EDB POSTGRES AI

Summer Academy

Postgres as Your AI Data Platform: Architecture, Vectors, and What to Watch Out For

01 02 03 04

Lucie Zeng - Sales Engineer
Tim Waizenegger - Principal Engineer

AGENDA

TODAY'S SESSION

Why AI stalls, five topics, two hands-on demos, one platform.

01 **PG/AI (AI Sidecar) Overview**
Architecture, RBAC/RLS data governance, how to integrate

02 **Hybrid Manager Features**
GPU model serving, Langflow visual builder

03 **AI Pipelines for RAG / Knowledge Base**
Auto retrieve → chunk → embed → index + DEMO 1

04 **PG Agents**
Memory, governance, semantic layer, Text-2-SQL + DEMO 2

05 **External Agents**
MCP, Langflow components, governance for tool calls

EDB POSTGRES AI

POLL

HAVE YOU ALREADY STARTED AN AI PROJECT WITH POSTGRES?

- Not started yet and don't know much about the possibilities
- I know pgvector and want to get started
- Already started with pgvector or similar extensions
- Already started with another vector database

INTRODUCTION · THE PROBLEM

WHY ENTERPRISE AI STALLS

It's not model failure — it's architecture, cost, and control.

40%+ of agentic AI projects will be canceled by 2027

Not model failure — escalating costs, unclear business value, and inadequate risk controls. (Gartner, June 2025)

01

NO AI INSIDE THE DATABASE

Engineers and DBAs have no native way to call models or agents from Postgres. Stepping outside means extra tooling, integration, and risk.

02

EVERY INFERENCE CALL = DATA TRANSFER

Data leaving your environment creates egress that needs controls and audit — securing data in motion adds massive overhead.

03

CLOUD AI COST IS HARD TO PREDICT

73% of enterprises say AI costs exceeded projections (FinOps Foundation, 2026). Uber burned its 2026 AI budget in four months.

04

DIY MEANS 7–8 SEPARATE TOOLS

Vector stores, embedding pipelines, agent frameworks, observability — each tool adds an operational burden.

SO WHAT

AI doesn't have to live outside your database — moving AI to the data is simpler, faster, and cheaper than moving the data to the AI.

EDB POSTGRES AI

SECTION 01

PG/AI (AI SIDECAR)

Section 01 — Architecture, features, and how to integrate.

THE VISION

ONE PLATFORM: OLTP + ANALYTICS + AI

AI is a fourth workload on the Postgres you already run — not a new stack.

01

OLTP

System of record

Mission-critical Postgres for core operations — the trusted source of truth.

02

ANALYTICS

Insight at scale

Analyze enterprise data at scale on the same platform. No separate warehouse, no copies.

03

AI

Intelligence in place

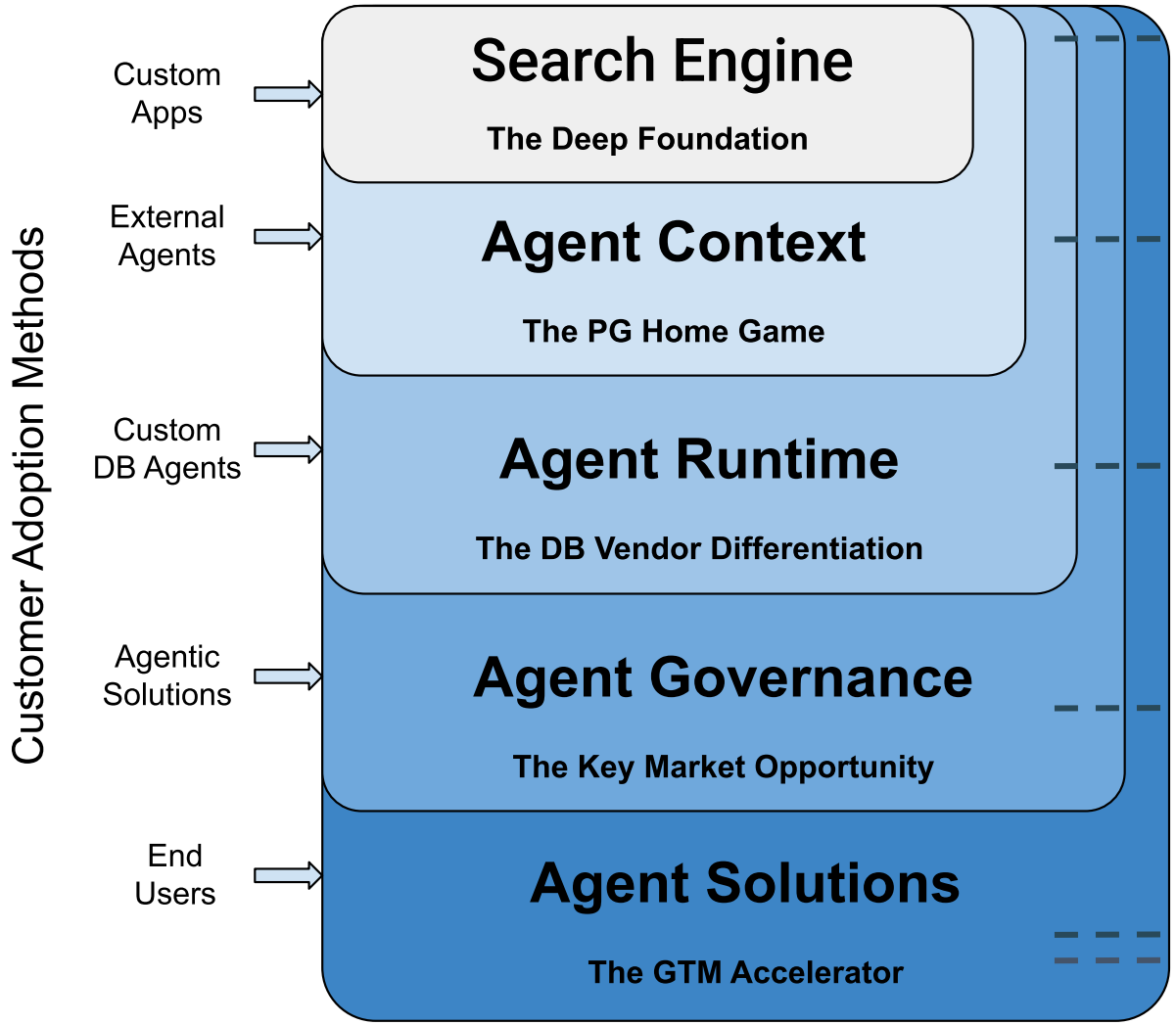
Models, search, and agents running inside the data through SQL, with governance built in.

