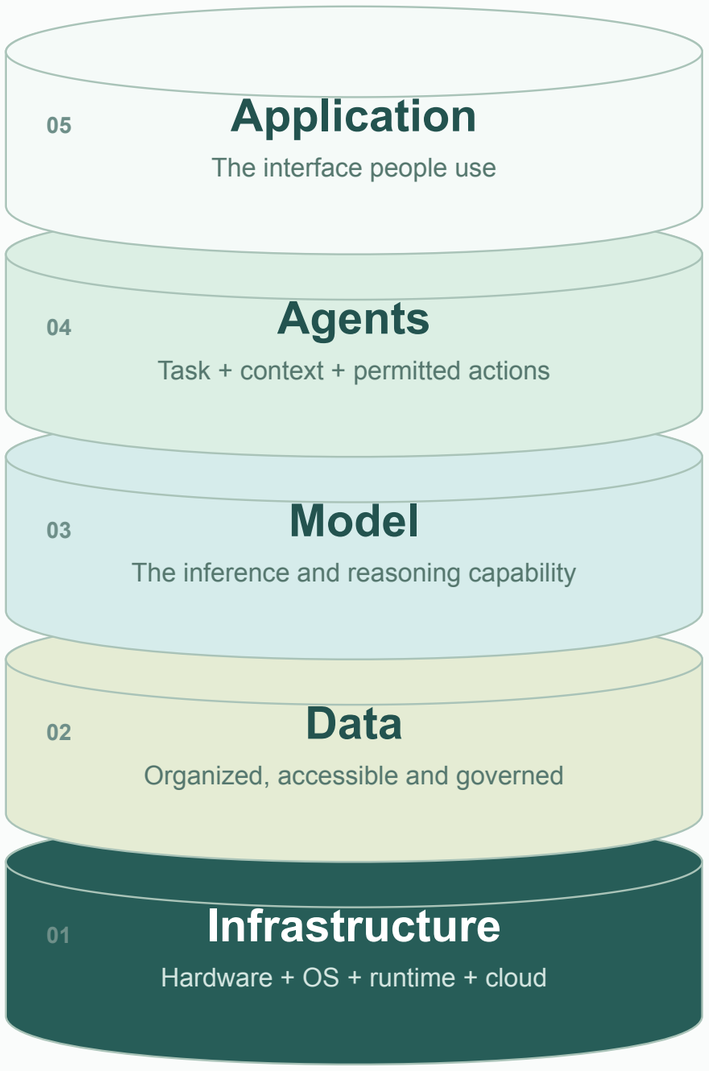
The AI capability stack

THE CAPABILITY STACK



WHAT EACH LAYER CONTRIBUTES

End-user experience / harness

Chat, workspaces, dashboards and approval screens turn intent into tasks.
The signed-in person is distinct from the agent and its service credentials.

ChatGPT · Claude · AionUI · Hermes / other harnesses

An accountable worker with a bounded task

Named owner + purpose + skills + instructions + semantic context + model.
Tools / APIs / MCP connect legacy and new apps; enforce permissions at execution.

Owner & scope · MCP / tools · Audit & evaluations

Choose a model for the input, task and deployment

Text LLMs, multimodal models, embeddings and specialized models.
Cloud APIs, self-hosted endpoints or edge models infer from permitted context.

Cloud / private / edge

Structured + semi-structured + unstructured evidence

Tables / SQL • JSON / events • documents / PDFs • images / audio / video.
Catalog, lineage, quality, freshness and access policies follow the use case.

