

Fx_{rl}

Research Collaboration Packet

Version: 1.5

Status: **Working Draft** (not to be shared without permission)

Date: August 2026

Prepared by: Functional Intelligence Research Lab (FIRL)

`firl.nl`

`colophon.firl.nl`

This is a working document, and it will keep changing as the project does. For current consent, ethics-review, and IRB status specifically, visit colophon.firl.nl/ethics rather than treating anything below as fixed, since that is the page we actually keep up to date.

This packet is an overview of what Functional Intelligence Research Lab currently studies, how Colophon and TWFF work, and what running a study together would actually involve.

1. FIRL

FIRL is a research group studying how people actually write when Generative AI (GenAI) is part of the process. Not whether they used it, but how: what they kept, what they rewrote, where a suggestion got accepted whole-sale and where it got argued with. That is a different question from the one most institutions are currently asking, and it is the one we think is actually answerable.

Colophon is the instrument we built to answer it. It is an open source Chrome extension that records the writing process: every edit, paste, and AI interaction as a log the author controls. We run it with real courses as case studies and publish what we find, to help academics produce better policy recommendations.

To see why this matters, it helps to look at what everyone else is doing.

2. The research problem

The dominant institutional response to generative AI (GenAI) in writing has been detection. Run a submitted document through a classifier or watermark detector to get the probability that GenAI was used. It is an appealing shortcut, but has been provably ineffective at best and harmful at worst. Detectors trained to guess authorship from finished text carry documented, structural biases, misclassifying non-native English writers at far higher rates than native speakers, and light rewriting is usually enough to evade them entirely. Worse, even when a detector is right, it gives back a number with nothing behind it. There is no trail a student can point to, no process a teacher can inspect, just a score to argue about.

We ask a different question: how was AI used during the work, not whether it was used at all?

Instead of reconstructing the writing process after the fact from the finished product, Colophon records it as it happens, timestamped and tamper-evident using the Tracked Writing File Format.

3. TWFF

TWFF, the Tracked Writing File Format, is the open specification behind what Colophon exports. It is a deterministic event log that includes session boundaries, edit blocks, copy-paste events, AI interactions, each one chained together with a hash so tampering after export is detectable rather than just theoretically possible to catch.

We have laid out the full argument for this approach, and for why detection-based alternatives fail structurally, in a working paper now available on SSRN:

papers.ssrn.com/sol3/papers.cfm?abstract_id=7272718

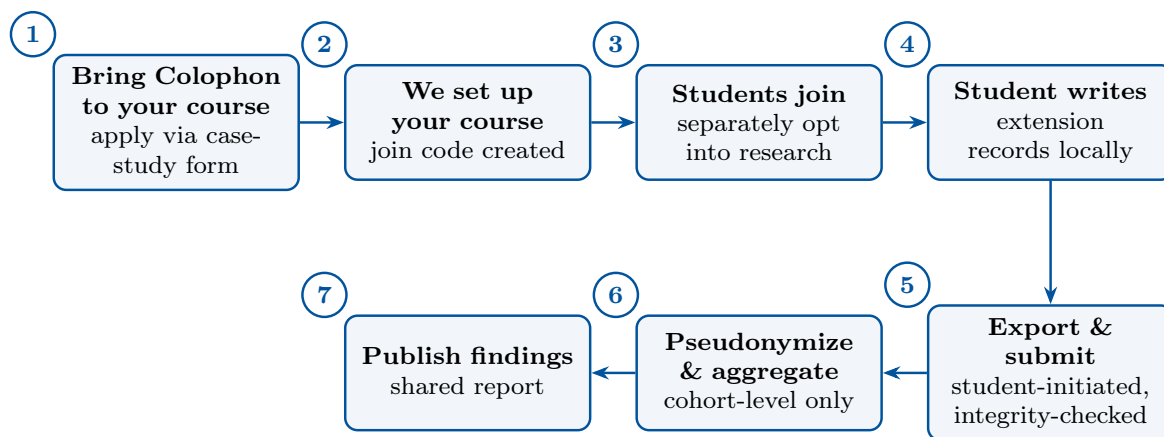
The format specification itself is publicly available, independent of Colophon or any single implementation:

github.com/Functional-Intelligence-Research-Lab/twff.

Reading the spec tells you what will be recorded in principle. What it looks like in an actual classroom is easier to show than describe.

4. Colophon

This is how a document's writing process actually gets from a student's screen to a researcher's dataset:



Full walkthrough, including illustrative sample data: colophon.firl.nl/demo.

5. Example study

Research question: How does students' use of generative AI change over a semester, and which observable behaviours provide evidence of meta-cognition?

`Edit_block` events give a proxy for engagement independent of GenAI use: how much work went into the document, how the pacing of writing changes over the term.

`AI-interaction` events give which model a student used, what they kept versus rejected, and, through a similarity score, roughly how much of the GenAI suggestion's original wording survived into the final text. Tracked across several assignments over a semester rather than one snapshot, the trend in `acceptance rate` and `kept-AI-share` starts to look like actual trajectory data.