ONE PLATFORM **No new vendor to integrate — data never moves between silos to create value.**

EDB POSTGRES AI

ROADMAP

THE AI SIDECAR VALUE STACK



Today: pgvector, VectorChord, VectorChord-BM25
September: edb_vectorplus, pg_textsearch (differentiated semantic + text search)

Today: - AI Pipelines with Knowledge Bases for RAG, Basic SemKB
September: agent memory + extended SemKB

Today: Langflow & models in HM, In-PG models, external MCP server
September: In-PG agents + tools + MCP

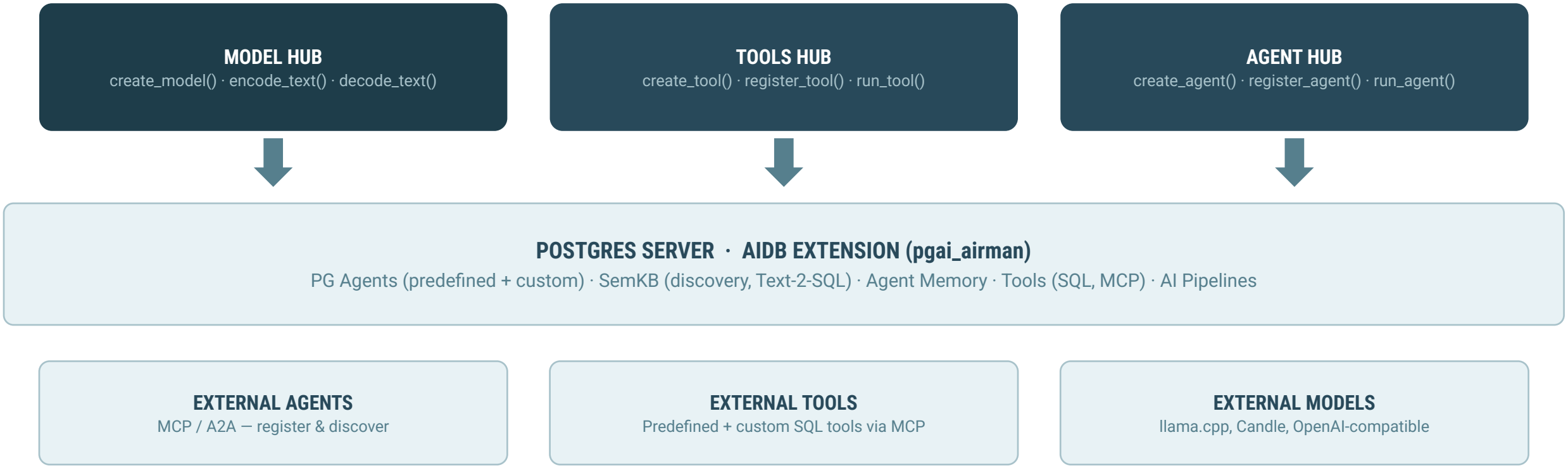
Today: Agent tracing & purpose enforcement in HM & Langflow -> preview only
September: In-PG agent tracing & purpose enforcement -> GA

Today: langflow patterns & demos for RAG & Text-2-SQL
September: agents, examples & blueprints for full agentic applications built on and in PG

ARCHITECTURE

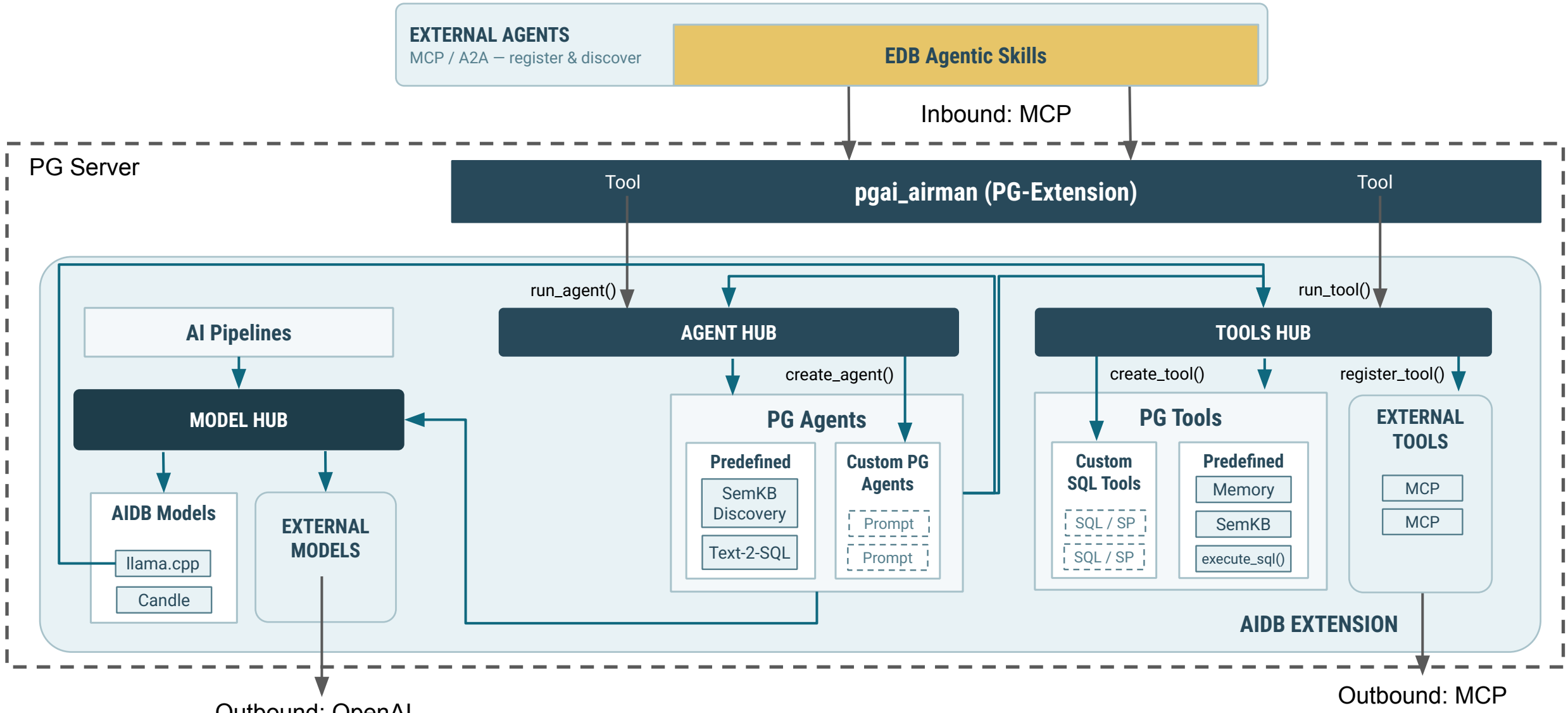
THE AI SIDECAR ARCHITECTURE

Turns Postgres into an agent runtime — models, agents, and tools, all governed in PG.



