ScamLingua

PUBLIC

PUBLIC SUMMARY

Tuteliq text-safety detection: Phase 1 independent evaluation

Prepared for	Tuteliq and its partners
Date	October 26, 2026
Status	Final, version 1.0
System evaluated	Tuteliq API 3.4.0, detector `tuteliq-detect-v5.20261015`, tested October 19, 2026
Author	Carl Lochstampfor (ScamLingua)
Second evaluator	Dr. Ayan Roy, Christopher Newport University

Public summary, published with Tuteliq's permission. Quote figures only as set out in the claims register.

> On October 19, 2026, an independent evaluation tested Tuteliq's text-safety detector on 1,344 synthetic cases in five harm categories. It detected 84% to 100% of harmful conversations per category, with false alarms on 0.8% to 15.9% of harmless ones. Detection held up on synthetic test conversations written to hide or disguise the harm. Social engineering was the weakest category, with the lowest detection and the most false alarms. The conversations were synthetic and in English text only, and the results cover only these five text detections, not Tuteliq's service as a whole.

Recommended citation: Lochstampfor, C. (2026). *Tuteliq Phase 1 independent evaluation: Public summary.* Second evaluator: Dr. Ayan Roy, Christopher Newport University.

Naming Tuteliq next to results (wording agreed with Tuteliq): "Results for Tuteliq (detector version `tuteliq-detect-v5.20261015`, tested October 19, 2026) from an independent evaluation by Carl Lochstampfor on synthetic conversations. Tuteliq reviewed the report for factual accuracy only; the scoring and conclusions are the author's."

About this evaluation

This is a black-box evaluation: we tested how Tuteliq's detector behaves from the outside, without access to its internals. Tuteliq describes its service as a content moderation API built around the harm categories of the U.S. Kids Online Safety Act, for platforms where minors interact online.

Independence. Carl Lochstampfor designed, ran and scored the evaluation, with Dr. Ayan Roy (Christopher Newport University) as second evaluator. Tuteliq provided access to its evaluation harness and reviewed the report for factual accuracy only. It had no role in the test design, the scoring or the conclusions. "Independent" describes who designed and scored the evaluation; it is not an accreditation, certification or audit of Tuteliq. Tuteliq did not see any test conversation before the sets were frozen, and had no part in writing, choosing or changing them. The frozen sets were placed in its evaluation repository only so its harness could run them, and every file is fingerprinted, so any change would be detectable. Tuteliq has not paid for this evaluation or for any of its content.

Funding. The evaluation took more than 175 hours from September 25, 2026, and was carried out unpaid. It builds on earlier unpaid work auditing and refining the COVA and COVA-X datasets. No payment has been received or agreed; if that changes, it will be disclosed.

Use of AI. Claude (Anthropic), an AI assistant, helped draft this summary, cross-check every figure against the scoring output, and write the generation and scoring software under the author's direction. The test conversations were generated by a separate open-weight language model run locally. All design decisions, labels and conclusions are the author's. The author's Claude account has the "Help improve our AI models" privacy setting turned off, so these sessions are not used to train Anthropic's models.

Data handling and retention. Only synthetic text was used and sent to Tuteliq; no real person's data was handled at any stage. Tuteliq's documentation describes its API as not storing submitted text between requests and not using it for training. Tuteliq's detailed responses are its confidential information: they were shared only with the second evaluator and the AI assistant (with Tuteliq's written consent), used only for this evaluation, and kept privately by the author only for later verification, with Tuteliq's written permission. Tuteliq also holds an archive copy. This summary contains aggregate figures only.

The full report, with the complete method, results and findings, is published alongside this summary with Tuteliq's permission.

What was tested

Five of Tuteliq's text detections were tested on 1,294 multi-turn conversations, about half harmful and half harmless, plus 50 single messages.

Category	Conversations	Harmful	Harmless
Romance scam	260	130	130
Social engineering	256	130	126
Coercive control	260	130	130
Bullying	259	129	130
Self-harm and emotional distress	259	129	130

These harms are costly. In 2025 the FBI's Internet Crime Complaint Center received over one million complaints with $20.9 billion in reported losses; people aged 60 and over lost $7.7 billion. Tech-support, romance and government-impersonation scams were among the costliest types (FBI IC3 2025 Annual Report, pp. 6–8, 44–46).

