

WHITE PAPER

Token Economics of Governed AI Work

How ISS Reduces Token Waste in Long Enterprise AI Work

Cognous · September 2026 edition, with benchmark results

CORE THESIS

In long, high-context AI workflows, significant token spend is wasted on re-explaining context, correcting drift, regenerating failed outputs, and reconciling inconsistent intermediate work. ISS shifts the operating model from repeated context reconstruction to governed state transition. A published benchmark now validates the core mechanism: bounded governed context reduced estimated input tokens by 53.9% on 100-turn conversations, 75.5% at 200 turns and 89.9% at 500 turns. We expect total-token reductions, including output and rework, of 30% to 50% in lengthy enterprise exchanges, and are now measuring those on live providers.

MEASURED INPUT-TOKEN REDUCTION AT A GLANCE

53.9%

at 100 turns

75.5%

at 200 turns

89.9%

at 500 turns

86.6%

across the full suite

Measured on synthetic conversations with a published, repeatable method (July 2026). Figures cover input tokens only; see section 02 for scope.

CONTENTS

What this paper covers

1 Executive Summary	3
Results at a Glance	3
2 Benchmark Results: Input-Token Reduction	4
How the Benchmark Works	5
Savings Grow as the Conversation Continues	5
Scope of the Benchmark	5
3 The Business Problem: Tokens Are Being Spent on Waste	6
4 Strategic Context: Token-Based Competition	7
5 ISS in Business Terms	7
The Four Business Functions of ISS	7
6 How ISS Reduces Token Waste	8
7 Estimated Reduction Range for Lengthy Exchanges	8
8 Illustrative Token Economics	9
9 Why This Matters to Executives	10
10 Validation Status and Next Steps	10
How the Next Benchmark Works	11
11 Conclusion	11
12 Sources	12

01 Executive Summary

Enterprise AI is entering a phase in which model access is no longer the scarce asset. The scarce asset is the ability to convert model output into reliable, repeatable, auditable business work. As AI usage expands, companies are discovering that the true cost of AI is not only the price of tokens. It is the organizational waste surrounding token use: repeated context loading, drift correction, regeneration, manual review, and fragmented reasoning history.

ISS addresses this problem by giving AI work a persistent operating structure. Instead of treating every prompt or agent step as a largely independent event, ISS maintains a stable frame around the work, anchors each step to prior state, checks whether the work remains within acceptable bounds, and preserves lineage across the full exchange. In business terms, ISS turns a long AI conversation from an accumulation of messages into a governed workflow.

MEASURED RESULT

A published benchmark confirms that ISS keeps the input sent to the model bounded. Against shipping the full conversation history, Cognous reduced estimated input tokens by 53.9% at 100 turns, 75.5% at 200 turns and 89.9% at 500 turns, and by 86.6% across the full suite. The savings grow with conversation length, because without Cognous every new turn re-sends everything that came before it.

EXPECTED NEXT

Across the whole workflow, including output, corrections and regeneration, we expect total token reductions of 20% to 40% in general enterprise workflows, 30% to 50% in long, high-context exchanges, and up to 60% where work is repetitive, correction-heavy or prone to drift. These figures are being measured on live providers now.

