ScamLingua

PUBLIC

INDEPENDENT EVALUATION REPORT

Tuteliq text-safety detection: Phase 1

Prepared for	Tuteliq
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

Published with Tuteliq's permission. Case-level data and Tuteliq's individual responses are not included. Quote figures only as set out in the claims register (Appendix E).

Document control

Version	Date	Change
0.9	October 22, 2026	Draft for Tuteliq's factual review; second evaluator copied
0.95	October 22, 2026	Revised for Tuteliq's first comments: scope and independence qualified, regulatory section tightened, weaknesses kept
0.96	October 23, 2026	Tuteliq's second comments: evasion wording scoped to the synthetic probes, data handling and retention table, claims register (Appendix E)
1.0	October 26, 2026	Final. Approved by Tuteliq for publication of both versions; dataset-use terms and results retention agreed in writing

Distribution	Role
Public	Published with Tuteliq's permission
Tuteliq	Evaluated party; factual review
Dr. Ayan Roy, Christopher Newport University	Second evaluator
Carl Lochstampfor	Author

This full report and the shorter public summary are both published with Tuteliq's permission. Case-level data and Tuteliq's individual responses are not published in either version.

Executive summary

On October 19, 2026 we tested Tuteliq's detector (`tuteliq-detect-v5.20261015`) on 1,344 synthetic cases across five categories: 1,294 multi-turn conversations, about half harmful and half harmless, plus 50 single messages. Every case returned a usable response. The conversations are synthetic and in English text only, and the results cover only these five text detections, not Tuteliq's service as a whole.

1. **Strong on romance scams, coercive control and self-harm.** Recall was 100%, 99.2% and 96.1%, with false alarms of 6.9%, 0.8% and 0.8%.
2. **Detection held up on our synthetic evasion and adversarial probes.** Recall on these probes, written by us to hide or disguise the harm, stayed close to baseline in every category. Evasion in real conversations may differ.
3. **Government-identity social engineering is the clearest gap.** Only 7 of 28 such scams were detected (25%), and harmless conversations on that route were flagged as often. Social engineering also had the highest false-alarm rate overall (15.9%).
4. **Smaller gaps.** Bullying that escalates gradually from teasing (65.2% detected) or mocks someone's abilities (45.5%), and genuine loan requests between people who trust each other flagged as romance scams (8 of 15).
5. **Act on flag for review and above, not on monitor.** Monitor is applied to many harmless conversations (25% to 57% in three categories); the levels above it are what separate harmful from harmless.

A second evaluator agreed with our labels on 41 of 50 conversations (82%; Cohen's kappa 0.64). The results describe synthetic conversations, not live platform traffic (section 5). Section 8 gives five recommendations.

> **How to read this report.** *Recall* is the share of harmful conversations detected; a *false alarm* is a harmless conversation flagged. Figures in square brackets are 95% intervals: wide brackets mean small numbers. Unless stated otherwise, "detected" means a recommended action above monitor. A glossary is at the end.

1. Purpose

This report is an independent black-box evaluation of Tuteliq's behavioral detection of online harms in text conversations. It supports Tuteliq's EIC application (deadline November 5, 2026).

Independence rests on two things: the test conversations were designed and generated by us, not by Tuteliq, and the scoring and conclusions are ours. Tuteliq reviews this draft for factual accuracy about its own system only. "Independent" describes who designed and scored the evaluation; it is not an accreditation, certification or audit of Tuteliq, and the author is not an accredited testing or certification body.

The focus is multi-turn detection: whether the system recognizes a harmful pattern as it builds across a conversation. A small set of single messages is included as a baseline. We report failures as plainly as successes.

Why these harms. In 2025 the FBI's Internet Crime Complaint Center received 1,008,597 complaints with $20.877 billion in reported losses; people aged 60 and over filed 201,266 of them and lost $7.748 billion (FBI IC3 2025 Annual Report, pp. 6, 44). The scam types we test are among the costliest:

IC3 crime type	Complaints, all ages	Losses, all ages	Losses, aged 60+	Closest category here
Tech/Customer Support	47,794	$2,134,675,818	$1,040,730,043	Social engineering: tech support
Confidence/Romance	23,159	$929,287,469	$584,032,745	Romance scam
Government Impersonation	32,424	$797,943,193	$413,206,251	Social engineering: government identity

Source: IC3 2025 Annual Report, pp. 7-8 and 45-46. IC3's categories are not ours: its Confidence/Romance also covers family and friendship schemes, its Government Impersonation requires a money request, and investment-style romance scams likely fall under its Investment category. These are complaints filed, not how often the harms occur. The report has no category for bullying or coercive control; it names cyberbullying once, among the drivers of 13,168 complaints from people aged 17 or younger (p. 34).

2. Methodology

All test conversations are synthetic text-message exchanges between two people, generated by an open-weight model (Gemma 3 12B) running locally on our own workstations. Carl Lochstampfor designed and ran the generation, with Claude (Anthropic) assisting on design, code and analysis. No real person's messages are used, quoted or imitated.

How a conversation is built

1. **Plan.** A seeded planner fixes each conversation's class, route, stages, length and turn of the key request before any text is written. Plans are reproducible from the seed.
2. **Templates.** Layered instructions tell each speaker what to do at each stage, never how to word it: shared texting rules, scenario briefs, stage directives, probe operators and personas.
3. **Generation.** Each speaker is a separate model call that sees only its own role and the conversation so far.
4. **Checks.** Automatic output checks reject a message that breaks a rule (for example a real-looking phone number or a stray link); the message is regenerated.

What each category set contains

Part	What it is	Why
Baseline positives	Harmful conversations written plainly	Can the system find the pattern at all?
Evasion probes	The same plan, written to hide the harm (euphemism, slow escalation, fragmentation)	Does detection survive realistic hiding?
Adversarial probes	The same plan, written to look legitimate or anti-scam	Does detection survive deliberate disguise?
Baseline negatives	Harmless conversations with the same setup, length and speaker pattern	False alarms on ordinary talk
Hard negatives	Harmless conversations that look most like the harm (real fraud alerts, real relatives on new numbers)	False alarms where it matters most

Each probe is a twin of one baseline plan, so a drop in detection from baseline to probe is attributable to the wording, not the scenario.

Quality control before a set is used

- **Balance checks.** Message length, turn count, who speaks first and who speaks most are compared across harmful and harmless conversations, so a detector cannot score well on length or shape alone.
- **Shortcut checks.** Phrases and topics that appear in only one class are listed per speaker, so we can see what a detector could latch onto besides the harm itself.
- **Diversity checks.** Openings, names and near-duplicate conversations are counted.
- **Label screens and human reading.** Harmless conversations are screened for any harmful request; flagged items and a sample of each class are read by Carl.
- **Freeze.** A passed set is frozen: every file is hashed into a manifest, and copies are verified against that manifest on two separate storage locations. Results are only ever computed on frozen sets.

Where a check fails and we accept the set anyway, the failure is disclosed in section 5 with its figure.

Each harmful pattern and kind of request in our sets is grounded in published research or official guidance. Appendix D lists the sources for each category.

3. Available detections

Tuteliq's public documentation lists 15 text detections in four groups; five of them are tested here. Image, voice, video and document analysis, and age and identity verification, are also offered and are outside this text-only evaluation.

Group	Detection	What it looks for (Tuteliq's description)	Tested here
Safety	Bullying	Verbal abuse, social exclusion, threats and cyberbullying patterns	Yes
Safety	KOSA harm signals (unsafe)	All KOSA harm categories, including self-harm and suicidal ideation and eating disorders	Yes (self-harm set)
Safety	Grooming	Multi-stage grooming tactics across a sequence of messages	No
Safety	Synthetic content	Signs of AI-generated text and synthetic identities	No
Fraud	Romance scam	Romance scam patterns, including love bombing and isolation	Yes
Fraud	Social engineering	Manipulation tactics, including authority impersonation	Yes
Fraud	App fraud	Payment coercion, account takeover and safe-account scams	No
Fraud	Money mule recruitment	Easy-money promises and requests for account access	No
Extended safety	Coercive control	Monitoring, isolation and related control patterns	Yes
Extended safety	Distress signals	Linguistic patterns historically associated with distress	No
Extended safety	Gambling harm	Chasing losses, desperation and borrowing	No
Extended safety	Radicalization	Ideological seeding and related indicators	No
Extended safety	Technology-facilitated gender-based violence	Image-based abuse and related patterns	No
Extended safety	Exploitation of vulnerable individuals	Age targeting and emotional exploitation	No
Analysis	Emotions	Dominant emotions and sentiment trajectory across a conversation	No

Every detection returns a risk score from 0 to 1, a severity level, the categories found, a short rationale and a recommended action: none, monitor, flag_for_review, block or immediate_intervention. Sources: Tuteliq documentation, documentation index, How it works.

4. Detections tested

Five categories are evaluated, and all five datasets are frozen: romance scam, social engineering, coercive control, bullying, and self-harm and emotional distress. Categories without a dataset are named below as not evaluated, with the reason.