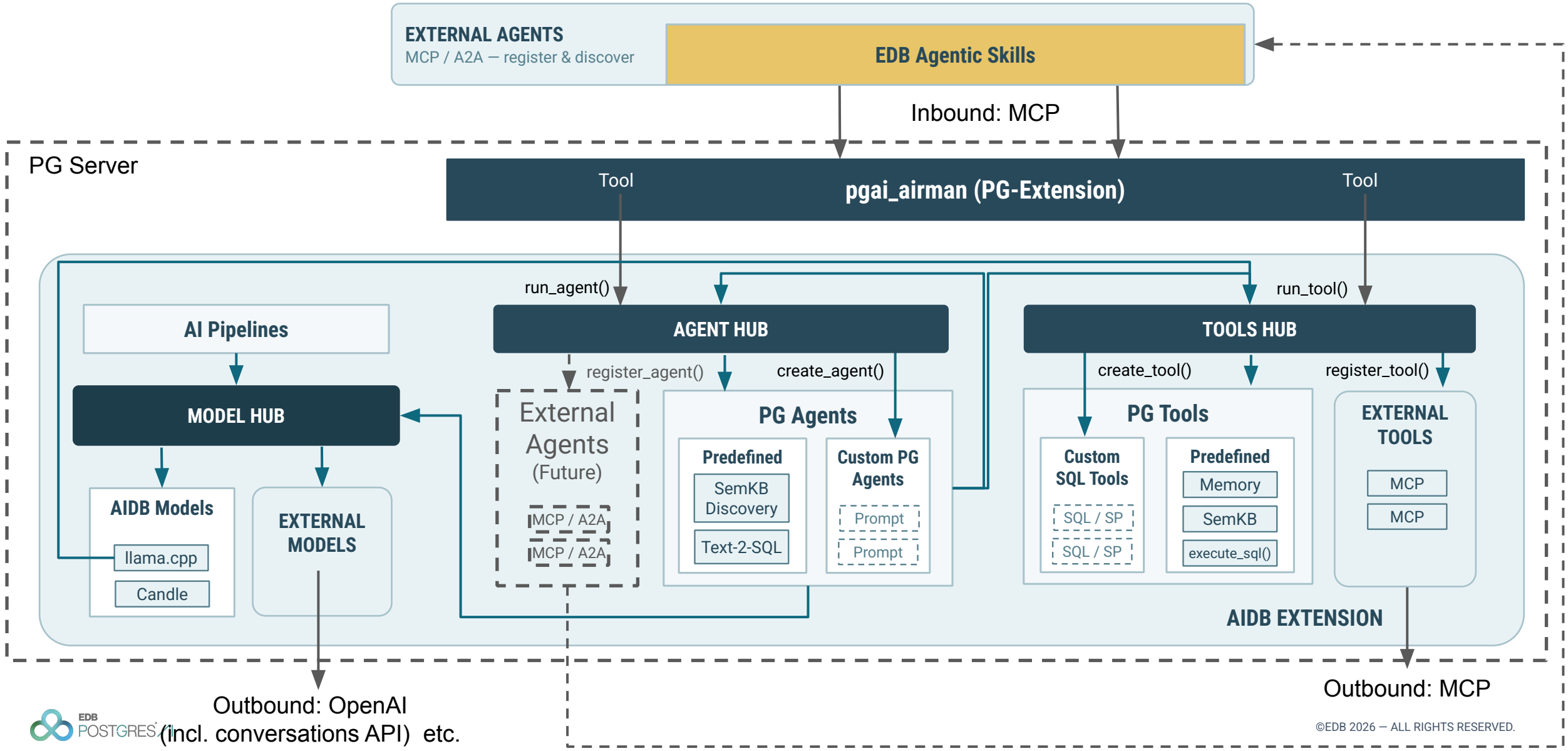
THE AI SIDECAR ARCHITECTURE

Turns Postgres into an agent runtime — models, agents, and tools, all governed in PG.



THE AI SIDECAR ARCHITECTURE

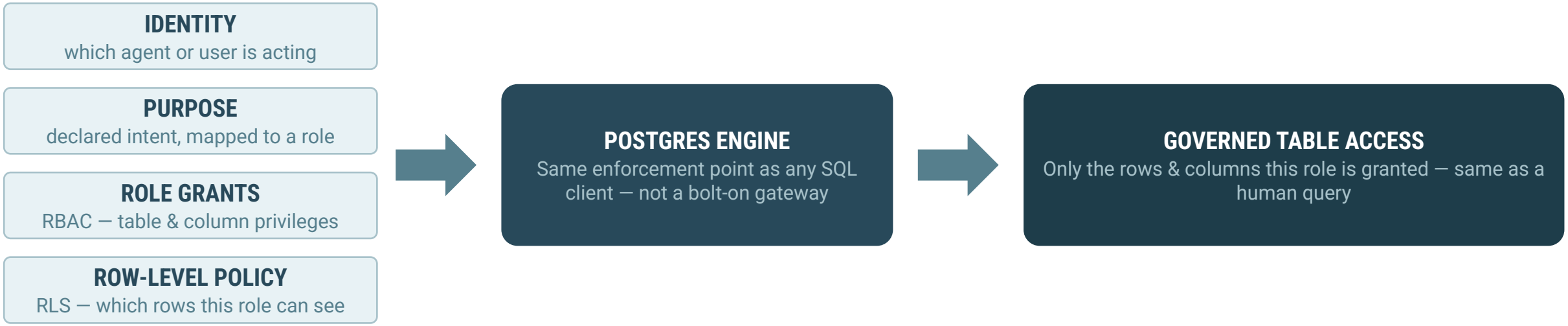
Turns Postgres into an agent runtime — models, agents, and tools, all governed in PG.