Operationally, a declining share of AI text kept verbatim over the term, or a rising rate of substantially modified acceptances relative to outright accepted ones, would be a defensible signal worth describing in advance, before you have seen the data, since that is what keeps it a pattern in recorded behaviour rather than a story fitted to the results afterward.

This is one design among several the file format and data enable. The full field-by-field breakdown, including a handful of other worked research questions if this one is not the right fit for your course, is in the appendix at the end of this packet.

6. Research participation

Workload is usually the first real question a prospective collaborator asks, so here what actually happens on each side.

An instructor creates a course, a handful of fields, name and dates, and gets a join code immediately. If the course is going to be part of a research report, which is not required in order to use the tool day to day, tagging it as research-included means reading and confirming that Colophon must never be used to grade or police students. Past that, sharing the join code with the class is the last thing an instructor has to actively do; there is no ongoing data collection and no spreadsheet to maintain.

A student follows the join link, reads a short consent page, and agrees or declines; declining costs them nothing, not their grade and not their course access. From there they install the browser extension and write normally (Google docs is the only fully supported platform for now). Exporting and uploading the finished `.twff` file is currently a manual step on the student's end, and an instructor reminding students to actually do it at the point an assignment is due is realistically what makes a study collect real data.

What comes back to the researcher is a pseudonymized CSV export per case study, course-level counts and rates, downloadable on demand once the study is running, with minimal to no manual cleaning required.

Here is an illustrative excerpt of what that export actually looks like:

document	date	typed_pct	ai_gen_pct	paste_pct	ai_para_pct	primary_ai_use
Consumer Behavior Research	2025-05-12	0	53	18	29	paraphrase
Market Analysis Draft	2025-05-05	35	36	10	19	paraphrase
Sustainable Packaging Proposal	2025-04-28	84	8	4	4	grammar
Executive Summary	2025-04-14	1	51	20	28	paraphrase
Implementation Plan	2025-04-07	42	31	11	16	brainstorm
Literature Review Draft	2025-03-10	92	5	0	3	grammar

Columns shown are illustrative (composition shares and primary AI-use category per document); the full export also carries course- and cohort-level aggregates.

None of this works, without participants who actually understand and agree to what is being recorded.

7. Ethics, consent & data

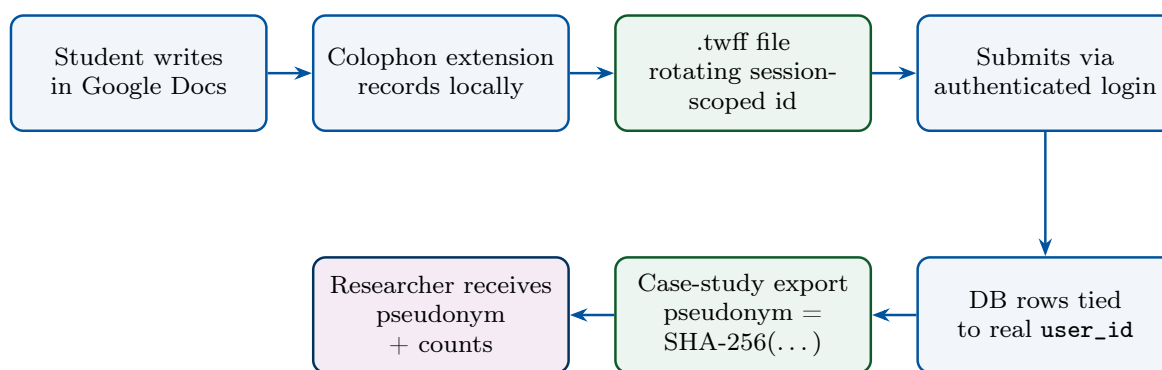
This section covers what’s actually collected, how consent works for both students and instructors, and exactly how pseudonymization protects participants.

Collected	Grouped edit changes, paste source , AI interaction model & acceptance category (modified or unmodified), periodic word-count checkpoints.
Not collected	Individual keystrokes, full prompts or AI outputs, screen recordings, mouse movements, biometric data, browsing activity outside the document.

What’s collected. Per writing session: grouped edit changes (a position, a character delta, and a backspace/delete count as a rough churn measure, never a keystroke-by-keystroke timeline), paste events (source and a short truncated preview, not the full pasted content), AI interactions (which model, a short output preview capped at 100–500 characters, an acceptance category, and a similarity score), and periodic word-count checkpoints.

Consent. Students and instructors give separate, individual consent, and it works differently for each. A student follows the join link, reads a consent page, and chooses *I understand and agree*, *No thanks*, or *Not now*; declining enrolls them in ordinary Google Docs access with no process logging, and never affects their grade, their course access, or their relationship with their instructor. An instructor separately agrees to include their course in a study, and confirms, as a required in-app step, that Colophon will not be used to grade, police, or surveil students.