PostgreSQL · Neon · object storage

The foundation that runs and connects the system

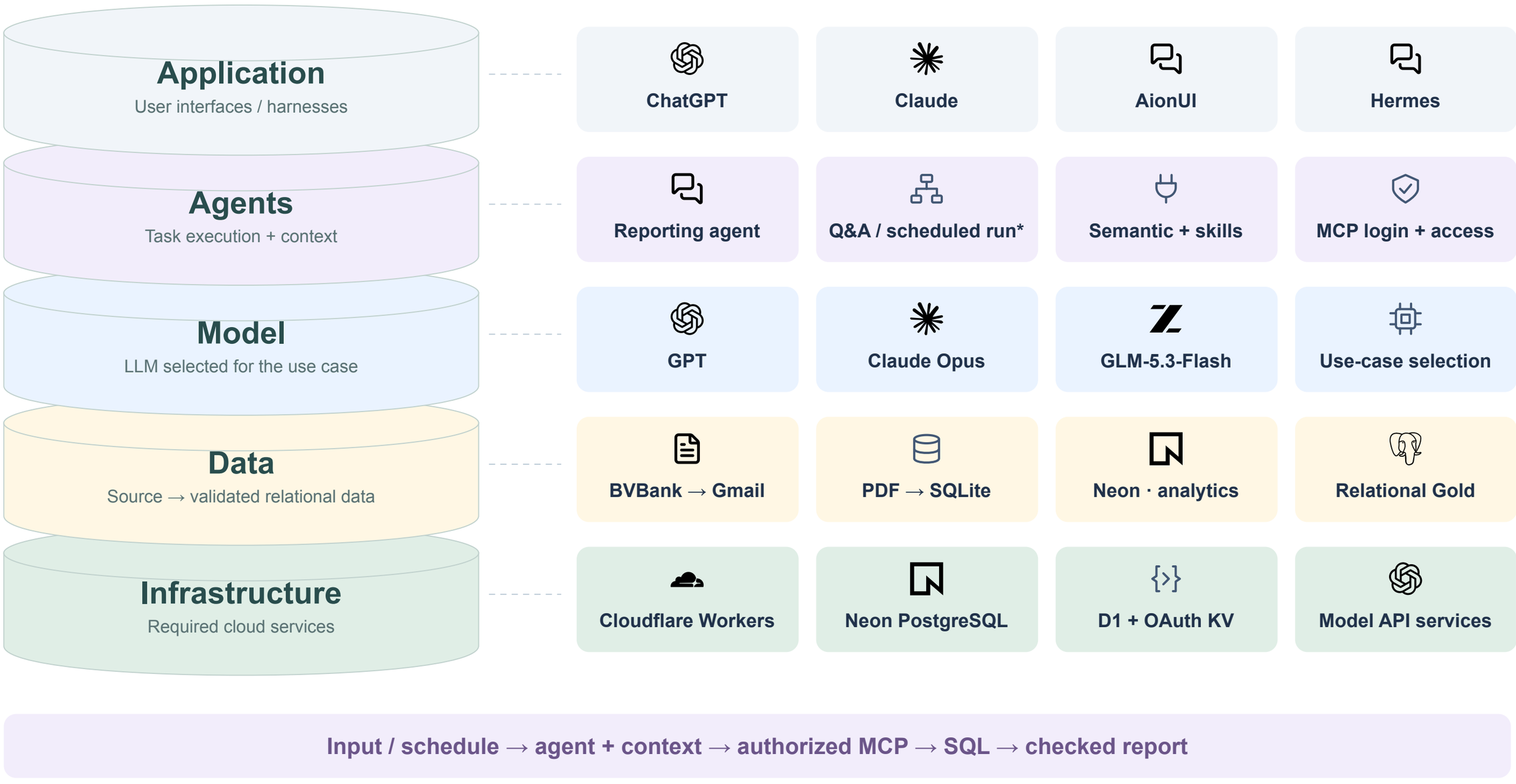
CPU / GPU / NPU • servers / edge • OS • containers • storage • networking.
Inference serving, scaling, secrets and telemetry support the layers above.

