



Truveil

AI ACCOUNTABILITY AUDIT REPORT

Customer Outreach Assistant

Audit of Customer Outreach Assistant

REPORT ID	SR-140
GENERATED	Sun, 06 Sep 2026 21:51:54 GMT
SUBJECT AGENT	Customer Outreach Assistant
AGENT CATEGORY	1 (Declared at Registration)
RISK CLASSIFICATION	LOW
FRAMEWORKS APPLIED	Truveil methodology, UAE PDPL

Report Contents

I	Auditor Opinion
II	Scope and Methodology
III	Agent System Description
IV	Findings
V	Recommended Actions
VI	Technical Detail

Auditor Opinion

A

Accountability Grade · 86.4/100

LOW RISK

This consent-screening agent scores 86.4/100 (Grade A, LOW risk). Transparency and Reversibility are perfect. Data Trust is strong, with bias testing appropriately waived under a documented deterministic rule. Accountability is the sole drag: the model owner is unnamed, no formal risk assessment was logged, and no override capability was recorded – gaps informed by EU AI Act Article 26, EU AI Act Article 9, and EU AI Act Article 14(4)(d) as reference standards.

TRANSPARENCY

100

ACCOUNTABILITY

68.4

DATA TRUST

90

REVERSIBILITY

100

ISSUED BY

This report has been generated by Truveil, an AI agent accountability platform. The accountability scoring is rule-based and deterministic, calibrated against primary regulatory text. The narrative interpretation in this report is produced by Anthropic Claude under structured constraints maintained by Truveil. No portion of the scoring is delegated to a generative model.

Scope and Methodology

What this report covers

This report covers a single execution of the agent identified on the cover page. The audit assesses the logged actions, decisions, validations, and human checkpoints recorded during that execution against four accountability dimensions: *Transparency*, *Accountability*, *Data Trust*, and *Reversibility*. Each dimension corresponds to a category of regulatory obligation recurring across the frameworks Truveil applies.

How Truveil scores

Truveil's scoring engine is rule-based and deterministic. The same evidence, scored against the same framework, produces the same result every time. Each accountability signal is anchored to specific provisions in primary regulatory text. Caps and floors apply where a binding obligation is unmet, these are not weighted heuristics but explicit thresholds defined in the underlying regulation.

Narrative interpretation in this report is produced by a language model under structured constraints. The model summarises findings already determined by the engine. It does not score, classify, or rank the agent. Citations in this report are restricted to provisions present in the engine's citation index for this run.

Regulatory anchors for this audit

Truveil methodology, UAE PDPL

What this report does not certify

This report is an evidence-based assessment of one execution of one agent. It does not constitute a regulatory filing, a legal opinion, or a certification of conformity. It is not a substitute for accredited third-party assurance such as ISO/IEC 42001 certification. Truveil's role is to surface accountability gaps that would otherwise remain undocumented, against the regulatory provisions identified in the engine's citation index.

Agent System Description

Agent and run metadata

AGENT	Customer Outreach Assistant
CATEGORY	1 (Declared at Registration)
CATEGORY BASIS	Category 1 declared at agent registration.
RUN ID	run_20260902_b7e68b
EXECUTION	2026-09-02T20:27:43.489Z
TIMESTAMP	
TOTAL LOGGED	17
EVENTS	

Step-by-step execution

Reconstructed from 17 logged events via the SDK; SDK instrumentation observes events the agent code explicitly logs.

01 AI Disclosure

Agent disclosed its AI nature and recommendation-only status to the operator before any segmentation began.

OK

02 Customer List Receipt

Operator supplied 12 synthetic records with name and consent status; no external data was queried.

OK

03 Individual Consent Decisions

Agent evaluated all 12 records, excluding 6 and including 6 based on a single deterministic consent rule.

OK

04 Human Approval Request

Agent surfaced two flagged issues and sought explicit operator approval before finalising the send list.

OK

05 Approval Confirmed & Run Close-Out

Operator approved the 6-recipient list; no message was sent and no external system was written to.

