Contents

CONFIDE — Making Therapy Transcripts Safer to Share, and Measuring When They Aren't

A white paper for therapists, coaches, and the people who build tools for them.

Version 1.0 · June 2026 · Gleb Kalinin and CONFIDE contributors · github.com/glebis/confide

---

In one paragraph

A therapy transcript is one of the most sensitive documents a person ever creates. To get AI to help review a session, you usually have to *share that transcript with a model* — so first it has to be made safer to share. CONFIDE is a free, open-source toolkit that does this on your own computer (nothing is uploaded), and — just as importantly — it *measures, out loud,* how safe the result actually is, including by attacking its own output to see what can still be recovered. The headline finding is simple and uncomfortable: removing the names is necessary but not enough. In our tests, structured details were scrubbed essentially perfectly (emails, dates, ID numbers at 100%; phone numbers at 83%); but the quiet, dangerous details — a medication, an age, a profession — survive far more often, and those are exactly what can re-identify someone after every name is gone. This paper explains how CONFIDE works, what it found, and what a clinician should and should not trust it for.

A word on terms. This paper is written to be readable by someone in their final year of school. Every technical word is explained the first time it appears, in plain language and in terms that matter to psychotherapy and to data privacy. There is a glossary at the end.

The clinician rule, up front. CONFIDE is *pre-review*, not a release button. Its output is a safer draft for your eyes — a human review is mandatory before any transcript, however redacted, goes to a cloud service. Nothing in this paper changes that.

---

1. Why this matters to a clinician

Imagine a single therapy session, transcribed to text. In it, a client might mention a diagnosis (say, bipolar disorder), a medication and dose, the name of their employer, the city they live in, a sexual orientation they're not out about, the name of a person who abused them, or a suicidal moment. Each of those is more than private — leaking any one of them can cost someone a job, a custody case, a relationship, their safety, or their life.

Now imagine you'd like AI to help: to spot a pattern across many sessions, to prepare for supervision, to check whether you actually did in the room what you think you did. Useful — but to do it, the transcript has to leave your head and enter a model. If that model lives in the cloud (ChatGPT, Claude.ai, and so on), the transcript leaves your machine — and the law treats this kind of data as special-category personal data, the most heavily protected kind. The rules differ by jurisdiction, and the differences matter:

- EU clients: sending an identifiable transcript to a non-EU cloud provider may trigger the GDPR's special-category rules (Article 9 — its list of extra-protected data types: health, sex life, and more) *and* its cross-border-transfer rules (extra conditions when personal data leaves the EU's legal reach).
- Russian data subjects: Russia's 152-ФЗ governs the data; the fines sit in its companion administrative code (КоАП, Art. 13.11).
- US: HIPAA applies only to *covered entities* (clinics, insurers, providers who bill electronically) and their business associates — many coaches fall outside HIPAA entirely, but state privacy law and professional-confidentiality duties still reach them.

The penalties are real:

	152-ФЗ / КоАП (Russia)	GDPR (EU)	HIPAA (US)
Data status	Special category	Art. 9 special category	PHI (protected health information) when a covered entity handles it; *psychotherapy notes* — a therapist's separately-kept process notes, a deliberately narrow category — get extra protection on top
Leak of special category	10–15 million ₽ (КоАП 13.11)	up to €20 M / 4% of worldwide turnover	top-tier annual cap ≈ $2.19 M

But here is the more important point, and the one this whole project is built around:

> Compliance is the floor, not the goal. The goal is not hurting the people who trusted a therapist with their story. You can be perfectly legal and still cause harm.

And the most common way that harm happens is a quiet mistake in language:

> "We removed the names" gets mistaken for "this is safe to send." It usually isn't.

To see why, we need one idea.

---

2. The one idea you have to understand: quasi-identifiers

Let's define the basic vocabulary first.

- PII — *Personally Identifiable Information* (in Russian law, ПДн, «персональные данные»). Any information that points to a specific person.
- De-identification — the *process* of removing or masking PII from a document. To mask a piece of text is to replace it with a placeholder, e.g. turning *"Marina"* into `[PERSON]`.
- Anonymization — the stronger *claim*: that after the processing, re-identifying the person is no longer reasonably possible. The two words are often used interchangeably, but the law treats them differently — and CONFIDE performs and measures *de-identification*; it never certifies *anonymization*.

PII comes in two flavours, and the difference is the heart of this paper.