AI SIDECAR · GOVERNANCE

DATA GOVERNANCE: ROLE-BASED ACCESS TO YOUR TABLES

The AI Sidecar introduces no new permission model — every agent runs through Postgres's own RBAC and RLS.



NATIVE PRIMITIVES

PG roles + row-level security — no separate policy engine to keep in sync with the schema.

DYNAMIC ROLE VENDING

Short-lived, purpose-scoped roles are vended per agent session — no standing over-privileged credentials.

ONE AUDIT TRAIL

Every table read or write — human or agent — lands in the same RBAC-governed log.

SO WHAT

An agent querying customer_feedback sees exactly what an analyst with that role would see — nothing more, nothing less.

EDB POSTGRES AI

GETTING STARTED

TWO WAYS TO INTEGRATE

Start on plain Postgres. Add Hybrid Manager when you need GPU scale and visual tooling.

WITHOUT HYBRID MANAGER

Add AI to your existing Postgres — install extensions, no new infrastructure.

› aidb extension — run AI directly from SQL
› pgvector — built-in vector search
› Built-in CPU models — start today

Best for: teams not ready for HM yet, fastest path to a first AI POC.

WITH HYBRID MANAGER

Everything above, plus enterprise-scale serving and visual tooling.

› KServe — GPU model serving at scale
› Langflow — build AI workflows visually
› Unified management across all nodes

Best for: GPU workloads, multi-cluster ops, teams who want a visual builder.

NO REWRITE
Moving from plain Postgres to Hybrid Manager later requires no rearchitecture.

EDB POSTGRES AI

SECTION 02

HYBRID MANAGER FEATURES

Section 02 — GPU model serving and the Langflow visual builder.

MODEL SERVING ON GPUs

When CPU inference isn't enough, Hybrid Manager serves models at scale on Kubernetes.

KSERVE

Model serving on Kubernetes

- Standard model-serving abstraction
- Autoscaling, canary rollout
- Works with any framework

BEST FOR Teams standardizing on K8s-native serving

NVIDIA NIM

NVIDIA-optimized inference

- Prebuilt, optimized containers
- Tuned for NVIDIA GPU hardware
- Fast path to production throughput

BEST FOR Maximum throughput on NVIDIA GPUs

vLLM

HuggingFace & BYO models

- Bring your own fine-tuned model
- High-throughput LLM serving
- Broad HuggingFace compatibility

BEST FOR Custom or open-source LLMs

HM · AGENT STUDIO

LANGFLOW — VISUAL AGENT STUDIO

Build agents and pipelines by dragging components — every flow ships as a production REST API.

A visual IDE for agents

Langflow lets teams compose data connectors, models, and tools into a working agent or RAG pipeline — without writing orchestration code.

Also used for our AI Pipelines and RAG demos later in this session.

Drag & drop

Connect a chat input, a knowledge base, and a model in minutes.

Ships as an API

Every flow is instantly callable over REST — no separate deploy step.

PG-native components

Built-in nodes for AIDB knowledge bases and Hybrid Manager model servers.

MCP server

Any flow can also be exposed as MCP tools for other agents to call.

EDB POSTGRES AI

POLL

WHAT KINDS OF DATA SOURCES ARE YOU PLANNING TO USE FOR RAG OR AGENTIC RETRIEVAL?

- Text in PG tables
- Text documents in S3
- Structured documents (json/csv) in S3
- MS office formats
- Sharepoint
- Local files
- Remote DB connections e.g. MongoDB, Elastic Search

SECTION 03

AI PIPELINES FOR RAG

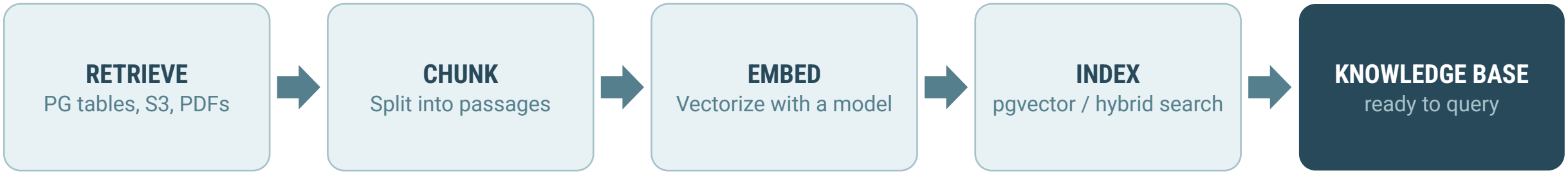
Section 03 — Automated retrieve, chunk, embed, and index — then Demo 1.

EDB POSTGRES AI

AI PIPELINES

NO MANUAL ETL

AIDB pipelines keep vector embeddings in sync with source data automatically.



Add or change a source file → the knowledge base updates automatically. No manual re-embedding, no stale search results.

SEMANTIC SEARCH

Search a Postgres knowledge base kept in sync with a source table via AIDB.

HYBRID RAG

Combine a Postgres table with unstructured PDFs from object storage in one pipeline.

AUTO-UPDATE

New documents trigger real-time knowledge-base updates on upload — no manual step.

EDB POSTGRES AI

HANDS ON

DEMO 1 — RAG PIPELINE IN AIDB + PYTHON FRONT-END

Live: build and query a knowledge base from Postgres data and PDFs.

What we'll show

1. Semantic search directly on a Postgres table (customer feedback), kept in sync by AIDB.
2. A RAG flow in PG/AIDB answering questions over that AIDB knowledge base.
3. A hybrid source flow — mixing the Postgres table with PDF product catalogs from object storage.

Stack

Postgres + AIDB extension · Python application: “ACME RAG”

Why it matters

The knowledge base updates itself when a document is uploaded — no manual re-embedding step, no stale answers.

