Utrecht University

Department of Information and Computing Science

Applied Data Science master thesis

Expanding the HORIZON
The Strategic Repurposing of Europe's Research Framework

First examiner:
Dr. P. (Maxigas) Dunajcsik

Second examiner:
Dr. Dennis Nguyen

Candidate:
Ruben Swarts

In cooperation with:
Open Future

July 7, 2025

The Strategic Repurposing of Europe's Research Framework

April 17, 2026

1 Abstract

This thesis investigates whether the EU's Horizon Europe program, with its €95.5 billion budget, is undergoing a strategic rhetorical shift, particularly concerning its climate-related research portfolio, in light of growing geopolitical instability and calls for 'strategic autonomy'. We introduce a computational framework utilizing large language models (LLMs), combining semantic similarity, transformer-based classification, and zero-shot inference to detect discursive drift from green language to strategic framing at the project level. A new metric, the Strategic Alignment Score, tracks this shift over time, weighted by funding volume, providing a data-driven measure of "green-to-strategic" drift, based on 53,000 project descriptions from the SEDIA database, focusing on 4,049 climate-aligned projects. Results show a statistically significant increase in funding-weighted strategic alignment between the Horizon 2014-2020 and Horizon Europe 2020-2027 programmes, but no such shift is evident within Horizon Europe itself, particularly not around the 2022 Russian invasion of Ukraine. Overall, this study offers a reproducible NLP based method for discourse auditing in large funding programs and provides critical insights into transparency, climate policy, and European research strategy. This thesis is supported by the developed `sedia-api-fetchers` Python package for automated data retrieval and processing, enhancing Open Science principles and contributing to EU funding transparency.

2 acknowledgements

I kindly thank P. Maxigas and Zuzanna Warso for their guidance, intellectual feedback, and critical engagement throughout the development of this thesis. Their perspectives helped sharpen both the methodological rigor and political relevance of this work.

I also wish to thank the researchers and developers behind the open-source NLP tools and datasets that made this analysis possible, as well as the maintainers of the Horizon Europe SEDIA database for ensuring transparent access to publicly funded project data.

3 Introduction – A Horizon Expanded: The *Zeitenwende* and the Rhetoric of Strategic Autonomy

"This is not only good for the planet; it also supports our open strategic autonomy." — Ursula von der Leyen, Industry Summit, 26 February 2025 [1]

Recent policy discourse suggests that the 30% of Horizon Europe's €95.5 billion budget allocated for the green transition, is increasingly framed through the lens of 'strategic autonomy' and 'security', rather than as a global public good. Without systematic auditing, this strategic drift risks undermining the EU's 55% emissions reduction target by 2030. Moreover, it has implications for budgetary credibility, democratic oversight, and transparency. This project aims to critically assess that shift on the level of individual research projects.

3.1 From Peace Dividend to Security Imperative

> *"We are in an era of rearmament. And Europe is ready to massively boost its defence spending."*
>
> *— Ursula von der Leyen, President of the European Commission (4 March 2025) [2]*

This statement marked a notable shift in strategic posture: from a peace-oriented union of sovereign states to an autonomous geopolitical actor with expanding military capabilities. The current security environment is arguably the most volatile Europe has faced since World War II. Defence spending is now framed not merely as a sovereign prerogative, but as a collective necessity. Support for this shift cuts across ideological lines, including parties historically opposed to EU military integration, reflecting a broad consensus on the need for a more unified and capable European defence framework.

3.2 Horizon Europe's Evolving mandate

This turning point, framed as a *Zeitenwende* by Chancellor Scholz in 2022[3], has implications beyond defence. It signals the potential strategic repurposing of existing civilian institutions and funding vehicles. This new frame recasts existing civilian funding vehicles as levers of strategic autonomy. Horizon Europe, the Union's €95.5 billion research instrument, has historically targeted global public goods: climate stability, biodiversity, digital transformation, public health and inclusive growth [4]. This is laid down in Regulation (EU) 2021/695 (Art. 54)[5], as it restricts the programme to *civil* applications only, while hard-security research and development (R&D) flows through the European Defence Fund [6].

Yet the 2024 Strategic Plan[4] shifts rhetoric toward technological sovereignty and resilience, reframing "green" and "digital" objectives as pillars of geopolitical strategy, addressing events like the Russian invasion in Ukraine.

> "The second half of the Horizon Europe programme period includes the former challenges that have been exacerbated by an unstable geopolitical context... This has also put the ambition of the EU's green and digital transitions under

> pressure... For these reasons, it is becoming even more urgent to take transformation measures towards supporting the objectives of the European Green Deal, the digital transition and boosting our industrial competitiveness." [4]

Despite adhering to the the strict civil purpose, issues traditionally associated with climate change are no longer framed solely in environmental terms, but increasingly in relation to the EU's geopolitical and strategic positioning. By March 2022, Charles Michel had already described decarbonization as a "geostrategic imperative," tying climate policy directly to Europe's security and sovereignty goals [7]. In this context, the green transition is no longer viewed solely as a moral or ecological necessity, but increasingly as a pillar of geopolitical strategy.

3.3 Societal Gap

As the rhetoric of strategic autonomy reshapes EU political priorities, a clear accountability framework at the level of individual funded projects remains absent. Policymakers increasingly conflate environmental and security objectives, yet no systematic metric exists to measure the extent to which green Horizon Europe's civilian R&D spending is being redirected toward strategic goals. In the absence of such benchmarks, democratic parliamentary institutions face structural limitations in enforcing democratic oversight. Likewise, civil society and independent watchdogs and think tanks, such as Open Future[8], lack the tools to assess alignment with the EU's Green Deal targets[9]. This research seeks to fill that gap by developing and applying project-level diagnostics to assess the evolving strategic orientation of green transition funding. In parallel, it aims to build a modular pipeline to support future research and transparency dashboards, facilitating ongoing monitoring and public accountability.

3.4 Scientific Gap

A systematic literature review was conducted on Scopus using the following queries to identify relevant research:

1. `TITLE-ABS-KEY ( "Horizon Europe" AND ( "strategic autonomy" OR ( "green transition" OR "climate policy" ) ) )`
2. `TITLE ( "Horizon Europe" AND ("text analysis" OR "text mining" OR "keyword analysis"))`
3. `TITLE-ABS-KEY ( "Horizon Europe" AND ("llm" OR "BERT" OR "zero shot" OR "topic modelling" OR "LDA"))`

These search queries capture the full thematic breadth of Horizon-Europe policy analysis research, while stratifying the depth of sophistication of the NLP techniques deployed to analyse it.

Existing research on European Union (EU) funding programs, particularly Horizon Europe, increasingly utilizes computational methods for textual analysis. Previous scholarly work has demonstrated the technical feasibility and relevance of analyzing EU project texts, as highlighted by the following key contributions. However, a notable gap persists in quantifying nuanced rhetorical shifts within project objectives and, critically, linking these shifts to corresponding budgetary allocations.

- **Thematic Understanding via Topic Modelling:** Esztergár-Kiss (2024)[10] successfully applied Latent Dirichlet Allocation (LDA) to 310 Horizon 2020 transport abstracts, identifying five latent themes. This work demonstrates the feasibility of large-scale text analysis for thematic insights. However, its "bag-of-words" approach is limited in detecting nuanced policy framing and lacks integration with financial data. Our approach, in contrast, leverages LLM embeddings and attention mechanisms to capture more subtle textual meanings and contextual understanding, which is crucial for identifying intricate shifts in strategic policy, and integrates financial linkages.
- **Improved Classification with Advanced NLP:** Rodella et al. (2025)[11] significantly advanced textual analysis of Marie Skłodowska-Curie Actions (MSCA)

evaluation summaries using Distil-RoBERTa, achieving a 14-point F1 score improvement over traditional TF-IDF methods and incorporating SHAP-based interpretability. Their findings, particularly regarding the influence of project-relevant terms on funding success, underscore the power of modern transformer models in grant evaluation. This thesis draws inspiration from their robust application of advanced NLP for classification and the emphasis on understanding influential language in funding documents.

- **Ethnographic approach into Policy Rhetoric and exploring the gap between rhetoric and reality:** Cerinšek and Podjed (2024)[12] conducted extensive ethnographic research on Framework Programme projects, revealing a disconnect between "innovation rhetoric" and "material outcomes." This qualitative approach provided rich insights into how terms like "innovation," "sustainability," and "impact" are used and internalized by project actors, often serving bureaucratic expectations rather than driving genuine change. Their qualitative methodology establishes the theoretical importance of "discourse drift" in EU policy. This thesis builds on their work that confirmed the concept, charting the gap between high level rhetoric and low level implementation, and arguing for complementary quantitative metrics to rigorously track these shifts.
- **Computational Cataloguing of Clean Energy Technologies:** Koretsky et al. (2021)[13] developed a multi-method approach to catalogue climate crisis mitigation technologies funded by Horizon 2020. While robust for identifying the presence of specific technologies, their methodology, which relies on deterministic keyword and fuzzy matching, is limited in detecting nuanced rhetorical shifts in policy framing. It primarily identifies *what* technologies are mentioned, not *how* they are discussed in relation to strategic objectives like "autonomy" or "security." Furthermore, their method introduces approximations in funding sums for projects with multiple technologies. Our thesis, in contrast, aims to measure rhetorical drift through a fine-grained, LLM-based metric that captures nuanced semantic changes beyond keyword presence, directly linking these linguistic shifts to financial allocations to quantify the strategic repurposing of climate funds.

Table 1: Comparison of Key Literature and Contributions to Research Gap

Study/Author	**Primary Method(s)**	**Key Contribution**	**Limitation(s) relevant to gap**	**How This Thesis Addresses**
Esztergár-Kiss (2024)[10]	LDA Topic Modelling	Thematic understanding of H2020 abstracts	Bag-of-words; lacks nuance; no financial linkage	LLM-based nuance; financial integration
Rodella et al. (2025)[11]	Transformer-based Classification (DistilRoBERTa)	Improved grant classification; interpretability of features	Focus on classification; not tracking rhetorical drift	Focus on rhetorical drift; new metric development
Cerinšek and Podjed (2024)[12]	Ethnographic Discourse Analysis	Theoretical grounding of policy rhetoric	Qualitative; no quantitative metrics for discourse drift	Quantitative, scalable LLM-based metric
Koretsky et al. (2021)[13]	Deterministic Keyword, Fuzzy Matching	Cataloguing clean technologies; alignment with goals	Presence-based (not nuance); financial approximations; no rhetorical linkage	Nuanced semantic change; precise financial linkage

While existing literature advances latent discovery, NLP techniques, and multi-step classification of climate projects, it lacks a unified, scalable method for linking rhetorical framing to financial allocation. This thesis addresses that gap by introducing a novel, LLM-based metric that tracks project level discourse shifts. In contrast to current dashboards, which rely on static keyword analysis and overlook rhetorical nuance, this approach quantifies the strategic drift from green rhetoric to geopolitically framed discourse, weighted by committed euros. With over €30 billion in climate funding at stake, this study offers a critically needed tool to assess how evolving policy narratives shape funding decisions, contributing original insights at the intersection of computational linguistics, policy analysis, and climate governance.

3.5 Research Question

RQ1 Rhetorical Shift
To what extent has climate-oriented Horizon Europe discourse adopted "strategic-autonomy" framing over time (2021 – 2025), is there measurable conflation between green and geopolitical language at the project-objective level?

RQ2 Budgetary Echo
After weighting each project by the euros actually committed by the EU, does EU funding increasingly concentrate in those green projects that register a high strategic-alignment score over time?

Rather than focusing on top-down policy rhetoric or shifts in funding criteria, this study adopts a bottom-up approach by analyzing the stated objectives of funded projects. The aim is to assess whether political reframing has influenced the nature of these objectives and the ways in which they are articulated, justified, or motivated by researchers themselves. By systematically reviewing project descriptions across the last two Horizon programmes, this research examines to what extent geopolitical strategy is being embedded within the discourse of civilian research initiatives.

Special attention is given to green-oriented projects, to investigate claims of their newfound strategic importance. To address whether the Horizon Europe's green research agenda reflects a strategic reframing rather than a substantive shift, gives rise to the following data science questions.

3.5.1 Data Science Questions

- **Can we detect "green" research projects independently of official EU classifications within the funding landscape?**

- **Can we quantify the emergence and conflation of "strategic" and "green" language in climate-oriented Horizon Europe project objectives?**

- **How can we assess the temporal dynamics of a funding-weighted "strategic score" to reveal budget allocation trends?**

Proposed Methods

- **For Q1**, we will develop an LLM driven classifier that combines embedding similarity thresholds with supervised learning to identify green-focused projects with high precision from objective texts alone, then benchmark its performance against annotated labels. Crucially, our classification aims for a nuanced understanding of 'greenness,' moving beyond simple keyword matching.

- **For Q2**, we will apply a zero-shot classification pipeline in project texts, turning "rhetorical shift" into a quantitative metric without requiring large labeled corpora.

- **For Q3**, we will weight each project's quantified strategic value by its committed EU budget. This will be used to derive a time series of funding-weighted strategic scores, directly linking discourse emphasis to monetary allocation.

3.6 Ethical and Responsible-AI Considerations

3.6.1 Data availability and personal information

All Horizon Europe data used in this project are publicly available through official EU platforms. The only personal information processed consists of the names of researchers listed as authors or coordinators of funded projects. These names are part of the public record and are disclosed in the interest of transparency and attribution under EU research policy.

This usage complies with the General Data Protection Regulation (GDPR), as the data concern individuals acting in a professional capacity and are made publicly available by the data subjects themselves or their affiliated institutions. Data collection and processing are conducted solely for academic research purposes, in line with the public interest basis outlined in the GDPR Article 6.1.e. No sensitive personal data are processed, and no profiling or identification is performed beyond the use of publicly available information. Importantly, no evaluative or normative claims are made about individual researchers.

3.6.2 AI

The LLM-based approach of this research, while powerful, will carefully consider potential biases and hallucination risks inherent in such models, aligning with responsible AI practices and the principles of the evolving EU AI Act. Given that the output of the LLM may influence perceptions of the real world among policymakers and the general public, it is of utmost importance to stress that the results of this study should be interpreted with caution. They do not constitute objective truths but rather model generated insights shaped by training data, prompting strategies, and contextual limitations. Transparency, traceability, and clear disclaimers will be integral to all dissemination efforts to mitigate the risk of misinterpretation or over reliance on the model's outputs. It is not the intention of this research to sway public discourse using biased outputs. Rather, all findings will be presented with appropriate contextualization and adhering to Open Science principles.

3.6.3 Language and Annotation

Bias may arise due to the predominance of English-language data. All data processing and analysis in this project were conducted in English, reflecting the limited scope and resource constraints of the study. While this approach ensures consistency, it does not consider content in other official EU languages. Future research could address multilingual processing to improve representativeness and inclusivity across the Union's languages and to address english bias. However, it is of importance to note that the EU advices that proposals are to be submitted in english for faster processing, thus english information is the primary source of data.

Due to the absence of labeled ground truth data, human annotation was used for validation. Two independent coders, with no conflict of interest, annotated a random subset of the dataset. Inter-coder agreement reached $\kappa = 0.82$, indicating substantial agreement.

3.7 Contributions and Structure

This thesis delivers three major contributions.

C1 First quantitative metric of "green-to-strategic" drift. We supply a data-driven measure that tracks how green focused Horizon-Europe projects are adopting strategic-autonomy language weighted by funding allocation. The metric gives policymakers, auditors, and the public a basis to monitor and hold the EU accountable for its spending.

C2 Python package `sedia-api-fetchers`[14] Along with the thesis, we release a scalable, open-source Python toolkit for automated retrieval and processing of EU funding data. This toolkit supported the current thesis and can be applied to audit data ingestion, LLM-based data science pipelines, and dashboard-ready outputs. It generalizes across funding instruments and endpoints (e.g., tenders, organizations), enhancing transparency of EU funding.

C3 Open Science and EU policy transparancy: Open data and software is provided, adhering to the overarching Open Science principles, according to the FAIR principles and in line with UU's commitments, of Utrecht University. Data retrieval code is available publically as mentioned, the data wrangling notebook and final dataset is published at zenodo under these references[15, 16].

The remainder of the document is organized as follows.

- **Section** 4 explains with the domain in which the data is embedded, explains how EU research funding is structured, defines the key terminology for readers unfamiliar with the field, and highlights an emerging policy shift that could guide follow-on projects building on this work.
- **Section 5** describes data acquisition, preprocessing, quality and exploratory analysis. It also introduces the accompanying Python package.
- **Section 6** outlines the methodology and data science tools used, and validates each intermediate step.
- **Section 7** presents the empirical findings.
- **Section 8** evaluates the robustness of the data-science approach (Subsection 8.1), assesses how well the study answers the research questions (Subsection 8.2), and outlines avenues for future work.
- **Section 9** summarizes the main results, weighs the strength of the supporting evidence, and discusses their broader scientific and societal implications.

4 Strategy, structure and organization of EU funding on research and innovation

This section explains the domain context and outlines the architectural framework for readers not familiar with the EU's biggest funding research vehicles.

Since the 1980s, the Framework Programmes (FPs), which follow a sequential seven year funding cycle, have supported thousands of research projects across Europe[17]. The most ambitious programme to date is Horizon Europe (2021–2027).

4.1 Horizon Europe summarized

The Horizon Europe Research and Innovation programme (2021–2027) is the EU's key funding programme for research and innovation, with a record budget of €95.5 billion, its largest to date[18].

The Strategic Plan serves as an interface between broad EU policy priorities and the Horizon Europe R&I activities, facilitating their implementation through the Horizon Europe work programmes[4].

The second Horizon Europe Strategic Plan (2025–2027) outlines three strategic orientations:

1. **Green Transition** Climate action will receive 35% of the overall budget, with 10% specifically reserved for biodiversity initiatives[4].
2. **Digital Transition** At least €13 billion will be invested in core digital technologies[4].
3. **A More Resilient, Competitive, Inclusive, and Democratic Europe**

An important addition to this plan are the two overarching principles guide all three orientations: Open Strategic Autonomy and Securing Europe's Leading Role in Developing and Deploying Critical Technologies.