Category	Status	Reason or note
Romance scam	Evaluated (260 conversations)	Priority category
Social engineering	Evaluated (256 conversations)	Priority category
Coercive control	Evaluated (260 conversations)	Priority category
Bullying	Evaluated (259 conversations)	Priority category
Self-harm / emotional distress	Evaluated (259 conversations)	Priority category
Grooming	Not evaluated	Deferred to a later round; it needs its own design, approvals and safeguards
Money-mule recruitment	Not evaluated	Dataset not ready in time for this round
Radicalization / extremism	Not evaluated	Not in scope this round
Exploitation of a vulnerable person, technology-facilitated gender-based violence, gambling harm, and the unsafe detection's other harm categories (for example eating disorders)	Not evaluated	No dataset this round
Images, video, child sexual abuse material	Never tested	Text only; no CSAM testing under any circumstances

Not-evaluated categories are outside this report's conclusions; no result here should be read as covering them.

5. Datasets

Five sets are frozen, 1,294 conversations in all: romance scam 260, social engineering 256, coercive control 260, bullying 259 and self-harm 259. Each set is about half harmful and half harmless, and every set reports its known weaknesses below with their figures.

The disclosures list cues a detector could use besides the harm itself. Because Tuteliq was not trained on this data, such cues matter less here than they would for training; we report them so readers can judge.

Set	Harmful	Harmless	Turns per conversation	Generated	Frozen	Manifest hash
Romance scam	130	130	20–26	Oct 15–16	Oct 16	`9df705e4…77440d`
Social engineering	130	126	12–15	Oct 10–12	Oct 12	`e84cdcd3…1683d5`
Coercive control	130	130	16–19	Oct 12–13	Oct 13	`110913d7…62813d`
Bullying	129	130	14–17	Oct 13–14	Oct 14	`fc765275…856c28`
Self-harm and emotional distress	129	130	14–17	Oct 14	Oct 15	`07ee5304…81dc4e`

All dates are 2026. Every set was generated with Gemma 3 12B and frozen with every file hashed into its manifest; copies are verified on two separate storage locations.

Romance scam

Part	Conversations
Baseline scams (classic romance 47, investment "pig butchering" 23)	70
Evasion probes	32
Adversarial probes	28
Baseline harmless	55
Hard negatives	75
Total	260

Generated with Gemma 3 12B on October 15–16, 2026, and frozen on October 16. Manifest hash `9df705e4…77440d`; copies verified on two separate storage locations. Carl Lochstampfor read 37 conversations across both classes before the freeze; all read as their labels.

Why this set replaces the first romance set. The first set (259 conversations, frozen October 8) passed its checks, but topics alone nearly separated its classes: a simple classifier on which topics each speaker raised reached an AUC of 0.932. Harmless conversations also never mentioned money. This second set fixes both, and the first set is kept but not used.

- Harmless conversations follow everyday storylines between partners, drawn from research on how couples maintain relationships (Stafford & Canary 1991), and include ordinary money talk such as splitting a cost or paying someone back.
- Both classes get the same pool of hobbies and topics, so topics no longer mark the class: the topic classifier's AUC fell to 0.671.

- Conversations run 20 to 26 turns, against 12 to 19 in the other sets. Real romance scams are long: in the two successful scams among 30 real scammer–victim chats, the conversation ran 742 to 1,274 lines (Lee & Ang 2026). The extra turns give the scam room to build rapport before the request, and give harmless storylines room to develop.

Disclosures. We accepted this set with the following known issues:

- **Harmless contacts ask more questions.** The gap is 0.110 question marks per message, just above our limit of 0.10; the share of messages with a question differs by 6.2 points, within its limit. The extra questions come from the setup and resolution steps of the harmless storylines.
- **Persuasion phrases on the scam side.** Forms of "it would mean a lot" and the phrase "important to me" each appear in about 24% of scams and almost never in harmless conversations. These are realistic persuasion language, so we treat them as part of the signal rather than suppress them.
- **Links on the scam side.** Scam contacts share a link in about 30% of scams ("the link https", 30.0%; "here's the link", 25.4%), and harmless contacts never do. Sending a link is realistic for a scam, but its presence alone marks the class.
- **Money talk is close but not equal.** Money comes up in 45% of scam contacts and 37% of harmless contacts. Every harmless money request states its amount; some disguised scam requests leave it out.
- **Some scam reasons are implausible**, for example a large gift-card request for a small household purchase. Real scammers' reasons are usually more believable.
- **Harmless-side habits.** In 13% of harmless conversations the partner signs off with a stock closing, and in 14 harmless conversations (10.8%) the contact asks for a phone number.
- **Message length is balanced.** The standardized difference in message length is 0.101 overall and 0.058 for the contact, within our limit of 0.2 (0.200 for the contact on baseline items only, at the limit).

Social engineering

Part	Planned	Delivered
Baseline scams (account security 23, tech support 17, government identity check 16, a relative on a new number 14)	70	70
Evasion probes	32	32
Adversarial probes	28	28
Baseline harmless	55	54
Hard negatives	75	72
Total	**260**	**256**

Generated with Gemma 3 12B on October 10, 2026; 18 harmless conversations were regenerated on October 12 (see below). Manifest hash `e84cdcd3…1683d5`; copies verified on two separate storage locations.

Disclosures. We accepted this set with the following known issues:

- **Four plans failed every attempt** (one baseline harmless, three hard negatives; six tries each), so the harmless class has 126 conversations. Their setups pulled the model toward money talk, which our checks reject in harmless conversations.
- **Five harmless conversations were replaced after a practitioner read.** In three, a bank or tech-support contact asked for a full account number; in two, the contact kept asking about the address after it was answered. A careless representative might do this, but a careful person would rightly be suspicious, so the harmless label would be unclear. The originals are kept, with their hashes in the set's record.
- **Message length is not balanced.** The standardized difference in mean message length is 0.349 overall and 0.339 for the contact (0.303 on baseline items only), above our balance limit. Turn counts are balanced (0.006).
- **Shape alone does not separate the classes well.** A classifier on length and turn-taking features reaches an AUC of 0.618 (0.567 on baseline items), within our limit. Harmless contacts ask more questions; that single feature reaches an AUC of 0.790.
- **Scam-only phrasing** is mostly the request itself: "a verification code" (34.6% of scams), "your social security number" and "date of birth" (20.8% each). This is the signal a detector should find. The urgency phrase "it's vital we" (30.8%) is a generator habit. Some scam messages also repeat wording from our own instructions word for word: the request for a social security number and date of birth (about 18–20% of scams) and "here as soon as" (17.7%). The request is the harm itself, but its identical wording is a shortcut a detector could use. This set was frozen before we added the check that catches such copying.
- **Harmless-only phrasing** comes from the harmless requests as designed: yes/no confirmation ("could you confirm if", 31.0% of harmless conversations), reassurance that the contact never asks for codes or passwords (23.0%), and scheduling a repair ("to schedule a", 22.2%).
- **Realism.** Messages from banks and services read more like phone conversations than the short, formal texts such organizations send. Harmless contacts never quote any digits of a customer's account or identity numbers; our checks forbid it to keep harmless conversations free of personal details. Real banks may refer to a card or account by its last four digits, but a genuine representative never volunteers digits of a customer's Social Security number: the customer gives them, and the representative may only repeat them back to confirm.
- **Diversity.** No near-duplicate conversations; openings and names passed the diversity check.

Coercive control

Part	Planned	Delivered
Baseline control (whereabouts and proof 23, after separation 23, phone access 16, isolation rules 16)	78	78
Evasion probes	26	26
Adversarial probes	26	26
Baseline harmless	39	39
Hard negatives	91	91
Total	260	260

Generated on October 12–13, 2026, and frozen on October 13. Manifest hash `110913d7…62813d`. Two plans failed the first run and succeeded on a second attempt; none was excluded.

Disclosures. We accepted this set with the following known issues:

- **Contact message length is not balanced.** The standardized difference is 0.272 for the contact, above our limit of 0.2 (0.419 on baseline items only). Overall length (0.193) and turn counts (0.000) are balanced.
- **Commas mark the controlling side.** Comma rate alone separates the classes at an AUC of 0.80 or more, because controlling messages are often conditional ("if you do this, then that"). The shape classifier as a whole reaches 0.693, within our limit.
- **Phrases in one class only are mostly the harm itself.** Controlling contacts say "I will have to" (31.5% of harmful conversations), "who you're with" and "you didn't mention" (20.8% each) and "from now on" (20.0%). These are the demands and consequences a detector should find.
- **Harmless contacts are warm.** "Kind of you" (23.1%), "that sounds like a" (16.9%) and "I saw a photo" (14.6%) appear only in harmless conversations, and harmless partners never push back, since there is nothing to push back on.
- **Partners appease on the harmful side.** "Try to be more" (16.2%) and "what you're saying" (14.6%) appear only in harmful conversations. This is realistic, but it is a cue on the other speaker's side.

Bullying

Part	Planned	Delivered
Baseline bullying (exclusion 26, teasing that turns 26, exposure 26)	78	78
Evasion probes	26	25
Adversarial probes	26	26
Baseline harmless	39	39
Hard negatives	91	91
Total	260	259

Generated on October 13–14, 2026, and frozen on October 14. Manifest hash `fc765275…856c28`. One evasion plan failed all six attempts on a style check, so the harmful class has 129 conversations; one harmless plan succeeded on a second attempt. None was excluded.

