We Killed AI's Biggest Cost.

Wednesday 3 June, 2026

A new AI memory engine restores what a model has already read on a different server, byte-for-byte identical and cryptographically proven. Long AI conversations run up to 21 times faster on consumer hardware.*

OOSTKAMP, Belgium, June 3, 2026.** Every time you ask an AI a question about a long document, the AI re-reads it from scratch. Every. Single. Time. Ask ten questions about a 100-page report and the AI has read a thousand pages. That hidden re-reading is the single biggest cost in today's AI bills, and the reason long conversations get slower and more expensive over time.

Corbenic AI today released the fix, along with the cryptographic receipts that prove it works. Its engine, called **Taliesin**, saves what the AI has already read and restores it on demand. Crucially, the restored memory is mathematically identical to a fresh re-read, down to the last bit. Not "almost." Not "close enough." Identical.

Two minutes becomes seven seconds.

On a $0.69-per-hour graphics card, the longest test contexts took an AI more than two minutes to read from scratch. Taliesin restored them in under seven seconds: a 21-times speedup, with no loss of accuracy.

Across GPU generations, token for token.

In a bidirectional hardware relay between an Ampere A6000 and an Ada Lovelace RTX 4090, Corbenic moved AI memory back and forth and generated 64 of 64 output tokens identical to what the originating card would have produced. Cross-architecture verification is seldom attempted in public: GPU generations round floating-point operations differently. Most disaggregated-serving systems (Mooncake, CacheBlend) accept that as approximation; Corbenic shipped the proof. The practical meaning: a context prefilled on a cheap card, served from a more expensive one, gives the user the same answer either way.

The proof anyone can check.

Corbenic published SHA-256 hashes, the same fingerprint type used to verify a software download. Anyone can run the test on three public AI models from Meta, Alibaba and Mistral and confirm the hashes match. 45 of 45 trials matched. A 60-process follow-up: 60 of 60. Cross-machine: two physically separate servers, matching in both directions.

The EUR 600 model, by design.

Corbenic also released **Galahad-0.5B**, an open-source AI model trained from scratch for EUR 600 ($677), the price of a mid-range laptop, while frontier AI models reportedly cost hundreds of millions to train. The same engine works on both. Galahad is not top-tier and loses to similar-sized open models on standard tests; its purpose is to make the full verification chain auditable, weight by weight.

The cost killer is a stack.

Taliesin pairs with **Merlin** (open-source, on arXiv), Corbenic's byte-exact deduplicator. Merlin cuts redundant tokens before compute, Taliesin eliminates recompute on every reuse. Together they attack recurring AI compute from both ends.

"You don't need a bigger brain. You need a better memory," said Sietse Schelpe, Founder & CEO of Corbenic AI. "I got feedback on our earlier deduplication work and wanted to prove it worked end to end. In my eagerness to prove it, I realised I had built Taliesin."

What this could mean for AI at scale.

Picture a library where every reader had to re-read every book before answering one question. That is roughly today's AI. With Merlin and Taliesin: 13.9 to 71 percent input reduction (Merlin) plus up to 21 times less compute per reused context (Taliesin). For reuse-heavy workloads, that compounds to well over 90 percent off the recurring bill.

Media:



Related Sectors:

Business & Finance :: Computing & Telecoms ::

Related Keywords:

AI :: Anthropic :: OpenAI :: Alphabet :: Corbenic AI :: NVIDIA :: Meta :: Anthropic ::

Scan Me:



About Corbenic AI

Corbenic AI is a Delaware-incorporated company headquartered in Oostkamp, Belgium. It builds infrastructure that makes AI cheaper to run without making it less accurate. The company has shipped Merlin (open-source data deduplication, on arXiv), Galahad (open-source 570-million-parameter model, trained for EUR 600), and Taliesin (proprietary memory engine, with public verification receipts). More at https://www.corbenic.ai.

Company Contact:

_

Corbenic AI

E. sietse@corbenic.ai
W. https://www.corbenic.ai

View Online

Additional Assets:
https://youtu.be/6zVSdBR7NLc

Newsroom: Visit our Newsroom for all the latest stories:
https://www.corbenic.pressat.co.uk