4.2 Horizon Europe architecture

Horizon rests on three pillars, each consisting of multiple Work Programmes (WP).

4.2.1 Pillar I : Excellent Science, a bottom up approach

Budget : 25/95.5 billion EUR

This pillar supports a bottom-up approach, topics are not set by the EU, but funding is based on 'the criterion of excellence'. The pillar is aimed to support frontier research, fellowships, doctoral networks, training and researcher exchange, and the development of research infrastructures across all Member States.[19] The main WPs constitute:

- **European Research Council (ERC) (16/25 billion EUR)**: Funds frontier science through annual work programmes. It provides ambitious, high-gain/high-risk research projects with long-term financial stability.

- **Marie Skłodowska-Curie Actions (MSCA) (6.6/25 billion EUR)**: The EU's reference programme for doctoral education and postdoctoral training. It funds excellent research and innovation while equipping researchers at all career stages with new knowledge and skills through mobility and interdisciplinary exposure. It enhances training, cooperation, networking, and global attractiveness.
- **Research Infrastructures (RIS) (2.4/25 billion EUR)**: Facilities providing resources and services for conducting research and fostering innovation. These can be single-sited, distributed, or virtual.

4.2.2 Pillar II: Global Challenges and European Industrial Competitiveness, a top down approach

Budget: 53/95.5 billion EUR

This pillar has a top-down approach, topics are set out by the EU and are divided in clusters, as can be seen below.

> "The pillar 'Global Challenges and European Industrial Competitiveness' (...) aims to maximize integration and synergies across respective thematic areas while securing high and sustainable levels of impact for the Union in relation to the resources that are expended. The Pillar aims to encourage cross-disciplinary, cross-sectoral, cross-policy and cross-border collaboration in pursuit of the Sustainable Development Goals (SDGs). This pillar aims to support the creation and better diffusion of high-quality new knowledge, technologies and sustainable solutions, (...) in particular in 'small and medium sized enterprises' (SMEs), start-ups and society to address global challenges"[20].

- Cluster 1: Health
- Cluster 2: Culture, Creativity and Inclusive society
- Cluster 3: Civil security for society
- Cluster 4: Digital, Industry and Space
- Cluster 5: Climate, Energy and Mobility
- Cluster 6: Food, Bioeconomy, Natural Resources, Agriculture and Environment
- Cluster 7: Non-nuclear direct actions of the Joint Research Centre

"Social sciences and humanities shall be fully integrated across all clusters, including specific and dedicated activities."[20]

4.2.3 Pillar III: Innovative Europe

Budget: 13.6 / 95.5 billion EUR

This pillar turns Europe's scientific excellence into market-creating innovation and strengthens an interconnected, inclusive innovation ecosystem across the Union.[21]

- **European Innovation Council (EIC) (10.1 / 13.6 billion EUR)** – Provides blended finance (grants and equity) plus bespoke acceleration services to high-risk, high-impact start-ups, SMEs and research teams via the *Pathfinder*, *Transition* and *Accelerator* schemes, helping breakthrough technologies reach the market.[21]
- **European Institute of Innovation and Technology (EIT) (3.0 / 13.6 billion EUR)** – Operates pan-European Knowledge and Innovation Communities (KICs) that integrate business, research and higher education to deliver entrepreneurial training, support start-ups and scale-ups, and develop innovative products and services in areas such as climate, digital, health and raw materials.[21]
- **European Innovation Ecosystems (EIE) (0.5/ 13.6 billion EUR)** – Funds actions that connect national and regional innovation actors, improve flows of knowledge, talent and capital, foster joint innovation programmes and stimulate innovation-friendly public procurement, complementing the EIC and EIT.[21]

4.3 Horizon in the Future: A Dual-Use Funding Vehicle for Defence and Civil Purposes?

To date, Horizon Europe projects have been explicitly restricted to civilian research under Regulation (EU) 2021/695, point 54. "Activities under the EDF must focus exclusively on defence research and development, while civil-only calls remain under Council Decision and the EIT, unnecessary duplication is avoided".[5]

The Zeitenwende has catalysed a fundamental reorientation: in its 19 March 2025 White Paper for European Defence, the Commission and High Representative mandate that Horizon must proactively channel civil research toward defence capabilities [22]. They call for a “European Armament Technological Roadmap” to steer investments into advanced dual-use technologies and adapt regulations “to be more conducive towards risk-taking” by defence innovators [22]. In sum, Horizon Europe is evolving from a purely civil R&I fund into a strategic instrument of European autonomy, blending frontier science with capability development.

To inform actionable funding decisions, computational analysis of large project datasets is essential, since existing HORIZON project data contain little explicit defence components, this paper focuses instead on strategic framing methods to detect emerging dual-use orientations, which could in the future be repurposed for detecting militaristic shift in funding.

5 Data

5.1 Data retrieval: The `sedia-api-fetchers` Python package, beyond official dumps

A robust computational analysis depends critically on a high-quality, comprehensive dataset. Furthermore, the ability to perform continuous, up-to-date analyses, rather than relying on a static snapshot, is essential when informing dynamic policy processes. To meet these demands, we developed `sedia-api-fetchers`, a Python package that encapsulates the core functionality of our automated data pipeline, providing standardized modules for authenticated API access, asynchronous request handling, and incremental change detection. This package aims to make open data more easily accessible for researchers.

`sedia-api-fetchers`[14] automates the retrieval, monitoring, and processing of records of EU funding data using APIs, addressing challenges such as pagination, rate limiting, retries, and state persistence out of the box. Each fetcher class can be configured to emit detailed logs that track added, modified, or deleted records. This design not only ensures scalability, reliability, and real time integration, but also enhances transparency and reproducibility by recording such metadata for every retrieval operation. Its modular architecture supports extension to new data sources and downstream applications, including interactive dashboards and advanced analytics. For detailed information about the package's design, API, and configuration options, please refer to its GitHub repository and PyPI page[14].

5.1.1 Data source

Initially, project data was sourced from the *Community Research and Development Information Service* (CORDIS) dataset[17], which provides details on project objectives, timelines, participants, funding, and research outcomes. However, CORDIS only contains data related to the Horizon framework programmes (e.g., Horizon 2020, Horizon Europe) and excludes other key funding instruments such as the European Defence Fund or the DIGITAL Europe Programme.

To address this limitation, the project also integrates data from the EU Funding & Tenders Portal, which serves as the operational platform for EU grants and tenders. Crucially, project data can be programmatically retrieved from the underlying Single Electronic Data Interchange Area (SEDIA) API. However, the publicly available documentation for the SEDIA API[23] exhibits several limitations that complicate its use:

1. **Opaque parameter codes.** Filter criteria such as `type`, `status` or `frameworkProgramme` use numeric reference codes (e.g. `31094501` for "Open" calls) without explanation or lookup tables, forcing developers to switch back to the web UI responses to resolve meanings. The only first party workaround is the Facet endpoint, which returns label to code look ups, but you must already know each field name to query it, and it not clear which filter criteria are available in the first place.
2. **Lack of machine-readable specification.** No OpenAPI/Swagger or JSON Schema is provided, preventing use of automated tools for endpoint discovery or schema validation.

3. **Sparse usage examples.** The documentation relies on Postman screenshots and partial JSON fragments, without language-agnostic examples (e.g. `curl`, Python `requests`, JavaScript `fetch`) or complete response payloads that illustrate pagination metadata.
4. **Minimal error and rate-limit guidance.** There is no description of HTTP status codes, error payload formats, retry strategies, or quotas, leaving integrators to guess best practices for reliable operation.
5. **Unversioned endpoints and poor navigation.** All APIs share a single `/prod/rest/` URL path with no explicit versioning or changelog, and the online docs lack a stable table of contents or anchor menu for quick reference. The existence of the apiVersion field in the response data is undocumented.

Access to the SEDIA API was gained by capturing the network traffic generated when filters are applied in the EU Funding & Tenders Portal in the front end. Recording the JSON payloads sent to `/search-api/prod/rest/` and their responses revealed

- the full request schema, including Boolean *must*/*should* clauses, nested `terms`/`term` filters, and date-range filters on `es_SortDate` (or `lastModified`).
- The complete set of filter fields, both documented *and* undocumented, such as `programmePeriod`, `crossCuttingPriorities`, `missionGroup`, `destination`, and `topicAbbreviation`
- the numeric reference codes used by the front end and exposed by the Facet service.
- The response layout, including pagination keys (`from`, `size`, `total`), project records and auxiliary metadata such as `apiVersion`.

This reverse-engineered specification enabled reliable scripted queries and the ingestion of more than 50 000 projects spanning multiple EU programmes, functionality unattainable with the published docs alone. Replicating the discovered endpoints in our pipeline unlocked metadata absent from CORDIS, including project keywords, cross-cutting priority tags (climate, biodiversity, AI), detailed participant data, calls, tenders, and institutional affiliations.

5.1.2 Comparative analysis between CORDIS and SEDIA

A comparative analysis of the data provided by CORDIS and the SEDIA API is presented in figure 1. This comparison considers only the project level data retrieved from the primary API endpoint, without incorporating additional metadata from other endpoints. The naming conventions and data structures differ significantly between the two sources. Broadly, CORDIS places greater emphasis on research outputs, whereas SEDIA provides more detailed information on the administrative and institutional context of EU funding. This latter focus makes SEDIA particularly well-suited for the objectives of this project.

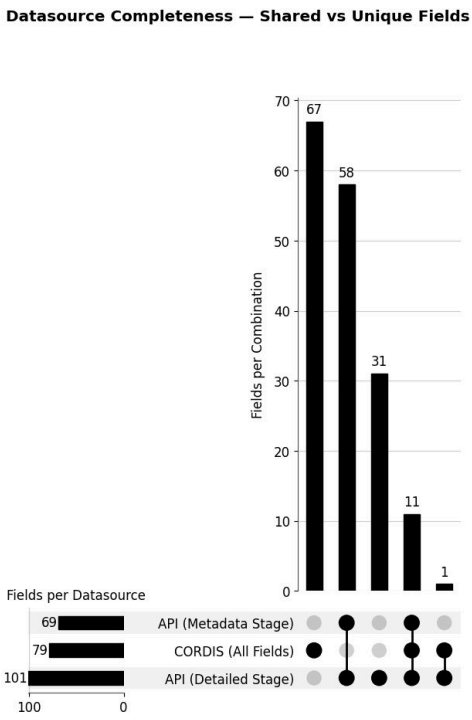


Figure 1: The figure shows the number of fields across CORDIS and two stages of the SEDIA API. The detailed SEDIA API, which includes fields from individual project-level calls, offers the most fields ($n = 101$), followed by CORDIS ($n = 79$) and the SEDIA metadata stage ($n = 69$). Overall, CORDIS emphasizes research data, while the SEDIA API adds administrative and institutional details useful for this project.

5.1.3 Adressing API limitations

While the SEDIA API offers a general endpoint for project-level data allowing for paginated calls, detailed metadata must be retrieved per project, significantly increasing acquisition time. With over 50k projects, full extraction can take hours. Additionally, paginated queries are limited to 100 items per page and capped at 100 pages, restricting retrieval to 10k records.

To overcome this, an adaptive temporal partitioning algorithm was implemented. Instead of issuing a broad query, the algorithm recursively splits the date range, based on the `es_SortDate` field (with `lastModified` as a fallback if unavailable), until each partition yields fewer than 10k records, thereby staying within API limits.

The algorithm operates as follows (see code snippet 11.4.1):

(1) Query broad date ranges (e.g., full programme periods).

(2) If results hit the page limit, split the range in half.

(3) Repeat for each sub-range until results are complete.

(4) Aggregate all partitions to reconstruct the dataset.

Partitioning decisions are made in real time based on actual data counts, not predefined rules. The algorithm dynamically adjusts to project concentration, ensuring efficient

scaling. Overall, this approach enables complete retrieval of large datasets with minimal extra execution time.

5.1.4 The final product

A Python package called `sedia-api-fetchers`[14] has been developed to encapsulate the entire SEDIA data-retrieval pipeline and support future projects. It can be installed via:

```
pip install sedia-api-fetchers==1.0.0
```

The code, examples, and documentation are available at the references[14].

The package provides modules to retrieve data from the SEDIA API, for example:

```python
from sedia_api_fetchers.EUFT_retrieve_projects import SEDIA_GET_PROJECTS

# Fetch multiple programmes by alias, returns a pandas.DataFrame
multi_programme_data = SEDIA_GET_PROJECTS.get(
    programmes=['h2020', 'horizon', 'digital'],
    save=True  # saves to CSV
)

# Fetch a single programme by numeric ID
edf_data = SEDIA_GET_PROJECTS.get(44181033, save=True)  # European Defence Fund
```

It also includes:

- A temporal partitioning algorithm (see code snippet 11.4.1) to split large queries into sub-ranges, avoiding the 10k record pagination cap.
- Fetchers for facets, projects, tenders, faq, participants/partners and topics.
- A demo Extract-Transform-Load (ETL) pipeline with change detection to flag additions, deletions or modifications between runs.
- JSON unnesting and pandas-friendly normalization routines, easing downstream database or dashboard integration.

In future work, `sedia-api-fetchers` could be embedded directly into interactive dashboards or connected to databases to power real time EU funding monitoring. This open source package aims to simplify access to detailed EU funding data, fostering transparency in the allocation of public resources. By enabling policymakers, researchers and citizens to monitor funding flows and outcomes, it supports evidence based decision making and strengthens public trust in research and innovation programmes.

5.1.5 Ethical considerations

In developing and deploying `sedia-api-fetchers`, we have been careful to respect data privacy and compliance by only retrieving and processing publicly available records and adhering to the EU's GDPR principles for personal data protection. Moreover, by making the code open source and documenting our reverse engineering methods transparently, we aim to mitigate risks of misuse and encourage ethical reuse and accountability in the analysis of public funding flows.

5.2 Data manipulation pipeline

The data retrieved from the SEDIA API is heterogeneous, deeply nested, and inconsistent. To transform these responses into a form suitable for academic analysis and policy evaluation, a data manipulation pipeline was designed. This pipeline addresses structural inconsistencies, standardizes data types to ensure analytical readiness.

- **Metadata Flattening and Structural Normalization**
  Nested metadata structures are flattened into top-level tabular columns. For example, `metadata.budget.amount` becomes `metadata_budget_amount`, simplifying data access by enabling column level operations and enabling compatibility with tools such as `pandas`.

```
  @staticmethod
    def flatten_project_data(input_json: dict) -> dict:
        """
        (...) (see github)
        """
        working = dict(input_json)
        flat_data = {}

        project_data = working.get("project_data", {}) or {} # Flatten
        ↪  project_data contents under project_data_XXX

        # Flatten project_data.metadata to project_data_metadata_XXX
        proj_meta = project_data.get("metadata", {}) or {}
        Functions._flatten_metadata_section(proj_meta, flat_data,
        ↪  prefix="project_data_metadata_")

        # Flatten other `project_data` keys to project_data_XXX
        for key, val in project_data.items():
            if key == "metadata":
                continue
            flat_data[f"project_data_{key}"] = Functions._unwrap(val)

        # Flatten top-level `metadata` to metadata_XXX
        top_meta = working.get("metadata", {}) or {}
        Functions._flatten_metadata_section(top_meta, flat_data,
        ↪  prefix="metadata_")

        # Preserve all other top-level keys exactly (with _unwrap on lists)
        for key, val in working.items():
            if key in ("project_data", "metadata"):
                continue
            flat_data[key] = Functions._unwrap(val)

        return flat_data
```

- **Data Unwrapping and Type Normalization**
  Scalar values that are incorrectly wrapped in single-element lists are unwrapped, and all fields are cast to consistent types, ensuring consistent typing across records, essential for joins and filters.

```
  @staticmethod
    def _unwrap(value):
        """
        If `value` is a single-element list, return its only element.
        If it's an empty list, return None.
        Otherwise, return `value` unchanged.
        """
        if isinstance(value, list):
            if len(value) == 1:
                return value[0]
            if len(value) == 0:
                return None
        return value
```

- **Duplicate Removal (with Unhashable Type Handling)**
  Duplicate entries, both hashable and non-hashable, are identified and removed without failing due to nested structures, preserving data integrity by avoiding double-counting, possibly addressing API-side duplication caused by pagination overlaps or timeout. It also reduces unnecessary memory and disk usage.

```
  #8-byte hexadecimal hash string representing the content of a Python object
  @staticmethod
    def _fingerprint(obj: Any) -> str:
        payload = json.dumps(obj, sort_keys=True, default=str).encode()
        return hashlib.blake2b(payload, digest_size=8).hexdigest()

  # applied to for example inside function normalise, for catching duplicates:
  def normalize(...)
    exploded["_fp"] = exploded["parsed"].apply(Functions._fingerprint)
    unique_items = exploded.drop_duplicates(subset=["_fp"]).copy()
  (...)
```

- **Empty Container Cleaning**
  Fields containing empty strings, lists, or dictionaries are cleaned and replaced with standardized null values (`NaN`), ensuring uniform representation of missing values across the dataset. This prevents false positives in value counts or categorical analysis and allows valid statistical treatment of missing data during modeling in the future.

- **Data Standardization**
  Dates and categorical fields can be normalized to ensure consistent formatting across the dataset. This facilitates joins and comparisons across multiple datasets. For example, this is important for supporting datetime operations in `pandas` and other libraries.

- **Relational Normalization**
  Certain fields, such as the `participants` column in the project data, contain dictionaries of entities (e.g., multiple participants associated with a single project). To im-

prove data structure and to fascillitate future migration to normalized database systems (e.g., PostgreSQL), this information is normalized into separate relational tables. For example, a distinct `participant` table may be created, and a `project_participant` join table maps the many-to-many relationships. This promotes data integrity and reduces redundancy by storing each participant once, it also allows for querying of relationships using relational joins. This cannot be captured in a code snippet, please refer to[14].

In summary, this data manipulation pipeline serves as a critical bridge between raw API data and analytical workflows. Its design ensures reliable quantitative analysis of complex European research funding data.

5.3 Data quality

5.3.1 Data Validation and Quality

Each dataset underwent a structured validation process to identify and exclude incomplete or malformed records prior to analysis. Validation rules targeted three critical dimensions: essential identifiers, financial data, and the presence of project objectives. Records failing any of these criteria were removed from the dataset.