Pseudonymization. The TWFF file’s own author identifier is a rotating, session-scoped value generated client-side, and by design is not built to support linking sessions together. Separately, a study-scoped pseudonym is computed only at the point data leaves the system for a specific study,



The file-level id (step 3) and the researcher-facing pseudonym (step 6) are deliberately different values with different lifetimes; the file's own id plays no role in computing the export pseudonym.

Withdrawal. A student or an instructor can withdraw at any point, without needing to give a reason. Withdrawal removes their data from any future export or analysis. Data already published in an aggregate report, where an individual's contribution is not separable from the cohort, cannot then be retroactively unpublished.

Retention. Data is retained for two years by default from last activity, configurable per institution, and erasable on request at any time by writing to hello@firl.nl, regardless of where that window is.

Risks

- Re-identification of a pseudonymized participant within a small class: pseudonymization protects against casual identification, not against someone who already knows which student wrote which document.
- Data breach affecting stored writing-process data: mitigated by authenticated-only access, and a tamper-evident hash chain on every submitted file, so a modified file after export is detectable rather than silently accepted.

There must be no risk to a student's grade or academic standing from declining or withdrawing

We are not going to claim a specific ethics-review status here to avoid the risk of it going stale the moment this document is out of our hands. colophon.firl.nl/ethics will be updated with information on where that stands right now. What we can say with more confidence is what a collaboration built on top of all this actually looks like.

8. Collaboration

FIRL provides the instrument (colophon), the web dashboard, pseudonymized data access, onboarding support, analysis support for the case study's data, and, at the end of a study, a published findings report. FIRL does not run a study independently of the researcher or instructor who leads it.

In return, we're asking for your course, your students' voluntary participation, and your own institution's ethics review and sign-off. Domain expertise FIRL does not have include curriculum design, pedagogical context, and what a given pattern in the data actually means for your students

Findings get published rather than held as a private asset, and the instructor or researcher who leads a study is a genuine collaborator on what comes out of it, not a data source we credit in passing. A study's lead gets to review findings before publication and raise concerns, though FIRL retains final editorial judgement on the data. Findings are not softened or reframed to suit a collaborator's preferred narrative.

If any of this sounds like something you would want to try with a reareains final editorial judgement on what the data actually showsl class, the fastest way in is to just write: Muhsin@firl.nl or hello@firl.nl.

Ordinary authorship or credit questions will be worked out directly between the collaborators involved, same as any research partnership. Allegations of plagiarism, improper AI use, or a claim that research data itself was faked or manipulated follow a separate, more formal process, disclosed to those involved if one is ever needed.

Demo

The full interactive pipeline walkthrough, with real (or seed) composition charts, is at <https://colophon.firl.nl>. The viewer, which opens any `.twff` file entirely client-side with nothing uploaded, is at <https://colophon.firl.nl/viewer>, including a sample file if you do not have one yet (see appendix).

Appendix: what the data actually looks like

A field-by-field summary of what Colophon captures today, organized by event type, since the public schema documents what's *allowed* and this is what's *actually there* in practice.

Table 1: TWFF event types

Event Type	Description	Key meta Fields
session_start	Start of a writing session	—
session_end	End of a writing session	—
edit_block	Human typing or deletion block	position_start, position_end, source
paste	Text pasted from clipboard	char_count, source, position_start, content_preview
paste_link	External or in-container link pasted	link_scope, position_start
image_upload	Image inserted into the document	filename, position_start
ai_interaction	AI assistant invoked, output reviewed	interaction_subtype, model, content_before, content_after, acceptance, similarity_score
ai_suggestion	Passive or autocomplete AI assistance	model, content_before, content_after, acceptance
chat_interaction	Multi-turn AI conversation	model, message_count, message_preview, source_file
focus_change	Editor focus lost or regained	direction, duration_ms
checkpoint	Periodic auto-save snapshot	char_count_total, word_count_total

Every event carries a timestamp, so within-session pacing and time-of-day patterns analysis, and cross-session analysis at the student level works through the account tied to each submission, not the file's own rotating identifier, which was deliberately not built for that purpose. A few worked examples of what these fields combine to answer: AI acceptance trends over a semester, editing behaviour as a proxy for engagement independent of AI use, AI reliance broken down by what a student actually asked for, and model preference where students have a real choice between tools. If your study needs something these fields do not obviously support, ask; some of what is possible depends on data that is already stored but not yet exposed through a self-serve export.

Technical appendix

- TWFF specification:
<https://github.com/Functional-Intelligence-Research-Lab/twff>
- TWFF preprint (SSRN):
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7272718
- Sample TWFF:
<https://colophon.firl.nl/viewer?sample=westinghouse>
- Colophon extension (Chrome Web Store): <https://chromewebstore.google.com/detail/colophon/manpdfemofccpcppkiepppbineggdnka>
- Public pipeline walkthrough:
<https://colophon.firl.nl/demo>
- Public ethics pages:
<https://colophon.firl.nl/ethics>