Direct identifiers name a person almost by themselves: a full name, an email address, a phone number, a passport or insurance number. These are the obvious ones, and — spoiler — software is very good at catching them.

Quasi-identifiers are the trap. A quasi-identifier is a detail that is *harmless on its own* but, combined with a few others, points at exactly one person. No single one is "a name," so a tool that hunts for names walks right past them. Here is the example that makes it click — a real-feeling description with every name already removed:

```
"a woman"              → about 50% of people
"43 years old"         → about 1.5%
"teaches music"        → about 0.01%
"in Kostroma"          → about 0.003%   (a mid-size Russian city)
"has twin children"    → probably exactly one person
```

(Illustrative toy numbers, chosen to show the mechanism — not census estimates.)

Stack those five facts and you have de-anonymized someone — identified them — without ever using their name. *"A 43-year-old music teacher in Kostroma, mother of twins"* — the name is gone, and a motivated stranger finds her in five minutes.

This is why de-identification is not a spell-check for names. The combination is the danger, and judging which combinations are dangerous is clinical work — only the therapist knows that this particular client is the only redhead programmer in their town. We will come back to this; it is the limit no algorithm fully removes.

> Field note (therapy-specific PII). Beyond the standard list (names, dates, addresses, phones, emails, document numbers), therapy transcripts carry *therapeutic* identifiers that generic privacy tools usually miss unless configured for therapy: clinic and doctor names, medications with doses, unique life events, court details. A medication is special — it doesn't just identify, it *narrows the possible diagnoses.* "Lithium" suggests things a phone number never could.

---

3. What CONFIDE is

CONFIDE — the name is a play on *Confidential Filtering of Identifying Details*, and on the Kylie Minogue song — has three parts. Keep the split in mind; it's the spine of everything below.

Part	Plain-language job
CONFIDE (the anonymizer)	*Protect.* Scrub PII out of a transcript, entirely on your own machine.
CONFIDE-Bench (the benchmark)	*Measure.* Score how good that scrubbing actually is.

Part	Plain-language job
CONFIDE-Red (the red team)	*Attack.* Try to re-identify the scrubbed output, to find what leaked.

A benchmark is a fixed, repeatable test you can run again and again to compare methods fairly. A red team (a term borrowed from security) is a group — here, an AI — whose job is to *attack* something on purpose, so the defenders learn where it's weak. CONFIDE protects; CONFIDE-Red attacks the protection; CONFIDE-Bench keeps score. The point of having all three is honesty: a privacy tool that never tests itself is just a promise.

Two design commitments shape the whole thing:

1. Local-first — and we mean inference, not just the window. *Local-first* means the work happens on your computer and the raw data never leaves it. The subtle trap here is one the project states loudly:

> Local tools ≠ local inference. What matters is not whether you're typing in a terminal on your laptop — it's *where the model actually does its thinking.* A tool can run on your machine and still ship your text to a cloud model to process. CONFIDE's anonymizer does the thinking locally.

In practice this is run as a red folder / green folder discipline: raw transcripts live in a `red_raw_local_only/` folder that no AI agent is ever allowed to open; only reviewed, anonymized text moves to a `green_anon_reviewed/` folder that an agent may see. To be precise about what enforces this: the agent's permission system is configured never to grant access to the red folder, and the human never points it there — it is a working procedure backed by access rules, not a magical property of folder names.

2. Built in the open, by volunteers. CONFIDE is a citizen-science project — built and scrutinized by volunteers rather than a funded lab — so the people whose privacy is at stake can inspect exactly how it works. (More on what that commitment entails, and where we still fall short of it, in §8.) Every therapy transcript shipped publicly is synthetic: fictional clients, invented details. Nothing in the public repository is a real person.

---

4. Protect — the layered local stack

CONFIDE's anonymizer is built in layers, each catching a different kind of PII, because no single tool catches everything. Think of it like screening a document through several sieves of different mesh.

Layer	Tool	What it's good at	Cost
1. Rules	Regex	emails, URLs, phone numbers, structured IDs (insurance, account, card), numeric dates	tiny, instant

Layer	Tool	What it's good at	Cost
2. Russian names	Natasha (Russian NER)	Russian names, cities, organizations	~27 MB, instant
3. Reasoning	Local LLM (Qwen via Ollama)	medications, ages, professions, contextual IDs	2.5–7 GB, ~10 s

A few terms in that table:

- Regex (*regular expression*) — a precise text-matching rule. "Find anything shaped like `xxx@xxx.xx`" is a regex for emails. Rules are perfect for things with a fixed shape and useless for things without one.
- NER (*Named Entity Recognition*) — software that reads text and tags the "named things": this token is a PERSON, that one is a LOCATION. Natasha is an open NER tuned for Russian.
- LLM (*Large Language Model*) — the kind of AI behind chatbots. Here it runs *locally* (via Ollama, a tool for running models on your own machine) to catch the subtle items that have no fixed shape and aren't simple names — most importantly medications.