A summary of validation outcomes across all datasets is provided in Table 2. While Horizon 2020 and Horizon Europe exhibit high data quality, the European Defence Fund dataset shows substantial incompleteness, particularly regarding project objectives.

Table 2: Summary of Record Validation and Retention

Dataset	**Initial Records**	**Excluded**	**Final Records**	**% Retained**
Horizon 2020	35,441	139	35,302	99.61%
Horizon Europe	17,904	45	17,859	99.75%
European Defence Fund	183	28	155	84.70%

Figure 2 illustrates the distribution of excluded records by validation rule across the three datasets. Most exclusions occurred in the Horizon 2020 dataset, with missing essential identifiers as the dominant cause. In contrast, project objective omissions were disproportionately high in the European Defence Fund records.

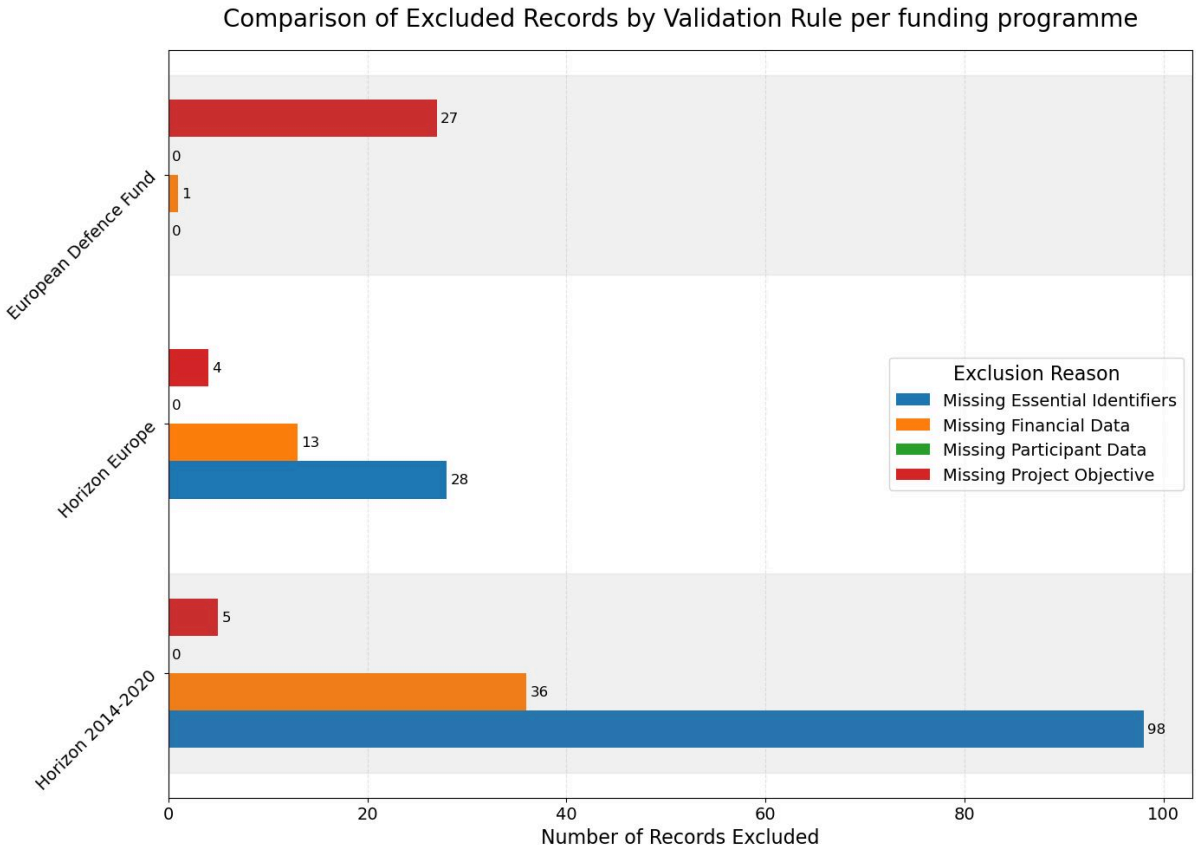


Figure 2: Comparison of Excluded Records by Validation Rule

5.4 Exploratory Data Analysis

5.4.1 Simple keyword analysis on green terms

Using green keywords (determined in Subsection 6.2), and strategic keywords (determined heuristically based on the strategic plans[4]), an exploratory data analysis is conducted to obtain an estimate of the amount of 'green' and 'green-strategic' projects. Figure 3 shows a large influx of green funding during the first years of the current Horizone programme. Figure 4 suggests this same influx of funding is true within these green projects towards projects that contain 'strategic' language. It is important to note that the analysis conducted here is rather superficial, it only considers raw counts of projects containing words, without consideration of the context of those words, making the tagging highly uncertain. Figure **??** indicates that there is no major change in the relative amount of green projects that contain strategic words, while more happens in terms of funding, there seems to be a slight influx during the first years of the current Horizon program and also a peak at the start of the previous Horizon program (presumably due to nuclear fission plant funding).

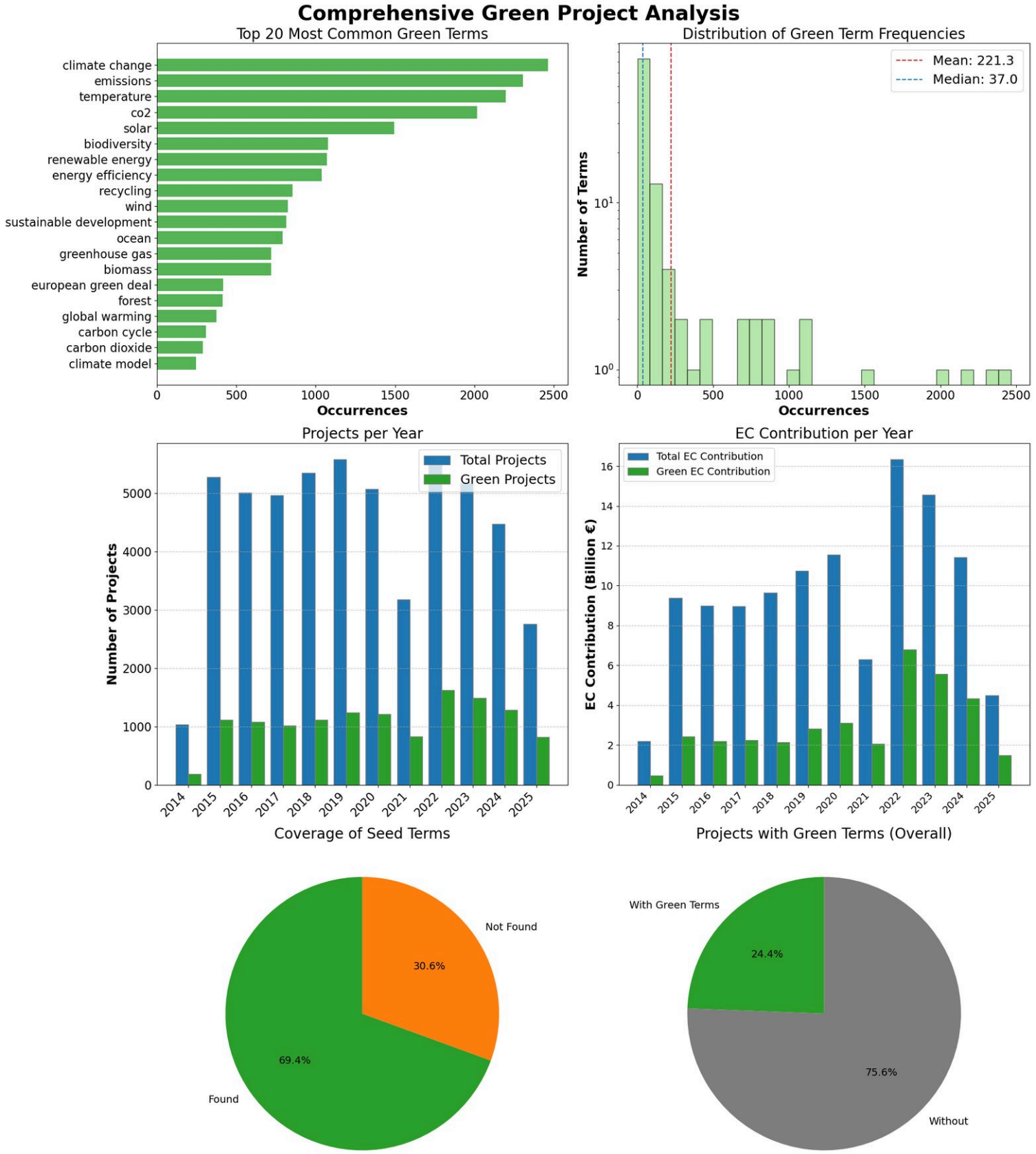


Figure 3: A dashboard into the distribution of green keywords across projects, and the distribution of green tagged projects and funding over time. A noticable influx of funding is present in the earlier years of the current Horizon Programme.

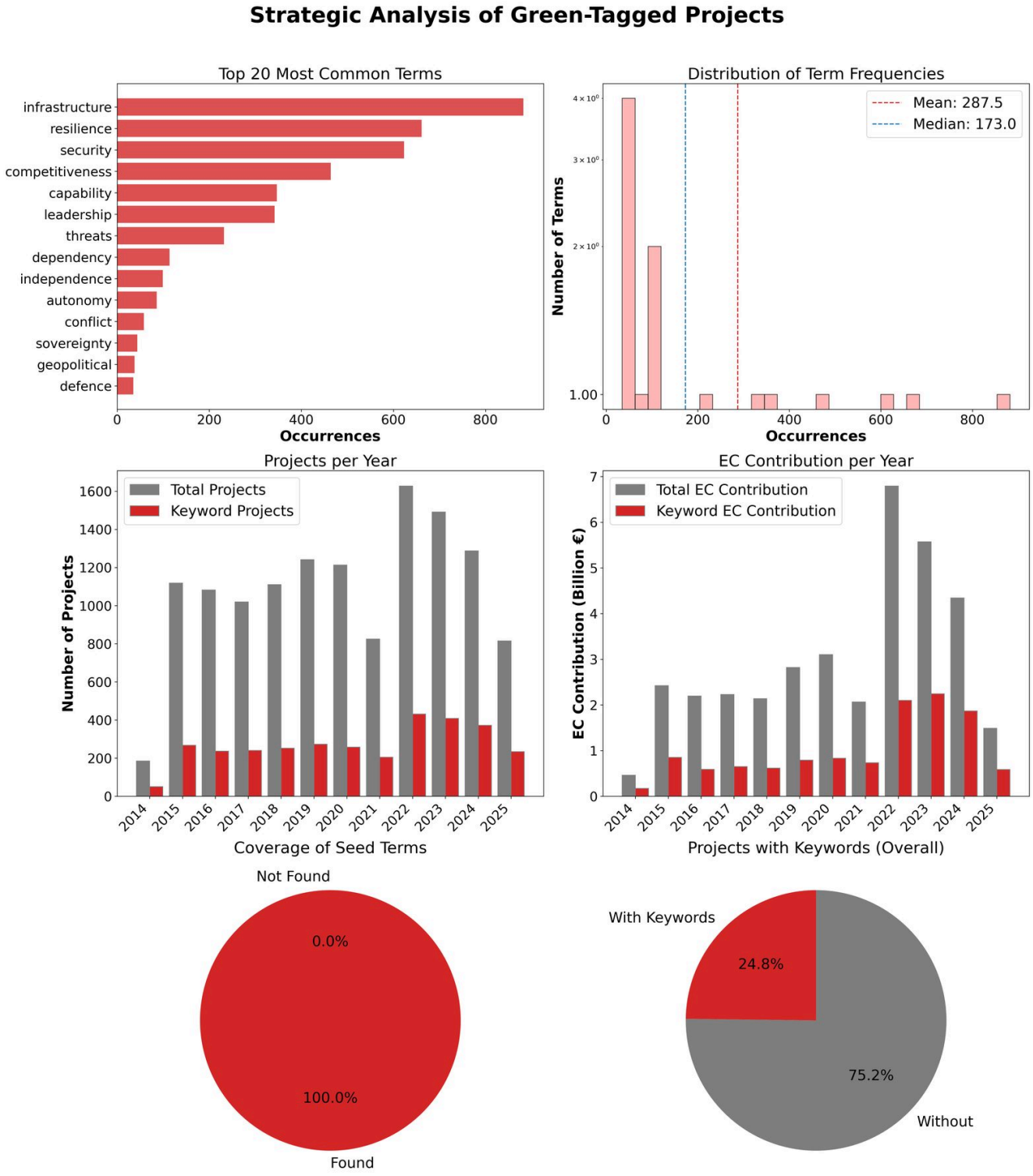


Figure 4: A dashboard into the distribution of strategic keywords among green-tagged projects, and the distribution of those projects and funding over time. A noticable influx of funding is present in the earlier years of the current Horizon Programme.

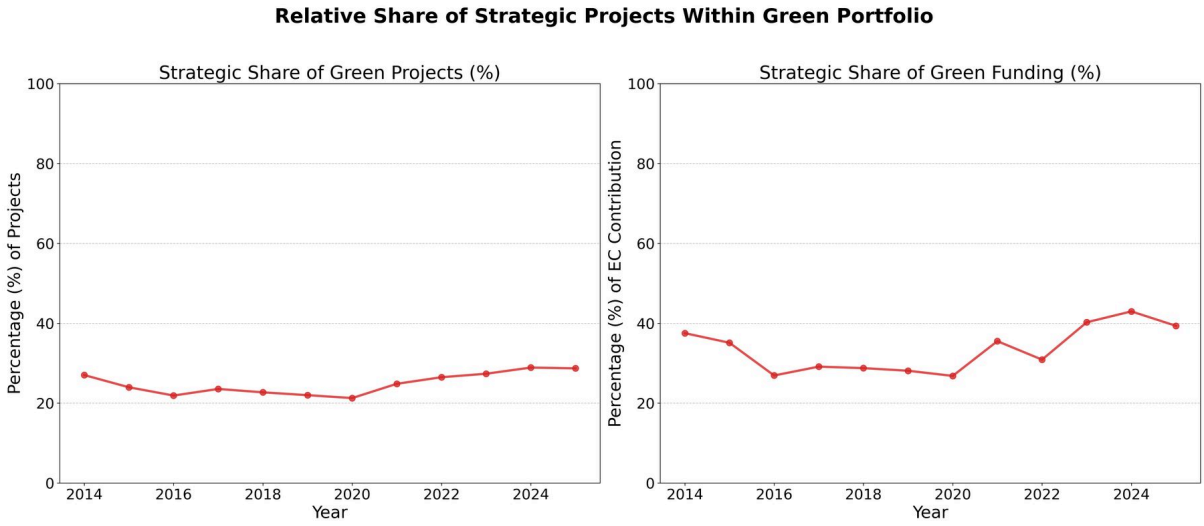


Figure 5: relative count of projects and funding of green projects that contain strategic words versus projects that do not contain those words.

6 Methodology

6.1 Determining the Strategic Alignment of European Green Research Projects: Vector Embeddings and Zero-Shot Inference

To address the research question: *"To what extent has green rhetoric become conflated with strategic-autonomy rhetoric in HORIZON-funded projects? Is there evidence of geopolitical greenwashing?"*, we employ a two-stage methodology.

First, we identify high confidence climate oriented projects by combining two complementary techniques: semantic similarity scoring via vector embeddings and binary classification. The former provides a continuous estimate of how closely a project objective aligns with climate related language, but it does not explicitly distinguish between "green" and "non-green" projects. Conversely, binary classification enforces a categorical distinction but lacks nuance in measuring the degree of alignment.

By analyzing the distribution of similarity scores across the two classification outcomes, we can heuristically identify a threshold that separates climate aligned from non-aligned projects. If such a separation exists, this threshold allows us to confidently select projects with a strong climate orientation.

This method deliberately prioritizes precision over recall in the 'Precision/Recall Trade-off', minimizing false positives by classifying projects as climate-aligned only when both semantic and categorical signals agree. While this may exclude some truly green projects, it avoids the more damaging risk of including a large sample of non-climate related projects, which could distort trend analyses, policy interpretations, or assessments of geopolitical greenwashing.

Next, we assess the strategic framing of these projects using zero-shot natural language inference (NLI) to estimate a project level "strategic alignment" score. Zero-shot natural

language inference (NLI) is employed due to the inherently nuanced and subjective nature of strategic alignment, coupled with the absence of annotated datasets that could support supervised learning in this domain.

6.1.1 Choice of LLM Models

For model selection, we prioritized models that perform well while fitting comfortably on a single NVIDIA RTX 4090 GPU.

- **Embedding model**: For embedding generation, we selected `Qwen/Qwen3-Embedding-0.6B`[24]. Benchmarked on MTEB[25], it punches above its weight at 0.6B parameters. This allows for good quality embeddings, efficient computation times while leaving GPU headroom. The Model selection is further illustrated in Figure 6, which visually represents the trade-offs between model performance and computational requirements.