— simple application logic in python: embeddings, retrieval, generation: all handled in the AIDB extension in PG

TAKEAWAY

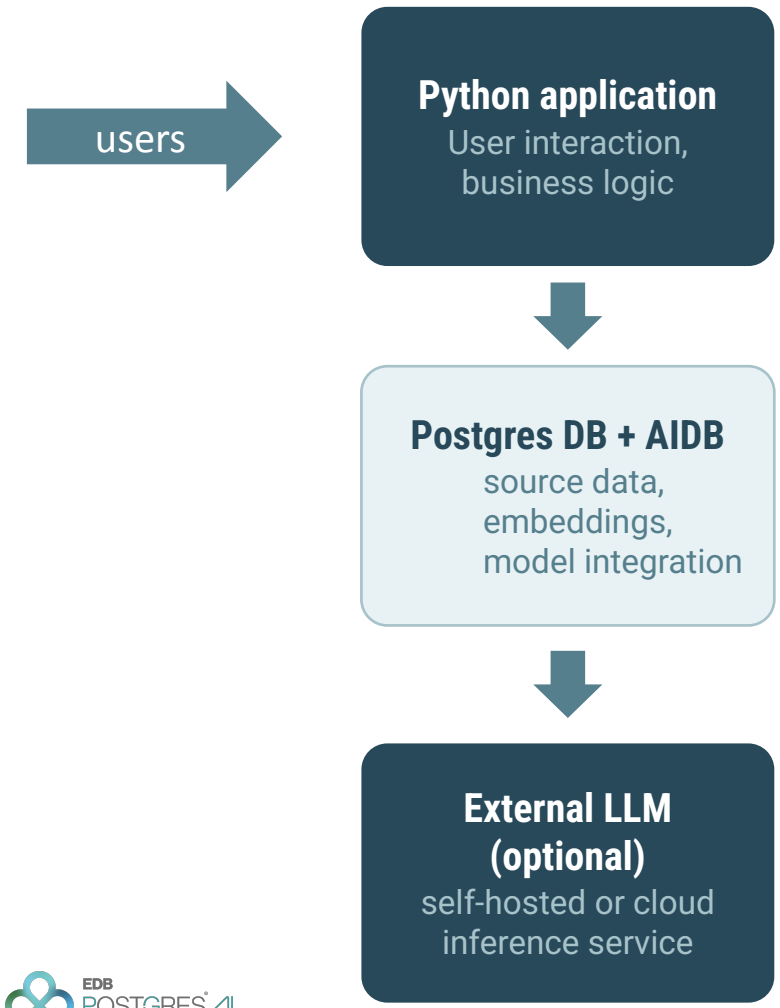
RAG in Postgres — one system for the data, the vectors, and the retrieval.

EDB POSTGRES AI

HANDS ON

DEMO 1 — RAG PIPELINE IN AIDB + PYTHON FRONT-END

architecture overview



Bring your own front-end / application

Choose any language/framework you want, it only needs a PG DB connector.

Focus on your user experience and business logic.

AIDB extension

Handles the AI logic: embeddings processing, retrieval/vector search, LLM API integration.

PG Database

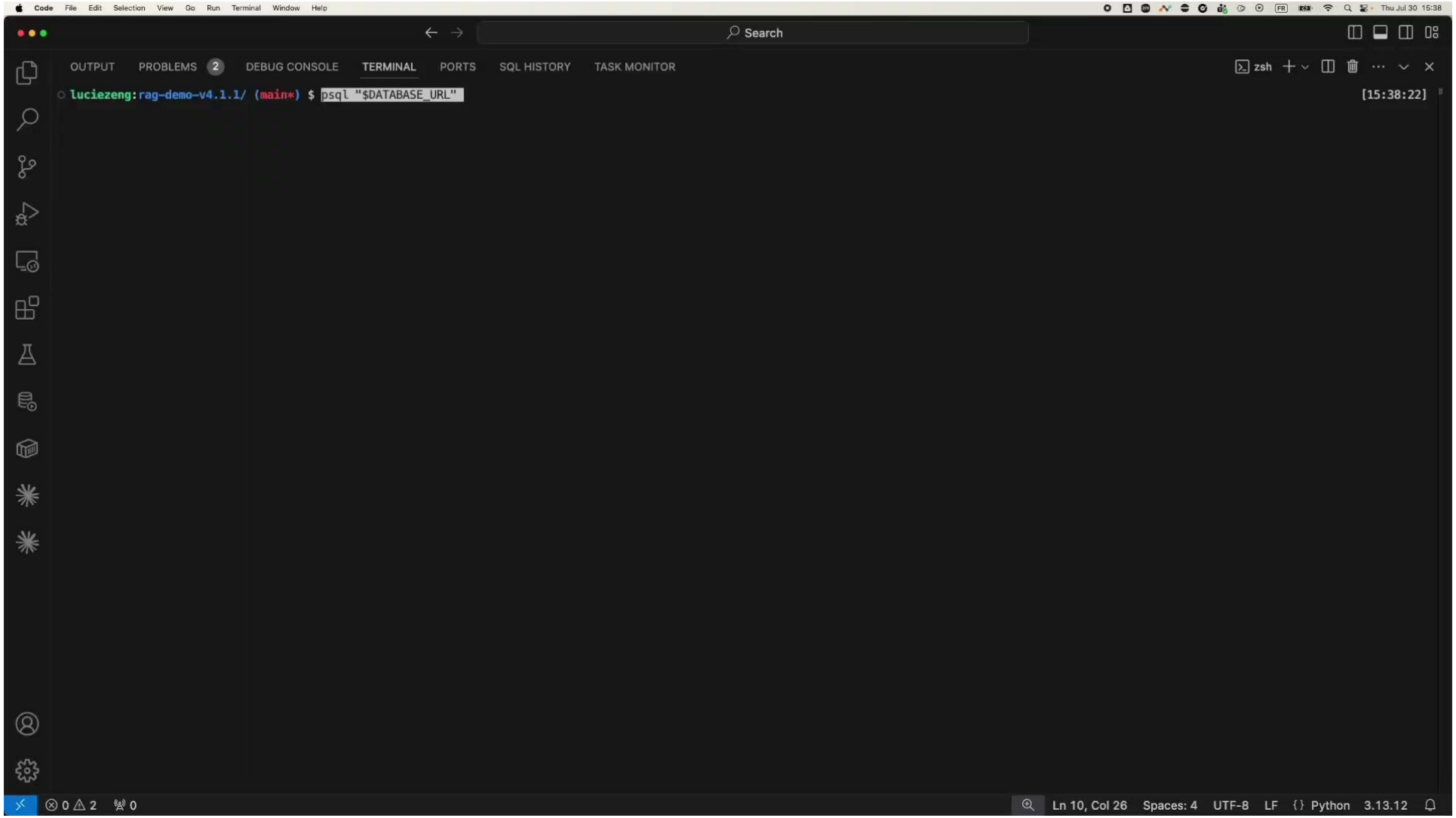
Holds your source data, embeddings and runs the AIDB extension.

optional external LLM

Connect AIDB to a wide range of external LLMs, or use one of the built-in models that run on CPU right on the database host.

