FundOS — AI Evaluation Report

Issued August 13, 2026 · latest eval run: July 16, 2026 (manual baseline run) · FundOS by VantedgeAI · fundos.vantedgeai.com/trust

Metric	Result	Detail
Documentation benchmark	78/80 (97.5%)	40 tasks, scored 0–2 each
Hallucinations detected	0	tripwired auto-fail on invented tools/endpoints
Tier-router golden evals	17/19 (89.5%)	live through the production LLM router
Deterministic calc gates	enforced per merge	waterfall / NAV / tenant isolation — blocking CI

Longitudinal trend begins accruing with the nightly unattended runs; every future trend point is a scheduled run — never a manual re-run.

Methodology

Doc-benchmark (nightly). A fixed 40-task question suite is answered by an LLM agent using only the published FundOS documentation. Each answer is scored 0–2 against golden expectations, with hallucination tripwires that auto-fail any answer inventing tools, endpoints, or permissions that don't exist.

Tier-router golden evals (nightly). Golden fixtures (fund distributions, reconciliation-break resolution, and core reasoning) run live through the production LLM tier router and are checked with deterministic verbs — the same route real workloads take, so provider model drift is caught within a day.

Deterministic gates (every merge). Waterfall and NAV golden-file mathematics, capital-account allocation, and tenant-isolation access probes run as blocking CI on every pull request. A regression cannot merge, independent of any LLM.

Data integrity. Trend points come exclusively from unattended scheduled runs — manual re-runs are tagged and excluded from the longitudinal series, so the trend cannot be cherry-picked.

This report is generated directly from the committed eval artifacts (docs/eval/) with no manual editing. Live version: fundos.vantedgeai.com/trust