The set covers one-to-one chats between teenagers. Group chats, adults, sexual harassment, identity-based slurs and threats of violence are out of scope.

Disclosures. This set passes every balance and shape check: message length 0.025 overall and 0.160 for the contact, turns 0.003, and a shape AUC of 0.605 with no single feature above 0.7. Known issues:

- **A repeated story.** Harmful contacts bring up "that time you" (20.9% of harmful conversations), and one image recurs as a generator habit (13.2%). Real bullies do return to the same incident, but not in the same words across different people.
- **Harmless-side habits.** Harmless contacts ask for a picture ("a picture of", 26.2%), and apology words appear in 30.0% of harmless contacts against 1.6% of harmful ones. Apologies are by design: in harmless teasing, the joke stops when the other person objects.
- **Exclusion words mark the harm.** They appear in 22.5% of harmful contacts and 1.5% of harmless ones. This is the signal a detector should find.
- **Some bullying is indirect.** In 36 of 129 harmful conversations, the contact uses none of our listed mockery, exposure or exclusion words after the request turn. These are subtler cases and are kept.

Self-harm and emotional distress

Part	Planned	Delivered
Baseline disclosures (suicidal thoughts 26, self-injury 26, warning signs 26)	78	78
Evasion probes	26	26
Adversarial probes	26	25
Baseline harmless	39	39
Hard negatives	91	91
Total	260	259

Generated on October 14, 2026, and frozen on October 15. Manifest hash `07ee5304…81dc4e`. One adversarial plan failed all six attempts because the model kept naming a method, which our checks reject, so the harmful class has 129 conversations. One conversation was kept after reading although a check flagged it: the flagged word was part of a musical phrase. A first generation run used the wrong model by mistake; it was set aside entirely and none of it is in this set.

Safety rules. Disclosures are clear but never detailed: no conversation names a method or gives instructions, and a check rejects any message that does. There are no farewell messages. The warning-signs route (feeling like a burden, hopelessness) has no acute marker by design, which is why Tuteliq's scoring rule for it differs (section 7).

Disclosures. We accepted this set with the following known issues:

- **Contact message length is slightly unbalanced.** The standardized difference is 0.204 for the contact, just above our limit of 0.2 (0.326 on baseline items only). Overall length (0.187) and turns (−0.010) are balanced.
- **Commas.** Comma rate alone separates the classes at an AUC of 0.734; the shape classifier as a whole reaches 0.608, within our limit.
- **A generator habit on the harmful side.** "It means a lot that" appears in 37.2% of harmful contacts and no harmless one. "Hard to explain" (23.3%) is ordinary disclosure language.
- **Harmless-side habits.** Harmless contacts ask for a picture ("send me a picture of", 35.4%) or to borrow something (30.0%); these come from the harmless request steps.
- **The listener's side.** Supportive replies such as "here if you want to" (23.3%) appear only in harmful conversations. This is realistic, but it is a cue on the other speaker's side.
- **Warning-sign words mark the harm.** Self-injury words appear in 27.9% of harmful contacts, burden in 21.7% and hopelessness in 11.6%, and none of them in harmless ones. This is the signal a detector should find.
- **Some disclosures are indirect.** In 27 of 129 harmful conversations, the contact uses none of our listed warning-sign words after the request turn. These are subtler cases and are kept.

The figures for coercive control, bullying and self-harm were computed from the frozen sets with the same checks used before freezing.

Label check by a second evaluator

Dr. Ayan Roy (Christopher Newport University) checks our labels blind. He reads 50 conversations, 10 from each set (5 harmful and 5 harmless, drawn at random with a fixed seed), shuffled together under neutral IDs. Each shows its category, so he can judge the harm in question, but no class, route or label. For each, he records whether the conversation shows the harm named for its category, how sure he is, and any note. We compare his answers with our design labels as percent agreement and Cohen's kappa, overall and by category. Every disagreement is read and discussed, not dropped.

Result. Dr. Roy agreed with our labels on 41 of 50 conversations (82%; Cohen's kappa 0.64). Agreement was 9 of 10 for bullying, coercive control and self-harm (kappa 0.80 each), 8 of 10 for romance (0.60) and 6 of 10 for social engineering (0.20). Eight of the nine disagreements were harmless conversations he judged harmful, six of them hard negatives built to look suspicious. The ninth was a self-harm evasion conversation he read as hard times without self-harm. In social engineering he judged four of the five harmless conversations to be scams, so our harmless contacts from banks and services can read as suspicious even to an expert reader.

On all nine disputed conversations, Tuteliq's main-measure result matched our label rather than his: no false alarm on the eight harmless ones, and the self-harm conversation was flagged for review. The disagreements therefore do not inflate Tuteliq's error counts, but they show that the line between harmful and harmless is genuinely hard to draw in places, most of all in social engineering. Appendix C gives the full breakdown.

Realism notes. Dr. Roy's comments also point to weaknesses of generated dialogue that we disclose: tone that turns hostile abruptly where real escalation is usually gradual, repeated openings, and sudden changes of topic within a single message. He also noted replies that do not follow up on the previous message. In text conversations this can be natural, since a thread may continue in person, but it can make some exchanges read as disjointed. These features may make some harm easier or harder to spot than in real conversations.

Limitations common to all sets

Coverage limitations. Because this evaluation was time-boxed, some areas were not covered and may warrant further testing: money-mule recruitment and grooming (section 4) and Tuteliq's other harm categories; languages other than English; images, audio and video; conversations longer than 26 turns; real platform traffic; and the optional pass that tells the detector the type of conversation.

- All conversations are synthetic and come from one generator model; real conversations are messier and more varied. A recent preprint comparing synthetic and real bullying dialogues found that synthetic text keeps the direction of the power gap between bully and victim but misjudges its size, and misjudges how intense, humorous and escalating real bullying is (Kazemi et al. 2026, preprint). In the terms of Suresh and Guttag's framework of harms in machine learning, this is evaluation bias: a benchmark that does not represent the population a system will meet in use (Suresh & Guttag 2021, p. 6). Our results describe performance on these conversations, not on real traffic.
- Our conversations are compressed. In the two successful romance scams among 30 real scammer–victim chats (27 from police files), the conversation ran 742 to 1,274 lines over 19 to 65 days (Lee & Ang 2026). Our romance conversations run 20 to 26 turns, which compresses that by a factor of about 30 or more; the other sets are shorter still. We test whether the pattern is recognized, not a full-length scam.
- Few studies analyze real written scam conversations turn by turn. Two recent studies support our stage orders: Ibañez (2024), on 10 failed scam chats posted by their targets, and Li et al. (2026), on 49 Chinese-language investment and romance scam chats that Telegram groups linked to scam operations posted as examples of success. Neither is a random sample, and some of our stage orders are still filled by design.
- English only.

6. System description and API calling

The run uses Tuteliq's own evaluation harness, the route Tuteliq uses for every external evaluator. Our frozen conversations go into Tuteliq's evaluation repository as eleven dataset folders. Carl Lochstampfor starts the run, Tuteliq approves the live job, and the harness saves the full response for every case, so each result can be traced to the exact conversation that produced it.

What Tuteliq is. Tuteliq describes itself as a content moderation API built around the harm taxonomy of the U.S. Kids Online Safety Act (KOSA), for platforms where minors interact online. Submitted text is normalized and its language detected. Specialized classifiers for each harm category then run, and a context engine weighs intent, relationship dynamics, platform norms and attempts to split harmful content across messages. Severity is calibrated to the user's age group; when no age group is sent, as in this evaluation, Tuteliq reads any sign of age from the conversation itself and otherwise does not treat the speakers as children (Tuteliq, in writing, October 17, 2026). Conversation-aware detections accept earlier messages as context. The documentation describes the service as stateless: submitted text is not stored between requests and is not used for training. The system version evaluated is recorded from Tuteliq's health check at the start and end of the run (How it works).

How we call it.

1. Each frozen conversation is sent once to the detection for its category. The run took place on Monday, October 19, 2026, as GitHub Actions run #42 (ID `39241363336`), started at about 07:00 ET from our branch (`cova-tp1-datasets`) in Tuteliq's repository, as Tuteliq asked; the pull request does not need to be merged. Tuteliq ran the same folders once on October 17 to check the setup (run #40); it is not used for results.
2. Each conversation is sent as one block of text, one line per message, in order. Speakers appear as "unknown", the harness's label when no speaker role is given, so the detector is not told which speaker is which. Bullying conversations carry the two speakers' invented first names instead.
3. No age group is sent, as Tuteliq asked.
4. Incident recording is switched off for every call, as Tuteliq asked for evaluation traffic.
5. The API accepts up to 10,000 characters per message. Our longest conversation block is 7,045 characters, so no case is near the limit.
6. A call that fails is retried. A call that still fails, or a response that cannot be read or has no recommended action, is logged as a failed call and reported separately; it is never counted as a miss.
7. Calls run one at a time, never in parallel.
8. Tuteliq's health check reported the same detector version (`tuteliq-detect-v5.20261015`), build (`71ff48a`) and configuration (`b45a9e613f13`) at the start of the run (06:53 ET) and at the end (12:48 ET), and its uptime shows the service did not restart in between. The build had changed since the check on October 17; the detector version and configuration had not. The run used commit `34419d8` of our branch.
9. Scoring is done offline on the saved responses, so no case is ever sent twice to get a better answer.