Results at a Glance

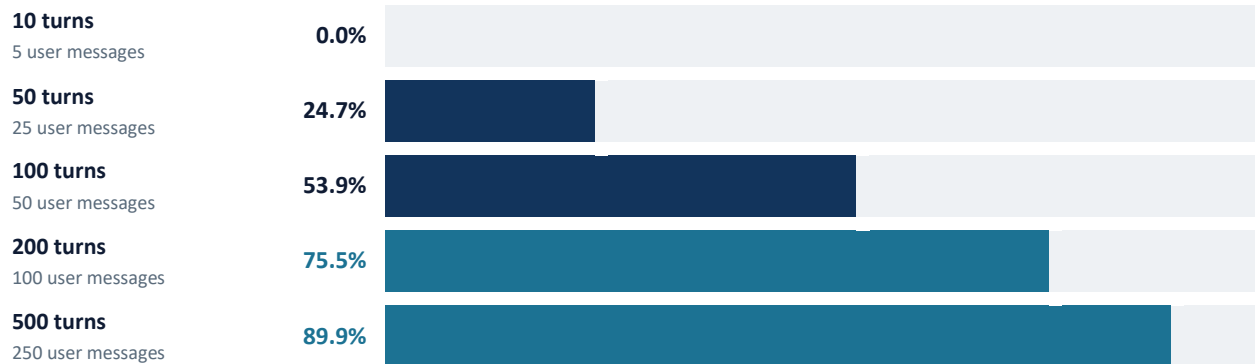
Result	Reduction	Status
Input tokens, 100-turn conversations	53.9%	✓ Measured
Input tokens, 200-turn conversations	75.5%	✓ Measured
Input tokens, 500-turn conversations	89.9%	✓ Measured
Input tokens, short conversations (20 turns or fewer)	None	✓ Measured
Total tokens, general enterprise AI workflows	20% to 40%	🕒 Expected
Total tokens, long, high-context exchanges	30% to 50%	🕒 Expected
Total tokens, correction-heavy workflows	Up to 60%	🕒 Expected

02 Benchmark Results: Input-Token Reduction

On a 100-turn conversation, Cognous reduces estimated input tokens by 53.9% versus shipping the full conversation history. At 200 turns the reduction is 75.5%, and at 500 turns it is 89.9%. Across all test conversations combined, 86.6% of input tokens are eliminated.

The savings accelerate as a conversation gets longer. Without Cognous, every turn pays again for every earlier turn, so costs climb steeply. Cognous keeps the last 20 messages word for word and a short summary of everything older, so cost per turn stays roughly flat. In short conversations of 20 turns or fewer, there is nothing to compress yet, so there are no savings.

INPUT TOKENS SAVED, BY CONVERSATION LENGTH



Conversation length	User messages	Without Cognous	With Cognous	Tokens saved	Reduction
10 turns	5	3,234	3,234	0	0.0%
50 turns	25	47,742	35,954	11,788	24.7%
100 turns	50	196,506	90,687	105,819	53.9%
200 turns	100	763,621	187,438	576,183	75.5%
500 turns	250	5,540,264	561,922	4,978,342	89.9%
All conversations	430	6,551,367	879,235	5,672,132	86.6%

Input tokens sent to the model over the whole conversation.

INPUT SENT TO THE MODEL ON THE LAST TURN OF A 500-TURN CONVERSATION



WHAT THIS MEANS FOR COST

Per-turn input cost stays roughly flat under Cognous — about 2,400 tokens whether a conversation is 50 turns or 500 turns long. On a 200-turn conversation, the same workload needs about a quarter of the input tokens. Because the full results are published, you can apply your own provider's price per token to the curve.

How the Benchmark Works

For every user message in each test conversation, we calculated the input sent to the model in two ways: with the full history, and with Cognous. Both use the same token estimate the product uses in production (about four characters per token), with the same settings Cognous runs in production: a 20-message window and a summary of about 150 tokens.

Approach	What is sent to the model
Without Cognous	Instructions + every earlier message + the new message
With Cognous	Instructions + a short summary of older history + the last 20 messages + the new message

Savings Grow as the Conversation Continues

Share of input saved at a given point in each conversation.

At turn	10-turn	50-turn	100-turn	200-turn	500-turn
10	0.0%	0.0%	0.0%	0.0%	0.0%
50	—	48.2%	46.8%	46.8%	52.5%
100	—	—	72.3%	72.2%	77.3%
200	—	—	—	86.7%	87.3%
500	—	—	—	—	94.5%

Scope of the Benchmark**WHAT IT SHOWS**

Cognous sends far less input to the model in long conversations, and the results are repeatable: the test conversations are fixed, so anyone can re-run the benchmark and get the same figures. The percentage savings hold even if the token estimate is slightly off, because any error applies equally to both approaches.