AWS · Cloudflare · On-prem / edge

ACROSS ALL FIVE LAYERS

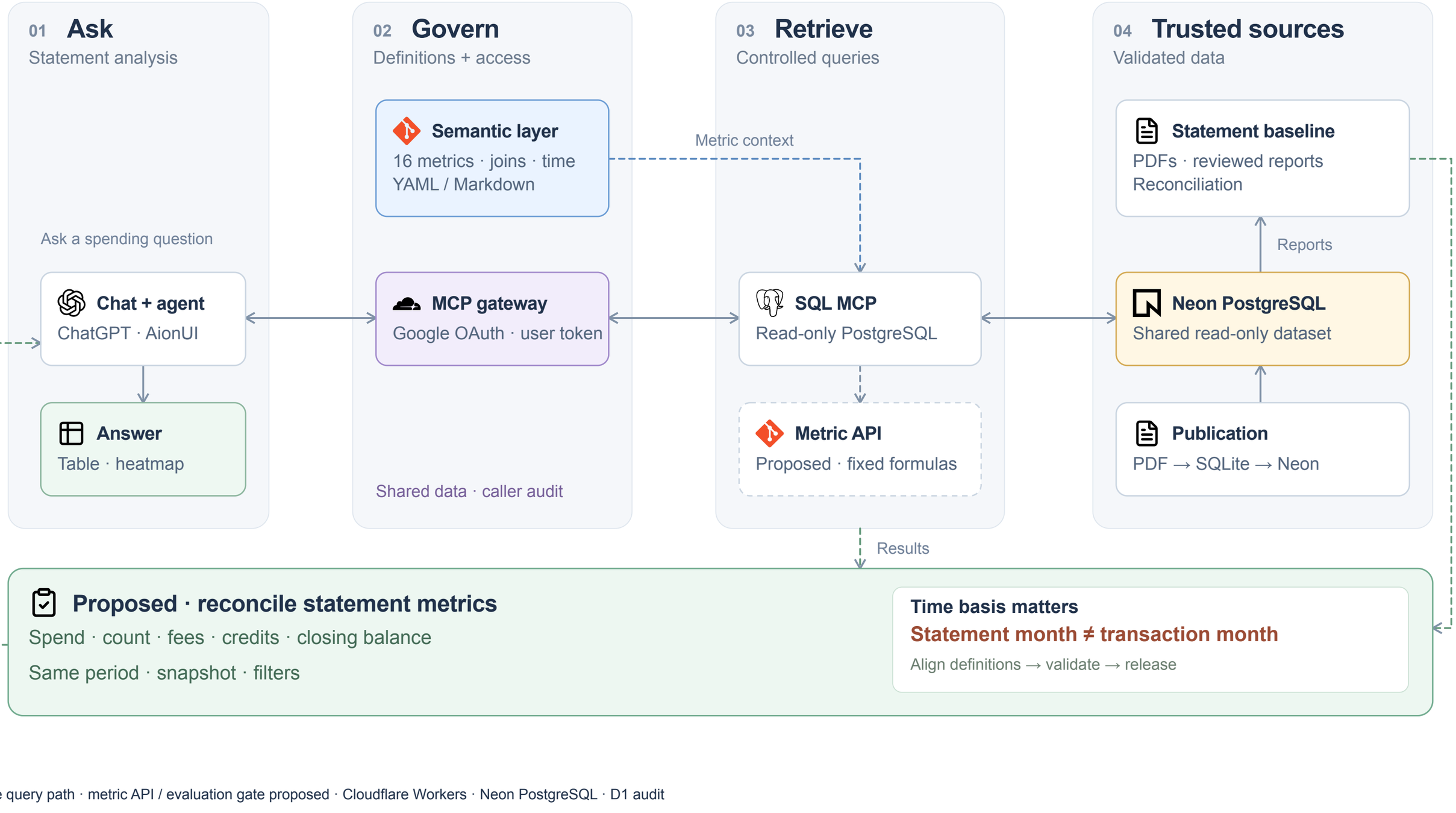
Ownership · identity · permissions · security · governance · audit · cost · evaluations

Our use case, mapped to the AI stack



Use-case map · scheduled reporting proposed · *Scheduled reporting: proposed · Model chosen per harness · Gold serving role = Neon analytics · Shared read-only data