HANDS ON

DEMO 1 — RAG PIPELINE IN AIDB + PYTHON FRONT-END



SECTION 04

PG AGENTS

Section 04 — Memory, governance, semantic layer, and Text-2-SQL — then Demo 2.

AGENT MEMORY — THREE TIERS, ONE DATABASE

Working, episodic, and semantic memory — joinable in a single query. No separate memory service.

WORKING MEMORY

Within a session

Previous prompts and agent output are carried-forward as context. Allowing multi-turn conversations with agents.

Compaction enables long-running interactions while saving on tokens.

EPISODIC MEMORY

Across sessions

Structured log of past runs — task, input, outcome, duration. SQL-queryable: “what did I do last time?”

SEMANTIC MEMORY

Long-term retrieval

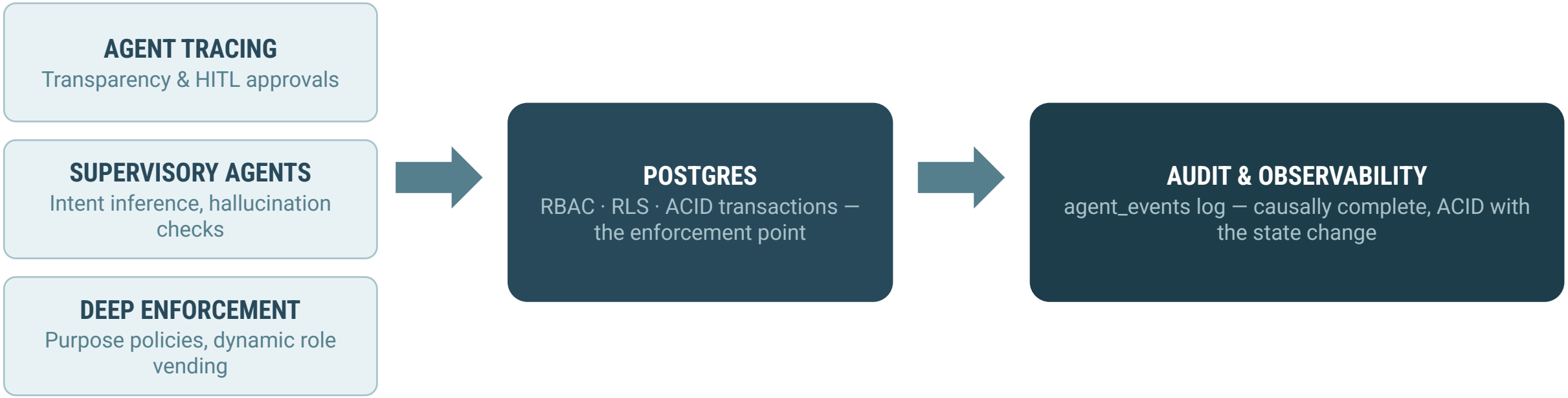
pgvector embeddings for similarity search, joinable with episodic memory and SQL filters in one query.

KEY INSIGHT

No ETL, no separate service — vector similarity + SQL filters + performance data in one JOIN.

GOVERNANCE — ENFORCEMENT, AUDIT, SUPERVISION

Every model, agent, and tool call is traced and policy-checked, inside the same transaction.



NO NEW TRUST BOUNDARY
Agents governed by the same roles and RLS as every human query.

SAME TRANSACTION
Policy violation → cancel task, revoke role, log alert — atomically.

ROADMAP: 1Q–2Q 2027
Transaction provenance and fine-grained rollback of agent actions.

SEMANTIC LAYER (SemKB) — WHY IT MATTERS

One indexed layer of business meaning saves tokens and raises accuracy for every agent.

WITHOUT SemKB

Every agent call re-sends full schema, column names, and business context in the prompt.

- Larger prompts, higher token cost
- Agent guesses at ambiguous table/column names
- Lower accuracy on domain-specific queries

WITH SemKB

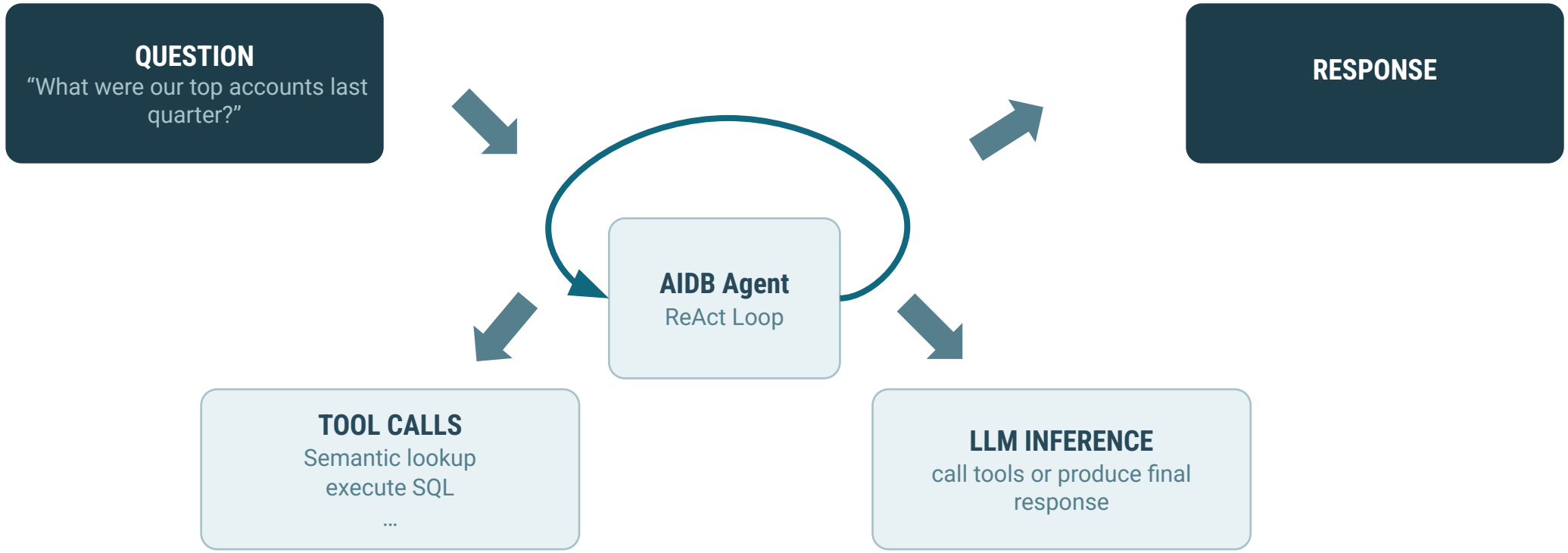
SemKB indexes table, view, and column semantics — plus curated “golden query” business SQLs.