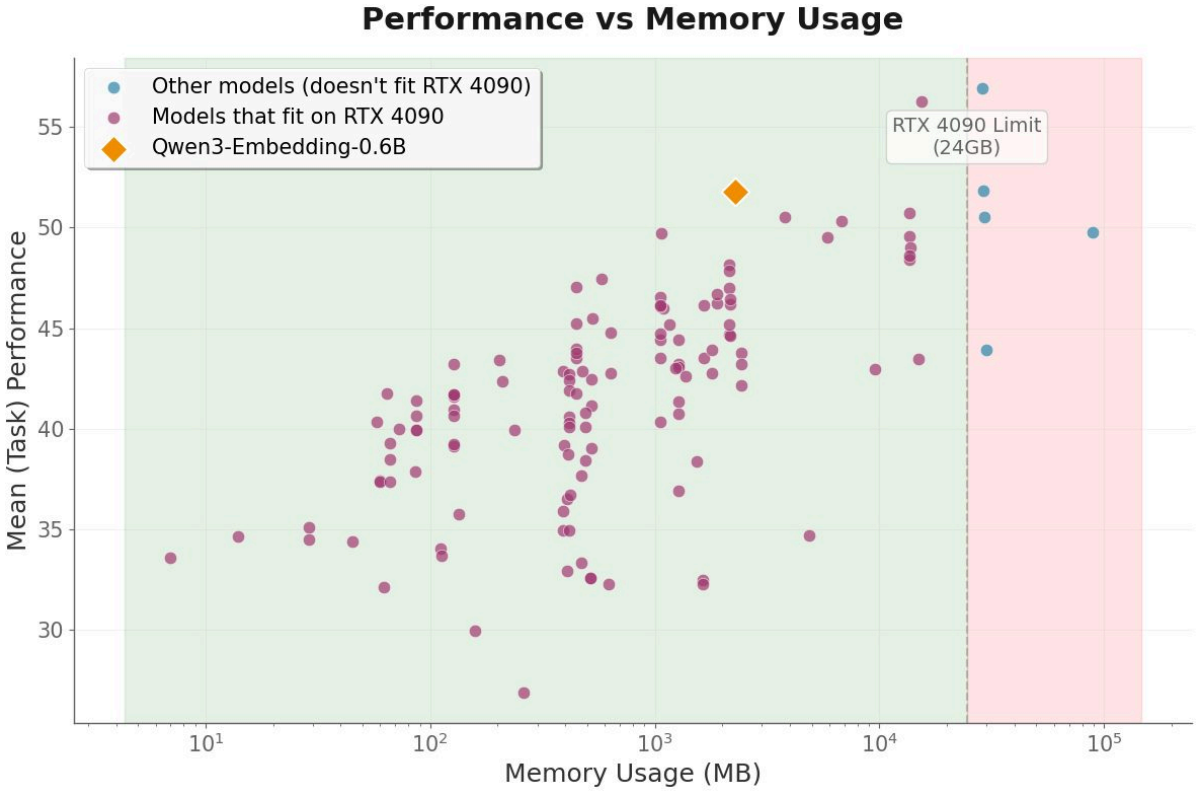


Figure 6: This figure allows for a visual assessment of model trade-offs, aiming for models that are both high-performing (higher on the Y-axis) and fit comfortably within the RTX 4090's memory limit (to the left of the dashed line, in the green zone). `Qwen3-0.6B` (orange diamond) delivers optimal performance given its size. Data is sourced from the mteb leaderboard.[25]

- **Zero-shot policy alignment**: To quantify "strategic alignment" without a labelled corpus, we deployed the `facebook/bart-large-mnli` model[26, 27, 28], which is an MNLI finetuned 406M-parameter BART encoder decoder LLM, meaning it has a strong performance on NLI benchmarks, making it suitable for classifying text into categories it hasn't been explicitly trained on.
- **Climate change text classification**: For climate classification, the

`climatebert/distilroberta-base-climate-detector` model[29] was used. Domain-specific pre-training on climate corpora yields better performance for sustainability texts while keeping the model lightweight for batch inference.

6.2 Methodology Step A: Finding high confidence climate-oriented projects

1. Define the project corpus $\mathcal{D}$. The project corpus that consists of 53k HORIZON project objectives obj_i, which are the descriptions of about 2000 characters long detailing the aim of the project.

$$\mathcal{D} = \{\text{obj}_1, \ldots, \text{obj}_N\}$$

2. Define the seed terms. These consist of key terms or short phrases curated from authoritative sources. In this project, the positive seed set S^+ contains terms t_i^+ aligned with climate research and policy discourse, sourced from the IPCC, UNSCC, and HORIZON glossaries. These are intended to reflect the "green-positive" axis. The negative seed set S^- contains terms unrelated to climate (e.g., *cancer*, *lyric poetry*, *philosophy of language*), used to anchor the opposite pole. The exact seed terms can be found in the appendix.

$$S^+ = \{t_1^+, \ldots, t_m^+\}, \quad S^- = \{t_1^-, \ldots, t_n^-\}.$$

3. **Green-signal vectorisation.**

   - *Seed-term embeddings:* Each term t is embedded using the aforementioned embedding model.
   $$\mathbf{e}_t = \text{Embed}(t) \in \mathbb{R}^d.$$

   - *TF–IDF weighting:* Let df_t be the document frequency of term t, i.e.,
   $$\text{df}_t = \left|\{i \mid t \in \text{obj}_i\}\right|.$$
   The inverse document frequency is computed as:
   $$\text{idf}(t) = \log\left(\frac{N+1}{\text{df}_t + 1}\right) + 1.$$

   - *Weighted centroids:* Compute the centroids of the positive and negative seed sets:
   $$\mathbf{c}^+ = \frac{\sum_{t \in S^+} \text{idf}(t)\, \mathbf{e}_t}{\sum_{t \in S^+} \text{idf}(t)}, \quad \hat{\mathbf{c}}^+ = \frac{\mathbf{c}^+}{\|\mathbf{c}^+\|_2},$$
   $$\mathbf{c}^- = \frac{\sum_{t \in S^-} \text{idf}(t)\, \mathbf{e}_t}{\sum_{t \in S^-} \text{idf}(t)}, \quad \hat{\mathbf{c}}^- = \frac{\mathbf{c}^-}{\|\mathbf{c}^-\|_2}.$$

   - *Semantic (green) axis:* Define the semantic axis vector:
   $$\mathbf{g} = \frac{\hat{\mathbf{c}}^+ - \hat{\mathbf{c}}^-}{\|\hat{\mathbf{c}}^+ - \hat{\mathbf{c}}^-\|_2}, \quad \|\mathbf{g}\|_2 = 1.$$

4. **Green-score computation.** For each objective obj_i, we compute its embedding:

$$\mathbf{o}_i = \text{Embed}(\text{obj}_i) \in \mathbb{R}^d,$$

and define the green-alignment score via cosine similarity:

$$\text{score}_i = \cos(\mathbf{o}_i, \mathbf{g}) = \frac{\mathbf{o}_i^\top \mathbf{g}}{\|\mathbf{o}_i\|_2 \|\mathbf{g}\|_2} = \frac{\mathbf{o}_i^\top \mathbf{g}}{\|\mathbf{o}_i\|_2}, \quad i = 1, \ldots, N.$$

5. Green-score **Interpretation.** A higher score_i implies stronger alignment of project i's objectives with the "green" semantic axis. The scores can range from -1 to 1, with -1 indicating a semantic meaning opposite to climate themes, 0 denoting neutral or unrelated content, and +1 reflecting near-perfect alignment with the semantic green axis.

6. Green-score **Verification**
To verify the scoring, projects from both ends of the green score distribution were manually inspected. Projects with high scores were found to be strongly climate-oriented, while those with low scores were clearly unrelated to green projects. Projects near zero appeared either semantically neutral with respect to green framing, or unrelated.

Table 3: Examples of project titles by green score range.

High Green Score Projects	
The Drought Impact on the Climate Benefit of Carbon Sequestration	0.315
Optimising forest management decisions for a low-carbon, climate resilient future in Europe	0.312
Carbon accounting in the forest-based industry - maximising the mitigation potential of wood products in Slovenia	0.306
Neutral Projects (Scores near 0)	
Molecular Ecology of Medieval European Landscapes	0.001
Inequality - Public Policy and Political Economy	-0.001
Light-modulated organic Electrolyte-gated Phototransistors	-0.001
Low Green Score Projects	
Heritage of Disease: The Art and Architectures of Early Modern Hospitals in European Cities	-0.299
Storytelling as Pharmakon in Premodernity and Beyond: Training the New Generation of Researchers in Health Humanities	-0.292
Reassessing Late Medieval Pharmacology: Logical and Metaphysical Tools in the Medical Context	-0.282
The Novel and the Heart: 1840–1940	-0.282

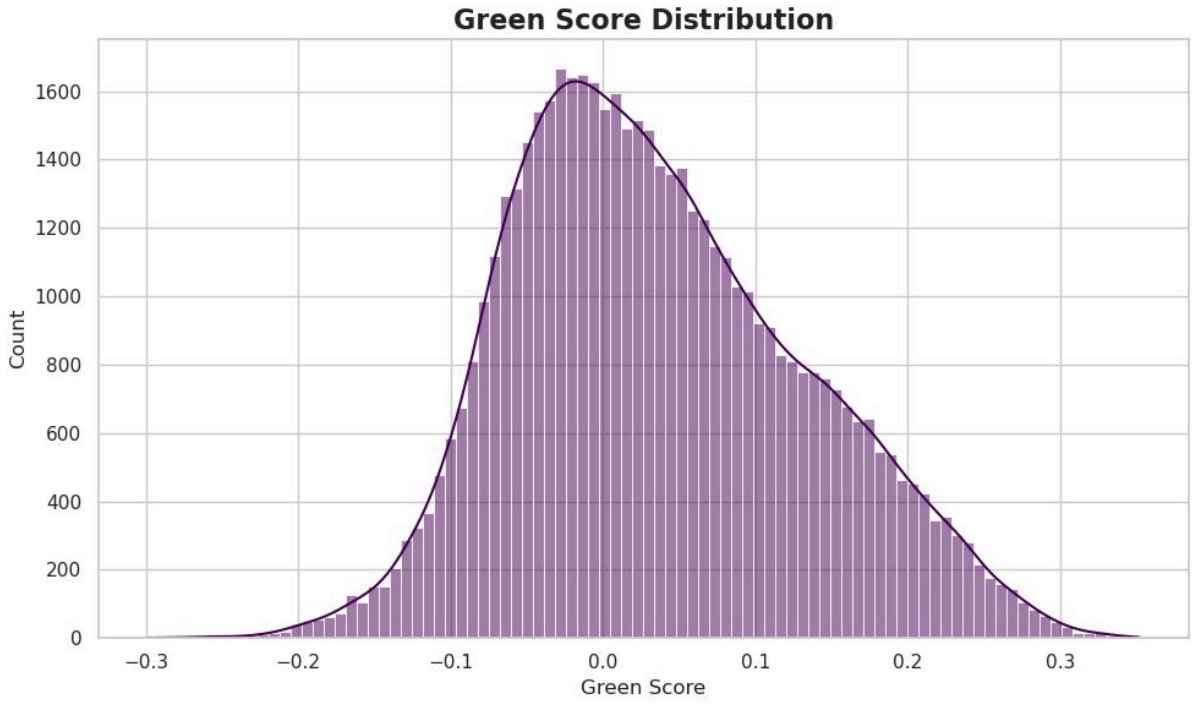


Figure 7: Distribution of Green Scores across the project corpus (range -1 to 1). Higher scores indicate stronger "green" semantic alignment, while values near 0 are neutral. The bimodal distribution suggests two populations: non-green projects clustering around 0 (left peak) and green-aligned projects tending towards 0.1-0.2 (right, heavier tail), confirming the score's effectiveness and raising suspicion of distinct 'green' and non-green populations.

7. **Climate classification**
   classification metrics and validation
   While the `climaText` dataset is primarily composed of short sentence samples. The model achieved an accuracy of 0.92 and a macro-F1 score of 0.91 on the held-out test set on the `climaText` dataset.

   However, in this project we are dealing with paragraph-level objectives, rather than sentence-level extracts. In order to dermine how well the model performs on this type of data in this domain, it is validated against 100 annotated projects (is green of type bool), generously performed by independent annotators. Before analysing the validation, another crucial step will be discussed.

8. **How to obtain a sample of purely green projects? The non-green percentile filter.**
   To ensure a high-confidence sample of climate-oriented projects from the SEDIA database of the HORIZON programme, we aim to minimize false positives in the green class predictions. As demonstrated in the baseline model results of the `BERTClimate` classifier (Figure 8), relying solely on classifier outputs introduces a substantial number of false positives.

   To address this, we propose a filtering strategy that combines the discrete class label with the continuous green cosine similarity score. The key assumption is that true green projects will exhibit higher cosine similarity with the green embedding

vector, while non-green predictions will show lower scores. If these distributions are sufficiently separable, a robust threshold can be defined. In our case, we use the 95th percentile of cosine scores from the non-green-classified distribution to filter out low-confidence green predictions. This ensures that any project classified as green must also exceed the similarity threshold, reducing the likelihood of false positives.

This method was validated using the annotated HORIZON dataset and yielded a clear shift in the precision-recall balance. As shown in the confusion matrices in Figure 8, the baseline model achieves full recall (TP = 15, FN = 0) but incurs many false positives (FP = 25), resulting in low precision and a mean cosine score of $\mu = 0.10$ in the FP quadrant. In contrast, the filtered model eliminates false positives entirely (FP = 0), achieving perfect precision, though at the cost of missing five true positives (FN = 5). Notably, the true positives retained by the filtered model exhibit a higher mean cosine similarity ($\mu = 0.19$) with reduced variance, indicating stronger alignment with the semantic embedding. A Mann–Whitney U test confirms that the cosine scores of retained true positives are significantly higher than those of removed false negatives ($p = 0.0007$), underscoring that our key assumption holds statistically true on the validation sample. The resulting sample is thus highly precise, with minimal noise, albeit at the expense of some recall, possibly introducing a small but acceptable bias.

Table 4: Extended Metric Comparison. Bold values indicate the superior metric. The table is grouped into threshold-dependent metrics, which vary between models, and threshold-independent metrics, which evaluate the underlying model scores and are unaffected by the filter.

Metric	**Baseline**	**Filtered**
Threshold-Dependent Metrics		
Accuracy	0.7500	**0.9500**
Balanced Accuracy	**0.8529**	0.8333
Precision (for "Green" class)	0.3750	**1.0000**
Recall (Sensitivity, TPR)	**1.0000**	0.6667
F1-Score (for "Green" class)	0.5455	**0.8000**
Specificity (TNR)	0.7059	**1.0000**
Negative Predictive Value (NPV)	**1.0000**	0.9444
False Positive Rate (FPR)	0.2941	**0.0000**
False Negative Rate (FNR)	**0.0000**	0.3333
Matthews Correlation Coefficient	0.5145	**0.7935**
Cohen's Kappa	0.4186	**0.7727**
Threshold-Independent Metrics (Unaffected by Filter)		
AUC-ROC	0.9537	
AUC-PR	0.7850	
Brier Score	0.2200	

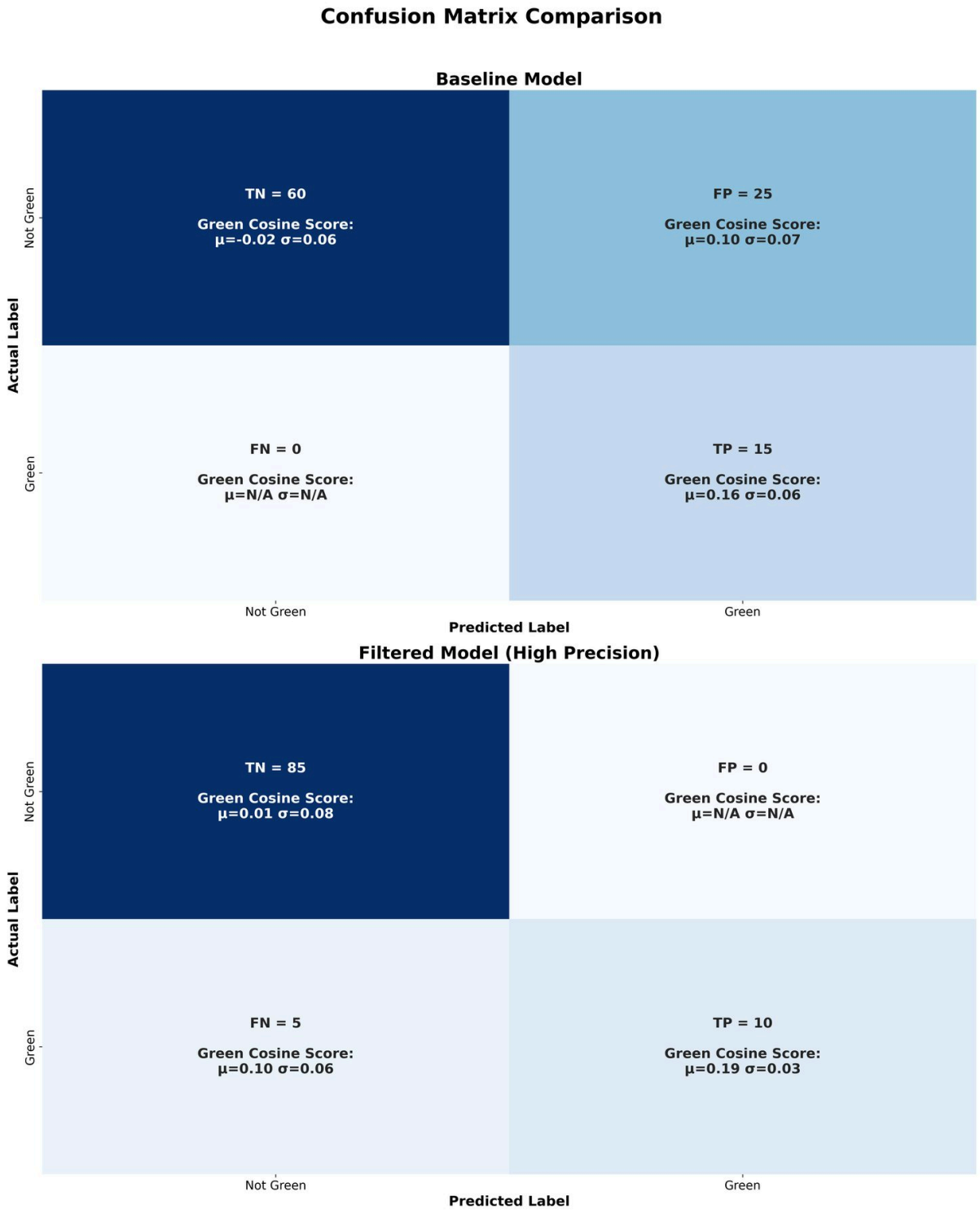


Figure 8: Confusion matrices comparing the **Baseline** and **Filtered (High Precision)** models, with cosine similarity scores for 'Green' predictions. The Baseline model achieves full recall but at the cost of many false positives, reflected in lower cosine similarity for incorrect predictions. The Filtered model eliminates false positives, improving precision and alignment with the embedding space, as confirmed by significantly higher cosine scores for retained true positives ($p = 0.0007$)

9. **Full Corpus Application** After validating the classifier and green percentile filter on the annotated validation set, we applied this methodology to the full corpus (N=53,345). Figure 9 displays the score distribution with the 95$^{\text{th}}$ and 99$^{\text{th}}$ percentile cut-offs derived from the non-green classified distribution. Summary statistics for the corpus are provided in Table 5. A clear visual distinction in green cosine scores between predicted categories is evident. To statistically validate this, a Welch's unequal-variance t-test was performed. Figure 10 shows the Q-Q diagnostics by class, to determine whether the distributions are Gaussian 'enough' to perform the t-test. Different percentile choices can be made for clear separation, a lower percentile increases the chance of false positives. Therefore, for all subsequent strategic analyses, we focus exclusively on the subset of 4,049 green-classified projects that exceed the 99th percentile threshold, representing a substantial dataset with a high degree of confidence in their green classification.