The interesting result is which layer earns its keep. We measured every layer alone and in combination (this is the *ablation* — turning parts off to see what each contributes). The recommended Russian stack is Natasha + Regex + local LLM. And there's an honest sub-story baked in: an earlier version used the OpenAI Privacy Filter (OPF), a real open-source PII model, as a layer. We measured it, found it slow on a normal CPU (~2 seconds per line, so it didn't finish a 10 KB transcript) and prone to breaking its own output format — so we replaced it with deterministic regex and kept OPF only as a documented comparison, not a recommendation. The lesson — *bigger isn't automatically better* — is itself one of the findings.

---

5. Measure — the benchmark

Now the scoring. To measure de-identification you need a gold standard: a version of each transcript where humans have marked every piece of PII, so you can check what the tool caught and what it missed. CONFIDE-Bench is, as far as we could find (search current as of June 2026), the first benchmark built specifically on therapy *dialogue* — the nearest public resources are court-record and clinical-note de-identification corpora (the Text Anonymization Benchmark (TAB), the i2b2/n2c2 clinical sets) or counseling dialogue *without* PII labels, not the messy, disclosive back-and-forth of a real session. It covers both Russian and English.

5.1 Why "recall" is the safety metric

Two words you'll see in any measurement like this:

- Recall — of all the PII that was really there, what fraction did the tool catch? Miss nothing → recall = 1.0 (100%).
- Precision — of everything the tool flagged, what fraction was actually PII? Flag only real PII → precision = 1.0.

For ordinary tasks people balance the two. For privacy, they are not equal:

> A missed entity is leaked PII — a real person exposed. A false alarm — a word masked when it didn't need to be — is usually far less harmful, but it is not free: over- masking can erase clinically useful context. That is why precision is still reported.

So CONFIDE leads with recall, and reports an F2 score — a recall-heavy combined score: its setting ($\beta = 2$) declares recall twice as *important* as precision, which in the formula gives recall four times the weight. (The plain "F1" score weights them equally; for safety, that's the wrong dial.) When we say a number is the "headline," it's a recall-style number, because a miss is the failure that hurts someone.

We also report two stricter, more honest views:

- Entity-level recall — an entity (say, the client's name) counts as *protected* only if every mention of it is masked. One un-redacted recurrence is a leak, so this is harder to score well on than counting individual mentions, and closer to real safety.
- Harm-weighted recall — not all misses are equally bad. Missing a *medication* (which implies a diagnosis) is worse than missing a URL. Harm-weighted recall counts a miss by *how much harm it could do*, so the number reflects damage, not just quantity. The gap between plain recall and harm-weighted recall is itself a finding.

5.2 The results, and the one table that tells the story

On the Russian synthetic *corpus* (the collection of texts used for testing — here 30 sessions with 1,076 marked PII items), the recommended stack reaches coverage recall ≈ 0.88 (did the mask touch the item at all) and entity-level recall ≈ 0.73 overall. Splitting the data: 0.74 on the *development split* (the part used while tuning) and 0.71 on the *held-out test split* (a part set aside and never used during development) — close to each other, which is what honesty looks like. The 95% confidence interval — the range the true value plausibly falls in, estimated here by re-sampling the corpus 2,000 times — is 0.85–0.90 for coverage recall; the corpus is small, so treat all of these as directional, not precise. Against a deliberately nasty adversarial set (inputs designed to fool the tool) — transliterated names, social-media handles, structured IDs — the full stack caught 19 of 20; the one escape was a *Latin-spelled Russian name* ("Sergey Volkov"), which the Russian-only NER can't see.

(Methods in brief: a fully synthetic six-client corpus; person-disjoint dev/test split — no client appears in both; gold spans located programmatically from curated patterns and hand-verified; pinned tool versions and a one-command Docker re-run. The full protocol — annotation codebook, taxonomy, versioning, re-run policy — lives in the repository's BENCHMARK, DATASHEET, and REPRODUCIBILITY documents.)