- Agent retrieves only the relevant slice of schema
- Fewer tokens per call, lower latency
- Higher accuracy — grounded in real business meaning

TODAY **Table/column semantics and Golden Queries are available now; relationship semantics and compound search are on the roadmap.**

AI SOLUTIONS: CONVERSATIONAL ANALYTICS (TEXT-2-SQL)

A business question in, a governed SQL query out — grounded by the semantic layer.



Every generated query runs under the caller's own PG role — Text-2-SQL inherits RBAC and RLS automatically, no separate permission model.

SO WHAT

Business analysts query Postgres in plain English, French, Spanish, … — no SQL, no separate BI tool required.

HANDS ON

DEMO 2 — CONVERSATIONAL ANALYTICS (TEXT-2-SQL AGENT)

Live: ask a business question in plain English, watch the agent ground it in SemKB and run SQL.

What we'll show

1. A natural-language business question against a real Postgres schema.
2. SemKB retrieval of the relevant tables, columns, and a matching golden query.
3. Generated SQL executing under the caller's own role — governed by RLS like any query.

Stack

AIDB SemKB · Text-2-SQL PG Agent · live Postgres database

Why it matters

Same accuracy work as a SQL developer, minutes instead of a support ticket — and the same governance a human query gets.

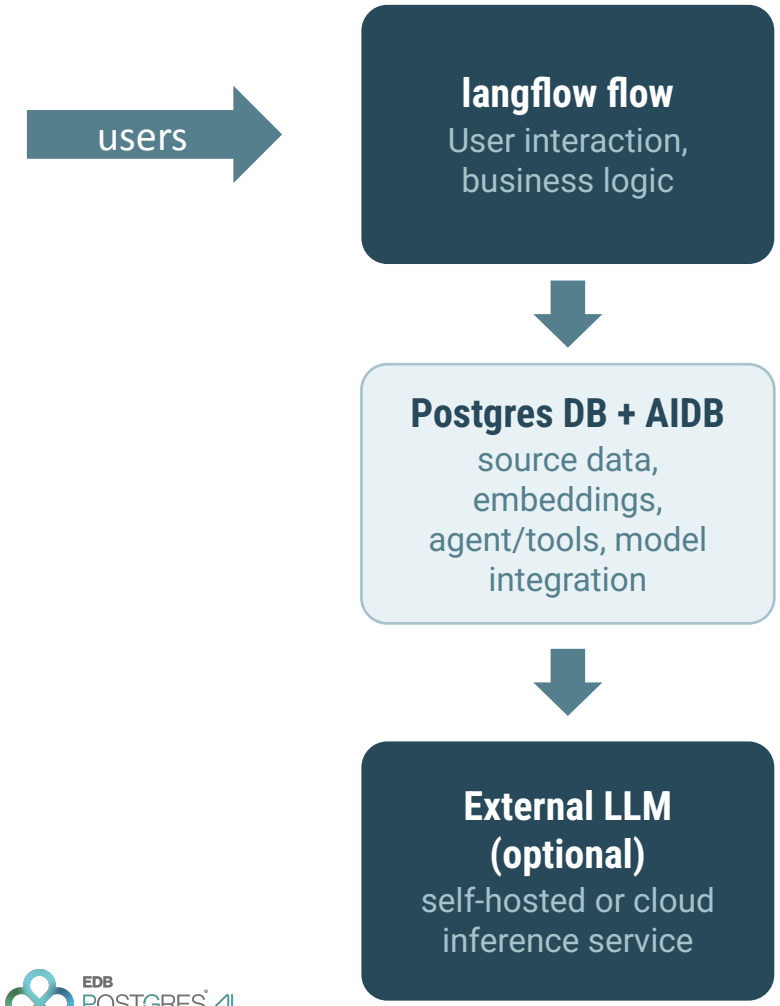
TAKEAWAY **Text-2-SQL is a PG agent, not a bolt-on — it inherits Postgres governance for free.**

EDB POSTGRES AI

HANDS ON

DEMO 2 — CONVERSATIONAL ANALYTICS (TEXT-2-SQL AGENT)

architecture overview



Bring your own front-end / application

Choose any language/framework you want, it only needs a PG DB connector.

Focus on your user experience and business logic.

AIDB extension

Handles the AI logic: agent loop, context handling, embeddings processing, retrieval/vector search, LLM API integration.

PG Database

Holds your source data, embeddings and runs the AIDB extension.