Grooming, money-mule recruitment and Tuteliq's other detections were not evaluated, and no result here covers them. Images, audio and video were out of scope, and no child sexual abuse material was used under any circumstances.

How it was tested

All test conversations are synthetic text-message exchanges, designed by us and generated on our own machines. No real person's messages were used, quoted or imitated, and only synthetic text was sent to Tuteliq. Each harmful pattern is grounded in published research or official guidance.

Each category set has five parts:

- **Baseline harmful:** the harm stated plainly.
- **Evasion probes:** the same harm, hidden or softened.
- **Adversarial probes:** the same harm, disguised as legitimate.
- **Baseline harmless:** ordinary conversations with the same setup and length.
- **Hard negatives:** harmless conversations that look suspicious on the surface.

Before use, each set passed balance, shortcut and diversity checks and was read by hand, then frozen with every file fingerprinted. Results were computed only on frozen sets.

The run used Tuteliq's own evaluation harness, the route it uses for every external evaluator. Each case was sent once, and scoring was done offline on the saved responses. A harmful conversation counts as detected when Tuteliq recommends an action above "monitor" (flag for review or stronger). Every rate carries a 95% Wilson score interval, because the sets are small.

Results

All 1,344 cases returned a usable response; none failed. Detection was highest for romance scams, coercive control and self-harm, and lowest for social engineering.

Table 1. Harmful conversations detected (recall) and harmless conversations flagged (false alarms), by category. Source: Tuteliq run of October 19, 2026 (detector `tuteliq-detect-v5.20261015`); our scoring.

Category	Recall	False alarms
Romance scam	100.0% (130/130) [97.1–100.0]	6.9% (9/130) [3.7–12.6]
Coercive control	99.2% (129/130) [95.8–99.9]	0.8% (1/130) [0.1–4.2]
Self-harm and emotional distress	96.1% (124/129) [91.2–98.3]	0.8% (1/130) [0.1–4.2]
Bullying	84.5% (109/129) [77.3–89.7]	4.6% (6/130) [2.1–9.7]
Social engineering	83.8% (109/130) [76.6–89.2]	15.9% (20/126) [10.5–23.2]

Figures in square brackets are 95% intervals. Precision is high at our mix of about half harmful (84.5% to 99.2% by category), but it would be much lower where harm is rare, as on real platforms; recall and false alarms are the figures to compare.

On 50 single messages sent without the rest of the conversation, 20 of 25 harmful messages were detected and 1 of 25 harmless messages was flagged. With five per category, these figures show direction only.

What the results show

Strengths

- **Romance scams, coercive control and self-harm** were detected in 96% to 100% of harmful conversations, with false alarms at or below 6.9%.
- **Detection held up on synthetic evasion and adversarial probes.** These probes hide or disguise the harm in ways we designed, and recall on them stayed close to baseline in every category. Evasion in real conversations may differ. A system relying mainly on keywords would be expected to drop sharply on these probes; that did not happen.
- **Whole conversations helped** in coercive control and self-harm, where the conversation was detected more often than its key message alone.

Weaknesses

- **Social engineering** had the lowest recall (83.8%) and the highest false-alarm rate (15.9%). Scams asking for identity data under a government pretext were mostly missed (7 of 28 detected), and harmless conversations of the same kind were flagged as often.
- **Bullying** recall was 84.5%; bullying that builds gradually from teasing was detected in 65.2% of cases (30 of 46), against over 90% for open exclusion or exposure.
- **Romance false alarms** (6.9%) came mostly from harmless conversations about money between people who trust each other.

The full report gives the figures behind each finding and Tuteliq's areas for improvement.

Validation

The results are complete, from one unchanged detector version, and checked against the scoring output.

- **Complete run.** All 1,344 cases returned a readable response with a recommended action.
- **Unchanged detector.** Tuteliq's health check reported the same detector version, build and configuration at the start and end of the run, with no restart in between.
- **Checked figures.** Every reported figure was checked by script against the scoring output, with intervals recalculated independently; no mismatches were found.
- **Second evaluator.** Dr. Ayan Roy labeled 50 conversations blind. He agreed with our labels on 41 (82%; Cohen's kappa 0.64). Most disagreements were harmless conversations built to look suspicious, which he read as harmful. On all nine, Tuteliq's result matched our label, so the disagreements do not inflate Tuteliq's error counts.