WHAT IT DOES NOT YET COVER

Output. The benchmark measures input only; the length of the model's answers is unchanged.

Live provider counts. Exact token counts on a live provider may differ by 10–20%, so dollar savings should be confirmed against actual provider usage.

Answer quality. The benchmark shows less is sent; a separate evaluation confirming answers are equally good is under way.

Other providers. The approach works with any chat-based model, but the cost translation depends on each provider's pricing.

Customer conversations. The test conversations are synthetic. Savings depend on the number of messages rather than their content, so the pattern should carry over to real use.

Benchmark methodology v1.0, July 2026. Full method and raw data available on request.

03 The Business Problem: Tokens Are Being Spent on Waste

Most enterprises initially think about token cost as a usage bill. That view is too narrow. In long AI workflows, token usage is not merely a measure of how much intelligence was consumed. It is also a measure of how much confusion, repetition, and correction the organization allowed into the process.

A long AI exchange often consumes tokens for productive reasons: drafting, analyzing, comparing, generating options, and producing final work. But it also consumes tokens for unproductive reasons: reminding the system what the project is, restating definitions, correcting terminology drift, repairing broken assumptions, regenerating outputs, and reconstructing why a prior step happened.

WHERE THE TOKENS GO IN A LONG EXCHANGE

PRODUCTIVE TOKENS

- Drafting
- Analyzing
- Comparing
- Generating options
- Producing final work

WASTE TOKENS

- Reminding the system what the project is
- Restating definitions
- Correcting terminology drift
- Repairing broken assumptions
- Regenerating outputs
- Reconstructing why a prior step happened

This waste grows nonlinearly. A short prompt may not need much structure. A long exchange does. Once the conversation spans multiple documents, revisions, constraints, and decisions, the absence of persistent structure causes the system to spend increasing amounts of token budget just trying to remain oriented.

04 Strategic Context: Token-Based Competition

The BCG article on token-based competition frames the macro shift clearly: advantage no longer comes simply from possessing intelligence, because AI makes intelligence more abundant and scalable. Advantage comes from applying AI to business problems more productively than competitors. BCG also argues that token consumption should be treated less like incidental IT cost and more like a strategic input into business output.

THE GAP

That framing is useful, but incomplete. It explains why tokens matter. It does not specify the operating infrastructure required to make token use efficient, consistent, and accountable. That is the gap ISS is designed to fill.

05 ISS in Business Terms

ISS should be understood as a governance layer for extended AI work. It does not merely ask the model to produce another answer. It keeps the work attached to a persistent structure so that each step inherits the right context, remains inside the intended bounds, and leaves a traceable record of how the work evolved.

The Four Business Functions of ISS

01 Persistent frame

Keeps the project identity, purpose, constraints, terminology, and operating assumptions stable across the exchange.

02 Continuity anchor

Ensures each new step inherits from the prior step instead of drifting into a related but wrong task.

03 Constraint surface

Rejects or redirects work that falls outside the allowed scope, meaning, quality threshold, or business rule.

04 Lineage-preserving update

Maintains a traceable history of how the work changed, so the organization can audit, reuse, and correct the workflow without restarting from scratch.

IN NON-TECHNICAL TERMS

ISS makes the AI remember what kind of work it is doing, why it is doing it, what it is not allowed to change, and how each intermediate result relates to the prior one.

06 How ISS Reduces Token Waste

ISS does not eliminate the tokens needed to perform real work. A long report, patent draft, due diligence review, or market analysis still requires output tokens. The savings come from reducing nonproductive tokens: repeated context, avoidable corrections, failed branches, and unnecessary regeneration.