Bankstatement today



Where the data comes from — and how it becomes Gold

01 SOURCE

BVBank → Gmail
Statement email attachments
PDF statements

Local source files
statements/YYYY/YYYY-MM_visa.pdf
Source reference retained

02 VALIDATE + ORGANIZE

Extract + normalize
PDF extraction · merchant labels
Validation gates

SQLite warehouse
db/statements.db
Statements · transactions · merchants

03 SERVE AS GOLD

Validated publication
Counts · keys · signed amounts
Content digest · atomic snapshot

Neon PostgreSQL
neondb.analytics
Relational SQL · SELECT-only

Semantic contract
16 metrics · relationships · date basis · SQL examples

MCP → reporting agent

Gold = reporting-serving role. Physical schema is analytics; no separate gold schema or medallion platform is deployed.

NEON POSTGRESQL / neondb.analytics

Gold serving layer · selected columns

statements

1 row / billing statement

PK statement_id

statement_date · due_date

opening_balance

closing_balance

min_payment

source_pdf

1 → many

transactions

1 row / extracted transaction

PK txn_id

FK statement_id

txn_date · post_date

amount_vnd

merchant_raw

merchant_normalized

category

is_payment · is_fee

is_foreign · is_installment

1 → many

Optional

merchant_cache

1 row / raw merchant token

PK merchant_raw

merchant_normalized

category

normalized_at

dataset_status

Snapshot · published time

Counts · content digest

Join rules

Join on merchant_raw

Cache link: logical, not FK

Use transaction labels

for historical reporting

Aggregate transactions before joining balances · Statement month ≠ transaction month · Missing data ≠ zero

Actual tables · analytics schema · Solid: statement relationship · Dashed: optional merchant lookup · dataset_status is snapshot metadata, not a transaction join