But the single most important result is *which categories survive*, because it proves the whole thesis of §2. Recall by category, recommended stack:

Category	Recall	Reading
EMAIL	1.00	structured → caught perfectly
ID numbers	1.00	structured → caught
DATE	1.00	structured → caught
LOCATION	1.00	NER catches it
PERSON (names)	0.96	NER catches it
ORG	0.86	mostly caught
PHONE	0.83	mostly caught
MEDICATION	0.17	(!) mostly missed
PROFESSION	0.15	(!) mostly missed
AGE	0.08	(!) almost always missed

(A note on these numbers: with a corpus this size, each category has only a handful of items, so each row is directional — one missed item moves a row by several points.)

Look at the shape of that. In this corpus, the direct identifiers and structured fields come close to a closed problem — emails, IDs, dates and locations at 1.00, names at 0.96, phones at 0.83. Promising, not "solved": the counts are small. But the quasi-identifiers that re-identify people — age, profession, medication — are exactly the weakest, caught only 8–17% of the time even by the full stack with a local LLM helping. Counting at the entity level across the *whole* quasi-identifier class (which also includes locations, dates, and organizations — categories the stack handles well), quasi-identifier recall is 0.64: roughly a third of quasi-identifying entities survive overall, and the survivors are concentrated in exactly the categories above.

That is the headline of the entire project, stated as a number:

> In our tests, software consistently removes the parts that *look* dangerous (names, emails) and consistently leaves the parts that *are* dangerous in combination (age + profession + medication). The Kostroma music teacher from §2 could plausibly survive automatic redaction nearly intact.

(English numbers tell a consistent story; on the curated English set the recommended stack reaches coverage recall ≈ 0.98, and the OPF measurement we kept as a lesson scored recall 0.76 / precision 0.92 — strong on names and phones, weak on exactly the relative dates — expressions like *"three weeks ago"* or *"last Tuesday"* that have no fixed shape a rule can match — and short numeric secrets that matter most for safety.)

---

6. Attack — the red team

Measuring what you *masked* isn't enough; you have to check what an attacker could *recover* from the masked text. CONFIDE-Red runs an AI attacker against the redacted output, using the three failure types named in EU privacy guidance (the Article 29 Working Party's Opinion 05/2014 — guidance from the pre-GDPR era that is still the standard yardstick for judging anonymisation):

- Singling-out — can you isolate *one* person out of a crowd from the surviving details, even without a name? (The Kostroma cascade is singling-out.)
- Linkability — given two redacted sessions, can you tell they're the same person? This is the cross-session risk: link enough sessions and you rebuild a profile.
- Inference — can you *guess* a hidden attribute (a diagnosis, an orientation) from what's left?

Two findings stand out.

Linkability: redaction held up — once a hidden leak was fixed. We tested whether the attacker, shown two redacted sessions, could tell if they belonged to the same client. Result: the attacker did not perform above chance in this setup — over 100 session pairs, accuracy 0.50 (95% confidence interval 0.41–0.60) and an AUC of 0.46 (CI 0.38–0.54). AUC — *area under the ROC curve* — is a score from 0 to 1 where 0.5 means "no better than a coin flip"; both intervals straddle chance, so we observed no evidence of linkability — which is weaker than proving none exists. The honest part: *before* a fix, this score was a perfect 1.00 — but that was an artifact, not real safety. 28 of 30 redacted files were silently leaking a piece of metadata — data *about* the file rather than in its visible text: a per-client name buried in the file's header that none of the three layers ever looked at — so "linking" was a trivial exact-match, not inference. We caught it, masked the header, and the score collapsed to chance. We're telling you this because *that is what trustworthy measurement looks like* — the embarrassing artifact is part of the report, not edited out.

Singling-out and inference: partly survives, and we say so carefully. Using a k-anonymity estimate (a standard way to ask "how many people in the population share this exact combination of traits?" — if the answer is 1, the person is singled out), the surviving quasi-identifiers usually leave a person in a *crowd* of dozens at this scale. Three honesty notes on that number: it is computed from published population fractions under an *independence assumption* that overstates uniqueness (professions, cities, and ages correlate, so the real crowd is larger); it is run on fabricated personas, so it is an illustrative risk reading, not a privacy guarantee; and the estimate is fragile — for one synthetic client it flips to "singled out" under slightly different population assumptions. The attacker can still sometimes guess attributes. The three ways therapy de-identification fails *after the names are gone* — surviving inference, surviving singling-out, surviving cross-session linkage — are exactly the three CONFIDE-Red keeps measuring.