Waste source	Typical baseline behavior	ISS effect
Repeated context loading	Long instructions, definitions, and constraints are repeatedly reintroduced.	Persistent frame carries the stable operating context.
Drift correction	The model moves away from canonical terminology or project scope.	Continuity anchor and constraint surface keep the work aligned.
Regeneration loops	Outputs must be recreated because earlier steps were unstable.	Invalid transitions are caught earlier and corrected locally.
Manual reconciliation	Users spend turns reconciling inconsistent versions.	Lineage records preserve how the work evolved.
Audit overhead	Reviewers cannot tell why the model made a change.	Provenance gives the workflow a traceable history.

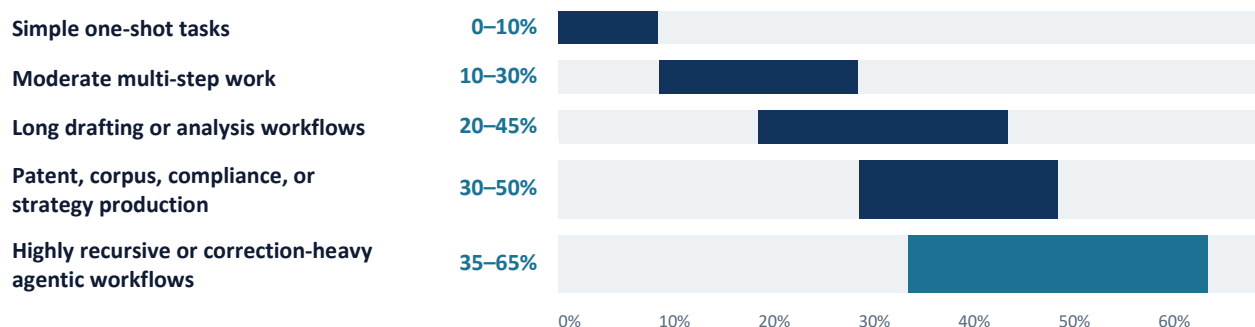
07 Estimated Reduction Range for Lengthy Exchanges

The practical estimate depends on the baseline. If the baseline workflow is already compact, well-instrumented, and low-drift, ISS may save relatively little. If the baseline workflow relies on repeated prompt scaffolding, manual corrections, and full regenerations, ISS can produce large savings.

HOW THIS RELATES TO THE MEASURED RESULTS

The ranges below are our expectations for total token reduction by type of workflow, including output, corrections and regeneration. The input portion is already measured (section 02) and runs well above these ranges in long conversations; the total-token ranges will be confirmed as output and rework are measured.

EXPECTED REDUCTION BY WORKFLOW TYPE



Workflow type	Expected reduction	Interpretation
Simple one-shot tasks	0% to 10%	Little context is repeated; ISS may add modest overhead.
Moderate multi-step work	10% to 30%	Some reduction from fewer corrections and less restatement.
Long drafting or analysis workflows	20% to 45%	Material savings from persistent context and fewer revisions.
Patent, corpus, compliance, or strategy production	30% to 50%	Strong fit where terminology, lineage, and constraint stability matter.
Highly recursive or correction-heavy agentic workflows	35% to 65%	Upside depends on disciplined implementation and rejection logic.

08 Illustrative Token Economics

The following example is illustrative, not a measured benchmark. It shows how token reduction can emerge in a long business workflow where the final output requirement stays constant but the surrounding waste is reduced.

ILLUSTRATIVE EXAMPLE — TOTAL TOKENS TO REACH AN ACCEPTABLE RESULT

Baseline	310,000 tokens	
ISS-governed	205,000 tokens	≈ 34% fewer tokens

Token category	Baseline	ISS-governed	Savings driver
Repeated canonical context	80,000	30,000	Persistent substrate frame
Task framing	20,000	20,000	No major change
Productive reasoning and output	120,000	120,000	Irreducible work
Corrections due to drift	50,000	20,000	Continuity and constraint gating
Regeneration and rework	40,000	15,000	Earlier rejection of bad transitions
Total	310,000	205,000	Approx. 34% reduction

THE CENTRAL ECONOMIC POINT

This is the central economic point: ISS does not need to make every model call shorter. It needs to reduce the total number of tokens spent reaching a stable, acceptable, reviewable result.