10. The conversation type was not sent. The optional second pass, which would have sent it as a direct message, was not run.
11. The single-message baseline is one more folder: 50 single messages, 5 harmful and 5 harmless per category, each cut by script from a frozen conversation and sent alone as plain text. The cut is the turn where the harmful request first appears, or the matching request in a harmless conversation, as fixed by the plan before generation. It is not chosen by hand, so some cut messages are milder than the conversation's peak.

Category	Detection (under /api/v1)	What is sent
Romance scam	`/fraud/romance-scam`	Whole conversation as one block, speakers unknown
Social engineering	`/fraud/social-engineering`	Whole conversation as one block, speakers unknown
Coercive control	`/safety/coercive-control`	Whole conversation as one block, speakers unknown
Bullying	`/safety/bullying`	Whole conversation as one block, speakers named
Self-harm and emotional distress	`/safety/unsafe`	Whole conversation as one block, speakers unknown

Run scope. The run's dataset selection was left at its default, so Tuteliq's harness also ran Tuteliq's own benchmark datasets in the same job. This did not change how our cases were sent or scored. The job reached its five-hour limit after all eleven of our folders had finished; the dataset still running at that point was one of Tuteliq's. All 1,344 of our cases returned a readable response with a recommended action: no failed calls and no case left unrun. We did not open the results for Tuteliq's own datasets.

Timeline.

Date (2026)	Step
September 25–27	Scope agreed; test design and generation method set
October 12–16	Datasets frozen: social engineering (Oct 12), coercive control (Oct 13), bullying (Oct 14), self-harm (Oct 15), romance (Oct 16)
October 17	Datasets placed in Tuteliq's evaluation repository; Tuteliq's setup check (run #40)
October 19	Live run (run #42) and scoring
October 20–21	Blind label check by the second evaluator
October 22	Draft sent to Tuteliq for factual review, with the second evaluator copied
About October 26	Final report

7. Results

Each figure gives the rate, the count and a 95% Wilson score interval in square brackets, because the sets are small and some rates sit near 0% or 100%. Tables show the main measure unless they say otherwise.

How we score. The main measure counts a harmful conversation as detected when Tuteliq's recommended action is above monitor. As a second measure, we also report any detection at all. For self-harm, conversations that show acute risk are scored on the main measure; conversations that show warning signs count as detected when Tuteliq flags self-harm, distress context or depression and anxiety, at any action level. Tuteliq confirmed these rules before the run. Recall is reported separately for baseline, evasion and adversarial conversations, so the cost of hiding or disguising the harm is visible. False alarms are reported separately for ordinary and hard harmless conversations. Precision is computed on the full set. Failed calls, including a response that cannot be read or has no recommended action, are listed apart and excluded from every rate.

Precision depends on the mix. Our sets are about half harmful and half harmless; in real messaging, harmful conversations are rare. The same detector with the same false-alarm rate has much lower precision when harm is rare, so we report recall and the false-alarm rate as the main figures, and give precision at our mix with a note on what it would be at lower prevalence. Accuracy is not reported as a headline figure for the same reason.

Naming Tuteliq next to results. Wording agreed with Tuteliq in writing on October 17, 2026; the detector version is the identifier the health check reports at the start of the run: "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."

Social engineering had the lowest detection and the most false alarms

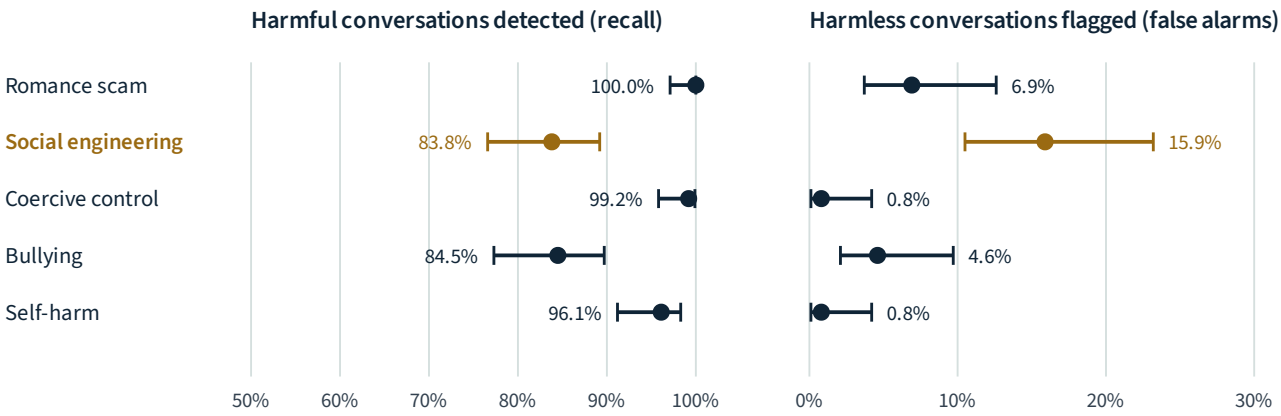


Figure 1. Source: Tuteliq run of October 19, 2026 (detector `tuteliq-detect-v5.20261015`); our scoring. n = 129–130 per category and class.

Figure 1 summarizes the main result; the tables below give every figure.

Detection of harmful conversations (recall)

Table 1. Harmful conversations detected, by category and test class (main measure). Source: Tuteliq run of October 19, 2026; our scoring.

Category	Baseline	Evasion	Adversarial	All harmful
Romance scam	100.0% (70/70) [94.8–100.0]	100.0% (32/32) [89.3–100.0]	100.0% (28/28) [87.9–100.0]	100.0% (130/130) [97.1–100.0]
Social engineering	81.4% (57/70) [70.8–88.8]	87.5% (28/32) [71.9–95.0]	85.7% (24/28) [68.5–94.3]	83.8% (109/130) [76.6–89.2]
Coercive control	100.0% (78/78) [95.3–100.0]	96.2% (25/26) [81.1–99.3]	100.0% (26/26) [87.1–100.0]	99.2% (129/130) [95.8–99.9]
Bullying	87.2% (68/78) [78.0–92.9]	76.0% (19/25) [56.6–88.5]	84.6% (22/26) [66.5–93.8]	84.5% (109/129) [77.3–89.7]
Self-harm / emotional distress	97.4% (76/78) [91.1–99.3]	92.3% (24/26) [75.9–97.9]	96.0% (24/25) [80.5–99.3]	96.1% (124/129) [91.2–98.3]

False alarms on harmless conversations

Table 2. Harmless conversations flagged, by category and type (main measure). Source: as Table 1.

Category	Baseline harmless	Hard negatives	All harmless
Romance scam	1.8% (1/55) [0.3–9.6]	10.7% (8/75) [5.5–19.7]	6.9% (9/130) [3.7–12.6]
Social engineering	20.4% (11/54) [11.8–32.9]	12.5% (9/72) [6.7–22.1]	15.9% (20/126) [10.5–23.2]
Coercive control	0.0% (0/39) [0.0–9.0]	1.1% (1/91) [0.2–6.0]	0.8% (1/130) [0.1–4.2]
Bullying	0.0% (0/39) [0.0–9.0]	6.6% (6/91) [3.1–13.6]	4.6% (6/130) [2.1–9.7]
Self-harm / emotional distress	0.0% (0/39) [0.0–9.0]	1.1% (1/91) [0.2–6.0]	0.8% (1/130) [0.1–4.2]

Precision

Table 3. Precision at our mix of about half harmful (main measure). Source: as Table 1.

Category	Precision	Harmful / all flagged
Romance scam	93.5% [88.2–96.6]	130/139
Social engineering	84.5% [77.3–89.7]	109/129
Coercive control	99.2% [95.8–99.9]	129/130
Bullying	94.8% [89.1–97.6]	109/115
Self-harm / emotional distress	99.2% [95.6–99.9]	124/125

These figures hold at our mix of about half harmful. For illustration only, if 1 conversation in 100 were harmful, the same recall and false-alarm rates would give precision of about 13% for romance, 5% for social engineering, 57% for coercive control, 16% for bullying and 56% for self-harm. These are calculated from our rates, not measured.

Second measure: any action

The second measure counts any recommended action other than none, including monitor.

Table 4. Recall and false alarms on the second measure (any action). Source: as Table 1.

Category	Recall, all harmful	False alarms, all harmless
Romance scam	100.0% (130/130) [97.1–100.0]	56.9% (74/130) [48.3–65.1]
Social engineering	100.0% (130/130) [97.1–100.0]	53.2% (67/126) [44.5–61.7]
Coercive control	100.0% (130/130) [97.1–100.0]	25.4% (33/130) [18.7–33.5]
Bullying	84.5% (109/129) [77.3–89.7]	4.6% (6/130) [2.1–9.7]
Self-harm / emotional distress	87.6% (113/129) [80.8–92.2]	0.8% (1/130) [0.1–4.2]

For romance, social engineering and coercive control, Tuteliq places many harmless conversations at monitor; what separates harmful from harmless there is the action level above monitor. For bullying the two measures give identical results. For self-harm, the second measure swaps the rule for warning-sign conversations (an action above monitor instead of a self-harm, distress or depression flag at any level), so it comes out lower than the main measure.

