

MAHA STRATEGIES LLC

Context-Control Evidence Assessment

Sample deliverable, produced from a frozen synthetic evaluation corpus

Field	Value
Prepared by	Maha Strategies LLC
Evaluation run	wso2-large-live-20260817-v1
Observed	2026-08-17
Corpus	20 synthetic workloads, 60 calls
Evidence artifact SHA-256	78a822dd3c85821a98ca01262b02c12b5c42bc3c1433515909d3a0c5554aa3f1

READ THIS FIRST

This document shows the shape and rigour of the evidence package a customer receives after a bounded Context-Control Evidence Assessment. Every figure in it comes from a frozen synthetic corpus. It is not a customer result, not a case study, and not a performance guarantee. Maha Strategies is not claiming WSO2 partnership, certification, approval, or customer validation.

1. Executive decision

What was evaluated

Three request paths through one WSO2 AI Gateway deployment, over 20 frozen synthetic workloads of 20K-100K estimated tokens, 60 calls in total:

1. Baseline (no compression)
2. WSO2 Prompt Compressor
3. WSO2 + Maha Context Compiler

Each path answered the same labelled questions against the same source documents, under a frozen configuration with no automatic retries.

Observed result

On this synthetic corpus, the Maha path forwarded 98.84% fewer provider input tokens than the baseline (1,602,704 tokens avoided) at 98.20% lower modeled cost (\$1.603584 avoided), while a path-blinded semantic adjudication scored 60 / 60 required facts answered - the same score the uncompressed baseline achieved, and against 0 / 60 for the gateway's native compressor. All 60 calls completed; none required a retry.

Recommended decision

PROCEED TO A BOUNDED EVALUATION ON A CUSTOMER-SHAPED WORKLOAD. DO NOT GENERALIZE THIS SYNTHETIC RESULT.

The observed result is a reason to run a real evaluation, not a substitute for one. Nothing here establishes behaviour on customer documents, at customer volume, or under a customer's own retention and citation requirements. The finding that would change a deployment decision - whether the reduction and the retention both survive real, messy source material - has not been measured and cannot be inferred from this corpus.

The result is stated above. The recommendation is stated here. They are deliberately separate: the first is a measurement, the second is a judgement about what to do next, and a reader should be able to reject the second while accepting the first.

2. Scope and configuration

The configuration below was frozen before any model call and is recorded in the committed reproduction manifest. Changing any of it invalidates comparison with these figures.

Setting	Value
Gateway product	WSO2 AI Gateway
Gateway version	1.1.0
Prompt Compressor version	0.9.0
Prompt Compressor retained ratio	0.55
Maha interceptor version	1.0.0
Maha interceptor fail-closed	Yes, both request and response phases
Model	claude-haiku-4-5-20251001
Temperature	0
Maximum output tokens	220
Automatic retries	0 (zero-retry rule)
Modeled input price	\$1.000000 per million tokens
Modeled output price	\$5.000000 per million tokens

Corpus

Property	Value
Workloads	20
Calls	60 (20 workloads x 3 paths)
Difficulty mix	6 easy, 7 medium, 7 hard
Labelled required facts	60
Expected citations	60
Nature	Synthetic. No customer data, personal data, or credentials.
Label-freeze digest	a6b95c81e981fcd20576046354105994bc8846e2b6da8a4e77c754a7085a8c9

Every required fact, expected citation and prohibited assertion was labelled and digest-frozen before any path was run. Changing an input or a label after seeing model output fails validation rather than silently moving the target.

3. Aggregate findings

All figures below are read from the committed evidence artifact, which carries every one of the 60 calls as an individual row and re-derives these totals from those rows. A hand-edited total fails validation rather than printing.

Path	Provider input tokens	Modeled cost	Latency p50	Latency p95
Baseline (no compression)	1,621,553	\$1.632963	2,867 ms	4,138 ms
WSO2 Prompt Compressor	1,489,323	\$1.505248	6,958 ms	16,301 ms
WSO2 + Maha Context Compiler	18,849	\$0.029379	1,701 ms	3,415 ms

Required-fact retention, by scorer

Two scorers were applied to the same answers. They disagree, and any retention figure is meaningless without naming which one produced it.

Path	Path-blinded semantic adjudication	Deterministic exact-span containment	Expected citations resolved
Baseline (no compression)	60 / 60	15 / 60	60 / 60
WSO2 Prompt Compressor	0 / 60	0 / 60	60 / 60
WSO2 + Maha Context Compiler	60 / 60	24 / 60	60 / 60

- Path-blinded semantic adjudication applies a fixed rubric to each answer with the path hidden from the reviewer, so a correct paraphrase counts as answered. This is the figure a business reader usually means by "did it keep the facts".
- Deterministic exact-span containment requires the labelled evidence span to appear literally. It is reproducible by anyone holding the corpus, and it under-counts: a correct paraphrase scores as a miss. That is why the baseline scores 15 / 60 under it while scoring 60 / 60 under adjudication.
- Reporting only the first figure would overstate the result. Reporting only the second would understate it. Both are published per workload in the artifact.