09 Why This Matters to Executives

The business value of ISS is not just cost reduction. Token reduction is the visible economic symptom of a deeper operating improvement: governed AI work. When an AI workflow preserves context, remains within scope, and records lineage, the organization gains more than lower consumption. It gains repeatability, accountability, and confidence that extended AI work can be trusted in business-critical processes.

Executive concern	ISS response
AI bills are rising without clear value attribution.	ISS makes long workflows more measurable by preserving state and lineage.
AI outputs are inconsistent across long projects.	ISS stabilizes context and constraints across steps.
Teams waste time correcting or restarting work.	ISS reduces drift and catches invalid transitions earlier.
Governance happens after output is produced.	ISS moves control into the workflow itself.
Leaders need ROI, not experimentation theater.	ISS supports a shift from usage metrics to productive token economics.

10 Validation Status and Next Steps

Input-token savings are now measured through a published, repeatable benchmark. The next phase measures total savings on live providers, using long workflows with multiple steps, changing intermediate work, constraints, revisions and review.

Result	Status	Evidence or next step
Input-token savings in long conversations	✓ Measured	Published benchmark, July 2026; repeatable
No input savings in short conversations	✓ Measured	Same benchmark: 0% at 10 turns
Token counts on a live provider	🕒 In progress	Re-run against actual provider usage
Savings on output and rework	🕒 In progress	Track correction turns and regenerations
Answer quality with Cognous	🕒 In progress	Side-by-side quality evaluation
Cost per accepted deliverable	🕒 Next	Combine tokens, quality and acceptance in one measure

How the Next Benchmark Works

- 1 Select three to five representative long workflows, such as patent drafting, research synthesis, compliance review, enterprise strategy drafting, or agentic task execution.
- 2 Run each workflow using the current baseline process and record total input tokens, output tokens, correction turns, regeneration events, elapsed time, and final acceptance rate.
- 3 Run the same workflows using an ISS-governed process with persistent frame, continuity anchor, constraint surface, and lineage record.
- 4 Compare total tokens to accepted output, not merely tokens per call.
- 5 Report both token savings and quality outcomes, because lower token usage is only valuable if the final work remains acceptable or improves.

THE EXECUTIVE METRIC

Tokens consumed per accepted business deliverable

That metric is more meaningful than raw token consumption because it captures the cost of reaching usable work.

11 Conclusion

Token savings from ISS are now measured, not just projected. A published benchmark shows that ISS cuts the input sent to the model by 53.9% at 100 turns and 89.9% at 500 turns, and the savings grow the longer the work runs. That confirms the core mechanism: ISS stops paying again and again for the same context.

Across the whole workflow, we expect total token reductions of 20% to 40% in general enterprise use, 30% to 50% in long, high-context work, and up to 60% where work is correction-heavy. We are measuring these on live providers now, alongside answer quality, so the result can be stated as cost per accepted deliverable.

THE LARGER STRATEGIC POINT

The larger strategic point is that ISS is not merely a token-saving mechanism. It is a way to make AI work structurally governable. Token savings are one measurable outcome of that governance.

12 Sources

This paper draws on Cognous's ISS architecture documentation, which describes how ISS keeps a persistent frame, continuity, constraints and a traceable history across multi-step work. It also uses the business framing from BCG's article on token-based competition, which argues that advantage is shifting from access to AI toward using tokens productively in business workflows.

Measured results come from the Cognous token-savings benchmark (methodology v1.0, July 2026). The full method and raw data are available on request.

WHAT IS MEASURED AND WHAT IS EXPECTED

Measured: input-token savings in long conversations, using a repeatable benchmark with a published method.

Expected: total-token savings, including output and rework, and the resulting cost per accepted deliverable. These are being measured on live providers now.