Results by route

Table 5. Recall and false alarms by route (main measure). Source: as Table 1.

Category	Route	Recall	False alarms
Romance scam	Classic romance	100.0% (87/87) [95.8–100.0]	—
Romance scam	Investment ("pig butchering")	100.0% (43/43) [91.8–100.0]	—
Social engineering	Account security	100.0% (42/42) [91.6–100.0]	23.7% (9/38) [13.0–39.2]
Social engineering	Tech support	100.0% (34/34) [89.8–100.0]	11.8% (4/34) [4.7–26.6]
Social engineering	Government identity check	25.0% (7/28) [12.7–43.4]	25.0% (7/28) [12.7–43.4]
Social engineering	A relative on a new number	100.0% (26/26) [87.1–100.0]	0.0% (0/26) [0.0–12.9]
Coercive control	Whereabouts and proof	100.0% (39/39) [91.0–100.0]	0.0% (0/39) [0.0–9.0]
Coercive control	After separation	100.0% (39/39) [91.0–100.0]	2.6% (1/39) [0.5–13.2]
Coercive control	Phone access	100.0% (26/26) [87.1–100.0]	0.0% (0/26) [0.0–12.9]
Coercive control	Isolation rules	96.2% (25/26) [81.1–99.3]	0.0% (0/26) [0.0–12.9]
Bullying	Exclusion	100.0% (41/41) [91.4–100.0]	4.8% (2/42) [1.3–15.8]
Bullying	Teasing that turns	65.2% (30/46) [50.8–77.3]	4.3% (2/46) [1.2–14.5]
Bullying	Exposure	90.5% (38/42) [77.9–96.2]	4.8% (2/42) [1.3–15.8]
Self-harm	Suicidal thoughts	97.6% (41/42) [87.7–99.6]	0.0% (0/42) [0.0–8.4]
Self-harm	Self-injury	90.2% (37/41) [77.5–96.1]	2.4% (1/42) [0.4–12.3]
Self-harm	Warning signs	100.0% (46/46) [92.3–100.0]	0.0% (0/46) [0.0–7.7]

Romance harmless conversations carry no route, so their false alarms appear only in the category tables. Notable results by the kind of request:

- **Social engineering, requests for identity data:** 7 of 28 detected (25.0% [12.7–43.4]). Every other kind of request (login credentials, one-time codes, remote access) was detected in all cases.
- **Romance, harmless requests for a loan between people who trust each other:** 8 of 15 flagged (53.3% [30.1–75.2]). Harmless requests to split a cost (0/38) or pay someone back (0/22) were never flagged.
- **Bullying, mockery of someone's abilities:** 10 of 22 detected (45.5% [26.9–65.3]). Threats to share something, demands to stay away, and removal from a group were detected in all cases.

Single-message baseline

Most messages in a harmful conversation are harmless on their own, so we expected detection on single messages to be lower than on whole conversations. This table shows how much the system relies on context. Each category has only five harmful and five harmless messages, so these figures show direction only.

Table 6. Single messages against the same items as whole conversations, five harmful and five harmless per category (main measure). Source: as Table 1.

Category	Detected as single messages	Detected as whole conversations (same items)
Romance scam	5/5 (false alarms 0/5)	5/5 (false alarms 0/5)
Social engineering	5/5 (false alarms 1/5)	5/5 (false alarms 1/5)
Coercive control	3/5 (false alarms 0/5)	5/5 (false alarms 0/5)
Bullying	4/5 (false alarms 0/5)	2/5 (false alarms 0/5)
Self-harm / emotional distress	3/5 (false alarms 0/5)	5/5 (false alarms 0/5)

Across all 50 single messages, 20 of 25 harmful messages were detected (80.0% [60.9–91.1]) and 1 of 25 harmless messages was flagged (4.0% [0.7–19.5]). Whole conversations did better for coercive control and self-harm. Bullying was the one category where the single message did better than the whole conversation.

Failed calls: none. All 1,344 cases returned a readable response with a recommended action, and no case was left unrun.

Validation

Checks that the results are complete and correctly reported:

- **Complete run.** All 1,344 cases returned a readable response with a recommended action; none failed and none was left unrun.
- **Unchanged detector.** Tuteliq's health check reported the same detector version, build and configuration at the start and end of the run, and the service did not restart in between.
- **Tested scoring.** The scoring script was tested on invented responses before the run, and scores only the saved responses, so no case was sent twice.
- **Checked figures.** Every figure in sections 7 and 8 was checked by script against the scoring output, with the intervals recalculated independently; no mismatches were found.
- **Second evaluator.** Blind labels from Dr. Roy agreed with ours on 41 of 50 conversations (Cohen's kappa 0.64; section 5, Appendix C).

8. Strengths and weaknesses

Each finding cites the figure in section 7 that supports it. All figures use the main measure unless stated, and they come from synthetic conversations; section 5 lists the limits of that data.

Table 7. Findings, with the evidence for each from section 7.

Finding	Strength or weakness	Evidence (section 7)
Romance scams were detected in every harmful conversation, classic and investment alike, including all evasion and adversarial probes	Strength	Recall 100.0% (130/130) [97.1–100.0]
Coercive control was detected almost without exception, with almost no false alarms	Strength	Recall 99.2% (129/130) [95.8–99.9]; false alarms 0.8% (1/130) [0.1–4.2]
Self-harm and emotional distress were detected reliably, with almost no false alarms	Strength	Recall 96.1% (124/129) [91.2–98.3]; false alarms 0.8% (1/130) [0.1–4.2]
Detection held up on the synthetic evasion and adversarial probes, which hide or disguise the harm; no category fell sharply from baseline to evasion or adversarial probes	Strength	Largest drop: bullying, baseline 87.2% to evasion 76.0%, with overlapping intervals
Social engineering that asks for identity data under a government identity check was mostly missed, and harmless conversations on the same route were flagged just as often	Weakness	Recall 25.0% (7/28) [12.7–43.4]; false alarms on the route 25.0% (7/28) [12.7–43.4]
Social engineering had the highest false-alarm rate, including on ordinary harmless conversations	Weakness	False alarms 15.9% (20/126) [10.5–23.2]; baseline harmless 20.4% (11/54); account security 23.7% (9/38)
Genuine loan requests between people who trust each other were often flagged as romance scams	Weakness	8 of 15 flagged, 53.3% [30.1–75.2]; split-cost (0/38) and pay-back (0/22) requests never flagged
Bullying that starts as teasing, and mockery of someone's abilities, was often missed	Weakness	Teasing that turns 65.2% (30/46) [50.8–77.3]; ability mockery 45.5% (10/22) [26.9–65.3]
Many harmless romance, social engineering and coercive control conversations were placed at monitor	For platforms to note	Second measure false alarms 56.9%, 53.2% and 25.4%; the action level above monitor is what separates harmful from harmless

Answers to the questions set before the run.

- **Does detection hold up when the harm is hidden or disguised?** Largely yes, on the synthetic evasion and adversarial probes we tested. Recall on evasion and adversarial probes stayed close to baseline in every category, and for social engineering it was slightly higher. A system that relied mainly on keywords would be expected to drop sharply on these probes; that pattern did not appear.
- **Does the system raise false alarms on harmless conversations that look suspicious?** Rarely for coercive control and self-harm (1.1% of hard negatives each), more often for bullying (6.6%) and romance (10.7%), where most romance false alarms were genuine loan requests. Social engineering is the exception: it flagged ordinary harmless conversations (20.4%) more often than hard negatives (12.5%).
- **Does it need the whole conversation?** Whole conversations did better for coercive control and self-harm (5 of 5 against 3 of 5 for single messages). Bullying went the other way (4 of 5 single messages against 2 of 5 whole conversations). With five items per category, these show direction only.
- **Does performance differ by route?** Within romance, no: classic and investment scams were both detected in every case. The clear differences are within social engineering, where the government identity route stands far below the others, and within bullying, where teasing that turns hostile is detected less often than exclusion or exposure.

Recommendations to Tuteliq

1. **Identity-data requests under a government pretext.** Review the 21 missed and 7 falsely flagged conversations on this route. The detector does not yet separate a scam asking for identity data from a legitimate identity check.
2. **Social engineering false alarms.** Review the 20 harmless conversations flagged above monitor, most on routine account-security and tech-support contact.
3. **Loans between people who trust each other.** Review the 8 flagged trusted-loan conversations. An ordinary request for money between people with an established relationship is not, by itself, a romance-scam signal.
4. **Bullying that escalates gradually.** Strengthen detection of mockery of ability and of teasing that turns hostile (16 of 46 missed on that route).
5. **Guidance on the monitor level.** Tell integrators that monitor is applied to many harmless conversations, so automated action is best tied to flag for review and above.

On request, we can give Tuteliq the case IDs behind each of these findings.

---

9. Privacy and AI ethics

No real person's data was used or sent anywhere in this evaluation: every conversation is synthetic, and only synthetic text reaches Tuteliq's API.