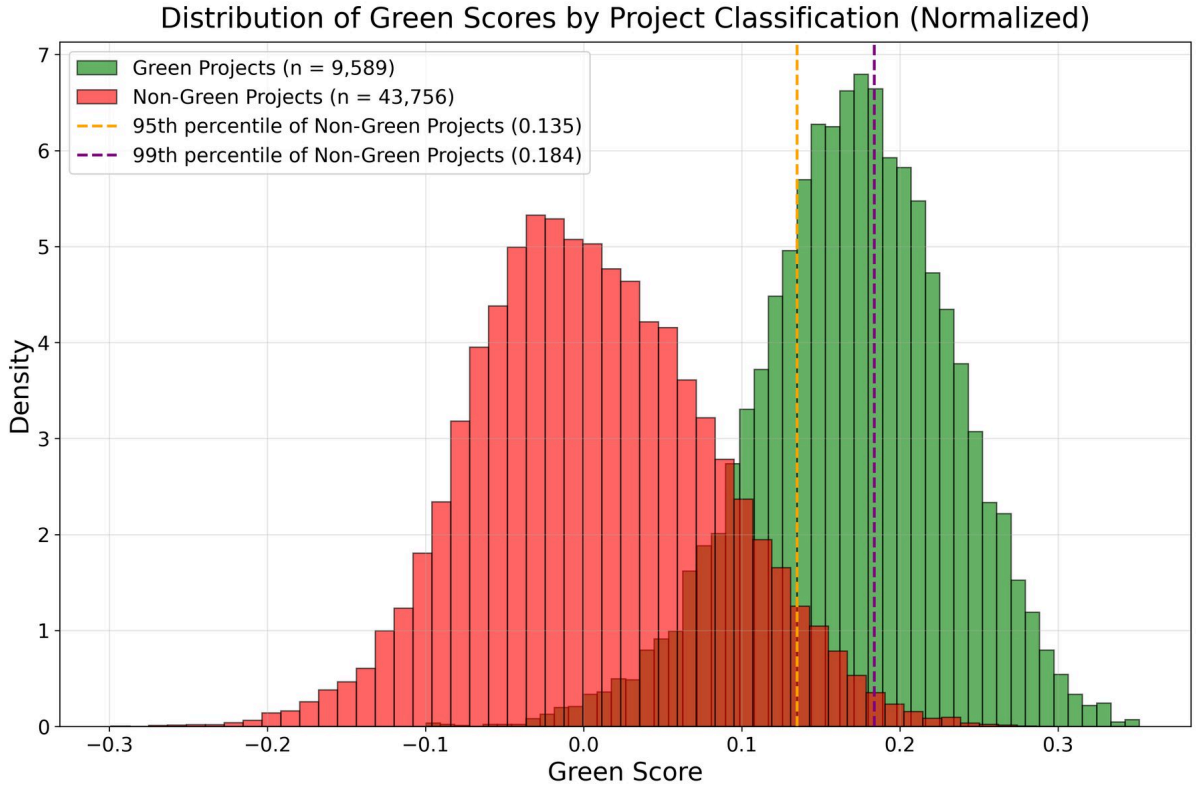


Figure 9: Density of the *green-score* over the entire corpus. Dashed orange and purple lines mark the 95$^{\text{th}}$ ($\tau_{0.95} = 0.135$) and 99$^{\text{th}}$ ($\tau_{0.99} = 0.184$) percentiles of the *non-green* reference distribution, respectively.

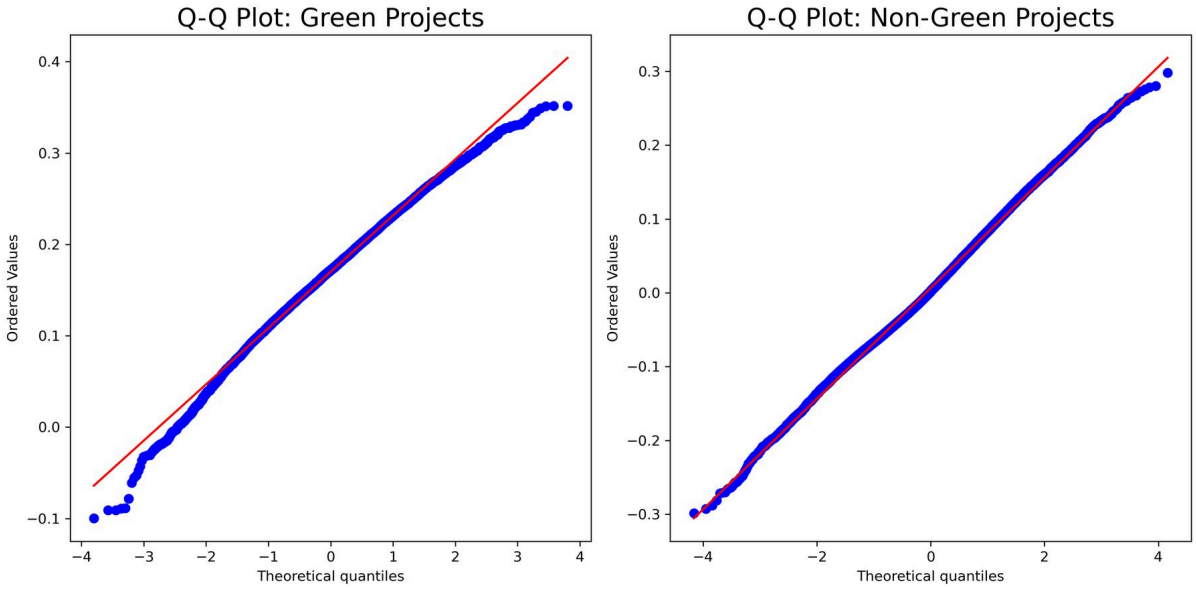


Figure 10: Q–Q diagnostics by class. The bulk of observations follow a distrubution close to a Gaussian distribution, heavier tails emerge for green projects. Given the sample sizes ($n \gg 30$), the Welch's t-test is legitimized.

Table 5: Corpus breakdown after full deployment of the classifier.

Metric	**Count**	**% of Total**
Total projects	53 345	100.0
Green Classified Projects	9 589	18.0
Non-Green Classified Projects	43 756	82.0
Green score $>95^{\text{th}}$ percentile	7 000	13.1
Green score $>99^{\text{th}}$ percentile	4 049	7.6

Welch's t-test comparing mean scores yields

$$t = 225.81, \qquad p < 10^{-300},$$

confirming a material divergence between green and non-green populations.

6.3 Methodology Step B - Zero-Shot Strategic Alignment

Zero-shot natural-language inference lets us score projects against policy goals without a single annotated label [27].

1. **Define the Green Corpus**
   Start with the *high-confidence climate subset*: cleaned project objective sentences destined for scoring.

2. **Define the prompts (labels):**
   Introduce `strategic_labels` that mirror EU priorities, extracted by heuristically determining key phrases sourced from the strategic plans [4]. Phrases containing jargon such as 'open strategic autonomy, competitiveness, energy security, geopolitical resilience, etc.'. Example: "This project enhances European energy security." All prompts can be found in the appendix (Subsection 11.3). These 11 labels were chosen to reduce common problems in zero-shot NLI, such as having too few examples, lexical-entailment bias and domain shift. This helps ensure the results stay accurate and aligned with EU strategic goals.

3. **Infer Alignment**
   Feed each *(objective, label)* pair into `facebook/bart-large-mnli`. Retain the entailment probability P_{ent}, ignore neutral and contradiction for this project.

4. **Aggregate and Normalise**
   Compute the *Strategic Alignment Index* as the mean P_{ent} across all labels for a project. Weight the scores using the relative funding weight. Several weighting mechanisms are tried to account for potential imbalances in money distribution, specifically focusing on the impact of high-value projects on the overall score (see Figure 11). These are:

   - **Max-weighted Score (Normalized):** This approach likely assigns a higher weight to the project with the maximum funding, aiming to see how the overall strategic alignment is influenced by the most significant financial investment. As seen in the histogram, this often results in a distribution skewed towards lower scores, meaning only a few projects with high funding contribute significantly, while others are down-weighted.

   - **Log-weighted Score (Normalized):** This method applies a logarithmic transformation to the funding weights. A logarithm dampens the effect of extreme values, meaning that while higher-funded projects still have more weight, the difference in weight between a very large project and a moderately large project is reduced. This helps to mitigate the disproportionate influence of a few extremely well-funded projects, leading to a more spread-out distribution of scores, as observed in the central part of the histogram.

   - **Rank-weighted Score (Normalized):** This approach assigns weights based on the rank of the projects by funding, rather than their absolute funding amount. This method ensures that all projects contribute to the weighted score, with higher-ranked projects having more influence. It effectively reduces the impact of the actual funding magnitude and focuses on the relative order of investment. This typically results in a distribution that is more evenly spread across the score range, similar to what's seen on the right side of the histogram, indicating a broader range of project contributions.

The goal of using these different weighting mechanisms is to understand how the definition of "strategic alignment" changes based on how funding is considered, particularly when there's a **significant disparity in the amount of money allocated to different projects**.

5. **Z-score normalization**: In order to isolate relative movement in the data set, standard z-score scaling is applied. This is then aggregated by year and the mean z score in the annual series makes distributional drift and any re-allocation effects explicit. Statistical analysis tools will be applied to these results.

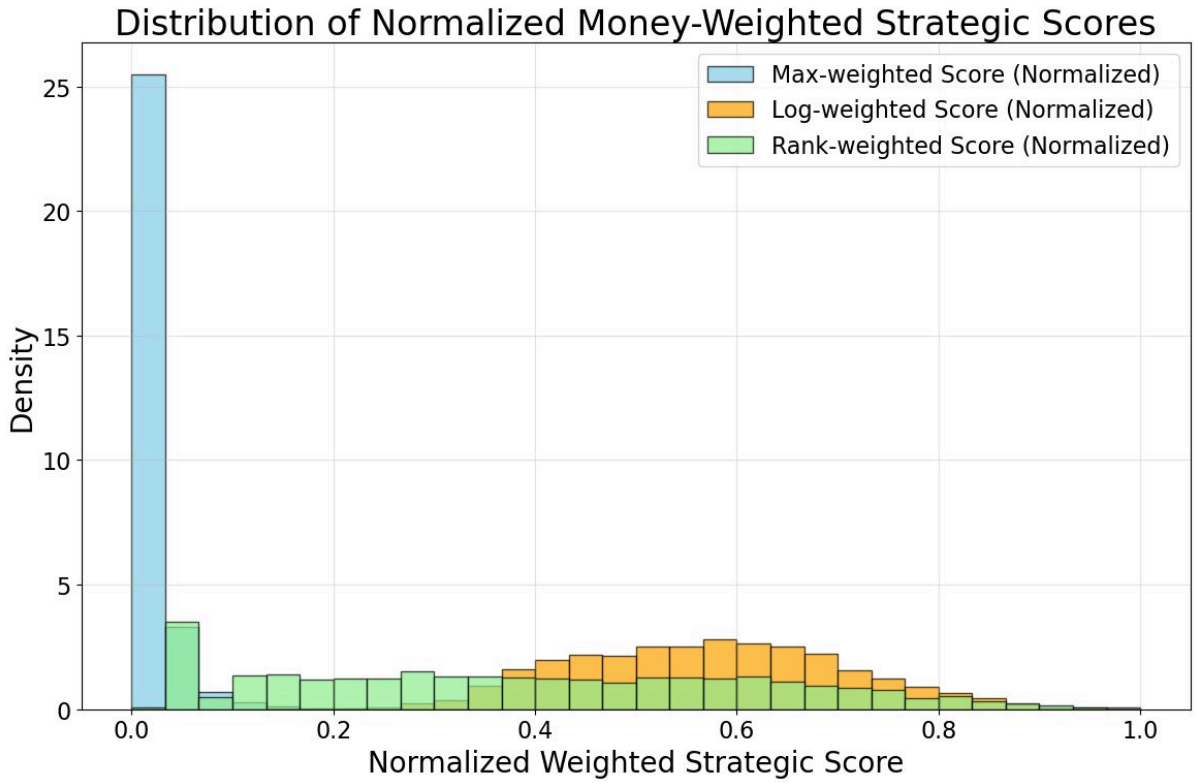


Figure 11: The figure displays the impact of different money-weighting techniques on the distribution of normalized strategic alignment scores, highlighting how weighting choice affects the score's aggregation.

Table 6: Examples of green-categorized projects with high strategic alignment

Project title	**Example of strategic alignment**	**Example of 'green' oriented objective**
Uptake of Solid Bioenergy in European Commercial Sectors – Bioenergy for Business	"the project will **contribute** to increase much-needed **security of energy supply** through lower dependence on fossil fuels from **politically volatile** sources."	"The goal of this project is to **promote the (partial) substitution of fossil fuels**"
Sustainable On-site and Innovative Technologies for Advanced Transport BioFuels from MicroalGae	"FuelGae will contribute to **advancing the European** scientific basis and **global technological leadership** in the area of renewable fuels, increase their technology competitiveness" "while supporting the EU goals for **energy independence**"	"FuelGae technologies will be further evaluated through life cycle assessment (LCA/LCC) to **confirm their lower environmental impact, use of resources, or GHG emissions**"
Rethinking the future of clean cooling through a revolutionary class of thermally-driven chiller based on a novel bio-based thermochemical material	"**securing our energy future in Europe**", "Europe will assert its **global research and innovation leadership**"	"In response to set targets for **reducing our carbon footprint**"

7 Results

The Russian invasion of Ukraine did not affect the project-level narrative or the funding allocation established following the strategic narrative shift outlined in the updated strategic plan [4] within the Horizon Europe (2020–2027) framework programme (see Figure 14). Figure 13 and the tests in Table **??** show a statistically significant uptick across several weighting schemes when comparing the two programme periods (before 2020 and after 2020). This indicates that the strategic shift had already been absorbed at the funding level, while the language dimension remains largely flat (unweighted strategic alignment). No notable shifts or trends are observed within the Horizon Europe programme itself when comparing pre- and post-invasion data, confirming that the major shift occurred between programmes, not within them.

Table 7: Two-sample t-tests were conducted on weighted and unweighted z-score means before and after the identified rhetorical policy shifts to assess whether the periods differ significantly in terms of strategic language and funding allocations. The funding shift seems significant between the two horizon programmes, but statistical evidence shows that within the current horizon programme no shift occured due to the Russian invasion. No shift occured in strategic language used in funded projects (see unweighted). Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Variant	**2014–2025 (Pre-2020 vs Post-2020)**	**2020–2025 (Pre-2022 vs Post-2022)**
max	$t = -2.45$, $p = 0.053$	$t = 0.03$, $p = 0.978$
log	$t = -3.45$, $p = 0.010^{**}$	$t = -0.11$, $p = 0.922$
rank	$t = -4.61$, $p = 0.003^{**}$	$t = -0.27$, $p = 0.801$
hybrid	$t = -3.90$, $p = 0.006^{**}$	$t = -0.17$, $p = 0.875$
unweighted	$t = 0.55$, $p = 0.598$	$t = 0.19$, $p = 0.865$
green	$t = -3.20$, $p = 0.012^{*}$	$t = -0.19$, $p = 0.869$

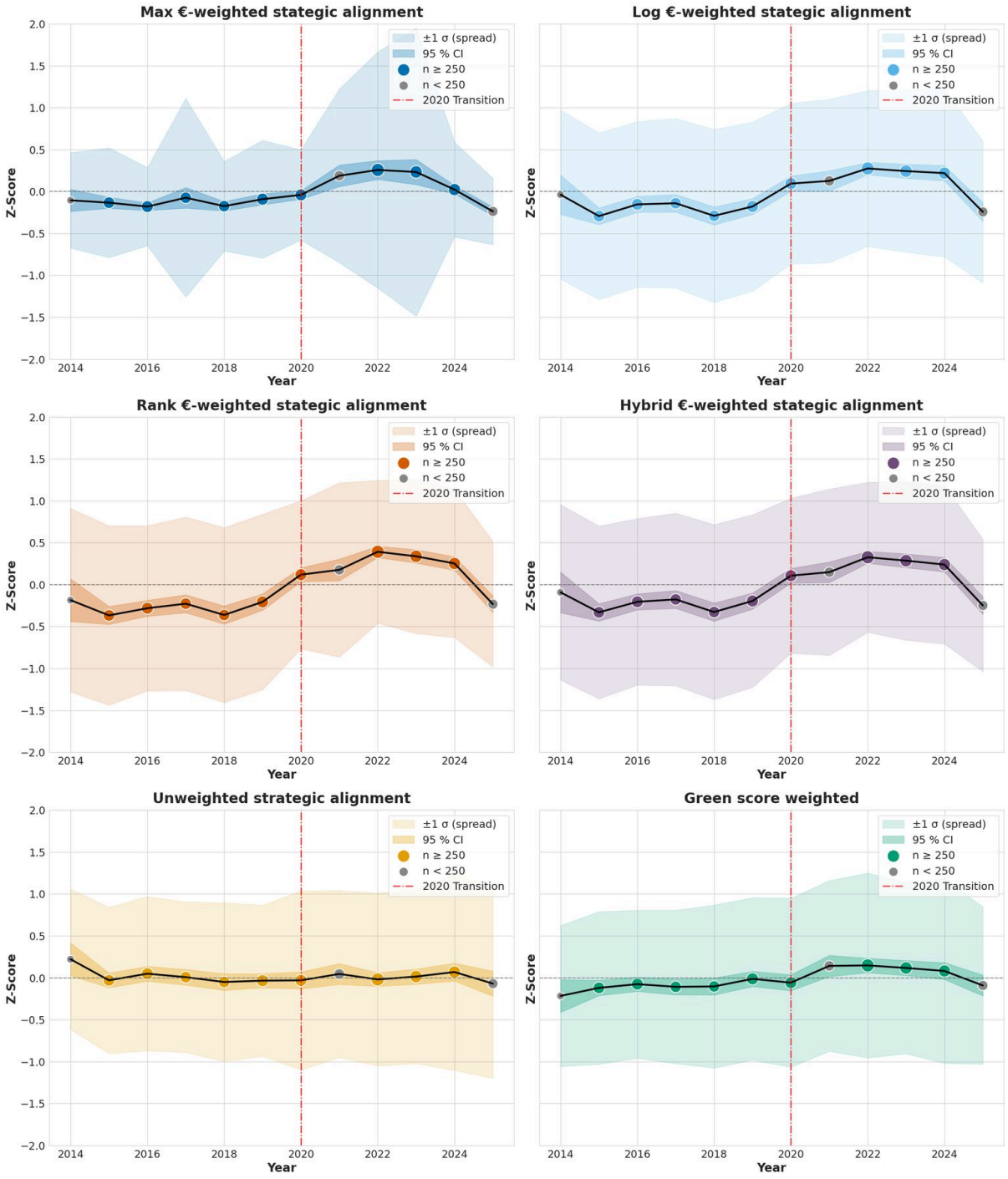


Figure 12: Results show that for each fund-weighting, there appears to be a peak during the first years the Horizon Europe (2020-2027) programme, that quickly descended over the last year. Table 7 indeed confirms that the programmes are statistically significantly different, indicating that strategic alignment has grown in importance in terms of funding between the two programmes. However, this same observation is not shown in the unweighted strategic alignment, which remains flat throughout time, indicating that the language of the projects itself did not change, only the funding allocation related to the strategic language did. No clear trends seem visible within the programmes, not clear shift is seen around the Russian invasion of Ukraine (2022).

