Limitations, uncertain cases and possible improvements

Registration section 6, item 5, which asks for this as its own deliverable rather than folded into the description document. `docs/description-document.md` section 6 points here; the substance is not in both places, because a limitation written twice is a limitation that drifts.

Every number below is quoted from a committed artefact, and `scripts/test_reported_figures.py` fails if any of them stops matching. The sources are `data/pipeline/evaluation-100-metrics.json` and `data/corpus/figures.json`. The worked examples in section 2 are read from the committed mapping table `data/corpus/mapping-table-0246412.csv`, so a reader can find every one of them without regenerating anything.

Written under `cordis-sdg-registration-v10.md`, sha256 `c5f2a26ce6049e801a7e757efc2cbe4253689987fad761f1d97cfb5b0ad68fb2`. That file is committed at `docs/cordis-sdg-registration-v10.md`, so this line can be checked rather than taken.

This document does not argue that the method is better than its numbers. The registration set no acceptance threshold, so a poor figure is reported with its explanation rather than with a defence.

---

1. Limitations

1.1 Recall is bounded at 0.442 by the method, and the bound is measured not estimated

The ceiling is what recall would be if every threshold were relaxed until any evidence at all counted. It is computed by asking, for each reference pair, whether the project's text contains *any* phrase the vocabulary maps to that target.

recall ceiling	**0.442**
with no evidence at any threshold	**53**
with some evidence	42

53 of 95 pairs are not recoverable by this method at any threshold. No tuning reaches them, because the link the labeller made is not present in the project's words at all. **This is the finding, not a defect to be apologised for**: it measures how much of the SDG-to-project relation is inference rather than vocabulary, and the answer is most of it.

1.2 No human labelled anything in this study

Every label in this study, on both sets, was made by a language-model seat. The registration provided for exactly one human labelling contribution, 20 of the held-out 100 hand-

labelled by the owner as an anchor, and **it was not made.**

Where this project reports agreement between two labellers, it means two model seats agreed. Where it calls a set the reference, it means a model's output is standing in for ground truth. The evaluation labeller was served by a different model family from the seat that built the pipeline, which weakens but does not remove the shared-priors concern, and does nothing at all about the fact that **no person read a project and said what it was about.**

1.3 The evaluation reference is one automated labeller, working blind

Two consequences follow, and each removes a figure rather than weakening one:

- **There is no inter-rater agreement figure on the evaluation set.** One labeller cannot disagree with himself. **No substitute is offered**, because comparing the labeller to the pipeline would be the pipeline's own accuracy reported a second time under a name implying two humans.
- **There is no human anchor**, so nothing in this study calibrates the model labellers against a person.

The adjudication rule was fixed in advance, before any number was in view: for the 20 the owner labelled, his call settles a disagreement; for the other 80, the model labeller's labels are the reference. **With no 20, that labeller is the sole reference for all 100** and this limitation covers the whole hundred rather than 80 of it.

The window for making the 20 honestly has closed. The registration requires an evaluation labeller not to see the pipeline output before sealing, and that output's hash is published. Twenty labels made now would carry that exposure as a stated limitation of their own; they would not repair this one.

1.4 Nearly a third of the corpus receives nothing

7,175 of 23,451 projects get no assignment at all, and 33 of the 169 targets are never assigned anywhere in the corpus. A further 6,882 projects reach a goal but no target, which the mapping table marks explicitly rather than leaving blank.

1.5 Most assignments sit in the weakest band

Of 14,776 target assignments: **9,228 low, 2,923 medium, 2,625 high.**

The band measures how much evidence was found, not whether the assignment is right. Section 2.1 below is a worked example of a wrong assignment carrying the **high** band, and that is the band behaving as defined rather than misbehaving. A reader who treats the band as a correctness score will be misled, so it is said here plainly.

1.6 Four fifths of the category vocabulary is deliberately unmapped

199 EuroSciVoc categories carry a crosswalk row and **87 of the 169 targets have none. That is the result, not a gap to be filled.** A category such as applied mathematics has no defensible one-line justification against any target's text, and inventing one would produce exactly the superficial mapping the competition's criteria penalise. A project in an unmapped category is not

thereby unmapped: it reaches targets through the lexical evidence that reads the project's own words.