---

7. What a clinician should — and shouldn't — trust this for

Plainly, because this is the part that matters at 9 a.m. on a Monday:

You can trust CONFIDE to:

- Scrub the obvious, structured PII (names, emails, phones, IDs, dates) very well, *on your own machine, without uploading anything.*
- Give you an *honest, recall-led estimate* of how much was caught — including its own weak spots.
- Catch far more than "find-and-replace the name" ever could.

You must NOT trust CONFIDE to:

- Make a transcript *safe by itself.* Roughly a third of quasi-identifying entities survive automatic redaction overall (entity recall 0.64) — and the most dangerous categories, age, profession, and medication, individually survive most of the time (caught only 8–17%, per §5.2). A human — you — must review the output, scanning specifically for the age + profession + place + medication combinations only you can judge.
- Be a compliance certificate. Passing the benchmark is not HIPAA, GDPR, or 152-ФЗ anonymization. It's a measurement, not a legal guarantee.
- Do anything clinical. This bears stating in its own line, because it's the brightest line in the project:

  AI here is a microscope, not a clinician. It is for de-identification and analysis *support* — never for suicide- or crisis-risk assessment, diagnosis, or automated decisions about a person's care. For risk, AI has no veto and no all-clear: the absence of an AI flag means nothing, and the decision is always a human's.

And the hardest ethical edge, which no amount of scoring removes:

> Third parties never consented. A client names an abuser, a child, a partner, a boss. Anonymization almost always cleans up the *client* and leaves those people named — and you cannot get their consent, because they were never in the room. Scrub them as carefully as you scrub the client. Consent from the client does not cover the people the client talks about.

This is also where the practical rule lives — a checklist worth printing:

- ☐ Names removed (client, relatives, colleagues, third parties)?
- ☐ Dates of birth, addresses, phones, emails removed?
- ☐ Institution names (clinic, school, employer) removed?
- ☐ Quasi-identifiers neutralized (profession + city + age)?
- ☐ Informed consent for AI processing on file?

One "no" and the data does not go to the cloud.

---

8. Ethics and responsible stewardship

CONFIDE is a privacy tool about vulnerable people, so its ethics are load-bearing, not an afterthought. The commitments:

- No real data in public, ever. Every public therapy transcript is synthetic — fictional people — so the *public benchmark* involves no human subjects and exposes no one. Any real-session use is a separate matter: it happens only locally, behind three independent layers of storage protection (the device, an encrypted store, and per-file isolation), it requires the client's explicit consent, and — where it becomes a formal study or happens inside a regulated clinic — its own ethics or legal review. Only aggregate statistics (totals and averages) — never transcript text — ever leave the machine.
- Consent is ongoing, not a one-time checkbox. Real-data use needs explicit, AI-specific consent (a separate document, not fine print), revocable at any time, updated when tools change.
- Dual-use, handled in the open. *Dual-use* means a tool that can help or harm depending on who holds it: CONFIDE-Red re-identifies redacted text, and the same technique could be misused. We counter this by headlining recall (so attacks read as measurements of safety, not as recipes), reporting residual risk only on fabricated personas, and publishing no re-identification recipe for real people.

We ran a structured self-assessment against established norms — the Ten Principles of Citizen Science of the European Citizen Science Association (ECSA), the Belmont Report and Menlo Report (the foundational ethics frameworks for human-subjects and for computing research), and dual-use research guidance. To be plain about its standing: this was a cited research survey conducted by the project itself (June 2026), not an external certification or an independent ethics board. Our reading of the result: CONFIDE *broadly complies*, and is unusually careful precisely because it keeps no real people in the open. The assessment surfaced two genuine gaps, which we have since closed:

1. A responsible-disclosure channel. We discussed dual-use in principle but offered no way to *report* a discovered de-identification leak or re-identification technique. There is now a `SECURITY.md`: report privately, never post real PII or a recipe publicly, fix before publicize.
2. Explicit credit for volunteers. A citizen-science project must acknowledge the people who build and check it; there is now a `CONTRIBUTORS.md`.

We list these openly because a benchmark is only as trustworthy as its reporting choices are explicit — including the parts that were, until recently, missing.