Our test data

- **No real people.** No message from a real victim or offender is quoted, paraphrased or imitated. Sources inform what happens at each stage, never the wording.
- **No real identifiers.** Links, phone numbers, account numbers and codes are fictional placeholders. Checks reject any message that writes out a real-looking password, ID number, date of birth or card number.
- **No real organizations.** Banks, agencies and companies are named only by kind or by invented names.
- **No child sexual abuse material,** and no sexual content involving minors, under any circumstances.
- **Local generation and controlled storage.** Conversations are generated on our own machines and stored on our own storage with verified copies; they are kept out of public code repositories. For the run, they are placed in Tuteliq's evaluation repository and kept there afterward under royalty-free terms agreed in writing on October 26, 2026: Tuteliq may use them for internal testing, improving detection and regression checks; the author keeps ownership and may publish them; Tuteliq does not share or publish them outside the company.

Tuteliq's system

From Tuteliq's public documentation, and facts Tuteliq gave us in writing:

- **Storage and training.** The documentation describes the service as stateless: submitted text is not stored between requests and is not used for training (How it works). Tuteliq told us in writing that processing takes place in the EU.
- **Explaining a flag.** Every detection returns a short rationale and a recommended action with its risk score, so a platform can see why a conversation was flagged and choose a proportionate response, from monitoring to blocking (section 3).
- **Age.** No age group is sent. Tuteliq reads any sign of age from the conversation itself and otherwise does not treat the speakers as children.
- **False accusations.** A false alarm on a harmless conversation can wrongly mark a person as a scammer or abuser. Our hard negatives are built to measure exactly this risk; the result is reported in section 7 and discussed in section 8. How a platform acts on a flag matters as much as the flag itself: Suresh and Guttag call the gap between how a system is evaluated and how people actually use its output deployment bias (Suresh & Guttag 2021, p. 6). A flag is best treated as a prompt for human review, not a verdict.

Results handling. Tuteliq's responses are its confidential information. They are stored only on the author's own computers and a private, encrypted cloud backup. They are shared only with Dr. Roy and the AI assistant used to draft this report (with Tuteliq's written consent of October 14), used only for this evaluation, and kept as a private copy by the author for later verification only, with Tuteliq's written permission of October 26, 2026. Tuteliq also holds an archive copy (see the table below).

Data handling and retention

No real person's data was handled at any stage. The table sets out what each kind of data is, who could see it and how long it is kept.

Data	Held where	Who could access it	Retention
Test conversations (synthetic text only)	Author's own storage, with verified copies; Tuteliq's evaluation repository for the run	Author; Tuteliq; Dr. Roy (the 50 in the label check); the AI assistant	Owned by the author. Tuteliq may keep them for internal testing, improving detection and regression checks, under royalty-free terms agreed in writing on October 26, 2026
Text sent to Tuteliq's API	Tuteliq's service (processing in the EU, per Tuteliq in writing)	Tuteliq	Not stored between requests and not used for training, per Tuteliq's documentation; incident recording switched off for every call
Tuteliq's responses and our scores per case	Run artifact in Tuteliq's repository; author's own computers and a private, encrypted cloud backup	Author; Dr. Roy; the AI assistant (Tuteliq's written consent of October 14)	Used only for this evaluation. The author keeps one private copy for later verification only, with Tuteliq's written permission of October 26, 2026; it is not shared or used for anything else. Tuteliq, which owns this data, also holds an archive copy
Second evaluator's labels	Author's computer, stored apart from Tuteliq's responses	Author; Dr. Roy	Kept with the evaluation record; they contain no Tuteliq data
Report figures	This report and the public summary	Readers of each version	Aggregate rates and counts only; no case-level data in either version

The AI assistant was used with the "Help improve our AI models" privacy setting turned off, so nothing it processed for this evaluation is used to train Anthropic's models.

10. Regulatory relevance

Both the proposed EU KIDS Act and the UK Online Safety Act push platforms to show that their services are safe for children. The last column notes where our results are relevant to each instrument. This section does not assess or validate compliance by Tuteliq or by any platform: compliance depends on how a platform deploys a tool and on its wider risk assessments and measures, which this evaluation did not examine. Our conversations do not state the speakers' ages, so the results describe detection in general, not for children specifically.

Instrument	Status	What it asks of platforms	Link to this evaluation
EU KIDS Act (European Commission proposal)	Proposed September 17, 2026; now before the European Parliament and the Council (Latham & Watkins)	Safety by design and a high level of privacy, safety and security for users under 18; certified age assurance; limits on account creation by children under 15 on some social networking and video-sharing services; very large platforms must submit compliance plans and commission an independent audit of them	Detection across whole conversations could be one input to a platform's safety-by-design measures: coercive control was detected in 99.2% of harmful conversations and bullying in 84.5%. False alarms stayed at or below 4.6% for coercive control, bullying and self-harm, which is relevant to proportionate moderation
UK Online Safety Act: illegal harms duties	In force since March 17, 2025 (CMS)	Risk assessment for illegal content, including fraud and child sexual abuse; measures against pathways to grooming; a channel for fraud experts to flag scams. Fraud (Schedule 7, para. 33) and controlling or coercive behavior (para. 11) are both priority offenses	Fraud: romance scams were detected in every case (130/130). Social engineering recall was 83.8% with 15.9% false alarms, and the government identity route (25.0%) is a gap a platform's fraud risk assessment would need to cover. Coercive control, also a priority offense, was detected in 99.2% of harmful conversations
UK Online Safety Act: protection of children codes	In force since July 25, 2025	Protect children from content harmful to them, including bullying content	Bullying recall was 84.5% with 4.6% false alarms; teasing that turns hostile (65.2%) and mockery of ability (45.5%) were detected less often

11. Honest overall assessment

On 1,344 synthetic cases across five categories, Tuteliq's detector (`tuteliq-detect-v5.20261015`) detected most harmful conversations at an action level above monitor, with few false alarms in most categories. It is strongest on romance scams, coercive control and self-harm, where recall is above 95% and false alarms are at or below 6.9%. Detection did not collapse on the synthetic evasion and adversarial probes, which hide or disguise the harm; evasion in real conversations may differ.

The clearest weakness is social engineering that asks for identity data under a government pretext: one in four such conversations was detected, and harmless conversations on that route were flagged just as often. Social engineering also has the highest false-alarm rate of the five categories. Smaller gaps are genuine loans between people who trust each other being flagged as romance scams, and bullying that escalates gradually from teasing.

How far our data supports these conclusions: the conversations are synthetic, generated by a language model and checked by script and by hand, with labels reviewed by a second evaluator (agreement 82%, Cohen's kappa 0.64; section 5). Routes and request types have tens of conversations each, so their figures have wide intervals. The sets are about half harmful; at realistic prevalence, precision would be much lower (section 7). The results describe this detector version on these conversations, not performance on real platform traffic.

Independence and funding

This evaluation was designed, run and scored independently by Carl Lochstampfor, with Dr. Ayan Roy 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. This evaluation is not an accreditation, certification or audit. 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. 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. After the evaluation, Tuteliq was given royalty-free permission to use the test dataset for internal testing and regression checks (section 9), so any future independent retest of Tuteliq should use new, unseen conversations.

Use of AI. Claude (Anthropic), an AI assistant, helped draft and edit this report, cross-check every reported figure against the scoring output, organize the sources, and write the generation and scoring software under the author's direction. The test conversations themselves were generated by a separate open-weight language model run locally (section 5). All design decisions, labels, judgments and conclusions are the author's. Tuteliq's responses were shared with the assistant under Tuteliq's written consent of October 14, 2026. 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.

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

Agarwal, S., Harvey, E., Mariconti, E., Suarez-Tangil, G., & Vasek, M. (n.d.). *"Hey mum, I dropped my phone down the toilet": Investigating Hi Mum and Dad SMS scams in the United Kingdom* [Unpublished manuscript].

Ai, L., Kumarage, T., Bhattacharjee, A., Liu, Z., Hui, Z., Davinroy, M., Cook, J., Cassani, L., Trapeznikov, K., Kirchner, M., Basharat, A., Hoogs, A., Garland, J., Liu, H., & Hirschberg, J. (2024). Defending against social engineering attacks in the age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 12880–12902). Association for Computational Linguistics.

Amirkhani, S., Krüger, J., Randall, D., Stevens, G., & Zytko, D. (2026). "When suspicion becomes detection": Folk deception cues and detection strategies in online dating romance scams. *Proceedings of the ACM on Human-Computer Interaction, 10*(5), Article MHCI8950. https://doi.org/10.1145/3821700

Annadorai, K., Krish, P., Shaari, A. H., & Kamaluddin, M. R. (2018). Mapping computer mediated communication theories and persuasive strategies in analysing online dating romance scam. *Journal of Xi'an Shiyou University, 14*(2), 37–47.

Archer, A. K. (2017). *"I made a choice": Exploring the persuasion tactics used by online romance scammers in light of Cialdini's compliance principles* [Master's research project, Regis University]. ePublications at Regis University.

