedin.com/in/ben-vivian-0782b0130](https://www.linkedin.com/in/ben-vivian-0782b0130).