BLITZVORTRAG · DeGEval 2026, FRANKFURT

Workflow for Transparent Coding Coding in Evaluations

Susan Steiner · rowsquared

Many documents, and a definition not written for coding

Close to 30 GIZ projects around the world

Each with whatever it had produced:

Project design	Annual reports	Evaluation reports
Theory of Change	Presentation	M&E framework

One judgement per project: Are project approaches…

gender-transformative?

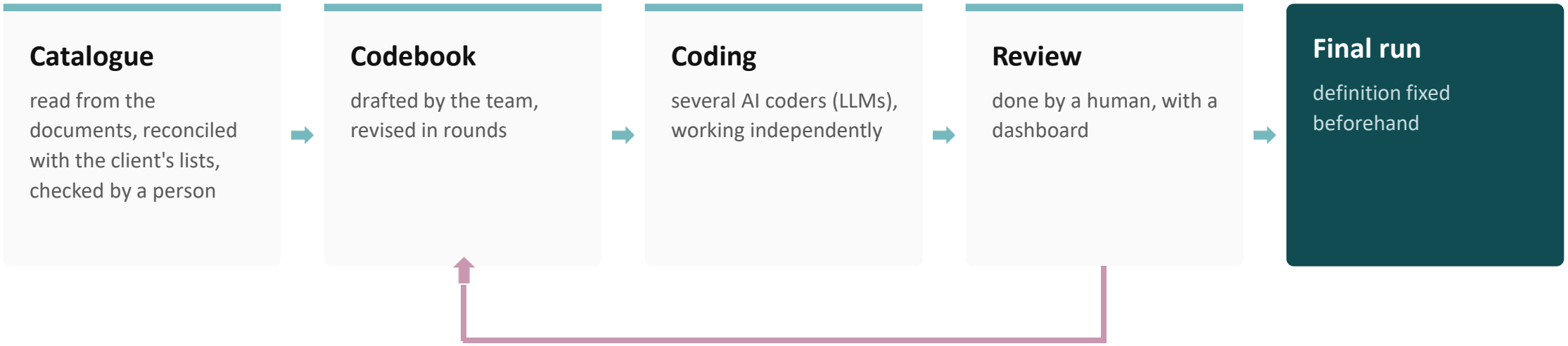
… with several sub-categories …

Definition ?

One existed — but never written with coding text in mind.

The workflow in one picture

The steps of a normal coding exercise, with one loop added at the end.



The catalogue

For every document: which programme or project, what kind of document, which country, which year, which language. With large volumes of documents, noone types that in.

What arrived

Endbericht_final_v3.pdf

Anlage_2.pdf

report_2019_final_FINAL.pdf

Jahresbericht.docx

…

What the machine built from it

Type · Sub-type	Programme · Project	Country	Year	Language
Evaluation · final	[Programme A] · [Project 1]	[Country]	2019	German
Annex · to design doc.	[Programme A] · [Project 2]	[Country]	2017	German
Annual report	[Programme B] · [Project 3]	[Country]	2019	English

type: read from the document

programme · project: reconciled with the client's lists

all: checked by a person

Housekeeping now — later it is what lets us look per programme, year or type of document, in order to tell the coders which kinds of document matter for which question and to see where the coding works less well.

The codebook, and a codebook you can revise

1 · The BMZ definition of gender transformation	*“The purpose of gender-transformative approaches is to bring about sustainable change in the gender inequalities that are the result of patriarchal power relations. To that end, gender norms and binarisms are critically analysed. The approaches aim to raise society's awareness of the drivers of inequalities in order, on this basis, to transform harmful norms, practices and stereotypes.”*
2 · The operational complement the team added	**Changes gender norms and power relations** (e.g. addresses the norm that assigns sole responsibility for unpaid care work to women). -> Distinguish from gender-responsive, which is about strengthening gender equality, for example by setting meetings at a place accessible to women and from gender-sensitive, which is about being aware of gender inequalities, for example by acknowledging that women and men have different time availabilities.

A code added in round two: done or promised — because we noticed we were partly judging interventions on their promises. Re-coding everything took a night. By hand it would have meant starting over.

Coding: not every page, and not by one coder

By structure. A question is only asked of the parts of a document that could answer it — for example, a recommendation is not a finding.

By substance. Each passage is asked whether it speaks to the question (addressing gender equality), or only mentions the topic (gender or women).

Then coded by several AI coders independently, each with the codebook and the catalogue entry in front of it.

Coded

“Our programme promotes gender equality as a fundamental human right and an essential to improve women's access to land.”

“Das "Vätergespräch" ist eine der Aktivitäten im Rahmen der Ernährungssensibilisierungskampagne, um eingefahrene geschlechtsspezifische Rollenkonzepte aufzuweichen.”

speaks to the question

Set aside

“Women contribute 78% of the workforce in agriculture.”

mentions the topic only

Where the human comes in

The reference point: two experts code a sample blind and independently first, so we know how well people agree on the material. Everything else is read against that.

The AI coders are treated as a panel of coders. Their agreement decides one thing only: **who looks at what**. It is never a claim that the answer is right.

Settled

The coders agree and the evidence is clear.

Sent to a person

The coders disagree, or the evidence is thin (for example, only one relevant text passage on an intervention). A person decides, passage by passage — seeing the passage, the coders' reason, the labels and the votes.

Checked at random

A sample of the settled cases is always reviewed too — otherwise the mistakes among the confident cases are never found.

What changes, and where it breaks

	Manual coding	AI-assist in coding software	This workflow
Who writes the codebook	The team	The team; the tool joins at coding, not at defining	The team, with the machine proposing wording, edge cases and positive and negative examples, each one accepted or rejected by the team
Who codes	One or two human coders, a sample double-coded if the budget allows	One pass by one coder	Several AI coders independently, then a person on the contested cases
Can it be revised	Frozen early — revising is unaffordable	Mostly no revison; if done, previous version lost	In rounds; every version kept
What the review looks at	A list of passages	No way to check what it did	Dashboard per programme and project, then the passage
What can be traced back	The coder's memory	The output	Passage, codebook version, decision — read against human coding

Where it breaks · Graded judgements are where the AI coders (like human coders) are weakest.

Scanned documents and tables are read badly — that can manufacture an absence of evidence.

A definition shaped by one set of documents may not fit the next.

The machine screens and argues. The evaluator defines and decides.