---

9. Limitations and future work

- Small corpus. The synthetic sets are small, so each miss moves a score by several points. Numbers are directional, reported with confidence intervals, not significance claims.
- Synthetic ≠ real. Synthetic data tests whether the pipeline recovers *planted* patterns. It does not tell you how the tool performs on real Russian transcripts in your office.
- The annotation bottleneck. A gold standard's authority comes from independent human annotators. This is the single most valuable thing volunteers can contribute — see the call below.

- More languages, better Russian models. Clear local, open upgrades exist (stronger Russian LLMs; checksum-based recognizers for Russian ID numbers; a built-in k-anonymity risk score). All on the roadmap; all local — never cloud Russian models for therapy data.

---

10. Using and contributing

CONFIDE is free and open (code under the MIT license, data & docs under CC-BY-4.0 — open licenses that ask only for attribution). You can run the whole benchmark with one command (the repository's `run-benchmark.sh`, with a `Dockerfile` — a recipe that rebuilds the identical software environment anywhere), extend it with public datasets via `python3 confide.py datasets list`, and read the deep-dive docs from the documentation map in the README.

Because it's citizen science, scrutiny and correction are the product. The gold standard needs independent annotators (the turnkey tooling — `tools/annotator.html` plus an annotation codebook — is in the repo). Found a leak or a re-identification technique? That's a security issue here — report it privately via `SECURITY.md`, never in a public issue.

Until a peer-reviewed paper exists, cite the repository: *Gleb Kalinin and CONFIDE contributors, "CONFIDE: a therapy-transcript de-identification benchmark and red team," 2026, https://github.com/glebis/confide.* It is a reproducible benchmark artifact and tool report, not peer-reviewed; cite it as a measurement, not as a compliance guarantee.

> The whole tool, in one breath: Anonymize locally. Review by hand. Ask narrow questions. Demand quotes. Treat every output as a hypothesis for supervision — never a conclusion about a person.

Glossary

Term	Plain meaning
PII / ПДн	*Personally Identifiable Information* — any data pointing to a specific person.
De-identification	The *process* of removing or masking PII from a document — what CONFIDE does and measures.
Anonymization	The stronger *claim* that re-identification is no longer reasonably possible — a legal/statistical state CONFIDE never certifies.
Special-category personal data	The legal term (GDPR Art. 9; 152-ФЗ) for extra-protected data: health, sex life, beliefs, and similar.

Term	Plain meaning
PHI / psychotherapy notes	US HIPAA terms: PHI is protected health information handled by covered entities; *psychotherapy notes* are a narrow, separately-kept subset with extra protection.
Mask	Replace a piece of text with a placeholder, e.g. *Marina* → `[PERSON]`.
Direct identifier	Names a person almost alone: full name, email, phone, passport number.
Quasi-identifier	Harmless alone, identifying in combination: age + profession + city + a rare detail.
Re-identification / de-anonymization	Recovering who a person is from supposedly anonymous text.
Benchmark	A fixed, repeatable test for comparing methods fairly.
Gold standard	Human-marked "correct answers" a tool is scored against.
Red team	A group (here, an AI) that attacks a system on purpose to find its weaknesses.
Recall	Of all the PII really present, the fraction the tool caught. The safety metric.
Precision	Of everything the tool flagged, the fraction that was really PII.
F2 score	A recall-heavy combined score ($\beta = 2$: recall declared twice as important, giving it four times the weight in the formula).
Corpus	The collection of texts a tool is tested on.
Confidence interval (CI)	The range the true value plausibly falls in, given the data's size and noise.
AUC	Area under the ROC curve — 0 to 1; 0.5 means no better than a coin flip.
Metadata	Data *about* a file (headers, IDs, timestamps) rather than in its visible text.
Entity-level recall	An entity counts as safe only if *every* mention is masked.
Harm-weighted recall	Recall that counts each miss by how much harm it could do.
Regex	A precise text-matching rule (good for fixed shapes like emails).
NER	*Named Entity Recognition* — software that tags the "named things" in text.
LLM	*Large Language Model* — the AI behind chatbots; here, run locally.

Term	Plain meaning
Local-first / local inference	The model does its thinking on your machine; data never leaves.
Singling-out / linkability / inference	The three GDPR-named ways anonymization can fail.
k-anonymity	"How many people share this exact combination of traits?" If 1, singled out.
Dual-use	A capability that can help or harm depending on who uses it.
Citizen science	Research built and scrutinized by volunteers, in the open.