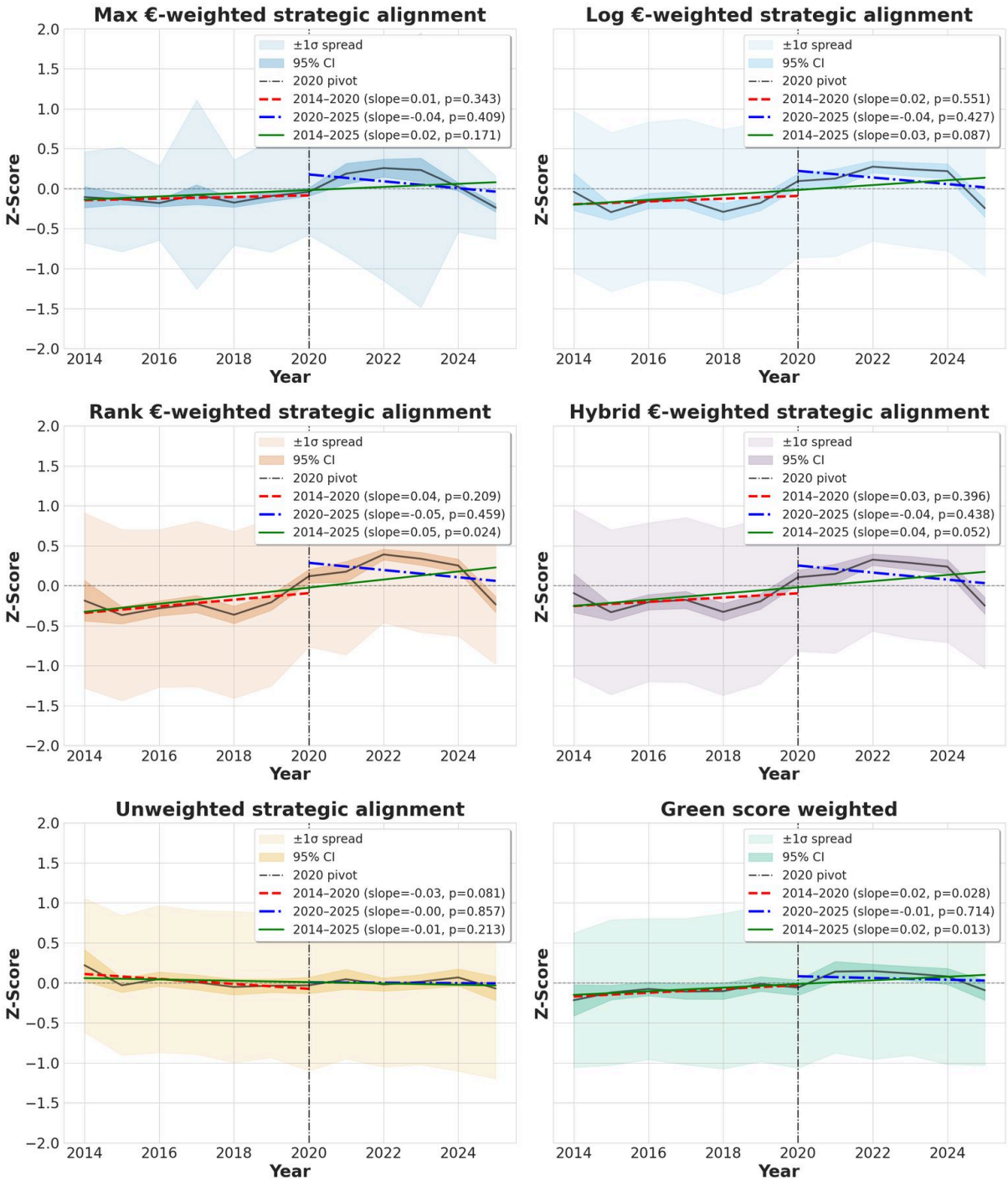


Figure 13: Lineair Regression analysis applied to the various weighted scores to inspect trends. Visually, we can determine a slight positive trend in the Horizon 2014-2020 programme, and a negative trend in the Horizon Europe (2020-2027) programme. The global trend appears to trend upwards, but this is only stastically significant for the ranked-fund-weighted strategic alignment ($p = 0.024 < 0.05$). None of the other trends are statistically significant ($p > 0.05$).

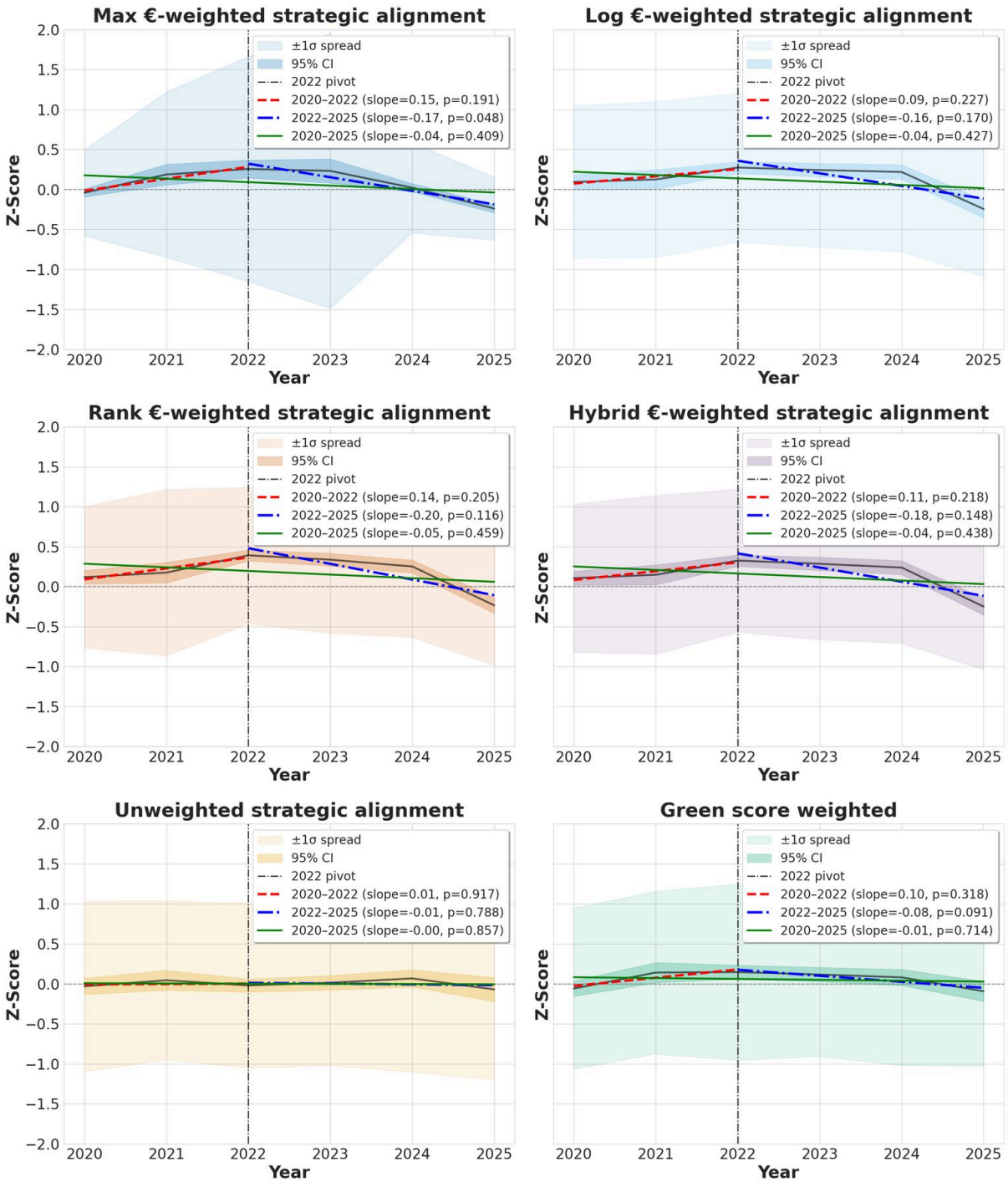


Figure 14: Regression analysis applied to the various weighted scores for the period 2020-2025, to observe impact of russian invasion of Ukraine. Visually, we can determine a positive trend before 2022, and a negative trend in the Horizon Europe after 2022. The global trend seems visually flat. None of the trends are statistically significant ($p > 0.05$).

8 Discussion

8.1 Data Science Approach

interpretability
This study aimed to identify green Horizon projects and assess their strategic framing independently of EU funding tags (Section 3.5.1). Our bottom-up approach extracted latent semantic features directly from project texts, which effectively mitigated 'policy biases' that would arise when adapting the EU architecture as a means of supplying classification labels. However, this model-driven approach introduced uncertainty because large language model (LLM) predictions reflect patterns detected in large sums of training data, rather than interpretable decision rules, complicating explanations for individual project labels.

Green alignment was assessed using two complementary techniques. First, we applied semantic matching via cosine similarity, comparing project embeddings to a predefined "green" template vector. This provided interpretability through relative semantic positioning. Second, we used ClimateBERT[29], a domain-specific model fine-tuned on climate-related texts, to classify projects as green or non-green. Despite robust validation performance (F1=0.80, precision=1.00; see Section 6.2), inherent ambiguity near classification boundaries remains. The conservative threshold we used effectively reduced false positives but increased false negatives. This reduced our sample size and amplified statistical uncertainty would especially problematic in niche thematic areas. Future implementations must carefully balance these trade-offs when adapting this methodology.

To enhance model interpretability and address these limitations, we propose applying BERTTopic modeling[30] separately to clearly green, clearly non-green, and borderline-classified projects. Such an analysis could clarify the implicit semantic distinctions driving model decisions, facilitating refinements of thresholds or prompts. Additionally, explainability tools like SHAP[31], as applied by Rodella et al. [11], could offer deeper insights into individual project-level model decisions, helping to mitigate the inherent opacity of LLMs.

Moreover, the abstract "funding-weighted strategic" and "green" scores provide valuable trend-level insights but remain opaque, which limits direct interpretability for policymakers and the general public alike, who are not familiar with advanced text analysis techniques and may have difficulty understanding what such a score exactly entails. To overcome these practical constraints, the context of the results should always be clear. This can be achieved by providing explicit examples, for example, by creating interactive plots that would allow one to see the individual texts when hovering over a visualization of the scores.

Limitations of the application of zero shot prompting For strategic alignment, our study used zero-shot natural language inference (NLI) guided by prompts that reflected geopolitical and policy framings. This approach allowed us to estimate strategic relevance without labeled data, but its interpretability remains limited due to inherent opacity and sensitivity to prompt wording. Minor lexical variations significantly affected predictions, although aggregating across thousands of projects and multiple prompts provided robustness at the trend level. Similar topic-modeling analyses, as proposed for green alignment, could also elucidate implicit strategic framings within this zero-shot

classification approach.

Moreover, by discarding neutral and contradictory probabilities from the NLI model the strategic alignment estimates are possibly inflated. Future iterations should integrate all probabilities, also those countering strategic alignment, to enable a more nuanced interpretation of alignment scores. Also, the general-purpose nature of the NLI models leads to diminished performance when confronted with specialized EU policy terminology. Targeted, lightweight domain adaptation using limited annotated datasets ($2,000$ premise-hypothesis pairs) is recommended to significantly improve classification accuracy in this specific context.

Limited validatoin
Finally, a crucial evaluation gap exists due to the absence of quantitative validation for the strategic alignment scores. Conducting targeted validation studies, such as expert annotation of a representative subset (approximately 300 projects), is essential for benchmarking. Addressing this gap will provide robust empirical validation, strengthening analytical rigor and enhancing policy relevance.

Note: Due to project timelines, these improvements were not implemented in this iteration, but are highly encouraged to be taken into consideration in future projects.

8.2 Research question

We addressed our research questions by applying text analysis techniques to quantify strategic drift via a funding-weighted strategic score, rather than raw counts of strategic language. The results indicate a clear temporal breakpoint in funding-weighted strategic alignment, with a statistically significant increase between the Horizon 2014–2020 and Horizon Europe 2020–2027 programmes. However, no such shift is evident within the Horizon Europe programmes itself, particularly not around the 2022 Russian invasion of Ukraine. This undermines claims that recent geopolitical events triggered a meaningful reorientation in funding priorities. Instead, the data suggest that the strategic repositioning was already embedded in the programme's initial design.

Importantly, while weighted strategic scores increased across several variants (ranked, hybrid, and logarithmic), the unweighted strategic alignment remained flat over time. This decoupling implies that the project language itself did not evolve in response to geopolitical developments; rather, projects with existing strategic framework were preferentially funded, especially in the early phases of Horizon Europe. Figure 13 illustrates this trend, with an initial peak in strategic alignment, followed by a tapering decline. Table 7 confirms this statistically, but also confirms the absence of any significant shifts pre/post-invasion (2022), reinforcing the conclusion that the transition occurred between programmes, not within them.

This divergence from the official narrative suggests a disconnect between high-level political framing as a response to major geopolitical events and actual resource allocation. Societally, this matters, as democratic stakeholders should be able to judge whether major geopolitical events have elicited adequate responses. Transparancy is crucial, especially in volatile times.

9 Conclusion

This study set out to assess whether green research under Horizon Europe has undergone a measurable strategic reframing in light of shifting geopolitical narratives (RQ1). Contrary to expectations raised by shifts in political rhetoric in response to major geopolitical events, like the Russian invasion in Ukraine, the project level objectives did not conflate with geopolitical language as a response. Neither have significant shifts and trends occurred in terms of language over the last decade. However, funding allocation (RQ2) did shift significantly between the two horizon programs, especially at the start of the latest Horizon program, significantly more money was allocated toward projects with objectives with strategically aligned language. However, within the programmes, no significant shifts or trends were observed, even in light of major geopolitical events.

A significant contribution of this project is the development of the `sedia-fetchers-api`[14] Python package. This package has the potential to support future work, contributing to open science and enhancing transparancy in EU policy.

10 Terminology

- EuroSciVoc : European Science Vocabulary.
- Horizon: The latest funding programme (FP9).
- CORDIS : Community Research and Development Information Service, "the European Commission's primary source for the results of projects funded by the EU's framework programmes for research and innovation, from FP1 to Horizon Europe." (https://cordis.europa.eu/about). CORDIS is basically a website hosting information on EU-funded projects and their results. It is itself funded by HORIZON Europe.
- **Framework Programmers**, or **Framework Programmes for Research and Technological Development**, abbreviated **FP1** to **FP9**, are funding programmes created by the EU/EC. starting 2014, the funding programmes were named Horizon. The latest funding programme is called Horizon Europe (FP9).
- **SME**: small and medium-sized enterprises.
- **SDG**: Sustainable Development Goal.
- **TRLs**: Technology Readiness Levels.
- **R&I**: Research and Innovation.
- **SEDIA**: Single Electronic Data Interchange Area, which is the underlying API for the EU Funding & Tenders Portal and serves as the operational platform for EU grants and tenders.

11 Appendix

11.1 Strategic scores per prompt analyzed over the years

Figures 15 and 16 display the temporal evolution of strategic scores for each individual prompt from 2014 to 2025. The scores for prompts such as “enhances energy security and independence” and “highlights the urgency of innovation and deployment driven by geopolitical instability” show a modest upward trend beginning around 2022. This increase is more pronounced in the most recent year, though the sample size in that period is relatively limited. Most other prompts display little to no systematic variation across time.

Some prompts, notably “promotes European open strategic autonomy” and “supports Europe’s energy sovereignty and resilience”, exhibit increases around the launch of the new Horizon funding programme. Meanwhile, prompts with overt military or security framing, such as “addresses armed conflict and geopolitical security challenges” and “is relevant to European security and defence”, remain consistently low throughout the entire period.

