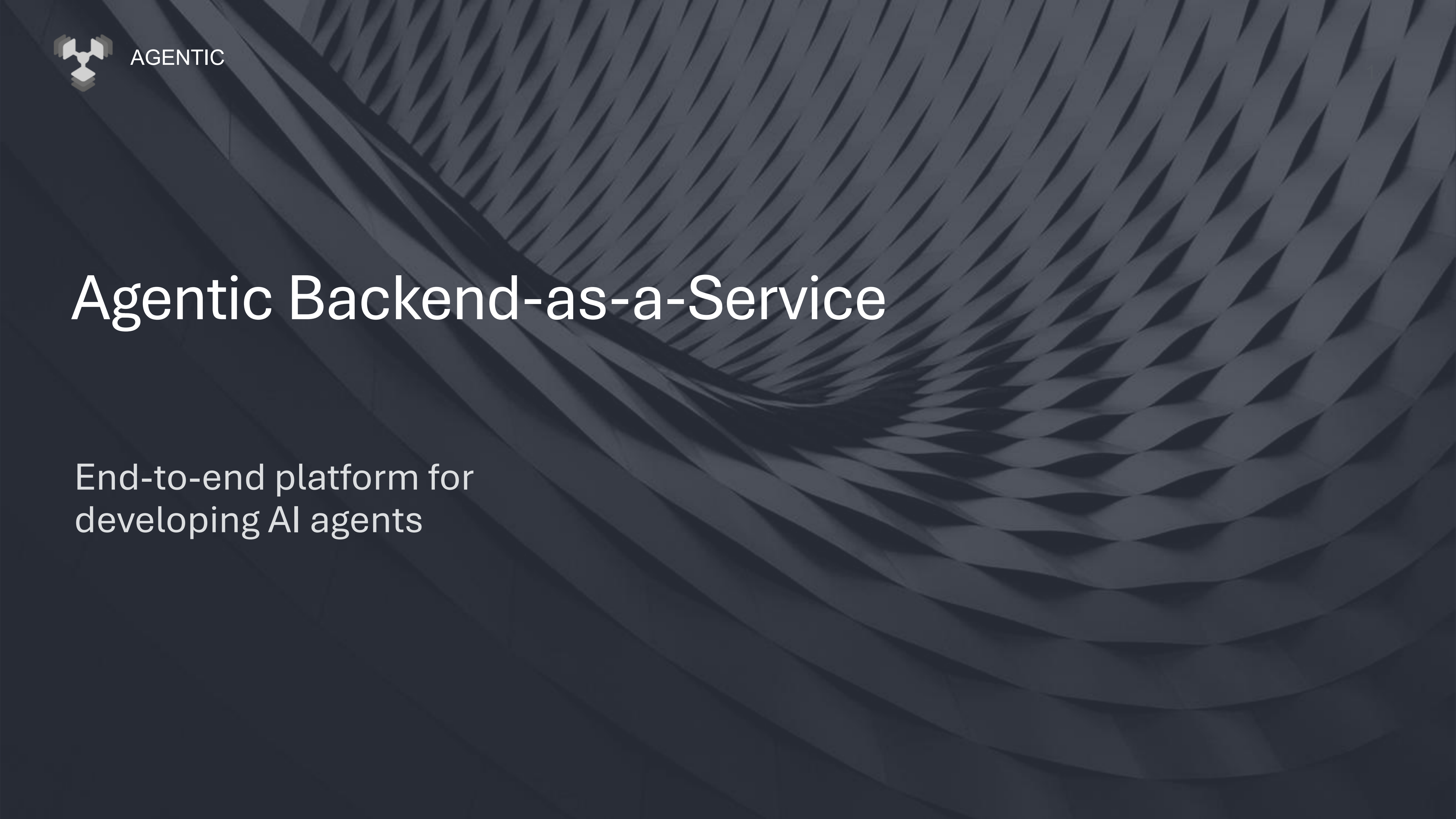
AGENTIC



Agentic Backend-as-a-Service

End-to-end platform for
developing AI agents

Key Features Listing

- Overview
- Baseline Infrastructure
- Model, Fine-Tuning, and Hosting
- Extract, Transform, and Load (ETL)
  - Baseline ETL
  - Advanced ETL
- Multi-Agent Orchestration
- Memory and Self-Learning
- Benchmarking, Observability, and Evaluations
- Tools

Agentic BaaS for AI-Native Applications

1

Platform

- Secure, scalable platform with **100+ pre-built vertical-specific AI agents**
- Deployable **on-premises or in private cloud** with built-in user-authentication and access controls
- Accompanied by pre-built **database tables** and **file storage**

Model

Supports any LLM (vendor agnostic)
Can self host open-source options
Custom fine-tuned (LoRA & QLoRA) LLMs available

Orchestration

Prebuilt, domain-specific agentic workflow templates
Combines deterministic chaining with dynamic AI supervisor routing
Gets better over time with **multi-layered memory** and **self-learning**

Data

Proprietary ETL pipeline
Custom fine-tuned VLM available for OCR
Knowledge graph construction support for complex data

Tools

Native MCP support
Assists with agentic quantitative reasoning
Structured data I/O

