TrustLabel.AI

Trust Labels for Scientific Knowledge through Hybrid AI

Programm / Ausschreibung	DST 24/26, DST 24/26, AI Ökosysteme 2025: AI for Tech & AI for Green	**Status**	laufend
Projektstart	01.09.2026	**Projektende**	31.08.2029
Zeitraum	2026 - 2029	**Projektlaufzeit**	36 Monate
Projektförderung	€ 674.605		
Keywords	Hybrid AI; Trust Labels; Visual Analytics; Research Knowledge Graph; Climate Communication		

Projektbeschreibung

Rasant wachsende Publikationsmengen und KI-getriebene Kommentare verschärfen wissenschaftliche und gesellschaftliche Herausforderungen. Wo verlässliche Analysen vorliegen, können Entscheidungsträger Trade-offs abwägen und Lösungen entwickeln, die dem Gemeinwohl dienen und die grüne Transformation beschleunigen. Fehlen glaubwürdige Informationen zu Klima und anderen Nachhaltigkeitsthemen, füllt Desinformation das Vakuum und demokratische Prozesse leiden. Zugleich beschleunigt KI das Publizieren, untergräbt Qualitätskontrollen und überfordert Forschende sowie Praktiker:innen, die für Entscheidungen auf gesicherte Erkenntnisse angewiesen sind.

TrustLabel.AI begegnet diesen Problemen, indem es den Zugang zu vertrauenswürdigem, nach FAIR-Prinzipien veröffentlichtem Wissen mit nachvollziehbarer Herkunft erleichtert. Es führt erklärbare „Trust Labels“ für wissenschaftliche Aussagen ein – analog zu Nährwertangaben. Auf Ebene einzelner Behauptungen zeigen die Labels zentrale Glaubwürdigkeitsindikatoren: Stärke und Art der Evidenz, methodische Qualität, Aktualität, Konsensgrad sowie Herkunft. So wird Verlässlichkeit auf einen Blick beurteilbar, mit klarem Pfad zu den Quellen. Die Labels helfen damit auch, falsche Ausgewogenheit zu vermeiden und aus dem Zusammenhang gerissene Aussagen von riskanten Entscheidungen fernzuhalten. Durch einen hybriden KI-Ansatz, der Bewertungen berechen- und erklärbar macht, fördert TrustLabel.AI offene, verantwortliche Wissenschaft.

Das Projekt verbindet die Flexibilität großer Sprachmodelle (LLMs) mit symbolischem Schließen über einen “Research Knowledge Graph” (RKG). Isoliert eingesetzt können LLMs halluzinieren, Bias übernehmen oder Vorgaben missachten. Dem begegnen wir mit Retrieval-gestützten Prompts, Guardrails und Konsistenzprüfungen mittels RKG. Dieser erfasst granulare Aussagen publikations- und quellenübergreifend. Die nachvollziehbare Herkunft der Daten stellt sicher, dass Preprints und graue Literatur anhand transparenter Qualitätskriterien bewertet werden können, bevor sie in Labels oder Zusammenfassungen einfließen.

Offener Zugang und Usability sind Kernprinzipien von TrustLabel.AI. Vorgesehen sind drei komplementäre Zugangsmechanismen: (1) ein Browser-Plugin für In-situ-Prüfung von Behauptungen, (2) ein interaktives Dashboard für Experten zur Analyse wissenschaftlicher Inhalte und der Beziehungen zwischen den einzelnen Aussagen, sowie (3) skalierbare Datenschnittstellen zur Nutzung in externen Anwendungen. Die Werkzeuge werden mit Wissenschaftler:innen, NGOs und Industrievertreter:innen co-designt und evaluiert, um reale Bedürfnisse zu adressieren, Wissenstransfer zu fördern und eine breite, sektorübergreifende Nutzung zu ermöglichen.

Das Projekt folgt den Prinzipien vertrauenswürdiger, nachhaltiger KI und wird auf strikte Konformität mit EU-AI-Act und DSGVO achten. Ein “Human-in-the-Loop” Ansatz wird in zwei Anwendungen verfolgt: Im ersten Use Case erhalten Forschende die Möglichkeit, den gesamten vertrauenswürdigen Informationsraum zu analysieren und zu visualisieren. Im zweiten Use Case unterstützt das System den Klimawandel-Wissenstransfer zu Unternehmen, NGOs und dem öffentlichen Sektor. Verdichtete, belastbare Aussagen reduzieren die Informationsüberlastung, schaffen einen effizienten Zugang zu verlässlichen Inhalten, stärken demokratischen Prozesse, und fördern evidenzbasierte Entscheidungen für die grüne Transformation.

Abstract

In today’s scientific and civic landscape, surging research outputs and AI-driven commentary intensify societal challenges. When reliable analyses are available, decision-makers can evaluate trade-offs and craft solutions that serve the public good and accelerate the green transition. When credible information about climate change and other sustainability issues is scarce, disinformation fills the vacuum and democratic processes suffer. The problem of identifying high-quality content is compounded by AI-accelerated publishing, which challenges quality control processes and leads to information overload for researchers and professionals who rely on scientific evidence for their decisions.

TrustLabel.AI aims to address these challenges by facilitating access to trusted knowledge, published under FAIR principles (Findable, Accessible, Interoperable & Reusable). It introduces explainable “trust labels” for scientific claims, akin to nutrition labels on food. At the level of individual assertions within articles, each label communicates credibility indicators (strength and type of evidence, methodological soundness, recency, degree of consensus, provenance). Reliability can thus be judged at a glance, with a transparent path to sources. Trust labels can help counter false balance and prevent an out-of-context presentation of findings. A Hybrid AI approach will make credibility assessments computable and explainable, enhancing resilience against such disruptions while promoting open and accountable science.

The project will fuse the language understanding capabilities of Large Language Models with symbolic reasoning over a provenance-rich Research Knowledge Graph (RKG). Used in isolation, LLMs can hallucinate, inherit biases, or overlook domain constraints. We mitigate these risks by grounding generation via retrieval-augmented prompts, guardrails, and consistency checks on the RKG. The RKG records fine-grained provenance to ensure that preprints and grey literature are evaluated against transparent quality criteria before being included in any label or summary.

Open access and usability are central pillars of TrustLabel.AI. The project will provide three complementary interfaces: a browser plug-in for in-situ claim checking, an interactive dashboard for professionals to explore the underlying evidence and claim relationships, and scalable data interfaces to enrich external applications. These tools will be co-designed and evaluated in collaboration with scientists, NGOs, and industry representatives to ensure the solution addresses real-world

needs, promotes knowledge transfer, and fosters broad adoption across sectors.

The project will adhere to the principles of trustworthy and sustainable AI, strictly compliant with the EU AI Act and the GDPR. Human experts remain in the loop and will be involved through two use cases. The first one will integrate trust labels into research data services to support scientists with a provenance-rich view of the entire trusted information space. The second use case targets climate knowledge transfer to NGOs, policy makers and the corporate sector, providing distilled, defensible evidence to inform strategy, compliance, and risk management. By reducing information overload and providing equitable access to credible scientific content, TrustLabel.AI enhances democratic deliberation and accelerates evidence-based decisions, which in turn support the green transition.

Projektkoordinator

- webLyzard technology gmbh

Projektpartner

- Modul University Vienna GmbH
- Universität Wien
- ORCA Evolution OG