OneUI — Comprehensive Executive Summary (2025/09/26)

Date: Sep 26, 2025

1) What OneUI is

OneUI is an **AI OS for Work**: a deterministic-first platform that unifies chat, tools, agents, and enterprise systems behind a single interface. Typed tools, rules, and SQL do the heavy lifting; **LLMs orchestrate, summarize, and explain**. The product streams **ChatBlocks** (markdown/table/chart/job/form) in real time, delivers **web-first answers with citations**, and provides the isolation, auditability, and cost controls enterprises require.

---

2) Why it matters

- Teams are drowning in fragmented SaaS and ad-hoc automations.
- Pure LLM chat is unreliable and unaudited; point tools are narrow and siloed.
- OneUI combines **deterministic workflows** with **agentic assistance** so orgs can ship faster while staying compliant and cost-aware.

---

3) Core Capabilities (v4)

A. Deterministic Core

- Typed tools, rules/DSL, and SQL routes.
- Real-time NDJSON streaming into ChatBlocks (markdown/table/chart/job/form).
- Hybrid retrieval (BM25 + embeddings) with **web-first** mode and clear citations.

B. Multi-LLM Router + Policy Engine

- **Router** across OpenAI, Anthropic, Gemini, Qwen/GLM, and OSS; supports A/B, failover, and small-model-first routing.
- **Policy Engine** with pre-prompt, tool, and post-gen hooks for redaction, allow/deny, human-review, and model/tool selection.
- Stable `trace_id` and final "receipts" for every response.

C. Durable Memory & Audit

- **Postgres + object storage** for long-term memory; **Kafka** topics for append-only logs.
- MemoryCards (≤120 tokens) for facts/preferences; per-session/agent working memory; versioned writes with replay.
- Full **Langfuse** tracing and cost accounting.

D. Agents, MCP, and Media

- **Deep Research Agent**: multi-step research with tool calling and evaluation harness (TTFC, citation precision, hallucination rate, cost, latency).
- **MCP Host & Client**: OneUI lists/controls cloud-hosted MCP servers (e.g., Unity, Blender, Filesystem, GitHub), proxies calls, and records traces.
- **Media pipelines**: FAL/ElevenLabs/Seedance; prompt-to-3D (Mixamo/Unity).
- **Developer VMs** (roadmap): orchestrate Claude/Gemini/Qwen coding copilots per VM with centralized observability.

E. Integrations & Data Workflows

- Google Workspace (Gmail/Drive/Sheets), Microsoft 365, SQL/BigQuery, Proxycurl → HubSpot, QuickBooks Online → Cloud SQL → Metabase, n8n automations.
- Sheets → Cloud SQL ingestion: schema inference, create/alter table, batched inserts, and query preview.
- **Runway** blueprints generate validated IaC bundles.

F. Tenancy, Security, and Governance

- Namespaced multi-tenant isolation (DB schemas, K8s namespaces, storage prefixes).
- Policy-gated tool access; budgets/quotas; PII tagging; egress allowlists; audit logs.
- OAuth2/SSO/SAML supported patterns.

G. Observability & Cost Controls

- Langfuse traces (prompt/response/tool spans), Prometheus/Grafana metrics, Elastic logs.
- Per-tenant budgets and rate limits; cost dashboards; model usage reports.

---

4) Current Product Surfaces

- **Chat** (streamed ChatBlocks) and **Streaming** views.
- **Agents**: MCP Host/Client UI; Deep Research agent + Eval UI.
- **Apps**: Gmail, Drive (New from Prompt), SQL/Sheets/BigQuery, HubSpot/Proxycurl, QBO→Metabase.
- **IDP Tab**:
  - **GitHub Control Room** (public repo browse live; PR assist mock; private flows via PAT/OAuth roadmap).
  - **SRE Operations Cockpit** (logs/metrics/traces/errors/SLOs/usage—adapters WIP).
  - **Runway** (Self-Service Blueprints) with `/api/runway/gemini/generate` returning strict JSON IaC bundles.

---

5) What's Live vs Mock (Sep 26, 2025)

Live

- GitHub read paths; top bar integration.

- Runway JSON generation endpoint (strict schema, correlation IDs, audit logs).
- Deep Research agent runs; partial eval harness with metrics.
- Memory: Kafka (`oneui.memory.v1`) + Postgres/pgvector; Langfuse traces.
- Deployments: Cloud Run + Cloud SQL; Metabase and n8n running; K8s/GitOps path (GH Actions → GHCR → Argo CD). Prometheus/Grafana app metrics.

Demo/Mock

- SRE dashboards (adapters to CloudWatch/Cloud Logging/OTel WIP).
- Runway apply/Open-PR path (scaffolded).
- Agent Evaluations UI (expanding coverage + rubric scoring).

---

6) KPIs & Targets

- **TTFC** (research) < 2s; **p95 latency** < 12s.
- **Citation precision** ≥ 0.90; **hallucination rate** ≤ 5%.
- **Cost/run** (research) ≤ $0.30 at p95.
- **Blueprint validity** ≥ 95% (fmt/tflint/OPA/OPA pass).
- **MCP connection success** ≥ 99% across registered servers.
- **Deterministic route p95** ≈ 800 ms.

---

7) 30/60/90 Roadmap

30 days

- MCP: SSE streaming; server registry/tags; per-server auth vault; policy gates; Langfuse spans per call.
- Eval: golden sets; rubric scoring; red-team checks; exportable reports.
- Runway: validate-bundle API; Open-PR; starter marketplace.

60 days

- Bind SRE to CloudWatch/Cloud Logging + OpenTelemetry; runbooks + ChatOps actions w/ approvals.

- Private GitHub flows; Issues/PR triage agent.
- Cross-agent memory and budget guardrails.

90 days

- Drift detection/auto-remediation; meta-agent reasoning layer.
- Developer VMs orchestration; Unity/Blender MCP servers in production.
- Customer onboarding (namespace provisioning + budgets + policy profiles).

---

8) Differentiation vs Generic LLM Chat

- **Deterministic-first** workflows, with LLMs used last for explanation/formatting.
- **Governance** (policies, approvals, audit) built in.
- **Observability** (traces, metrics, cost) by default.
- **Enterprise isolation** (namespaces, per-tenant budgets) and **web-first citations**.

---

9) Risks & Mitigations

- **Upstream model/API changes** → circuit breakers, retries, backoff, feature flags.
- **Eval drift** → frozen golden sets; versioned prompts; scheduled re-runs.
- **Cost creep** → budgets, small-model routing, caching.
- **Multi-tenant data leakage** → strict namespaces, egress allowlists, least-privilege creds, audits.

---

10) Quick Demo Script (3–5 min)

1. Run Deep Research on a market/tech query; show citations + metrics.
2. Open IDP → Runway; generate a validated IaC bundle (JSON receipt).
3. Browse a public repo in GitHub Control Room; preview PR assist (mock).

4. Switch to MCP Host; connect a sample MCP (demo server); call a tool; show Langfuse trace.
5. In Apps → Sheets→SQL, ingest a Google Sheet and query it; render a chart ChatBlock.

---

11) Packaging & Deployment

- Cloud Run + Cloud SQL baseline; optional K8s/GitOps deployment.
- Tenant provisioning: creates DB schema + storage prefix + K8s namespace (K8s path); budgets + policy profile applied.
- Secrets via cloud KMS/Secret Manager.

---

12) Ask

- Design partners for agent evaluations and Runway templates.
- Early adopters for MCP servers (Unity/Blender/GitHub) and SRE cockpit adapters.
- Feedback on KPIs; proposals for shared golden sets.