Semantic layer — what it contains

Business meaning → consistent SQL

13 Owner + version · Admin-owned · YAML / Markdown

01
Scope
BVBank credit-card statements
Not all personal spending

02
Entities + grain
Statement: one billing statement
Transaction: one extracted row

03
Data dictionary
amount_vnd → signed VND
txn_date → transaction date

04
Relationships
transactions → statements
Join using statement_id

05
Metric definitions
Gross spend: positive purchases
Exclude payments, fees, installments

06
Dimensions
Day · merchant · category
Group and compare consistently

07
Time rules
txn_date → calendar month
statement_id → billing month

08
Business filters
is_payment · is_fee
is_installment · is_foreign

09
Values + missing data
Preserve signs and currency
NULL / missing period ≠ zero

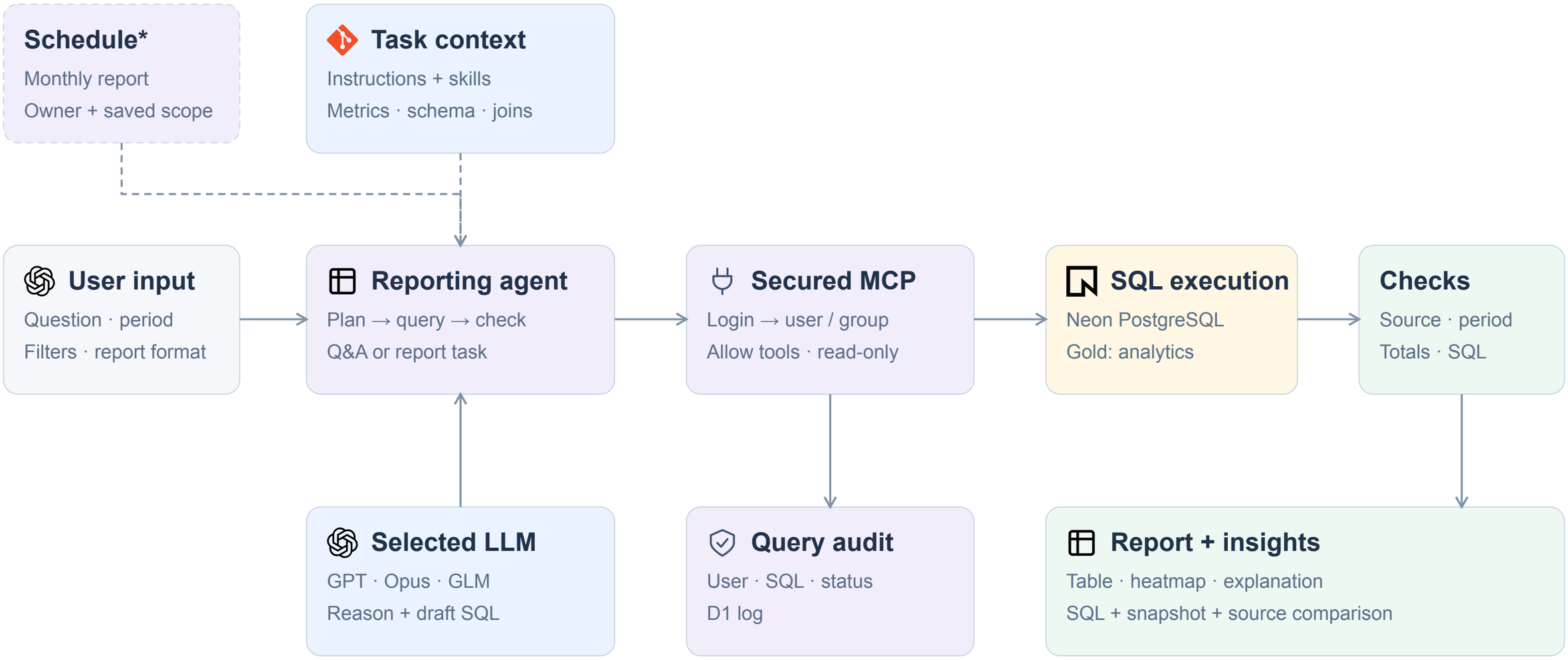
10
Lineage + freshness
source_pdf → original evidence
dataset_status → snapshot time

