Friday, April 17, 2026

TReC 2026 Project Proposal Submission Form

Submit your project proposal for the 7th TRAIL Research Camp (August 24th - September 4th, 2026, Lausanne, Switzerland). Please complete all required sections and submit your proposal before April 30th, 01:00 PM (CET).

Administrative Data

Full Name of Team Leader Oleh Bohatov

Contact Email oleh.bohatov@ulb.be

Project Information

Project Title From Misrepresentation to Recourse: Verification, Traceability, and Correction of LLM Representations of Organizations

Profile of the Team Leader(s) & Expected Team Composition

Oleh Bohatov holds a Specialized Master in Data Science & Big Data from ULB (September 2025). He combines training in data science with more than 20 years of experience in journalism and communications, which gives him a strong practical interest in how AI systems describe real organisations and shape public understanding.

He is currently developing VeriHub, a practical initiative focused on auditing and improving how public LLMs represent organisations in multilingual contexts, especially where inaccurate or inconsistent outputs can create reputation, informational, or service-related risks. This proposal builds on that experience, but is framed as a research-first and interdisciplinary project rather than a startup pitch.

The project is submitted as an open proposal seeking collaborators across TRAIL partner universities. The ideal team would combine three complementary profiles: a technical profile in NLP / LLM evaluation, a socio-technical or governance-oriented profile in areas such as law, ethics, HCI, policy, or STS, and an applied domain profile able to anchor the work in a concrete public-interest sector such as public administration, health, finance, mobility, or media.

The goal is to build a small interdisciplinary team able to connect technical experimentation with questions of accountability, usability, and real-world organisational impact.

Have you already identified potential team members for your project? No

Domain of Application Education & Public Policy

Scientific Theme Trustworthy AI

Proposal Content

Abstract

This project investigates how AI-generated representations of organisations can be made more understandable, contestable, and correctable. At the intersection of LLM evaluation (STEM) and

accountability/information governance (SSH), the team will design and test methods and protocols that support transparent claims, traceability, and correction workflows. The project is intentionally interdisciplinary and open to collaborators across TRAIL partner universities, with the goal of delivering a validated prototype and reusable artefacts -evaluation protocol, benchmark, and TRAIL Factory brick - with clear societal and organisational impact.

Background Information & Problem Statement

LLMs are increasingly becoming an important interface through which citizens and customers access information about organisations, including services, procedures, deadlines, and contact points. Yet for the organisations themselves, this creates a structural blind spot: they invest in accurate official communication but often lack systematic visibility into how public LLMs represent them, any reliable correction mechanism, and any clear way to measure whether interventions work.

The problem operates on three interconnected levels.

The technical level: LLM outputs are non-deterministic, vary by language and query phrasing, and resist systematic audit. Existing benchmarks address general factual accuracy or reasoning; the specific problem of organisational information accuracy - tested across languages, tied to official sources, with correction feedback - remains a gap in the evaluation literature (Bommasani et al., 2021; Weidinger et al., 2022).

The structural level: LLM providers have commercial and editorial interests that shape outputs in ways that are proprietary and opaque (Bender et al., 2021).

The governance level: the EU AI Act introduces transparency and accuracy obligations, but practical correction mechanisms at the organisational level remain undefined (European Parliament, 2024). The concept of recourse — the right of affected parties to understand, contest, and seek correction of automated outputs — provides the normative foundation for this project's intervention design (Wachter et al., 2018)

In Belgium, FR/NL language drift (the same question yielding materially different answers by language) adds a further dimension of structural inequality, documented in VeriHub's existing audit data across Belgian service organisations.

This project addresses that gap through a practical pilot-case methodology. Rather than discussing black-box AI in the abstract, it asks a concrete question: **how can an organisation understand, verify, trace, challenge, and improve the way it is represented in one or several LLM systems?**

Project Objectives & Concrete Implementation

The central aim is to develop and test a minimum viable framework for improving the way a pilot organisation is represented in one or several LLM systems.

