OneUI — Technical Summary (Beta)

What OneUI is

OneUI is a developer-first **AI app development platform** that unifies chat, typed tools, agents, and enterprise systems behind a single interface. It pairs a **deterministic core** (rules/DSL, SQL routes, typed tools) with **LLM orchestration** for reasoning and explanation. The UI streams structured **ChatBlocks** (markdown/table/chart/job/form) and delivers web-first answers with citations; every run produces a **receipt** with a stable `trace_id` for audit and cost accounting. Multi-tenant isolation, policy gates, and budgets are built in.

Why it matters

Teams want governed AI apps without rebuilding "AI plumbing." OneUI combines deterministic workflows with agentic assistance so engineering teams can **ship faster** while staying **compliant and cost-aware** (policy, audit, budgets).

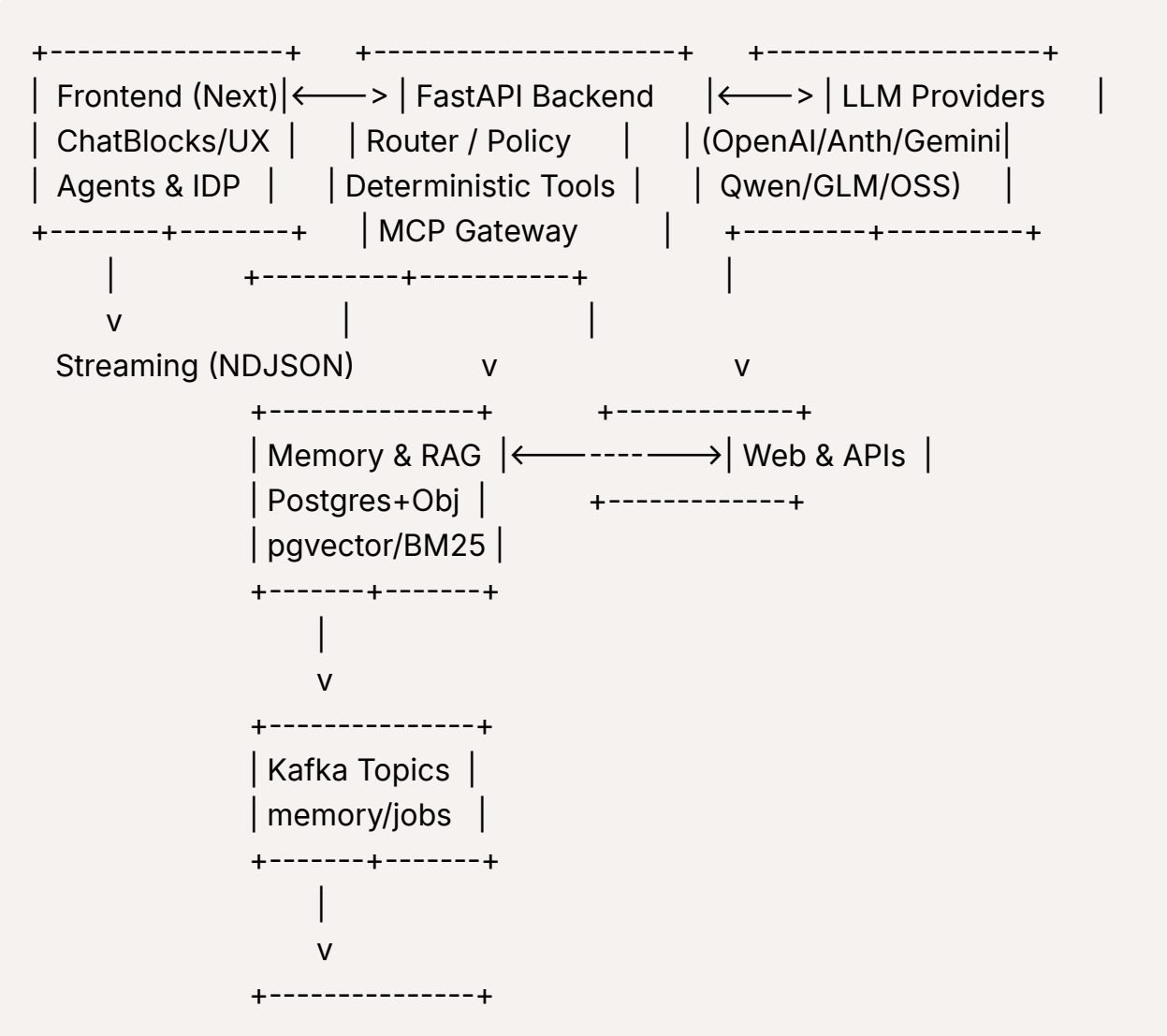
Core capabilities (v4)