11
Validation + evidence
Opening + net movement vs closing
Heatmap sum = filtered spend

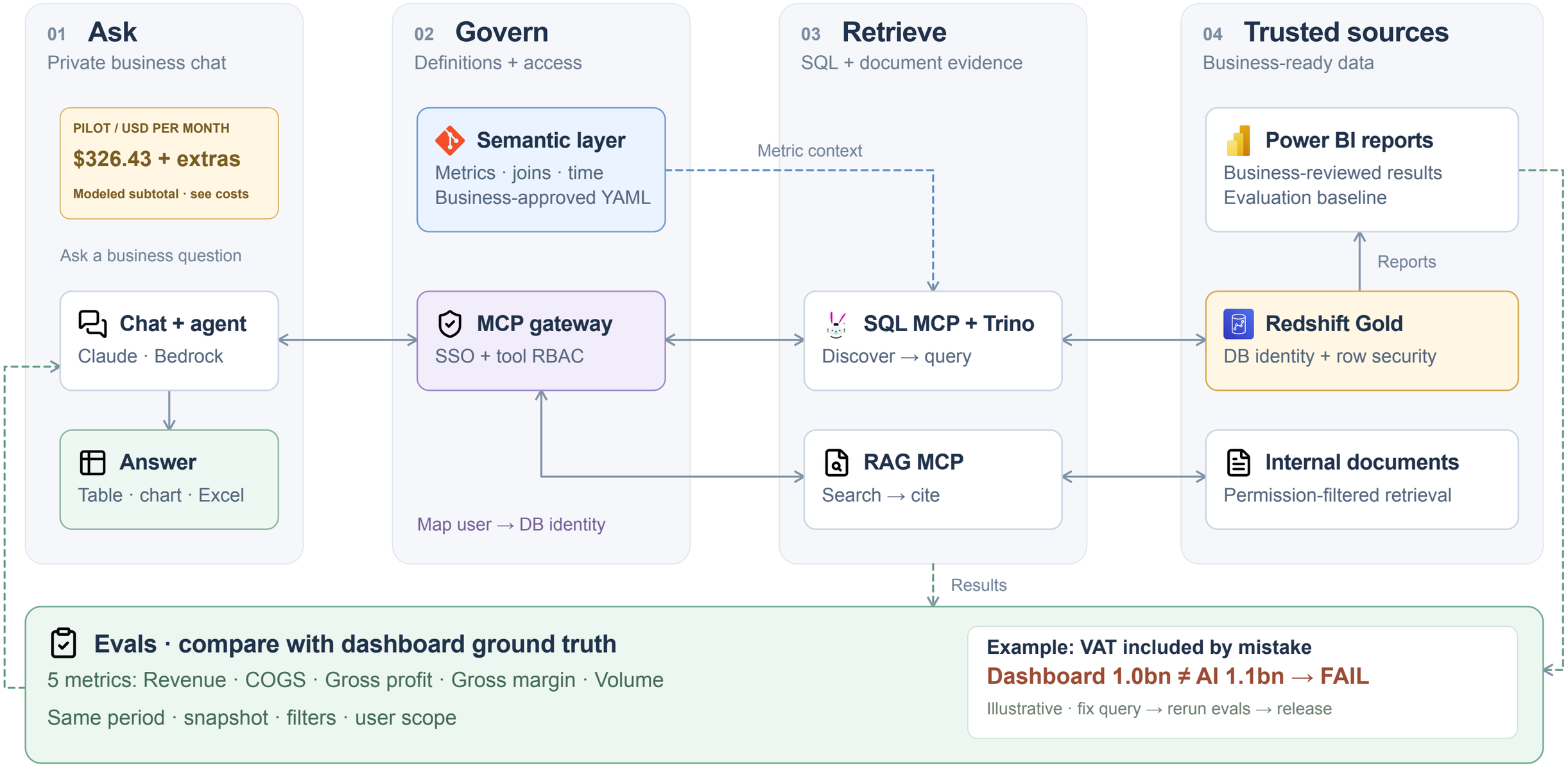
12
SQL patterns + limits
Canonical queries · safe joins
No credit limit → no utilization

Example: “February spending heatmap”
Purchase metric + transaction month + exclusions → daily SQL → check totals against the same filtered data

From a question or schedule to a checked report



*Scheduled trigger + automated report gate: proposed. Existing source validation and read-only MCP remain in place.


01 Ask
Private business chat
PILOT / USD PER MONTH
$326.43 + extras
Modeled subtotal · see costs
Ask a business question
Chat + agent
Claude · Bedrock
Answer
Table · chart · Excel
02 Govern
Definitions + access
Semantic layer
Metrics · joins · time
Business-approved YAML
MCP gateway
SSO + tool RBAC
Map user → DB identity
03 Retrieve
SQL + document evidence
Metric context
SQL MCP + Trino
Discover → query
RAG MCP
Search → cite
04 Trusted sources
Business-ready data
Power BI reports
Business-reviewed results
Evaluation baseline
Reports
Redshift Gold
DB identity + row security
Internal documents
Permission-filtered retrieval
Results
Evals · compare with dashboard ground truth
5 metrics: Revenue · COGS · Gross profit · Gross margin · Volume
Same period · snapshot · filters · user scope
Example: VAT included by mistake
Dashboard 1.0bn ≠ AI 1.1bn → FAIL
Illustrative · fix query → rerun evals → release

Illustrative pilot · USD · US East (N. Virginia) · 730 server hours / month

1,000 tasks / month × 10k input + 1k output tokens across all calls per task · No credits, tax or discounts

01 LLM · Bedrock Sonnet 4.6

$45.00 / month

10M input × $3 + 1M output × $15
Global cross-region; standard uncached inference
Retries, more context or another model change this cost.