Objective 1 — Baseline diagnosis
The team will document how the pilot organisation is represented across selected public LLMs and prompts, identifying issues such as factual inaccuracies, outdated information, unsupported claims, entity confusion, omissions, or overconfident phrasing. This phase will establish the baseline that an ordinary user might encounter when asking public-facing LLMs about a real organisation.

Objective 2 — Verification and traceability protocol
The project will design a small but reusable protocol for checking organisation-related AI claims against trusted evidence, categorising likely failure modes, and identifying which parts of the output are externally actionable. The goal is not only to detect that an answer is wrong, but also to understand how that wrong answer can be verified, traced, and explained.

Objective 3 — Test correction pathways
The team will compare several realistic intervention strategies, depending on team capacity and technical feasibility. These may include source restructuring, schema.org markup, FAQ reformulation, multilingual content alignment, evidence-grounded prompting, structured correction prompts, and optional human-in-

the-loop validation.
The key question is not only whether outputs improve, but **which interventions are realistically available to an external organisation and which ones work best in practice.**

Objective 4 — Practical correction road map
The project will produce a case-specific road map showing:

- which failure modes were observed;
- which claims could be verified and traced;
- which interventions were tested;
- which interventions improved outputs;
- which problems appear model-specific or remain difficult to correct externally.

Objective 5 — Reusable TRAIL Factory brick
The pilot-case lessons will be translated into a reusable workflow for verifying, tracing, and correcting problematic LLM representations of organisations. This workflow will include an evaluation protocol, an error taxonomy, an intervention matrix, and a practical recourse template.

Expected outputs
By the end of the camp, the project aims to deliver:

1. a baseline evaluation of one pilot organization across selected LLMs;
2. a verification / traceability protocol;
3. a misrepresentation taxonomy;
4. a comparison of correction strategies;
5. a practical correction roadmap for the pilot case;
6. a reusable workflow / prototype suitable for further development as a TRAIL Factory brick.

Do you plan to deliver, as an outcome of your project, a reusable "brick" for the TRAIL Factory (https://factory.trail.ac/en/home_page) that could later be transferred and converted into a company process?

Yes

Briefly describe what the brick would be and its intended users.

A reusable audit-to-correction workflow for problematic LLM representations of organisations. The brick would combine a compact evaluation protocol, an error taxonomy, a verification and traceability layer, and an intervention matrix to help diagnose misleading outputs, test correction strategies, and produce a practical recourse road map.

Intended users:

- **Public and private organisations** (e.g., public administration, service providers, NGOs) seeking practical mechanisms to audit, challenge, and improve how they are represented by public AI systems.
- **AI auditors and researchers** needing a grounded, real-world methodology to evaluate organisation-level accuracy and mitigate hallucinations.
- **Governance and compliance professionals** looking for actionable frameworks to bridge the AI accountability gap and align with EU AI Act transparency obligations.

Project Dataset

The dataset is designed for a two-track implementation, with a standalone public-data track as the guaranteed baseline and a partner-supported pilot case as an optional extension.

Track A - Public data (baseline)
Publicly accessible Belgian service organisations in sectors such as public administration, health, mobility, finance, insurance, or utilities. Official websites, FAQs, reports, contact pages, and public statements will serve as attributable reference sources. This track is operational from Day 1 and does not depend on external data access agreements.

Track B - Partner-supported pilot case (optional extension)
If a non-affiliated public-interest organisation is identified with TRAIL partners during scoping, the same audit-to-correction protocol will be applied to that case in greater depth, using its official materials and validated factual ground truth.

Across both tracks, the dataset will contain five layers:

1. a prompt set of organisation-related user queries;
2. a reference layer built from trusted and attributable sources;
3. a model-output layer from selected LLMs under different conditions;
4. an annotation layer covering failure type, severity, verifiability, traceability, actionability, and correction status;
5. an intervention layer documenting which correction strategy was applied and whether it improved the representation.