- **Deterministic Core + Streaming.** Typed tools and SQL routes; NDJSON streaming into ChatBlocks (markdown/table/chart/job/form).
- **Multi-LLM Router + Policy Engine.** Route across OpenAI, Anthropic, Gemini, Qwen/GLM, and OSS backends with A/B, failover, small-model-first; pre/tool/post policy hooks (redaction, approvals, allow/deny, model/tool selection).
- **Durable Memory & Audit.** Postgres + object storage; optional pgvector/BM25 hybrid retrieval; MemoryCards for short factual memory; append-only logs via Kafka.
- **Agents & MCP.** Deep Research agent with evaluation harness; **MCP Host/Client** to register and call external servers (e.g.,

GitHub/Filesystem/Unity/Blender).

- **Integrations.** Google Workspace (Gmail/Drive/Sheets), Microsoft 365, SQL/BigQuery; HubSpot/Proxycurl; QuickBooks→Cloud SQL→Metabase; n8n automations.
- **Tenancy & Governance.** Namespaced isolation; SSO/SAML/OAuth2 patterns; PII tagging/redaction; egress allowlists; budgets/quotas; audit logs.
- **Observability & Cost.** Langfuse traces (prompt/tool spans), Prometheus/Grafana metrics, Elastic logs; per-tenant budgets and usage reports.

Architecture at a glance



```
                    | Workers/Jobs  |
                    +---------------+
```

(Shape matches the design doc’s high-level system.)

Product surfaces

Chat & Streaming (ChatBlocks with citations), **Agents** (Deep Research + Agent Eval; MCP Host/Client), **Apps** (Gmail/Drive/Sheets, SQL/BigQuery, HubSpot/Proxycurl, QBO→Metabase), and **IDP** (GitHub Control Room, Runway IaC JSON bundles, SRE Operations Cockpit).

What’s live vs. mock (Sep 26, 2025)

Live: GitHub read paths; Runway JSON generation (strict schema, correlation IDs); Deep Research agent (partial evals); Memory (Kafka + Postgres/pgvector) with Langfuse traces; Cloud Run + Cloud SQL baseline; Metabase + n8n; Prometheus/Grafana app metrics. **Demo/Mock:** SRE dashboards adapters (CloudWatch/Cloud Logging/OTel) and Runway apply/Open-PR; expanded Agent Eval UI.

Deployment options

Cloud Run baseline (app + workers, Cloud SQL, object storage, Secret Manager/KMS, Prometheus/Grafana) or **Kubernetes path** (GitOps via GH Actions→GHCR→Argo CD; per-tenant namespaces; autoscaling workers; blue/green rollouts). Tenant provisioning creates DB schema + storage prefix (and K8s namespace if applicable) and applies budgets, policy profile, and keys.

Security & governance

Tenant-scoped schemas/namespaces; policy gates (pre/tool/post) with approvals; role-based access; PII redaction; egress allowlists; audit logs; budgets/quotas; secrets via cloud KMS/Secret Manager.

Observability & KPIs

Langfuse traces and cost accounting; Prometheus/Grafana metrics; Elastic logs. **Targets:** TTFC ≈ 2s; research p95 ≈ 12s; deterministic routes p95 ≤ 800ms; citation precision ≥ 0.90; hallucination ≤ 5%.

Near-term roadmap (30/60/90)

30d: MCP SSE + registry; policy gates & approvals; Langfuse spans per MCP call; eval golden sets + rubric + red-team; Runway validate-bundle + Open-PR; starter marketplace.

60d: SRE adapters (CloudWatch/Cloud Logging/OTel), runbooks, ChatOps approvals; private GitHub flows; issues/PR triage; budget guardrails.

90d: Drift detection/auto-remediation; meta-agent reasoning; Developer VMs orchestration; Unity/Blender MCP servers prod; onboarding flows (namespace provisioning, budgets, policy profiles).

Example 2-week pilot (DevOps/Platform/SRE)

W1: Deploy; SSO; ship 1 governed template (Runway or GitHub Control Room); receipts end-to-end. **W2:** Add a cloud adapter (CloudWatch/Cloud Logging), dashboards + evals; exit readout with metrics + next-step backlog. (Derived from product surfaces and deployment model.)

Availability & licensing

Available now under a **perpetual, permissive commercial source license** (one-time fee; early-adopter price **$10k** — 80% off). Use/modify/derive and offer SaaS/compiled apps; **no open-sourcing or re-selling the source**. (Your agreed licensing approach.)