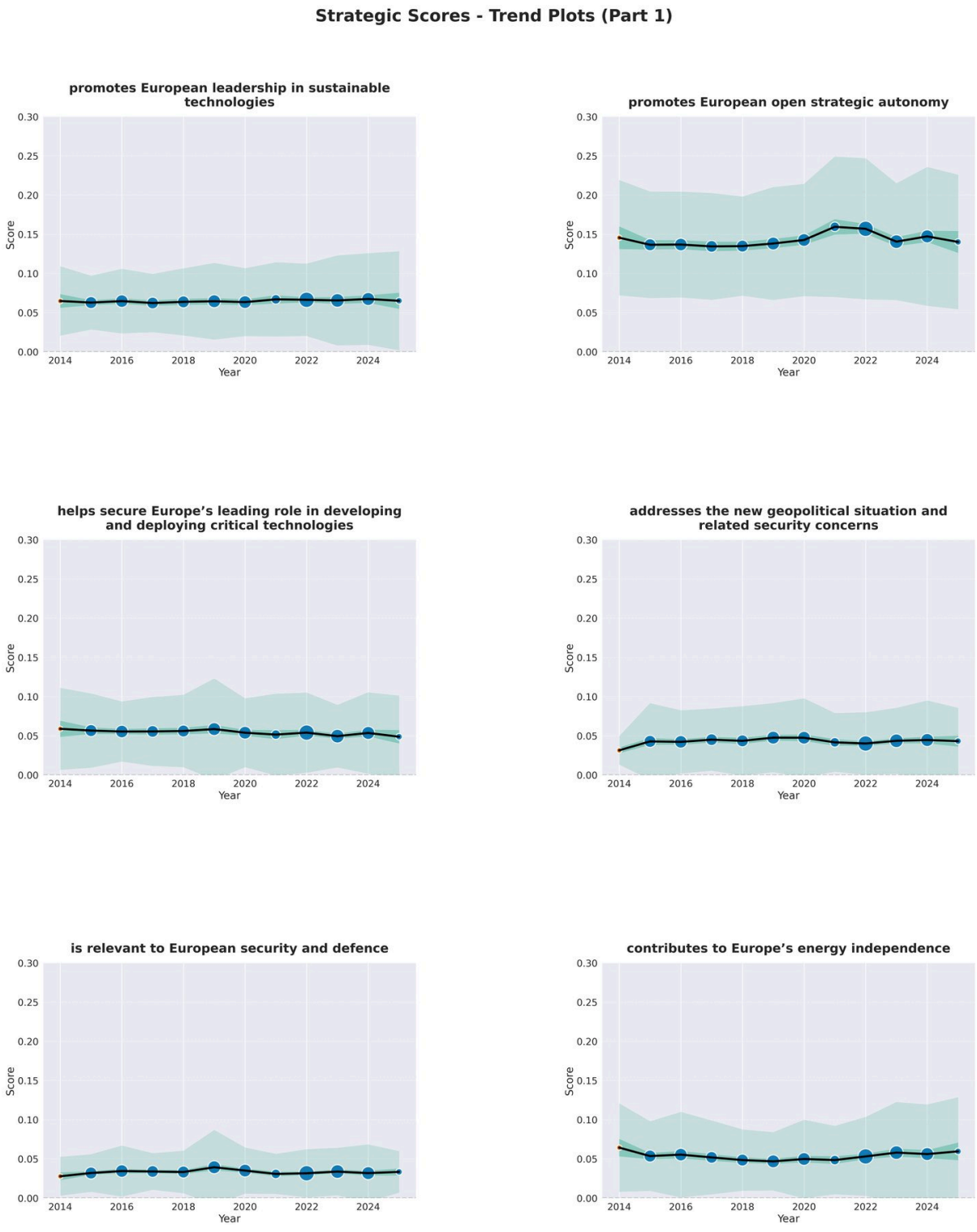


Figure 15: Enter Caption

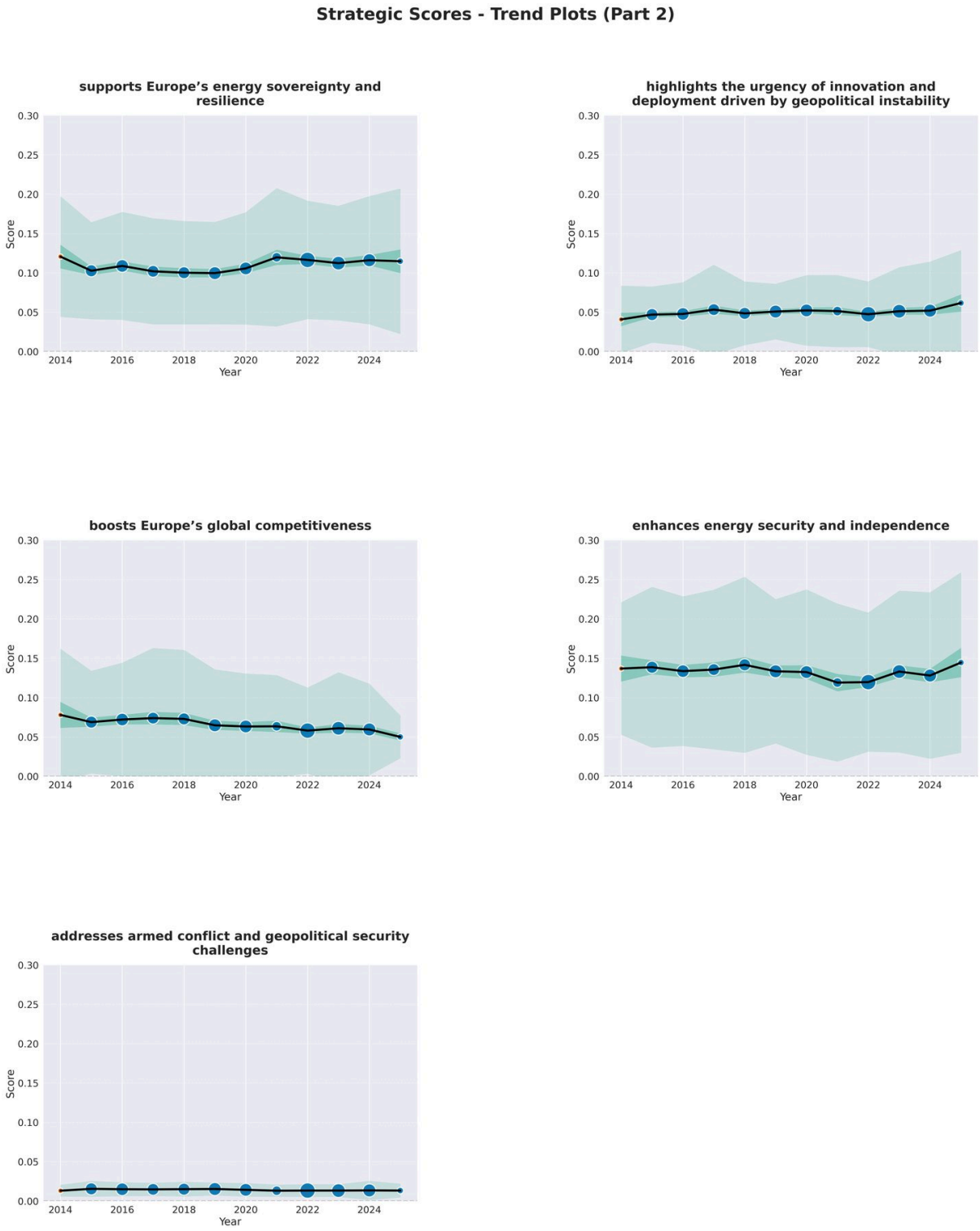


Figure 16: Enter Caption

11.2 Examples of high-confidence green projects, 5 projects with highest green scores, and 5 with lowest.

#	Title	Green Score	Objective
1	Exploring CO_2 activation in heterogeneous catalytic CO_2 hydrogenation		Carbon dioxide (CO_2) released into the atmosphere because of burning fossil fuels and industrial production processes has increased in concentration and enhanced the "greenhouse effect", causing climate change and extreme weather. Converting CO_2 into energy-rich fuels is one of the sustainable solutions to mitigate carbon emission challenges, while also remedying the over-reliance on fossil fuels as well as contributing to the European Green Deal. A reliable and affordable CO_2 conversion process remains a grand challenge because thermodynamic and kinetic constraints limit the reactions; hence novel approaches are needed to increase reaction rates by altering kinetics to reduce activation energy. This project will focus on CO_2 activation and hydrogenation to valuable products via energetic catalysts and plasma or electromagnetic activation to create "hot spots" at low temperature. Supported metal catalysts with oxygen vacancy sites will be designed for efficient, selective hydrogenation under mild conditions, deepening understanding and aiding industrial valorisation toward 90% CO_2 cuts by 2040 and climate neutrality by 2050.

#	Title	Green Score	Objective
2	Photo-electro Integrated Next-Generation energy technologies		To combat global warming, achieving net-zero or net-negative CO_2 emissions is imperative. The EU targets a 40% emissions reduction and 32% renewables by 2030. The PHOENIX project integrates photovoltaic-electrolyzer (PV–EC) and photoelectrochemical (PEC) technologies to convert CO_2 into CO and propanol and simultaneously recycle PET waste into glycolic acid. It will develop a tandem photovoltaic system (>2 V), novel electrocatalysts for CO→n-PrOH and PET→glycolic acid, and efficient photoelectrodes. Lab-scale demonstration (TRL 3–4), LCA, and recyclability studies will validate the environmental impact, delivering a high-risk/high-return breakthrough.
3	Minimum environmental impact ultra-efficient cores for aircraft propulsion	0.9984	Building a climate-neutral aviation future is critical to stay below +1.5 °C. MINIMAL unites European engine OEMs, atmospheric physics experts, and propulsion researchers to tackle non-CO_2 radiative forcing (contrails, NO_x) and CO_2 emissions through composite-cycle engines. Targets: 80% contrail reduction, 52% NO_x cut, and 36˘100%CO_2 reduction (fuel-dependent) by 2035. Numerical (TRL 2) and experimental (TRL 3) proof-of-concept of low-NO_x opposed-piston constant-volume combustion with hydrogen micromixing underpin an aggressive roadmap toward product readiness by 2035–2040.

#	Title	Green Score	Objective
4	Sustainable Water-Injecting Turbofan Comprising Hybrid-electrics		SWITCH addresses climate-neutral short-medium-range transport via a dual-spool hybrid-electric turbofan with water injection and waste heat recovery. Goals: 20% fuel burn saving and 50% reduction of NO_x & contrail climate impact versus state-of-the-art. Combines megawatt-class motors, low-emission combustor, and lightweight structures. Electric taxiing enhances local air quality. Compatible with 100% SAF and hydrogen. Phase 1 (2026) TRL 5 ground demo; Phase 2 (2030) TRL 6 flight test & TRL 5 waste-heat demo. Foundation for market entry by 2035, supporting 2050 neutrality.
5	Development of an innovative Gas Turbine Chemical Looping Combustor for Carbon Negative Power Generation		Negative emissions are essential to meet climate targets. This project develops a chemical looping combustor (CLC) for biofuels, enabling power and heat with inherent CO_2 capture. Using oxygen carriers in a circulating fluidized-bed reactor, the CLC yields a pure CO_2 stream for sequestration. The combustor will be tested on multi-metal oxide carriers, piloted in a 50 kW dual fluidized bed, and scaled with industrial partners to deliver a low-cost, rapid-deployment carbon-negative power plant.

Table 8: Top 5 Projects by Green Score (Full Objectives)

#	Title	Green Score	Objective
1	Transforming Weather & Water data into value-added Information services for sustainable Growth in Africa		Provide currently unavailable geo-information on weather, water, and climate for sub-Saharan Africa by enhancing satellite data with innovative in-situ sensors and developing information services tailored to African stakeholders and the GEOSS community. A feedback loop will reciprocally validate in-situ measurements and satellite data across 500+ citizen-science stations.
2	Hybrid Power for General Aviation (HyPoGA)	0.9941	Develop a superefficient 200 kW hybridized aircraft engine (140 kW combustion + 60 kW electric) to cut fuel consumption 30% and operating costs 27%. Certification per EU/US regulations will enable retrofits in 3 000 annual replacement units. A market, technological, and financial feasibility study (SWOT, design, marketing, certification, financial plan) will culminate in a business plan.
3	Precise and smart nanoengineered surfaces: Impact resistance, icephobicity and dropwise condensation		Design thermodynamically guided metallic surfaces (<10 nm morphology) with controlled stiffness for passive icing prevention and sustained dropwise condensation. Embed polymers/suspensions to enhance abrasion and chemical resistance, achieve icephobicity to –30 °C, and maintain condensation at 50–100 m/s vapor speeds. Applications include aircraft anti-icing, heat-pump evaporators, and industrial condensers.

#	Title	Green Score	Objective
4	Emissions SooT ModEl		ESTiMatE will develop CFD-based soot prediction models for aero-engines using detailed kerosene surrogate kinetics and advanced combustion/spray models validated against experiments. The methodology couples soot particles with gas-phase dynamics for high-fidelity, large-scale simulations, enhancing predictivity and reliability of soot forecasts in aeronautics.
5	AgroBiogel International Scale-up		Agrobiogel is a lignin-based hydrogel soil improver that stores and releases water/nutrients, saving up to 40% irrigation, protecting against drought and erratic rainfall, reducing input costs, and converting non-productive soils (e.g., sand) into arable land. Scaled production valorizes pulp-and-paper biorefinery lignin, cutting greenhouse gas emissions and expanding industrial portfolios.

Table 9: Bottom 5 Projects by Green Score (Full Objectives)

11.3 Prompts used for zero prompt classification

.

P1 This project promotes European leadership in sustainable technologies.

P2 This project promotes European open strategic autonomy.

P3 This project helps secure Europe's leading role in developing and deploying critical technologies.

P4 This project addresses the new geopolitical situation and related security concerns.

P5 This project is relevant to European security and defence.

P6 This project contributes to Europe's energy independence.

P7 This project supports Europe's energy sovereignty and resilience.

P8 This project highlights the urgency of innovation and deployment driven by geopolitical instability.

P9 This project boosts Europe's global competitiveness.

P10 This project enhances energy security and independence.

P11 This project addresses armed conflict and geopolitical security challenges.