Limits of this evaluation

The results describe this detector version on these synthetic conversations, not performance on real platform traffic.

- **Synthetic data.** All conversations come from one generator model. Real conversations are messier and more varied. A recent preprint found that synthetic bullying dialogue misjudges how intense and escalating real bullying is (Kazemi et al. 2026). A benchmark that does not represent real use is a known source of evaluation bias (Suresh & Guttag 2021).
- **Compressed conversations.** Our longest conversations run 26 turns. Real romance scams can run to over a thousand lines across weeks (Lee & Ang 2026).
- **Small subgroups.** Each category has about 130 harmful and 130 harmless conversations; subgroups within a category are smaller, so their intervals are wide.
- **Ages not stated.** Our conversations do not state the speakers' ages, so the results describe detection in general, not for children specifically.
- **Coverage.** English text only. Grooming, money-mule recruitment, Tuteliq's other harm categories, images, audio and video were not tested and may warrant further evaluation.
- **No compliance assessment.** The evaluation does not assess or validate regulatory compliance by Tuteliq or any platform.

Claims register

Any figure from this evaluation may be quoted only with its scope. Every quote must carry the standard qualifier and, for a category's detection rate, that category's false-alarm rate.

Standard qualifier: "In an independent evaluation by Carl Lochstampfor on synthetic, English-language text conversations (detector `tuteliq-detect-v5.20261015`, tested October 19, 2026). Not an accreditation, certification or audit."

Permitted claim	Must be quoted with
Romance scam: 100% of harmful synthetic conversations detected (130/130)	False alarms 6.9% (9/130)
Coercive control: 99.2% detected (129/130)	False alarms 0.8% (1/130)
Self-harm and emotional distress: 96.1% detected (124/129)	False alarms 0.8% (1/130)
Bullying: 84.5% detected (109/129)	False alarms 4.6% (6/130); gradual bullying from teasing 65.2% (30/46)
Social engineering: 83.8% detected (109/130)	False alarms 15.9% (20/126); government-pretext identity scams 7 of 28 detected
Across the five text detections: 84% to 100% detected, false alarms 0.8% to 15.9%	Social engineering named as the weakest category
Detection held up on synthetic evasion and adversarial probes	"Probes designed by the evaluator; evasion in real conversations may differ"

Precision may be quoted only with the test mix (about half harmful) and the statement that precision is much lower where harm is rare, as on real platforms.

Not permitted:

- any figure without its category, its false-alarm rate or the standard qualifier;
- a single overall detection or accuracy rate combining the categories;
- any claim about real platform traffic, other languages, images, audio or video, or Tuteliq's other detections;
- describing the evaluation as a certification, accreditation, audit or compliance assessment, or implying compliance with the EU KIDS Act, the UK Online Safety Act or KOSA;
- quoting the strengths in an overall summary of Tuteliq's performance without the weaknesses.

Where to find more detail

The full report, published alongside this summary with Tuteliq's permission, gives the complete method, results by type of scam or harm and by kind of request, and Tuteliq's areas for improvement. Case-level data and Tuteliq's individual responses are not published in either version.

The datasets are not public. For dataset access, contact requests@scamlingua.org.

About the Author and ScamLingua

Carl Lochstampfor is an undergraduate student pursuing a Bachelor of Science in Cybersecurity at Old Dominion University. ScamLingua is his research project on detecting scams and online harm in conversations. Website: scamlingua.org.

Contact: contact@scamlingua.org

References

Federal Bureau of Investigation. (2026). *Internet crime report 2025.* Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports

Kazemi, A., Qadeer, H., Asci, S., Wagner, J., & Davis, B. (2026). *Do social patterns hold in synthetic data? Analyzing cyberbullying dynamics in LLM-generated and authentic dialogues* (arXiv:2609.17549) [Preprint]. arXiv. https://arxiv.org/abs/2609.17549

Lee, K.-F., & Ang, C.-S. (2026). Unveiling scammer-victim interactions in online romance scams: Insights from conversation analysis and discursive manipulation theory. *Deviant Behavior.* Advance online publication. https://doi.org/10.1080/01639625.2026.2651287

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21).* Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953