02 Servers · Fargate + Trino

$176.20 / month

4 × Fargate tasks: $36.04 · chat / gateway / SQL / RAG
Trino: 2 × EC2 m6i.large × $0.096/h = $140.16
Fargate per task: 0.25 vCPU + 0.5 GB; pilot sizing.

03 Warehouse · Redshift Serverless

$45.00 / month

4 RPUs × 30 active hours × $0.375/RPU-hour
Usage example, not a guaranteed monthly cap.
Storage, scaling and extra active time are additional.

04 Network · supporting services

$60.23 / month + variable

1 ALB base: $16.43 · 1 NAT gateway base: $32.85
3 public IPv4 addresses: $10.95 (2 ALB + 1 NAT)
EKS option: add $73/month control plane, plus nodes.

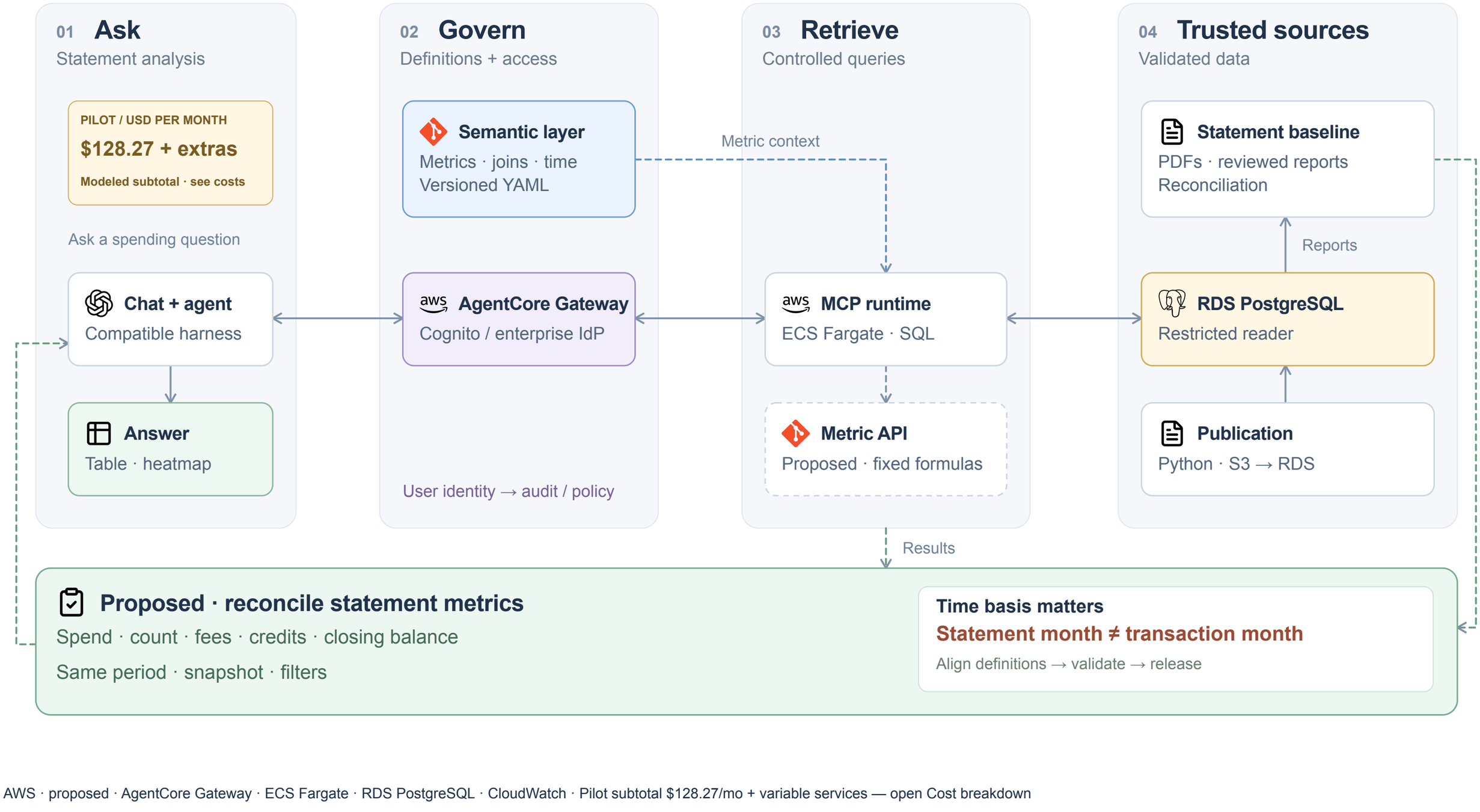
Modeled subtotal: $326.43 / month + unpriced usage / services

