

ONTBO - The Cognitive Context Layer for AI Assistants

From fragmented memories to a structured cognitive layer

AI assistants today remember fragments : isolated facts dumped into vector databases, forcing the host system to “connect the dots.” The result is shallow context, wasted tokens, and brittle performance. Ontbo changes that. It’s more than a memory layer. It’s a cognitive context layer.

What is it?

An API that normalizes and structures user historic data into a dynamic, reasoning-ready context layer to improve AI agent responses.

Proof in numbers

METRICS	ONTBO Best Performance	ONTBO Low Latency	Best of SoTA
Recall	90%	70%	70%
Token cost	-95% vs full-data	-99%	-90%
Latency (P50)	400ms	40ms	400ms
Model creation speed	+200% vs SoTA		N/A

Why it wins ?

Context Layer Multi-agent, reasoning-based context retrieval model that autonomously exploits user data to craft highly relevant context for the host agent’s task at hand.	Facts Lifecycle Management The user model is dynamically updated with new data. Automated conflict detection and resolution ensure consistency and integrity at scale.
Autonomous CoT Orchestration Self-directed context navigation and retrieval → fewer tokens, faster responses, higher accuracy.	Transparent Data Governance White-box approach to data, with inference traceability for result explainability, fine-grained user data control, and policy-by-design.

Built for scale

- Deployment models: plug-and-play SaaS or on-premises. Air-gapped options for critical infrastructure.
- Double data encryption.
- Modular architecture designed for scalability.
- Four retrieval methods (“Best Performance”, “Chain-of-Thought”, “Balanced” and “Low Latency”), for optimal balancing between performance and resource consumption.

Strategic benefits

- **Product:** drops into any agent/LLM stack to unlock sticky, compounding personalization.
- **P&L:** cuts inference spendings via context slimming; boosts conversion/CSAT.
- **Moat:** cognition-based context retrieval for more relevant agent replies with white-box, finely grained data control.
- **Distribution:** easy integration to assistants, copilots, search and workflow apps.