Other measured properties

- Prohibited assertions across all 60 calls: 0.
- Calls completing without retry: 60 of 60.
- Maha non-expansion bypass engaged on 0 of 20 workloads (see section 5).

4. Representative three-path trace

ILLUSTRATIVE ONLY

This is one workload, one call per path. It is not evidence for the aggregate in section 3.

Property	Value
Trace	wso2-sanitized-release-evidence-rag
Workload	release-evidence-rag (easy)
Source documents	3
Total source bytes	40,989

Path	Input tokens	Output tokens	Latency	Modeled cost
Baseline (no compression)	28,192	118	2,839 ms	\$0.028782
WSO2 Prompt Compressor	26,305	153	3,420 ms	\$0.027070
WSO2 + Maha Context Compiler	967	92	1,242 ms	\$0.001427

The trace carries no source document text, no compiled context, no request headers, and no credential. On the Maha path it carries the pack identifier and the input and output hashes, so the selection that produced that answer can be identified without republishing the material it selected from.

- The workload is synthetic and does not establish performance on customer data.
- Request bodies are structural reconstructions with source text replaced by lengths and digests.
- Prompt Compressor internal context is not observable at this measurement boundary.
- Deterministic fact scoring can send semantically correct paraphrases to manual review.
- One trace is illustrative; aggregate conclusions must use the complete frozen corpus.

5. Failure and boundary evidence

Reduction is only useful if the component fails safely. The behaviours below were exercised separately from the measurement run, with no provider credential and 0 live provider calls (maha-wso2-failure-paths-v1, 2026-08-19). The deployable policy pins `passthroughOnError: false` on both the request and response phases, with a 20000 ms timeout. Fail-closed is the point: an invalid or absent evidence seal must not become a successful response.

Condition	Layer	Observed	Forwarded upstream?
missing-interceptor-configuration	maha-interceptor	HTTP 503 (<code>interceptor_not_configured</code>)	No
invalid-interceptor-credential	maha-interceptor	HTTP 401 (<code>invalid_interceptor_credential</code>)	No
oversized-decoded-input	maha-interceptor	HTTP 413 (<code>payload_too_large</code>)	No
interceptor-timeout	wso2-interceptor-service-v1	HTTP 500	No
interceptor-unavailable	wso2-interceptor-service-v1	HTTP 500	No

The gateway-side behaviours were verified against the WSO2 Interceptor Service v1 policy implementation itself, not a Maha reimplement of it. Tests passed: `TestOnRequestBody_ErrorPassthroughOrFail`, `TestOnRequestBody_TimeoutPassthrough`, `TestMahaUnavailableInterceptorFailsClosed`, `TestMahaRepeatedPolicyLatency`. Each scenario was measured 9 times.

Non-expansion and minimum-size bypass

The interceptor does not substitute a compiled pack when the rendered whole-document input is below the minimum-size threshold, and above it still compares the compiled and original contexts and forwards the original whenever compilation would be the same size or larger. The response identifies the decision in its headers.

On this corpus the bypass engaged on 0 of 20 workloads, because every workload was 20K-100K tokens and compilation reduced all of them. That means the bypass path is present and declared but not exercised here - its behaviour on small or non-reducing payloads is unmeasured by this run.

Explicitly unmeasured

- Behaviour on customer documents of any kind.
- Behaviour at production concurrency or sustained volume.
- Latency as a distribution: section 3 reports p50 and p95 across 20 single observations, not repeated runs of the same workload.
- The bypass path on small or non-reducing inputs.
- Any deployed-gateway network overhead beyond the measured call latency.
- Recovery behaviour after a partial or ambiguous settlement in production.

6. Limitations

These are carried from the evidence artifact and the run records. None is rhetorical; each one bounds a claim above.