Pilot estimate, not an AWS bill or production capacity recommendation. LLM API usage is separate from chat subscriptions.

Additional: EBS / Redshift storage · S3 · RAG index + embeddings · logs · secrets · identity · Power BI licenses

Both exclude ALB LCUs, NAT processing, data transfer, tax and operational labor. Network design may change the baseline.

Rates checked 18 Nov 2026 · EC2 uses the cited AWS baseline · Open “Rates & assumptions” for sources and calculations.


01 Ask
Statement analysis
PILOT / USD PER MONTH
$128.27 + extras
Modeled subtotal · see costs
Ask a spending question
Chat + agent
Compatible harness
Answer
Table · heatmap
02 Govern
Definitions + access
Semantic layer
Metrics · joins · time
Versioned YAML
AgentCore Gateway
Cognito / enterprise IdP
User identity → audit / policy
03 Retrieve
Controlled queries
Metric context
MCP runtime
ECS Fargate · SQL
Metric API
Proposed · fixed formulas
04 Trusted sources
Validated data
Statement baseline
PDFs · reviewed reports
Reconciliation
Reports
RDS PostgreSQL
Restricted reader
Publication
Python · S3 → RDS
Results
Proposed · reconcile statement metrics
Spend · count · fees · credits · closing balance
Same period · snapshot · filters
Time basis matters
Statement month ≠ transaction month
Align definitions → validate → release

Illustrative pilot · USD · US East (N. Virginia) · 730 server hours / month

1,000 tasks / month × 10k input + 1k output tokens across all calls per task · No credits, tax or discounts

01 LLM · Bedrock Sonnet 4.6

$45.00 / month

10M input × $3 + 1M output × $15
Global cross-region; standard uncached inference
Retries, more context or another model change this cost.

02 MCP server · ECS Fargate

$9.01 / month

1 task × 0.25 vCPU + 0.5 GB × 730 hours
Rates: $0.04048/vCPU-hour + $0.004445/GB-hour
Single task pilot; replicas and ingestion jobs add cost.

03 Database · RDS PostgreSQL

$13.98 / month

Single-AZ db.t4g.micro: $0.016/h × 730 = $11.68
20 GB gp3 × $0.115/GB-month = $2.30
Extra backups, CPU credits and capacity add cost.

04 Gateway + network

$60.28 / month + variable

AgentCore: 10k calls × $0.005/1k = $0.05
ALB $16.43 + NAT $32.85 + 3 IPv4 $10.95
Tool search / indexing and data processing add cost.

Modeled subtotal: $128.27 / month + unpriced usage / services

Pilot estimate, not an AWS bill or production capacity recommendation. LLM API usage is separate from chat subscriptions.

Additional: Cognito / SSO · CloudWatch · Secrets Manager · S3 · backups · ingestion jobs · DNS · harness subscriptions

Both exclude ALB LCUs, NAT processing, data transfer, tax and operational labor. Network design may change the baseline.

Rates checked 18 Nov 2026 · EC2 uses the cited AWS baseline · Open “Rates & assumptions” for sources and calculations.