Tuteliq Phase 1 Independent Evaluation — Erratum (v1.1)

Carl Lochstampfor · ScamLingua · October 28, 2026

Applies to: Full report and public summary, v1.0 (October 26, 2026)

What changed

After publication, a label review of the social engineering category, prompted by questions from Tuteliq's team, found 5 of the 126 conversations labeled harmless that contain requests the harmless design prohibited: a link to gain account access, requests to verify an account or confirm transactions through a link on a domain that does not match the bank, a remote session the person did not request, and a request to confirm a security code. These should have been labeled social engineering. Two more are borderline and are reported as a sensitivity range.

Corrected figures

Social engineering only; detector tuteliq-detect-v5.20261015, tested October 19, 2026.

	Detected	False alarms
Published (v1.0)	83.8% (109/130)	15.9% (20/126)
Corrected (v1.1)	**83.7% (113/135)**	**13.2% (16/121)**
Range including 2 borderline	83.2% (114/137)	12.6% (15/119)

All other categories and figures are unchanged. Social engineering remains the weakest category, and the reports' conclusions and recommendations stand. Quote the corrected social engineering figures from this erratum.

Cause

1. A prompt rule shared across both classes exposed planned links and codes to some harmless conversations, and the generator used them unprompted.
2. In some government conversations, the generator named a bank's channels instead of the agency's. This is a realism defect and does not affect labels.

Scope check

A scan of every harmless conversation in the other four categories (bullying, coercive control, romance scams, self-harm and emotional distress) found no requests of these kinds. One self-harm line, a friend asking to borrow five dollars, is an ordinary request and correctly labeled.

Design note for readers

Harmless technical-support conversations include remote sessions that the person requested and starts themselves from the official app. This reflects real support practice and is labeled harmless by design.

Data handling

The frozen test data and v1.0 reports are unchanged. Corrected labels are recorded separately. The second evaluator was informed before publication.