Baseline Infrastructure

- Premade Postgres database tables
- Built-in vector stores
- Static file storage
- User authentication management with Row Level Security
- Extensible Postgres modules (queues, Cron, triggers)
- Ready-to-use APIs

Model, Fine-Tuning, and Hosting

- Supports connections to all popular proprietary LLMs out of the box
  - API key(s) can be configured at the org level or agent level
  - Can pick different vendors (e.g., OpenAI vs. MS Azure) for the same LLM
- Offers forward-deployed engineering (FDE) support for fine-tuning open-source LLMs
  - We have unstructured data from 150,000+ users collected from previous SaaS products usable for domain-specific fine tuning
  - Additional data can be provided by the client or scraped from public sources
  - Fine-tuning approach uses LoRA or QLoRA
- Supports managed hosting for open-source or fine-tuned models
  - Hardware requirements are proportional to model size and expected usage volume
- Supports popular embedding models and managed hosting for open-source embedding models
- Maintains stateful backend and supports managed hosting of related backend operations
- Hosting and associated services are SOC II, ISO 27001, and GDPR compliant

Baseline ETL (Extract, Transform, and Load)

- Supports common enterprise data types
  - PDF, Word, PPT, Excel (table only, no formulas), Markdown, TXT, CSV, Images, Audio (require separate speech-to-text model)
  - URL (or sitemap) crawler / scraper
- Streamlined data extraction and preparation
  - Methods include standard OCR (computer vision-based) and VLM (supports both proprietary and open-source)
  - Output defaults to Markdown with accompanying JSON metadata
- Chunking and metadata enrichment
  - Fixed-size, recursive, semantic chunking algorithms
  - Custom chunk size and size bin management (optional for addressing uneven chunking bias)
  - Metadata enrichment for chunk location indexing (section, page, document associations)
- Multimodal indexing for lossless retrieval
  - Index extracted text for semantic search but retrieve original image for multimodal LLM input
- Managed vector store (default is Postgres; more scalable options available)
- Managed knowledge base (chunks, metadata, full text, Agent associations, etc.)

Advanced ETL Options

- Agentic document picker
  - Define which documents to index based on file metadata patterns (e.g., filename_vfinal, last_modified_date, etc.) to avoid duplication and enforce better version control
  - Can use AI agent to prepopulate recommendations based on user’s natural language criteria
- Relational graph-based chunking algorithm
  - Agentic metadata enrichment for establishing cross-chunk links (useful for documents with chunks that explicitly reference each other based on document / section names)
  - Retrieved chunks include top k chunks from hybrid search + their n-degree neighbors
- Retrieval optimization
  - User query enrichment (useful for improving user intent interpretation and subsequent search)
  - Hybrid search (semantic + keyword-based)
  - Post-retrieval filtering based on chunk metadata (useful if user only wants results from specific documents or sections)

Advanced ETL Options – Continued

- Lossless document scanning algorithm
  - AI agent scans across all chunks of a document sequentially regardless of document length and iteratively refines its response
  - Ensures exhaustive document review rather than lossy RAG
- Unstructured data > structured JSON
  - AI agent scans across raw document clusters sequentially then populates predefined JSON template with extracted values
  - Domain-specific templates available

Multi-Agent Orchestration and Workflow Management

- Agent system architecture: Orchestration > Agent > Session (can be user-specific or agnostic) > Query
- Single supervisor agent with specialist agents as tools
- Intent-based multi-agent routing
- Passively monitor session state and extract data to store in session variables in real time
- Token distribution management (session memory, long-term memory, RAG, function response, prompt, input, reasoning, output).
- Add / Update / Delete / Fetch, etc.
- Exponential fallback, error logging, and automatic error handling

- Workflow architecture: Workflow > Orchestration / Agent / Code / Control Flow Nodes > Runs
- APIs for accessing Workflows

Memory and Self-Learning

- Multi-fidelity session-specific memory
    - Most recent AI exchanges are stored verbatim, but older exchanges are summarized to save tokens when inputting to LLM
- Multiple memory layers to enable advanced personalization and continuous learning
- (Optional) User-specific personalization
    - Insights derived from recent user activity across multiple sessions
    - Store in user-specific long-term memory
- (Optional) Org-wide insights
    - Self-learning AI agent periodically reviews all tracked users' user-specific long-term memory and distill out the most insightful snippets
    - Store in org-wide long-term memory
- Admin can review and edit long-term memory

Benchmarking, Observability, and Evaluations

- Full visibility and automatic change tracking for
    - Input and output to/from every AI agent
    - Agent reasoning process
    - Every session state, token count, and session variable values
    - RAG context, short & long-term memory, tool calling history and response
    - Execution time breakdown
- Detailed logging and error tracking
- Human alerts

- Controlled evaluations on Agent output vs. human-supplied reference answers.
    - Statistical analysis on aggregate AI performance

Tools

- Custom functions can be added to the MCP server
- Common tools
  - Deep research (requires certain third-party APIs for internet search)
  - All AI agents are aware of current date / time
  - Get day-of-week
  - Hybrid search (full text retrieval)
- Database read / write wrapper functions (SQL)
- Various domain-specific tool templates