Existing VeriHub examples and multilingual question-set templates may be used as seed material for protocol design, but the camp outputs will be generated and validated within the project itself. A curated and anonymized subset of prompts, outputs, annotations, and interventions may be prepared for release as a compact benchmark for future multilingual LLM recourse research. No sensitive personal data is required.

Detailed Work Plan

(10 working days, weekends excluded)

Days 1–2
The team will finalise the project scope, confirm the pilot-case selection strategy, define the question set and annotation rubric, and prepare the initial source collection. By the end of this stage, the project should have a compact multilingual benchmark design and a clear baseline evaluation protocol.

Days 3–5
The team will run the baseline audit across selected public LLMs, classify errors, and map likely causes. In parallel, the SSH component will analyse accountability and governance dimensions, including which kinds of misrepresentation are externally actionable and which remain provider-dependent. This stage will produce an issue map, a cause-mapped error taxonomy, and an initial accountability-gap analysis.

Days 6–8
The team will design and test correction interventions, such as structured source improvements, FAQ reformulation, schema markup, multilingual content alignment, and other practically realistic changes. Outputs will be re-evaluated after each intervention in order to measure whether the representation improves and under which conditions.

Days 9–10
Findings will be synthesised into two final outputs: first, a practical correction road map for the pilot case; second, a reusable audit-to-correction methodology, including protocol, taxonomy, and intervention logic, suitable for further development as a TRAIL Factory brick.

Minimum viable deliverable by the end of the camp

- one validated pilot-case analysis;
- one benchmark with prompts, evidence, and annotations;
- one tested verification / traceability protocol;
- one comparison of correction pathways;
- one reusable artefact suitable for further development as a TRAIL Factory brick.

Bibliographic References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Proceedings of FAccT '21, 610–623.

Bommasani, R., et al. (2021). *On the Opportunities and Risks of Foundation Models.* arXiv:2108.07258.

Ji, Z., et al. (2023). *Survey of Hallucination in Natural Language Generation.* ACM Computing Surveys, 55(12), 1–38.

Mökander, J., & Floridi, L. (2022). *Auditing AI Systems — A Problem Statement.* Ethics and Information Technology, 24, 33.

Raji, I. D., et al. (2020). *Closing the AI Accountability Gap.* Proceedings of FAccT '20, 33–44.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.* Harvard Journal of Law & Technology, 31(2), 841–887.

Weidinger, L., et al. (2022). *Taxonomy of Risks Posed by Language Models.* Proceedings of FAccT '22, 214–229.

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act).* Official Journal of the European Union.

Eligibility & Evaluation

Does the project include multidisciplinarity between STEM & SSH?

Yes

How?

This project is explicitly interdisciplinary. **On the STEM side**, it develops and tests a practical evaluation pipeline for LLM representations of organisations: structured prompting, multilingual benchmarking, verification against trusted sources, error taxonomy design, traceability analysis, and comparison of correction interventions such as source restructuring, schema markup, FAQ reformulation, and retrieval-informed support. **On the SSH side**, it examines the same problem through the lenses of accountability, information governance, usability, and recourse: who is affected by misleading outputs, what kinds of errors are socially or institutionally significant, which interventions are externally actionable, and how organisations can realistically challenge problematic AI-generated representations.

The bridge between STEM and SSH is central to the project design. Technical evaluation alone can show that an output is wrong, but it does not explain why that matters, for whom it matters, or which correction pathways are meaningful in practice. Conversely, governance and ethics discussions often remain too abstract without concrete experimental evidence. This project connects both perspectives by using one pilot case to study not only whether public-facing LLMs misrepresent an organisation, but also how such misrepresentations can be verified, traced, interpreted, and improved in ways that are technically grounded and socially useful.

We confirm that the Team Leader will be present for the full duration of TReC'26 if the project is selected (August 24th - September 4th, 2026, Lausanne, Switzerland)

I/We agree and confirm