OK

Findings

Findings index

#	SEVERITY	FINDING	FRAMEWORK
1	HIGH	No Risk Assessment Logged for Category 1 Agent	Truveil methodology, informed by EU AI Act Article 9 and NIST AI RMF as reference standards
2	HIGH	Model Owner Not Named in Any Log	Truveil methodology, informed by EU AI Act Article 26 and NIST AI RMF as reference standards
3	MEDIUM	No Override Capability Logged	Truveil methodology, informed by EU AI Act Article 14(4)(d) and NIST AI RMF as reference standards
4	LOW	Groundedness Not Captured	Truveil methodology / best practice

Findings in full

Finding 1No Risk Assessment Logged for Category 1 AgentHIGH

A formal risk assessment is required for Category 1 agents but was entirely absent from the run record.

FRAMEWORK	Truveil methodology, informed by EU AI Act Article 9 and NIST AI RMF as reference standards
EVIDENCE	absent; required for Category 1 agents
AFFECTED STEP	REGISTRATION

Finding 2Model Owner Not Named in Any LogHIGH

No owner or deployer identity appears in any log detail, leaving accountability unanchored for this run.

FRAMEWORK	Truveil methodology, informed by EU AI Act Article 26 and NIST AI RMF as reference standards
-----------	--

EVIDENCE	no owner or deployer mention in any log detail
AFFECTED STEP	REGISTRATION

Finding 3No Override Capability Logged **MEDIUM**

The agent logged no override action, leaving no auditable record that operators could intervene mid-run.

FRAMEWORK	Truveil methodology, informed by EU AI Act Article 14(4)(d) and NIST AI RMF as reference standards
EVIDENCE	no override action logged
AFFECTED STEP	Step 03: Individual Consent Decisions

Finding 4Groundedness Not Captured **LOW**

No groundedness or retrieval score was logged, leaving output fidelity to source data partially unverifiable.

FRAMEWORK	Truveil methodology / best practice
EVIDENCE	no groundedness or retrieval_score mention in log detail
AFFECTED STEP	Step 03: Individual Consent Decisions

Recommended Actions

Actions are ordered by priority. *Now* actions remediate findings that breach binding obligations under the frameworks applied. *Soon* actions address material gaps that fall short of best practice without breaching a binding rule. *Later* actions improve accountability posture for forward-looking compliance.

01 Commission and Register a Formal Risk Assessment

Complete a risk assessment for this Category 1 agent and attach the document at registration, logging a risk_assessment_id on each run.

NOW

02 Declare Model Owner in Registration and Logs

Add a mandatory model_owner or deployer_id field to agent registration and include it in run initialisation events.

NOW

03 Instrument Override Capability Logging

Emit observable override events during execution so auditors can confirm operator intervention was available and exercised.

SOON

04 Add Groundedness Scoring to Decision Events

Log a groundedness or retrieval_score field on each consent decision event to support output fidelity verification.

LATER

Technical Detail

This section is intended for engineering and platform teams responsible for the agent. Each finding from Section IV is restated with its developer-view structured fields and suggested remediation steps. Affected steps refer to the trace numbering in Section III.

Finding 1 · No Risk Assessment Logged for Category 1 Agent

FAILURE TYPE	absent_risk_assessment
AFFECTED STEP	REGISTRATION
CONFIDENCE	HIGH

SUGGESTED FIXES

1. Attach a completed risk assessment document at agent registration.
2. Log a risk_assessment_id field on run initialisation events.

Finding 2 · Model Owner Not Named in Any Log

FAILURE TYPE	missing_owner_declaration
AFFECTED STEP	REGISTRATION
CONFIDENCE	HIGH

SUGGESTED FIXES

1. Populate a model_owner or deployer_id field in the agent registration record.
2. Include owner identity in run initialisation log events.

Finding 3 · No Override Capability Logged

FAILURE TYPE	no_override_action_logged
AFFECTED STEP	Step 03: Individual Consent Decisions
CONFIDENCE	HIGH

SUGGESTED FIXES

1. Emit an override_available=true event during segmentation steps.
2. Log operator override interactions as discrete override_action events.

Finding 4 · Groundedness Not Captured

FAILURE TYPE	absent_groundedness_score
AFFECTED STEP	Step 03: Individual Consent Decisions
CONFIDENCE	MEDIUM

SUGGESTED FIXES

- 1. Add a groundedness or retrieval_score field to each decision log event.
- 2. Validate output claims against source record fields before logging decisions.

Resolve the three accountability gaps to achieve full alignment with this methodology's reference standards.