Live state-validity benchmark on Databricks + TfL

A real-time test of whether previously derived AI state can be safely reused, selectively re-grounded, or held as source data changes.

71.7%	43	17	0
MODEL CALLS AVOIDED	LIVE PRESERVE	LIVE REGROUND	FALSE-SAFE / FALSE INVALIDATIONS

WHAT WAS TESTED

- Real live Transport for London arrival predictions across four busy Underground stations.
- Source state written to and read back from a real Databricks Lakebase trial workspace.
- Frozen Persistara v1.5.2 Transition Contract engine kept private on the local laptop.
- GPT-4.1 control path recomputed on every scored live revision; Persistara recomputed only when the verified validity boundary was crossed.

FLOW

TfL live data → Databricks Lakebase → Persistara v1.5.2 → PRESERVE / REGROUND / HOLD → Evidence back to Lakebase

AUDITED RESULT

- 60/60 live decisions independently reproduced from the recorded before/after source values.
- 60/60 Transition Contract rules independently reproduced.
- 60/60 recompute flags were consistent with the decision.
- 43 of 60 control-path model calls avoided; 71.5% fewer tokens and ~72.0% lower estimated model cost in this benchmark.

EXAMPLE BEHAVIOUR

444s → 194s	**WAITING → WAITING PRESERVE — no model call**
324s → 45s	**WAITING → ARRIVING REGROUND — recompute**

SCOPE Designed benchmark using real live external data; the 180-second ARRIVING/WAITING boundary is synthetic. This validates Databricks Lakebase integration, not yet Managed Agent Memory or an independent customer production workload.

NEXT STEP: 14-DAY SHADOW PILOT • **No production control** • **Minimal integration** • **Confidential validity / recomputation report**

Evidence: run tfl-20260927-072708-03dde0 | Independent audit PASS | 60 live scored revisions