A second thing about the crosswalk, which is a defect rather than a design choice. `score()` adds every crosswalk row for every category a project carries, and only `contributing` rows are held to one per target. So a project carrying a category **and one of its own descendants**, where both map to the same target, scores that target twice from a single classification.

mapped EuroSciVoc categories	**199**
ancestor pairs reaching the same target	**46**
of those, pairs that would sum if a project carried both	**41**
pairs the contributing rule already neutralises	5
target assignments in the mapping table reached by such a pair	**0 of 14,776**
projects carrying a category and one of its own descendants	**0 of 23,451**

The zero is measured, not assumed, and it is not an artefact of sparse classification. 20,161 projects carry a category list and 14,562 carry more than one. **CORDIS classifies with the most specific term in a branch and does not also list the ancestors**, so the pairs exist in our crosswalk and never meet in a project.

No row of the submitted table is inflated. The reason is a property of CORDIS's classification, not of our rule. `score()` has no guard against an ancestor pair, and a snapshot that listed a parent beside its child would double count silently with nothing in the pipeline to say so. **That is why this is recorded here rather than fixed under a re-freeze**: no output byte moves, so a re-freeze would change nothing a reader could check, while the exposure is real and belongs on the record.

Both halves are re-runnable rather than asserted. `scripts/measure_ancestor_exposure.py` refuses to report until it has rebuilt all 14,776 scores and bands from the rows' own evidence cells and matched the committed columns. `scripts/measure_ancestor_structure.py` refuses until it has detected a planted ancestor pair, because a zero from an instrument nobody has seen find anything is indistinguishable from an instrument that cannot.

1.7 The explanation pass is parked

No model credential is set in the build environment and the registration forbids one in the repository or the room, so the pass records its prompts with their hashes and **writes nothing that pretends to be a model's work.** The `explanation` and `prompt_sha256` columns say so on every row rather than being blank.

1.8 Scope

Horizon Europe only. FP7 and Horizon 2020 are excluded for size, and the notebook says how the method extends to them. **No personal data is read**; participating organisations and their countries are, because an organisation is not a person.

---

2. Uncertain cases, with the evidence that produced them

The disagreement has a shape, and the shape is lopsided.

projects where pipeline and reference disagree	**64**
recall-only, the pipeline found nothing the labeller found	**46**
precision-only, the pipeline found something the labeller did not	**4**
both directions	14

The failure is silence, not noise, by an order of magnitude: the pipeline far more often finds nothing the labeller found than finds something the labeller did not. The counts are in the table above rather than repeated here. Four mechanisms account for what is there, and each is shown below with a real project, its assignment, and the evidence cell the pipeline recorded.

2.1 A word in the wrong sense, at the highest confidence

Project 101113215, CoFutures Literacy. The pipeline assigns **target 4.6**, adult literacy and numeracy, and the mapping table records `4.6 score 8.0 band high` for it. The reference assigns nothing.

The evidence cell records the phrase `Literacy`, matched in the title, the objective and the editorial description, as both a key term and a subject synonym. **Every one of those matches is real, and the assignment is wrong.** "Futures literacy" is a term of art for the capacity to imagine alternative futures; it is not reading ability. **The pipeline has no sense disambiguation and cannot acquire one from counting**, and because the phrase occurs in three fields it accumulates evidence from each, which is what pushes a wrong assignment into the strongest band.

2.2 The same word across two domains

Project 101203848, MORFE, on the molecular regulation of iron. The pipeline assigns **target 5.2**, violence against women and trafficking, on the phrase `trafficking` found in the title and objective, and the mapping table records `5.2 score 5.5 band medium` for it.

In this project the word is *intracellular iron trafficking*, molecular transport. **The phrase is a correct string match and a category error**, and no amount of context window inside the matching rule would fix it, because the rule is a literal match with word-boundary lookarounds by design, and that design is what makes the run reproducible and auditable.

2.3 A category crosswalk carries the whole category

The same project is also assigned **target 2.2**, malnutrition, from a single piece of evidence, and the mapping table records `2.2 score 3.0 band low` for it. That evidence is its EuroSciVoc category `medical and health sciences/health sciences/nutrition`, via a direct crosswalk row. The reference instead assigns 2.4, sustainable food production.