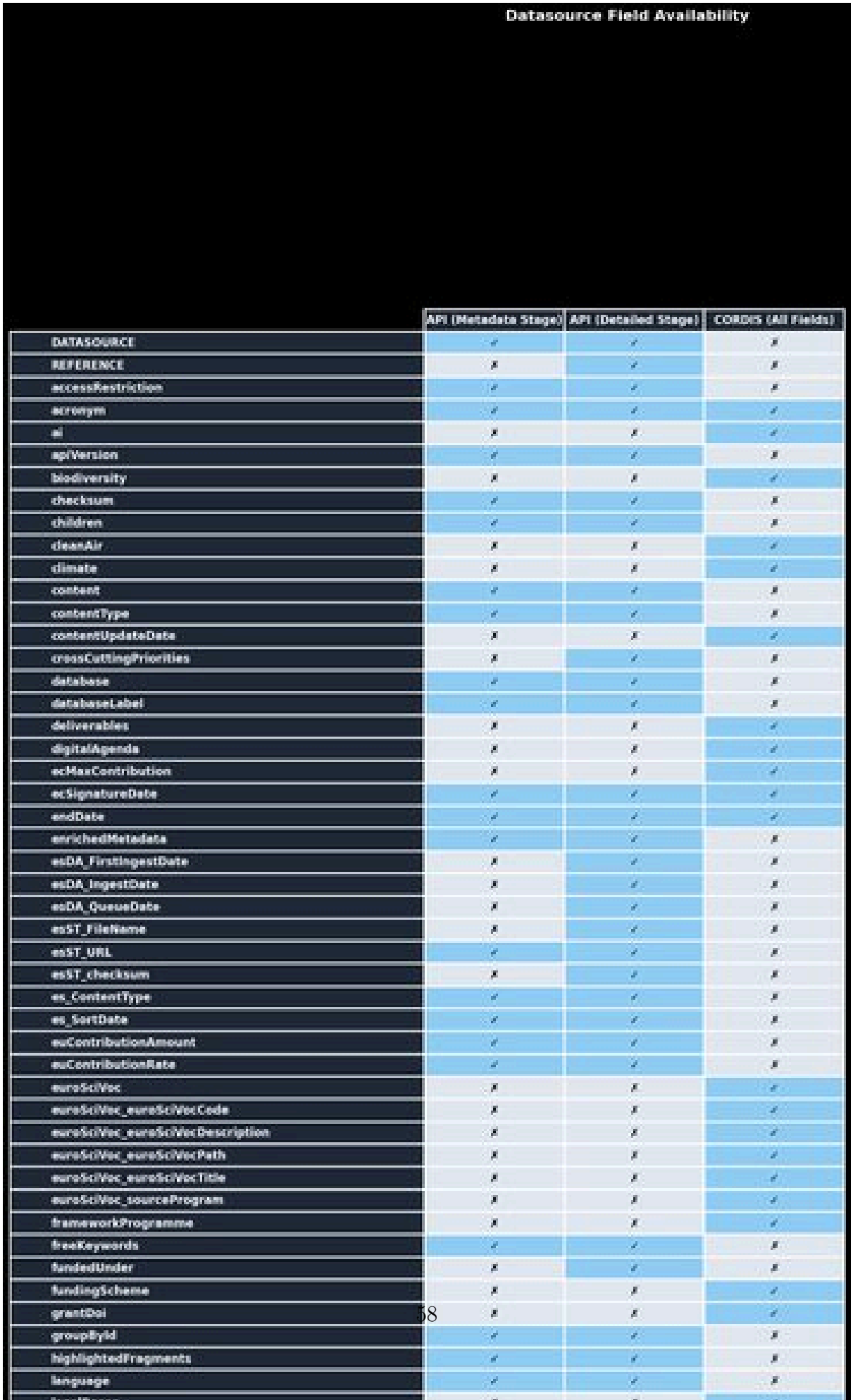
11.4 Code snippets

11.4.1 Temporal partitioning algorithm

```
def _fetch_with_date_partitioning(self, base_query: dict, sort: dict, total_records: int,
                                  min_date_str: str, max_date_str: str) -> pd.DataFrame:
    """
    Fetch data using recursive date range partitioning.
    This is the core algorithm that handles datasets > 10k records.
    """
    all_dfs = []
    date_format = "%Y-%m-%dT%H:%M:%S.%f"

    try:
        min_date = datetime.strptime(min_date_str.split('+')[0], date_format)
        max_date = datetime.strptime(max_date_str.split('+')[0], date_format)
    except ValueError as e:
        print(f"Error parsing dates: {e}")
        print("Falling back to non-partitioned fetch")
        return self._fetch_paginated_chunk(base_query, sort, min(total_records, self.API_FETCH_LIMIT))

    ranges_to_process = [(min_date, max_date)]
    self.pbar = tqdm(total=total_records, desc="Overall Progress", unit="rec")

    while ranges_to_process:
        start_date, end_date = ranges_to_process.pop(0)

        # Create date-constrained query with temporal filter
        range_query = {
```

```
            "bool": {
                "must": base_query["bool"]["must"].copy() + [
                    {
                        "range": {
                            "es_SortDate": {
                                "gte": start_date.strftime(date_format)[:-3] + "Z",
                                "lte": end_date.strftime(date_format)[:-3] + "Z",
                            }
                        }
                    }
                ]
            }
        }

        # Get count for this date range
        count_for_range, _, _ = self._get_metadata_with_date_range(range_query)

        if count_for_range == 0:
            continue

        # RECURSIVE PARTITIONING: If still > 10k, split the date range
        if count_for_range <= self.API_FETCH_LIMIT:
            print(f"Fetching chunk: {count_for_range:,} records ({start_date.date()}
            ↪  to {end_date.date()})")
            chunk_df = self._fetch_paginated_chunk(range_query, sort, count_for_range)
            all_dfs.append(chunk_df)
        else:
            print(f"Splitting large chunk: {count_for_range:,} records
            ↪  ({start_date.date()} to {end_date.date()})")
            # Split the date range in half and add both halves back to processing
            ↪  queue
            mid_point = start_date + (end_date - start_date) / 2
            ranges_to_process.append((start_date, mid_point))
            ranges_to_process.append((mid_point + timedelta(microseconds=1),
            ↪  end_date))

    self.pbar.close()

    if not all_dfs:
        print("No data was fetched.")
        return pd.DataFrame()

    print("Concatenating all chunks...")
    final_df = pd.concat(all_dfs, ignore_index=True)

    print(f"Fetch complete: {len(final_df):,} records retrieved (expected
    ↪  ~{total_records:,})")
    if abs(len(final_df) - total_records) > (total_records * 0.05):  # 5% tolerance
        print(f"Warning: Record count difference > 5%. Expected: {total_records:,},
        ↪  Got: {len(final_df):,}")

    return final_df
```

11.5 Field availability compared between datasources



11.5.1 green terms

- **UNFCCC Glossary Seed (UNFCCC_GLOSSARY)**: Terms sourced from the UNFCCC Glossary of Terms.
  - ◇ Afforestation
  - ◇ Biosphere
  - ◇ Carbon cycle
  - ◇ Carbon dioxide
  - ◇ Carbon equivalent
  - ◇ Carbon sequestration
  - ◇ Carbon sinks
  - ◇ Carbon storage
  - ◇ Climate feedback
  - ◇ Climate lag
  - ◇ Climate sensitivity
  - ◇ Climate system
  - ◇ Earth system
  - ◇ Co-control benefit
  - ◇ Cooling Degree Days
  - ◇ Deforestation
  - ◇ Emissions coefficient
  - ◇ Emissions factor
  - ◇ Emissions
  - ◇ Energy conservation
  - ◇ Energy efficiency
  - ◇ Enhanced greenhouse effect
  - ◇ Forcing mechanism
  - ◇ energy balance climate system
  - ◇ Unintended gas leaks Fugitive emissions
  - ◇ Global Warming Potential
  - ◇ Global Warming
  - ◇ Greenhouse effect
  - ◇ Greenhouse gas

- ◇ GHG greenhouse gas
- ◇ Nonbiodegradable
- ◇ Renewable energy
- ◇ recycling
- ◇ Temperature

- **IPCC Glossary Seed (IPCC_GLOSSARY)**: Terms derived from the IPCC Glossary.
  - ◇ climate variability
  - ◇ Antarctic Circumpolar Current
  - ◇ AOGCM
  - ◇ Biodiversity
  - ◇ Biodiversity Hot Spots
  - ◇ Biofuels
  - ◇ Biomass
  - ◇ Biomass Energy
  - ◇ Carbon Cycle
  - ◇ CO2
  - ◇ burning fossil fuels and biomass
  - ◇ anthropogenic greenhouse gas
  - ◇ Earth's radiative balance
  - ◇ Carbon Flux
  - ◇ climate average weather
  - ◇ Climate Change
  - ◇ Climate Model
  - ◇ Coupled atmosphere ocean sea-ice General Circulation Models
  - ◇ Climate Prediction
  - ◇ Climate Projection
  - ◇ climate forecast
  - ◇ response of the climate system to emissions
  - ◇ Climate Scenario
  - ◇ Climate Variability
  - ◇ Desertification

- ◇ El Niño-Southern Oscillation
- ◇ ENSO
- ◇ Ecosystems Services
- ◇ Ecotone
- ◇ Eustatic Sea-Level Rise
- ◇ sea level rise
- ◇ Extreme Weather Event
- ◇ Forest
- ◇ General Circulation Model
- ◇ GCM General Circulation Model
- ◇ Glacier
- ◇ Greenhouse Effect
- ◇ Ice Cap
- ◇ Ice Sheet
- ◇ Ice shelf
- ◇ climate impact
- ◇ Kyoto Protocol
- ◇ La Niña
- ◇ Microclimate
- ◇ climate mitigation
- ◇ reduce greenhouse gas source
- ◇ enhance sinks of greenhouse gases
- ◇ Net Biome Production
- ◇ Net Biome Production NPB
- ◇ Net Ecosystem Production NEP
- ◇ Net Ecosystem Production
- ◇ Particulates
- ◇ Permafrost
- ◇ Reforestation
- ◇ sequestration
- ◇ Sustainable Development

- **EUROSCIVOC**: Terms from the EUROSCIVOC thesaurus.

- ◇ climatology
- ◇ climate science
- ◇ climate study
- ◇ climate extreme
- ◇ arctic oscillation
- ◇ climate oscillation
- ◇ climate variability
- ◇ climate crisis
- ◇ climate change adapation

- **HORIZON_INFO**: Terms related to Horizon programmes and European policy documents.
  - ◇ UN Sustainable Development Goals
  - ◇ UN SDG
  - ◇ SDG
  - ◇ Paris Agreement
  - ◇ climate change adapation
  - ◇ climate-neutral and smart cities
  - ◇ A climate resilient Europe
  - ◇ climate disruptions preparedness
  - ◇ climate resilience
  - ◇ climate neutral cities
  - ◇ clean hydrogen
  - ◇ clean aviation
  - ◇ zero-emissions vehicles
  - ◇ zero-emission road transport
  - ◇ zero-emission waterborne transport
  - ◇ clean energy transition
  - ◇ Green transition
  - ◇ transition to climate neutrality
  - ◇ transition to zero emissions
  - ◇ transition to low emissions
  - ◇ transition to negative emissions

- ◇ reducing environmental footprint
- ◇ European Climate Law
- ◇ European Green Deal
- ◇ Fit for 55 package
- ◇ REPowerEU
- ◇ Circular Economy action plan
- ◇ zero-pollution action plan
- ◇ Sustainable and Smart Mobility Strategy
- ◇ Net-Zero Industry Act
- ◇ green jobs
- ◇ climate-neutral society
- ◇ climate-resilient society
- ◇ global emissions pathways
- ◇ impacts on climate change
- ◇ impacts of climate change
- ◇ climate services as response strategies
- ◇ climate-resilient, low-emission development
- ◇ achieving climate neutrality by 2050
- ◇ zero-pollution ambition
- ◇ energy and resource efficiency of cities
- ◇ renewable energy and decarbonisation
- ◇ solar
- ◇ wind
- ◇ geothermal
- ◇ ocean
- ◇ hydropower
- ◇ biomethane
- ◇ advanced biofuels
- ◇ synthetic renewable fuels
- ◇ Net-Zero Industry Act
- ◇ heat pumps
- ◇ renewable energy system integration

- ◇ Climate attribution
- ◇ anthropogenic climate change
- ◇ global climate models reference simulations
- ◇ climate change impact
- ◇ climate action

- **EU_STRATEGY**: Terms related to broader EU strategic objectives.
  - ◇ maintain the autonomy and competitiveness of the EU energy supply.
  - ◇ autonomy
  - ◇ competitiveness
  - ◇ strategic
  - ◇ leadership
  - ◇ global forefront
  - ◇ security
- **GENERIC**: General terms that indicate scientific inquiry or solutions.
  - ◇ actionable solutions
  - ◇ challenges
  - ◇ Advancing science
  - ◇ closing major knowledge gaps
  - ◇ synergies
  - ◇ trade-offs
  - ◇ other policy objectives
  - ◇ Progress
  - ◇ artificial intelligence solutions can be used to accelerate research in these crucial topics
  - ◇ effective response strategies
  - ◇ fostering synergies
  - ◇ boosting scientific excellence
- **ANTI_SEED**: Terms used to exclude projects from climate-related categories (e.g., humanities, health).
  - ◇ lyric poetry
  - ◇ renaissance art
  - ◇ cultural heritage

◇ media literacy

◇ gender studies

◇ philosophy of language

◇ classical philology

◇ social innovation

◇ public administration efficiency

◇ digital humanities

◇ ancient manuscripts

◇ ethnography

◇ linguistics

◇ medieval literature

◇ ancient history

◇ art history research

◇ health

◇ cancer

◇ healthcare

◇ healthcare system

◇ Health Emergency Preparedness

◇ human health

◇ mental health

◇ non-communicable

◇ disease

◇ covid

◇ covid-19

◇ pandemic

◇ epidemic

◇ epidemiology

◇ epidemic preparedness

◇ pharmaceutical

◇ pharmaceutical industry

◇ pharmaceutical research

◇ cell biology

◇ eye health

◇ brain

◇ human neural

References

[1] European Commission – President von der Leyen. *Opening address by President von der Leyen on the Clean Industrial Deal at the European Industry Summit*. 2025. URL: https://ec.europa.eu/commission/presscorner/detail/en/speech_25_628.

[2] European Commission – President von der Leyen. *Press statement by President von der Leyen on the defence package*. 2025. URL: https://ec.europa.eu/commission/presscorner/api/files/document/print/et/statement_25_673/STATEMENT_25_673_EN.pdf.

[3] Bundesregierung Deutschland – Olaf Scholz. *Policy statement by Olaf Scholz, Chancellor of the Federal Republic of Germany and Member of the German Bundestag, 27 February 2022 in Berlin*. 2022. URL: https://www.bundesregierung.de/breg-en/service/archive/policy-statement-by-olaf-scholz-chancellor-of-the-federal-republic-of-germany-and-member-of-the-german-bundestag-27-february-2022-in-berlin-2008378.

[4] European Commission and Directorate-General for Research and Innovation. *Horizon Europe strategic plan 2025-2027*. Publications Office of the European Union, 2024. DOI: 10.2777/092911. URL: https://data.europa.eu/doi/10.2777/092911.

[5] European Parliament and Council. *Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013 (Text with EEA relevance)*. Official Journal of the European Union, L 170, 1–68. Consolidated version in force: 01/03/2024. 2021. URL: https://eur-lex.europa.eu/eli/reg/2021/695/oj.

[6] European Commission – Directorate-General for Defence Industry and Space. *A Powerful Instrument to Boost Industrial Defence Cooperation throughout Europe*. June 2021. URL: https://defence-industry-space.ec.europa.eu/system/files/2021-06/DEFIS%20_%20EDF%20Factsheet%20_%2030%20June%202021.pdf.

[7] European Council – President Charles Michel. *Speech by President Charles Michel on the main challenges facing Europe, at Sciences Po, Paris, 28 March 2022*. 2022. URL: https://www.consilium.europa.eu/en/press/press-releases/2022/03/28/intervention-du-president-charles-michel-lors-de-la-conference-sur-les-grands-enjeux-europeens-a-sciences-po-paris.

[8] *About Us*. Open Future Foundation. URL: https://openfuture.eu/about/ (visited on July 7, 2025).

[9] European Commission and Directorate-General for Communication. *The European Green Deal – Delivering the EU's 2030 climate targets*. Publications Office of the European Union, 2023. DOI: 10.2775/783179. URL: https://data.europa.eu/doi/10.2775/783179.

[10] Domokos Esztergár-Kiss. "Horizon 2020 Project Analysis by Using Topic Modelling Techniques in the Field of Transport". In: *Transport and Telecommunication Journal* 25 (June 2024), pp. 266–277. DOI: 10.2478/ttj-2024-0019. URL: https://doi.org/10.2478/ttj-2024-0019.

[11] I. Rodella, A. Sciandra, and A. Tuzzi. "Textual Analysis of the Marie Skłodowska-Curie Actions Evaluation Summary Reports. Assessing Strengths and Weaknesses of Funded and Non-Funded Proposals". In: *Italian Journal of Sociology of Education* 17.1 (2025), pp. 247–266. DOI: 10.25430/pupj-IJSE-2025-1-12. URL: https://pupj.eu/journal/ijse/article/view/12.

[12] Gregor Cerinšek and Dan Podjed. ""Parole, parole": Unveiling the narrative framework of EU research and innovation projects". In: *Politics & Policy* 52.6 (2024), pp. 1210–1226. DOI: 10.1111/polp.12636. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/polp.12636.

[13] Zahar Koretsky et al. "A qualitative-computational cataloguing of the EU-level public research and innovation portfolio of clean energy technologies (2014–2020)". In: *Current Research in Environmental Sustainability* 3 (2021), p. 100084. ISSN: 2666-0490. DOI: 10.1016/j.crsust.2021.1000608. URL: https://www.sciencedirect.com/science/article/pii/S2666049021000608.

[14] R.A.J. Swarts. *sedia-api-fetchers: Python utilities for European Commission SEDIA endpoints*. Software package. Source code at https://github.com/ajruben/sedia-api-fetchers. Install with `pip install sedia-api-fetchers==1.0.0`. License: MIT. Utrecht University and Open Future, July 2025. URL: https://pypi.org/project/sedia-api-fetchers/1.0.0/.

[15] R. A. J. Swarts. *Data Wrangling*. Zenodo; accessed 2025-07-07. 2025. DOI: 10.5281/zenodo.15832371. URL: https://doi.org/10.5281/zenodo.15832371.

[16] R. A. J. Swarts. *Data Set Result*. Zenodo; accessed 2025-07-07. 2025. DOI: 10.5281/zenodo.15832371. URL: https://doi.org/10.5281/zenodo.15832371.

[17] Publications Office of the European Union. *CORDIS - EU research projects under HORIZON EUROPE (2021-2027) [Dataset] and CORDIS - EU research projects under Horizon 2020 (2014-2020) [Dataset]*. Accessed 2025-05-01. 2020. DOI: 10.2906/112117098108/20. URL: https://cordis.europa.eu/data.

[18] European Commission and Directorate-General for Research and Innovation. *Horizon Europe, the EU research and innovation programme (2021-27) – For a green, healthy, digital and inclusive Europe*. Publications Office of the European Union, 2021. DOI: 10.2777/052084. URL: https://data.europa.eu/doi/10.2777/052084.

[19] European Commission and Directorate-General for Research and Innovation. *Horizon Europe, pillar I - Excellent science – Driving scientific excellence and supporting the EU's position as a world leader in science*. Publications Office of the European Union, 2021. DOI: 10.2777/456952. URL: https://data.europa.eu/doi/10.2777/456952.

[20] European Commission and Directorate-General for Research and Innovation. *Horizon Europe, pillar II - Global challenges and european industrial competitiveness*. Publications Office of the European Union, 2021. DOI: 10.2777/881197. URL: https://data.europa.eu/doi/10.2777/881197.

[21] European Commission, Directorate-General for Research, and Innovation. *Horizon Europe, pillar III - Innovative Europe – Supporting and connecting innovators*

across Europe. Publications Office of the European Union, 2021. DOI: doi/10.2777/90204. URL: https://data.europa.eu/doi/10.2777/90204.

[22] European Commission and Directorate-General for Defence Industry and Space. *Introducing the White Paper for European Defence and the ReArm Europe Plan - Readiness 2030*. Mar. 2025. URL: https://defence-industry-space.ec.europa.eu/eu-defence-industry/introducing-white-paper-european-defence-and-rearm-europe-plan-readiness-2030_en (visited on July 3, 2025).

[23] European Commission, Directorate-General for Research and Innovation. *Funding & Tenders Portal Application Programming Interfaces (APIs)*. 2023. URL: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-results (visited on July 7, 2025).

[24] Yanzhao Zhang et al. "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models". In: *arXiv preprint arXiv:2506.05176* (2025).

[25] Kenneth Enevoldsen et al. "MMTEB: Massive Multilingual Text Embedding Benchmark". In: *arXiv preprint arXiv:2502.13595* (2025). DOI: 10.48550/arXiv.2502.13595. URL: https://arxiv.org/abs/2502.13595.

[26] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL]. URL: https://arxiv.org/abs/1910.13461.

[27] Wenpeng Yin, Jamaal Hay, and Dan Roth. *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach*. 2019. arXiv: 1909.00161 [cs.CL]. URL: https://arxiv.org/abs/1909.00161.

[28] Meta AI. *facebook/bart-large-mnli*. https://huggingface.co/facebook/bart-large-mnli. BART-large checkpoint fine-tuned on MultiNLI; accessed 2025-06-01. 2020.

[29] Julia Bingler et al. *How Cheap Talk in Climate Disclosures Relates to Climate Initiatives, Corporate Emissions, and Reputation Risk*. Working paper. Available at SSRN 3998435, 2023.

[30] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022). arXiv: 2203.05794. URL: https://arxiv.org/abs/2203.05794.

[31] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.