- The corpus is synthetic. This run does not establish performance on a WSO2 customer workload.
- This is one execution on 2026-08-17. Latency is a single observation per call, not a percentile over repeated runs.
- Two fact scores are reported per row and they disagree. The deterministic score is exact-span containment and under-counts correct paraphrases; the adjudicated score applies a path-blinded semantic rubric to the same answers. Any published retention figure must say which one it is.
- The Prompt Compressor result is specific to WSO2 AI Gateway 1.1.0, Prompt Compressor 0.9.0 and a 0.55 retained ratio, and must not be generalized before WSO2 or a customer confirms that configuration is the intended production setup.
- Provider cost uses the declared pricing assumption applied to observed token counts; it is not a provider invoice.
- The primary checkpoint and adjudication files are not committed and are not currently available locally. The public artifact is checkable from its rows to its aggregates; full source-to-row regeneration requires a retained copy of the digest-identified primary files, or an explicitly authorized new run.
- These are local contract tests; they do not claim a deployed WSO2 environment was exercised.
- The timeout test uses 100 ms to keep verification bounded; Maha production evaluation configuration pins 20,000 ms.
- Repeated latency measures the local WSO2 policy boundary, not network, deployed gateway, model-provider, or end-to-end request latency.
- No Anthropic request was made, so upstream non-forwarding is established by immediate-response control flow rather than provider logs.
- The Prompt Compressor configuration used here (version 0.9.0, retained ratio 0.55) has not been confirmed by WSO2 or by a customer as the intended production setup. Its result must not be generalized until it has been.
- Costs are modeled: the declared price assumption applied to observed token counts. They are not provider invoices and they are not a savings guarantee.
- The answer-bearing primary evidence - the checkpoint and the path-blinded adjudication - is not committed to the repository and is not distributed with this package, because both retain the model's answer text for every call. The public artifact is independently checkable from its rows to its aggregates. The primary files are not currently available locally; full source-to-row regeneration requires a retained copy matching the listed digests, or an explicitly authorized new run.
- This document describes a compatibility evaluation. It is not a regulatory certification, an accreditation, an endorsement by WSO2, or a statement about production reliability.

7. Recommended customer evaluation

A fixed-scope engagement designed to answer one question: does the result above survive your material?

1. You supply one sanitized, representative workflow.

A single document set or RAG export that resembles what your system actually reads, with the facts and citations that must survive named in advance. No production credentials and no personal data.

2. Configuration and spend are frozen before anything runs.

Gateway and policy versions, compressor ratio, model, temperature, output ceiling, an exact provider-spend ceiling, and the zero-retry rule are agreed and digest-recorded. Your labels are frozen at the same time, so the scoring target cannot move after results are seen.

3. The same three paths run against it.

Baseline, your gateway's native compressor, and Maha - identical inputs, identical labels, no automatic retries, a durable checkpoint after every call.

4. You receive a private evidence package.

The structure of this document: per-workload rows, both retention scorers, latency, modeled cost, failure-path behaviour, and every limitation that applies to your run.

5. Maha states a recommendation: proceed, revise, or stop.

Including stop. An evaluation that concludes the component does not help your workload is a successful evaluation, and it is delivered as plainly as the alternative.

8. Technical appendix

Verify the figures in section 3

```
shasum -a 256 content/integrations/wso2-live-evaluation-evidence.json
```

```
npm run validate:wso2-live-evidence
```

```
npm run reproduce:wso2-evaluation
```

The first prints the artifact digest below. The second re-derives every aggregate from the per-workload rows and fails on any inconsistency. The third is the frozen-corpus dry run: it contacts no gateway and makes no provider call.

Verify this document

```
npm run validate:context-control-sample-assessment
```

Regenerates the document from the committed evidence and fails if any figure in it differs from the artifact.

Digests

Artifact	SHA-256
Evidence artifact (published)	78a822dd3c85821a98ca01262b02c12b5c42bc3c1433515909d3a0c5554aa3f1
Frozen corpus label freeze	a6b95c81e981fcd20576046354105994bc8846e2b6da8a4e77c754a7085a8c9
Reproduction manifest	dd2e40deedfe80d746322ac7d1686223b91aa7cf2804795b6c0d47285dd53aff
Failure-path evidence	deb70a6fd203c5895269bec23e1e3fe3dfe8afb72c0d892f7f17ab35fab8ee2d
Source checkpoint (not published)	sha256:40cb6956b0b732918c1bf6d63852d1076505c0348dd195a86384a5a935f4d0f3
Source adjudication (not published)	sha256:a77a0e01f86fa1130159e29deb5a6a66eff8e2ea65c41f96404bb02c477dc68d

Referenced materials

- Evaluation policy bundle: content/integrations/wso2-policy-bundle/ - secret-free proxy template, compatibility manifest with artifact digests, create-only installer, confirmation-gated uninstaller. Validate with `npm run validate:wso2-policy-bundle`.
- Frozen reproduction manifest: content/integrations/wso2-reproduction.json.
- Sanitized representative trace: content/integrations/wso2-sanitized-three-path-trace.json.
- Failure-path evidence: content/integrations/wso2-failure-path-result.json.
- Technical integration notes: docs/integrations/wso2-context-interceptor.md.
- Evidence recovery record: docs/integrations/wso2-live-evaluation-evidence-recovery.md.

Data handling in this document

No model answer text, source document, prompt, credential, request body, response body, private file path, or customer-identifying content appears anywhere in this package.