The crosswalk is deliberately coarse. It maps a category, so every project in that category inherits the link. For a molecular study of iron regulation filed under nutrition, that produces a defensible-looking assignment from no project-specific evidence whatsoever. The low band is the signal that this happened, and the evidence cell names the mechanism rather than hiding it.

2.4 Generic institutional language

Project 101162110, NIGHTMADRID, a public-engagement night for researchers. The pipeline assigns **target 4.3** on `vocational`, `university` and `vocational education`, and **target 11.2**, public transport, on `children` and `public transport`. The reference assigns nothing.

Outreach and mobility boilerplate appears in project descriptions for administrative reasons, and the pipeline cannot tell a project's subject from its logistics.

2.5 The silent majority, evidence below threshold or none at all

Two examples stand for the 46:

- **Project 101111215, ALIVE**, a non-invasive liver-disease diagnostic platform. The pipeline finds evidence at **goal** level and nothing above the target threshold, so the table records `goal only, score 2.4` for it. The reference assigns target 3.8, universal health coverage. **The pipeline was looking in the right goal and could not reach the target.**
- **Project 101208351, TNT**, on neutrophil extracellular traps and glial interferon. The evidence cell reads `(no phrase matched)` and the assignment level is `none`. The reference assigns target 3.4, premature mortality from non-communicable diseases.

Neither project's text contains target vocabulary. Reaching 3.4 from neuroinflammation requires knowing that the disease class is non-communicable, which is domain knowledge and not a phrase. **This is section 1.1's ceiling seen one project at a time.**

---

3. Possible improvements

Each is named with what it would fix, what it would cost, and what it would not fix. None was adopted, because the registration fixed the method before any number was in view, and moving it after seeing the numbers is the thing this project exists not to do. They are the next registration's business.

3.1 Sense disambiguation on a small, named list of ambiguous phrases

Would fix sections 2.1 and 2.2, the homonyms, which are visible in the four precision-only cases and in most of the fourteen two-way ones.

Shape: a phrase like `literacy` or `trafficking` earns its match only when a second phrase from the same target is also present, or is refused when a listed domain phrase co-occurs. This is the existing negation-pattern mechanism pointed at ambiguity rather than at explicit denial, so it costs a rule file and no new machinery.

Would not fix recall, which is 46 of the 64 disagreements. **An improvement that addresses four cases out of sixty-four should be described as what it is.**

3.2 Lift the goal-level cases to target level

Would fix part of section 2.5, projects like ALIVE where the goal is already right.

Shape: among the 6,882 goal-level-only projects, the goal narrows the candidate targets from 169 to a handful, and a second pass could score only those, at a threshold set on the development set.

Measurable in advance: the ceiling computation already says which reference pairs have any evidence, so the gain is boundable before the pass is written.

Would not fix the 53 pairs with no evidence at any threshold.

3.3 Model-assisted assignment, evaluated against the same held-out 100

Would fix the ceiling itself, which is the only thing that can. The gap is inference, and inference is what a model does.

Cost, and it is the real one: it replaces an auditable decision path with one that cannot be reproduced from a hash. The current pipeline is pure standard library, so a judge can rerun it and get the same bytes. **That property is worth more to this submission than the recall it costs**, which is why the explanation pass is a separate, parked, clearly marked stage rather than part of the decision path.

3.4 A human anchor, made properly and in advance

Would fix section 1.2, which no amount of engineering touches.

It cannot be retrofitted to this run for the reason in section 1.3: the pipeline output is published. It belongs in the next registration, before any output exists, with the labeller's sequence fixed in advance.

3.5 Widen the vocabulary from the projects rather than from the taxonomy

Would fix part of the ceiling, and is the one improvement the data already supports.

Shape: the key terms are extracted from target text under a written rule. A second source, phrases that recur across projects the labellers linked to a target, would add the words practitioners use rather than the words the United Nations used.

The hazard is exactly the one this project is built to avoid: fitting vocabulary to labels is fitting the method to the evaluation. It would have to be built from the development 50 only, with the held-out 100 untouched, and the registration would have to say so before the first phrase was added.