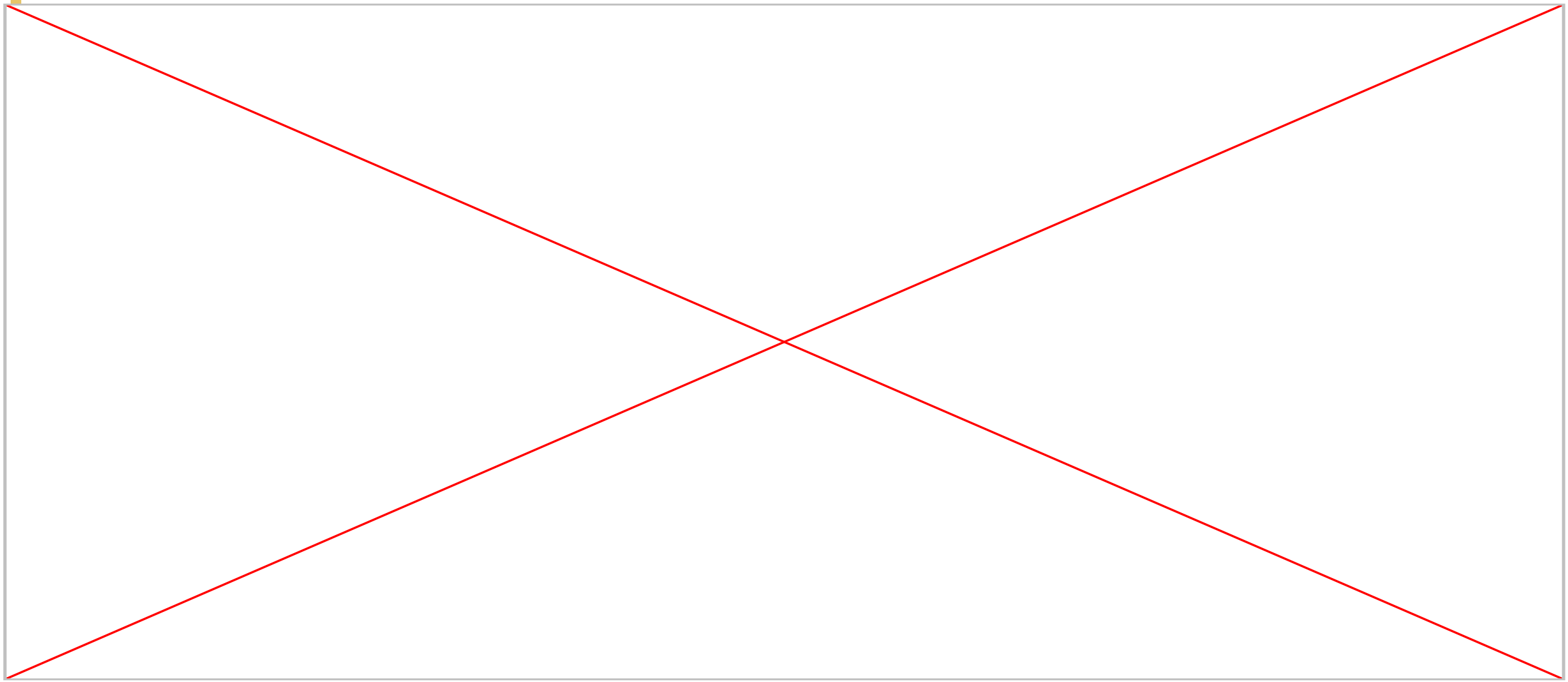
optional external LLM

Connect AIDB to a wide range of external LLMs, or use one of the built-in models that run on CPU right on the database host.

EDB POSTGRES AI

HANDS ON

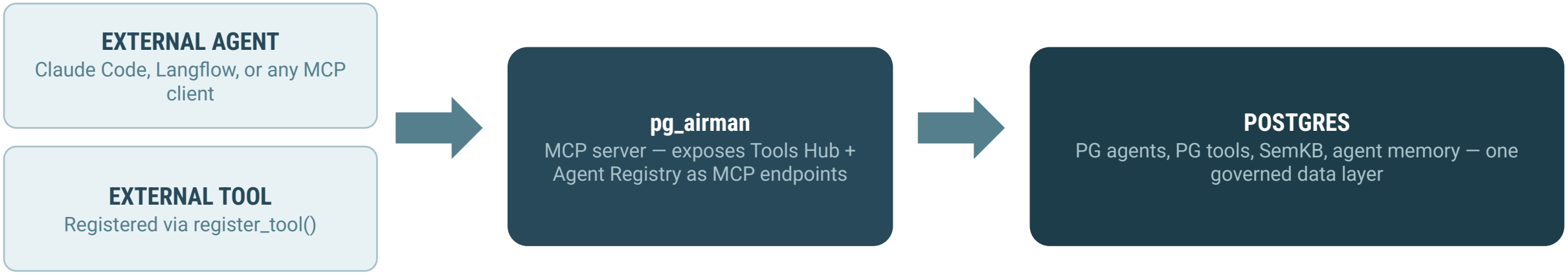
DEMO 2 — CONVERSATIONAL ANALYTICS (TEXT-2-SQL AGENT)



EXTERNAL AGENTS

MCP — POSTGRES AS THE AGENT GATEWAY

External agents (any framework) reach PG tools, models, and agents through one MCP surface.



REGISTER, DON'T REBUILD

External MCP servers register a filtered tools list — no custom gateway to build.

PG AGENTS AS SUB-AGENTS

PG agents can delegate to each other — used like tools, with agent-specific prompts.

LANGFLOW = MCP CLIENT & SERVER

A Langflow flow can call PG via MCP, or itself be published as an MCP tool.

EXTERNAL AGENTS · GOVERNANCE

GOVERNANCE FOR TOOL CALLS

External and internal calls hit the same enforcement point — Postgres roles and policies.

01
PURPOSE ASSIGNMENT

Each MCP connection maps to a declared purpose, mapped to a PG role.

02
ROLE VENDING

Short-lived, scoped roles are vended per agent session — no standing over-privileged credentials.

03
POLICY MAPPING

RLS, views, and security labels enforce what each purpose is allowed to see or change.

04
OBSERVABILITY

Every call — internal or external — lands in the same audit log, causally linked to its trigger.

RULE OF THUMB **An external agent is never more trusted than the purpose it was vended a role for.**

EDB POSTGRES AI

LOOKING AHEAD

2026 ROADMAP: DEEPER INTO THE DATABASE

Every item extends the Postgres you already run — one platform, not a new one.

IN-DATABASE AGENTS

Agents run inside Postgres, governed by the same roles, RLS, and audit as every human query.

→ *Zero new trust boundary*

IN-DATABASE MCP SERVER

An expanded MCP surface exposes governed data, search, and DB expertise to any agent framework.

→ *Works out of the box with any custom stack*

AGENT MEMORY

Durable, queryable long-term memory in Postgres — governed by existing RBAC and audit.

→ *Smarter, cheaper, more reliable agents*

AGENT GOVERNANCE

Tracing, HITL approvals, dynamic role vending, supervisor agents, transaction rollback.

→ *Trustworthy by construction*

ONE PLATFORM **Agents, memory, and knowledge all live in the Postgres you already govern.**

EDB POSTGRES AI

WRAP-UP

KEY TAKEAWAYS

What to remember from today.

Data never leaves

AI Sidecar runs models, agents, and search inside Postgres — no round trip to a cloud API.

Start simple, scale up

Begin with the aidb extension on plain Postgres; add Hybrid Manager for GPU serving and Langflow when you need it.

Governance is native

Every agent and tool call inherits Postgres RBAC, RLS, and audit — not a bolt-on policy engine.

The semantic layer pays for itself

SemKB cuts tokens and raises Text-2-SQL accuracy by grounding agents in real business meaning.

THANK YOU

QUESTIONS?

NEXT SUMMER ACADEMY SESSION

STATEFUL ON KUBERNETES: RUNNING POSTGRES THE RIGHT WAY

14:00-15:00 UTC+2
13TH AUGUST

FREE TRAINING

EDB POSTGRES AI DATABASE ESSENTIALS

EDB POSTGRES AI