Benedetti, N., & Queralt, S. (2026). Forensic linguistic evidence of psychological abuse and discursive manipulation in cyberstalking cases. *Journal of Language Aggression and Conflict*. Advance online publication. https://doi.org/10.1075/jlac.00152.ben

Buglass, S. L., Abell, L., Betts, L. R., Hill, R., & Saunders, J. (2021). Banter versus bullying: A university student perspective. *International Journal of Bullying Prevention, 3*, 287–299. https://doi.org/10.1007/s42380-020-00085-0

CMS. (2025). *Online Safety Act: Illegal content duties are now in force*. https://cms.law/en/gbr/legal-updates/online-safety-act-illegal-content-duties-are-now-in-force

Covington & Burling. (2026). *The EU KIDS Act proposal: Towards a new regulatory framework for the protection of children online*. Inside Privacy. https://www.insideprivacy.com/childrens-privacy/the-eu-kids-act-proposal-towards-a-new-regulatory-framework-for-the-protection-of-children-online/

Crosby, A. E., Ortega, L., & Melanson, C. (2011). *Self-directed violence surveillance: Uniform definitions and recommended data elements* (Version 1.0). Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Cross, C., & Holt, T. J. (2025). Does age matter? Examining seniors' experiences of romance fraud. *Security Journal, 38*, Article 46. https://doi.org/10.1057/s41284-025-00486-0

Dutton, M. A., Goodman, L., & Schmidt, R. J. (2006). *Development and validation of a coercive control measure for intimate partner violence: Final technical report* (NCJ 214438). U.S. Department of Justice.

Edwards, M., Peersman, C., & Rashid, A. (2017). Scamming the scammers: Towards automatic detection of persuasion in advance fee frauds. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1291–1299). Association for Computing Machinery. https://doi.org/10.1145/3041021.3053889

Ellyson, A. M., Adhia, A., Lyons, V. H., & Rivara, F. P. (2021). Prevalence, age of initiation, and patterns of co-occurrence of digital dating abuse behaviors nationwide. *Children and Youth Services Review, 122*, Article 105921. https://doi.org/10.1016/j.childyouth.2020.105921

European Commission. (2026a). *Proposal for EU KIDS Act: EU Keeping Internet Digital Spaces Accountable and Trustworthy*. https://digital-strategy.ec.europa.eu/en/library/proposal-eu-kids-act-eu-keeping-internet-digital-spaces-accountable-and-trustworthy

European Commission. (2026b, September 17). *EU KIDS Act: Helping children navigate a safer online world*. https://commission.europa.eu/news-and-media/news/eu-kids-act-helping-children-navigate-safer-online-world-2026-09-17_en

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

Financial Crimes Enforcement Network. (2026, September 3). *FinCEN alert on money laundering activity associated with digital asset investment scam centers* (FIN-2026-Alert005). U.S. Department of the Treasury.

Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., & Dell, N. (2018). "A stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Paper 667). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174241

Gladden, R. M., Vivolo-Kantor, A. M., Hamburger, M. E., & Lumpkin, C. D. (2014). *Bullying surveillance among youths: Uniform definitions for public health and recommended data elements* (Version 1.0). Centers for Disease Control and Prevention; U.S. Department of Education.

Havard, T. E., & Lefevre, M. (2020). Beyond the Power and Control Wheel: How abusive men manipulate mobile phone technologies to facilitate coercive control. *Journal of Gender-Based Violence, 4*(2), 223–239. https://doi.org/10.1332/239868020X15850131608789

Home Office. (2023, April 5). *Controlling or coercive behaviour: Statutory guidance framework*.

Hossain, I., Puppala, S., Alam, M. J., & Talukder, S. (2026). AI-in-the-Loop: Privacy preserving real-time scam detection and conversational scam-baiting by leveraging LLMs and federated learning. *Proceedings on Privacy Enhancing Technologies, 2026*(1), 87–114. https://doi.org/10.56553/popets-2026-0006

Ibañez, K. R. C. (2024). Scammer strategies and social actions in online Filipino transactions. *Southeastern Philippines Journal of Research and Development, 29*(1), 43–75. https://doi.org/10.53899/spjrd.v29i1.287

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

Kazemi, A., Qadeer, H., Wagner, J., Hosseini, H., Natarajan Kalaivendan, S. B., & Davis, B. (2026). *SynBullying: A multi-LLM synthetic conversational dataset for cyberbullying detection* (arXiv:2511.11599) [Preprint]. arXiv. https://arxiv.org/abs/2511.11599

Khadka, K., Ullah, A. B., Ma, W., Martinez Marroquin, E., & Alem, Y. (2023). A survey on the principles of persuasion as a social engineering strategy in phishing. In *2023 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (pp. 1631–1638). IEEE.

Khadka, K., Ullah, A. B., & Martinez Marroquin, E. (2026). Unmasking persuasion in phishing: A content analysis of principles of persuasion in emails and subject lines. *Information & Computer Security, 34*(1), 104–121. https://doi.org/10.1108/ICS-12-2024-0321

Kopp, C., Layton, R., Sillitoe, J., & Gondal, I. (2015). The role of love stories in romance scams: A qualitative analysis of fraudulent profiles. *International Journal of Cyber Criminology, 9*(2), 205–217. https://doi.org/10.5281/zenodo.56227

Latham & Watkins. (2026). *EU KIDS Act: EU Commission publishes proposal on child protection online*. https://www.lw.com/en/insights/eu-kids-act-eu-commission-publishes-proposal-on-child-protection-online

Lawson, P., Pearson, C. J., Crowson, A., & Mayhorn, C. B. (2020). Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. *Applied Ergonomics, 86*, Article 103084. https://doi.org/10.1016/j.apergo.2020.103084

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

Li, D., Zheng, R., Liu, X. F., Ma, B., Sun, H., & Jiang, L. C. (2026). Linguistic dynamics of online scam conversations: A multi-stage analysis based on the COLD framework. *Humanities and Social Sciences Communications, 13*, Article 698. https://doi.org/10.1057/s41599-026-07052-y

Menini, S., Palmero Aprosio, A., & Tonelli, S. (2021). *Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection* (arXiv:2103.14916) [Preprint]. arXiv.

Menon, M., Vinayaga-Sureshkanth, N., Schon, A., Hemberger, K., & Jadliwala, M. (n.d.). *Linguistic hooks: Investigating the role of language triggers in phishing emails targeting African refugees and students* [Preprint].

Montañez Rodriguez, R., & Xu, S. (2022). Cyber social engineering kill chain. In C. Su et al. (Eds.), *Science of Cyber Security (SciSec 2022)* (Lecture Notes in Computer Science, Vol. 13580, pp. 487–504). Springer. https://doi.org/10.1007/978-3-031-17551-0_32

Oak, R., & Shafiq, Z. (2025). "Hello, is this Anna?": Unpacking the lifecycle of pig-butchering scams. In *Proceedings of the 21st Symposium on Usable Privacy and Security (SOUPS 2025)*. USENIX Association.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions, 2*(2), 183–188. https://doi.org/10.1016/j.invent.2015.03.005

Ofcom. (2025). *Time for tech firms to act: UK online safety regulation comes into force*. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/time-for-tech-firms-to-act-uk-online-safety-regulation-comes-into-force/

Online Safety Act 2023, c. 50, Schedule 7 (UK). https://www.legislation.gov.uk/ukpga/2023/50/schedule/7

Orygen. (2023). *#chatsafe: A young person's guide to communicating safely online about self-harm and suicide* (2nd ed., US English edition).

Pritom, M. M. A., Sanjari, S. M., Mia, M., Shibli, A. M., Hossain, S. M. M., Ismail, M., & Xu, S. (2026). *Short Message Service (SMS) phishing attacks and defenses: A systematic review* (arXiv:2604.11429) [Preprint]. arXiv.

Schokkenbroek, J. M., & Snaphaan, T. (2026). Love as bait: A scoping review and crime script analysis of online romance scams. *Trauma, Violence, & Abuse, 27*(4), 1877–1893. https://doi.org/10.1177/15248380251361046

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology* (pp. 25–36). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0603

Stafford, L., & Canary, D. J. (1991). Maintenance strategies and romantic relationship type, gender and relational characteristics. *Journal of Social and Personal Relationships, 8*(2), 217–242. https://doi.org/10.1177/0265407591082004

Steer, O. L., Betts, L. R., Baguley, T., & Binder, J. F. (2020). "I feel like everyone does it": Adolescents' perceptions and awareness of the association between humour, banter, and cyberbullying. *Computers in Human Behavior, 108*, Article 106297. https://doi.org/10.1016/j.chb.2020.106297

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

Tuteliq. (n.d.). *How it works*. Tuteliq documentation. Retrieved October 2026, from https://docs.tuteliq.ai/how-it-works.md

Van der Zee, S., Clayton, R., & Anderson, R. (2019). *The gift of the gab: Are rental scammers skilled at the art of persuasion?* (arXiv:1911.08253) [Preprint]. arXiv.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLOS ONE, 13*(10), Article e0203794. https://doi.org/10.1371/journal.pone.0203794

Wang, F., & Kelsay, J. D. (2026). The prevention of online romance scams using a crime script analysis from the victim's perspective. *International Review of Victimology, 32*(2), 407–430. https://doi.org/10.1177/02697580251377100

Weekes, C. J., Storey, J. E., & Pina, A. (2025). Cyberstalking perpetrators and their methods: A systematic literature review. *Trauma, Violence, & Abuse*. Advance online publication. https://doi.org/10.1177/15248380251333411

White & Case. (2025). *UK Online Safety Act: Protection of children codes come into force*. https://www.whitecase.com/insight-alert/uk-online-safety-act-protection-children-codes-come-force

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

Wood, I. D., Kepkowski, M., Zinatullin, L., Darnley, T., & Kaafar, M. A. (2024). An analysis of scam baiting calls: Identifying and extracting scam stages and scripts. In *Network and Distributed System Security (NDSS) Symposium 2024*. Internet Society.

World Health Organization & International Association for Suicide Prevention. (2023). *Preventing suicide: A resource for media professionals* (2023 update). World Health Organization.

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 656–666). Association for Computational Linguistics.

Glossary

Term	Meaning in this report
Recall	The share of harmful conversations the system detected
False alarm	A harmless conversation the system flagged; the false-alarm rate is the share of harmless conversations flagged
Precision	Of everything flagged, the share that was actually harmful. It depends on how common harm is (prevalence)
Prevalence	How common harmful conversations are in the traffic being checked
Main measure	A recommended action above monitor: flag for review, escalate, block or immediate intervention
Second measure	Any recommended action other than none, including monitor
Baseline	A harmful conversation in which the harm is stated plainly
Evasion probe	A harmful conversation in which the harm is hidden, softened or spread out
Adversarial probe	A harmful conversation written to steer the detector away, for example by framing the harm as harmless
Hard negative	A harmless conversation that looks suspicious on the surface
Route	A sub-type within a category, such as investment romance scams or proof-of-whereabouts control
Request type	The specific thing asked for in a conversation, such as a one-time code or a loan
Single-message baseline	The one message where the request first appears, sent on its own without the rest of the conversation
Failed call	A request with no usable response; never counted as a miss
Frozen set	A dataset fixed before the run, with a fingerprint (manifest hash) so any later change would show
Wilson score interval	A 95% range for a rate that stays accurate for small sets and for rates near 0% or 100% (Wilson 1927)

Appendix A: Run record

Item	Value
Date	Monday, October 19, 2026
Repository and workflow	Tuteliq's evaluation repository, benchmark workflow, run #42 (ID `39241363336`)
Branch and commit	`cova-tp1-datasets` at `34419d8`
Settings	Full datasets; incident recording off; no age group; speakers unknown (bullying: named)
Detector	`tuteliq-detect-v5.20261015`, build `71ff48a`, configuration `b45a9e613f13`, unchanged from start (06:53 ET) to end (12:48 ET)
Cases	1,344: 1,294 conversations and 50 single messages, in 11 folders
Failed calls / not run	0 / 0
Results artifact	results-`39241363336`, SHA-256 ffcdb91c…dcfdbef
Scoring	Offline, by our script (commit e3b61e4), on the saved responses

Appendix B: Scoring rules

- Each case is scored once, from the saved response, with no resending.
- Main measure: detected when the recommended action is above monitor. Second measure: any action other than none.
- Self-harm: conversations showing acute risk use the main measure; conversations showing warning signs count as detected when Tuteliq flags self-harm, distress context or depression and anxiety, at any action level. Tuteliq confirmed these rules before the run.
- Harmless self-harm conversations placed at monitor with distress context are not counted as false alarms (none occurred).
- Failed calls, including unreadable responses or responses with no recommended action, are excluded from every rate (none occurred).
- Intervals are 95% Wilson score intervals.

Appendix C: Label check by the second evaluator

Dr. Ayan Roy labeled 50 conversations blind on October 20–21, 2026: 10 per category, 5 harmful and 5 harmless, sampled at random with a fixed seed. Our design labels are the reference.

Category	Agreement	Cohen's kappa
Bullying	9/10 (90%)	0.80
Coercive control	9/10 (90%)	0.80
Self-harm / emotional distress	9/10 (90%)	0.80
Romance scam	8/10 (80%)	0.60
Social engineering	6/10 (60%)	0.20
All	**41/50 (82%)**	**0.64**

Our label	Dr. Roy: harmful	Dr. Roy: harmless
Harmful (25)	24	1
Harmless (25)	8	17

By test class: baseline 23 of 25, evasion 4 of 5, adversarial 3 of 3, hard negatives 11 of 17. By his confidence: 32 of 38 where he was certain, 6 of 8 where fairly sure, 3 of 4 where unsure.

Disagreements: four social engineering harmless conversations (two baseline, two hard negatives), two romance hard negatives (in one he saw a scam but was unsure it was a romance scam), one coercive control hard negative, one bullying hard negative, and one self-harm evasion conversation. Tuteliq's main-measure result matched our label in all nine.

Appendix D: Sources by category

The published research and official guidance each dataset is grounded in, and what each source informed.

Category	Source	Type	Informed
Romance	Wang & Kelsay 2026, *International Review of Victimology*	Peer-reviewed article	Scam tactics
Romance	Schokkenbroek & Snaphaan 2026, *Trauma, Violence, & Abuse*	Scoping review	Scam tactics; request types
Romance	Oak & Shafiq 2025, SOUPS	Conference paper	Investment scams
Romance	Kopp et al. 2015, *International Journal of Cyber Criminology*	Journal article	Scam tactics
Romance	Amirkhani et al. 2026, *Proc. ACM HCI*	Conference paper	Scam tactics; look-alike harmless conversations
Romance	Cross & Holt 2025, *Security Journal*	Peer-reviewed article	Request types
Romance	FinCEN 2026, Alert FIN-2026-Alert005	Government alert	Payment methods; investment scams
Romance	Archer 2017; Annadorai et al. 2018	Thesis; journal article	Supporting only
Romance	Stafford & Canary 1991	Peer-reviewed articles	Everyday harmless conversations
Romance	Lee & Ang 2026, *Deviant Behavior*	Peer-reviewed article	Conversation length
Social engineering	Wood et al. 2024, NDSS	Preprint	Scam tactics; request types
Social engineering	Agarwal et al., "Hi Mum and Dad" SMS scams in the UK	Study (year and venue not printed)	Impersonation of a relative; request types
Social engineering	Edwards et al. 2017, WWW Companion	Conference paper	Scam tactics
Social engineering	Ai et al. 2024, EMNLP	Conference paper	Scam tactics; look-alike harmless conversations
Social engineering	Montañez Rodriguez & Xu 2022, SciSec	Conference paper	Scam tactics
Social engineering	Pritom et al. 2026, SMS phishing review	Preprint	Scam tactics; request types
Social engineering	Van der Zee et al. 2019, arXiv	Preprint	Scam tactics
Social engineering	Ibañez 2024, *SPJRD*	Peer-reviewed article	Scam tactics
Social engineering	Khadka et al. 2023, TrustCom; Khadka et al. 2026	Journal article; study	Request types

Category	Source	Type	Informed
Social engineering	Menon et al.; Hossain et al. 2026, PoPETs	Preprint; conference paper	Request types
Social engineering	Lawson et al. 2020, *Applied Ergonomics*	Peer-reviewed article	Look-alike harmless conversations
Coercive control	Havard & Lefevre 2020, *Journal of Gender-Based Violence*	Peer-reviewed article	Controlling behaviors; request types
Coercive control	Home Office 2023, statutory guidance (England and Wales)	Government guidance	Definition; controlling behaviors
Coercive control	Dutton, Goodman & Schmidt 2006 (NIJ report)	Government-funded measure	Controlling behaviors; look-alike harmless conversations
Coercive control	Benedetti & Queralt 2026, *Journal of Language Aggression and Conflict*	Peer-reviewed article	Controlling behaviors
Coercive control	Ellyson et al. 2021, *Children and Youth Services Review*	Peer-reviewed article	Controlling behaviors; look-alike harmless conversations
Coercive control	Freed et al. 2018, CHI	Conference paper	Controlling behaviors
Coercive control	Weekes, Storey & Pina 2025, *Trauma, Violence, & Abuse*	Systematic review	Controlling behaviors
Bullying	Gladden et al. 2014, CDC uniform definitions	Government definition	Definition; bullying behaviors
Bullying	Buglass et al. 2021, *International Journal of Bullying Prevention*	Peer-reviewed article	Bullying behaviors; look-alike harmless conversations
Bullying	Steer et al. 2020, *Computers in Human Behavior*	Peer-reviewed article	Banter and bullying
Bullying	Van Hee et al. 2018, *PLOS ONE*	Peer-reviewed article	Bullying behaviors
Bullying	Kazemi et al. 2026 (two preprints)	Preprints	Conversation dynamics; limits of synthetic data
Bullying	Menini et al. 2021; Xu et al. 2012, NAACL	Preprint; conference paper	Look-alike harmless conversations
Self-harm	Crosby, Ortega & Melanson 2011, CDC uniform definitions	Government definition	Definitions and labels; safety rules
Self-harm	Shing et al. 2018, CLPsych	Conference paper	Disclosures
Self-harm	#chatsafe 2023 (Orygen)	Safe-communication guidance	Safety rules; supportive responses
Self-harm	WHO & IASP 2023, guidance for media professionals	International guidance	Safety rules; look-alike harmless conversations
Self-harm	O'Dea et al. 2015, *Internet Interventions*	Peer-reviewed article	Look-alike harmless conversations

Appendix E: Claims register

The same register appears in the public summary. 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 (section 7).

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.