# Extraction to Coherence

using the GEA Framework

# Transfer Document


Alethea.
Veneta Angelova
Aadira Das
Kailen Roa Bernando

# Version History

| Version | Date | Author | Remarks / Changes |
|---|---|---|---|
| V0.1 | 23-01-2026 | Veneta Angelova, Kailen Roa Bernando, Aadira Das | • Initial setup of document and chapters |
| V1 | 24-01-2026 | Kailen Roa, Veneta Angelova, Aadira Das | • All chapters were completed |


**Version History** ..... **2**
Introduction ..... 4
Document Reading Guide ..... 4
Files Reading Guide ..... 5
Core Files ..... 5
1. Current Situation and Desired Future Outcome ..... 6
2. Research Overview ..... 7
Research Overview ..... 7
Key Findings ..... 7
Implications and Recommendations for the Next Group ..... 8
3. Data ..... 8
3.1 Data Overview and Structure ..... 8
3.2 Data Dictionary ..... 9
4. Data Cleaning ..... 10
4.1 Initial data inspection ..... 10
4.2 Cleaning and Processing ..... 11
5. Data Analysis ..... 12
5.1 Exploratory Observations & Key Findings ..... 12
6. Discovery Phase (Candidate URL Generation) ..... 13
Key Design Decisions ..... 13
6.1 Initial Design Intent ..... 13
6.2 Practical Challenges Encountered ..... 13
6.2.1 Inconsistent Website Structures ..... 14
6.2.2 Poor extraction Quality for certain GEA Statements ..... 14
6.3. Scope Reduction Decision ..... 15

6.4. Final Web Scraping Objective..... 16
6.5. Link Discovery & Selection Logic (Final Version)..... 16
6.6. Output of the Webscraping Phase..... 17
7. PDF Statement Extraction..... 17
7.1. Statement Filtering..... 18
7.2. Training & Model Selection..... 19
7.3. Statement Structuring..... 20
7.4. Modelling & Decision Logic..... 21
7.5. Role Within the Overall GEA Framework..... 21
8. Keyword Extraction..... 22
8.1 Method Chosen: YAKE..... 22
9. Classification and Modeling..... 23
9.1 Purpose..... 23
9.2 Semantic Representation via Sentence Embeddings..... 23
9.3 Natural Language Inference (NLI)..... 23
9.4 Hybrid scoring model..... 24
10. Coherence Matrix..... 25
10.1 Structure and dimensions..... 25
10.2 Pairwise evaluation..... 25
10.3 Scoring..... 26
10.4 Multi matrix..... 26
11. Demo (Interactive Interface)..... 27
11.1. Webscrapper..... 28
11.2. PDF extraction..... 30
11.3. Measure coherence..... 32
12. Work Done..... 34
13. Limitations..... 34
14. Future Project Concept..... 34
Final Note..... 35

# Introduction

This document serves as a comprehensive transfer document for the Extraction to Coherence project conducted during the advanced AI semester. GEA is an enterprise architecture methodology that analyzes the coherence within an organization by examining core elements such as mission, vision, core values, strategy, and goals. A GEA framework makes this coherence explicit, measurable, and manageable, thereby providing managers with a powerful tool to improve decision making and organizational performance. Traditionally, building such a framework requires intensive manual effort, as relevant statements must be gathered from documents, memos, or interviews with management.

The project focuses on measuring extracting statements such as: mission, vision, core values, goals and strategy and determining the strategic coherence between those statements. Throughout the semester research was conducted to determine the best possible way to achieve this within the scope of the semester.

The goal of this document is to enable a new researcher, developer or project team to fully understand what was done, why specific technical and methodological choices were made, and how the project can be continued or extended without prior involvement where future groups can improve or replace components

# Document Reading Guide

This document goes over the design implementation and evaluation of the extraction to coherence pipeline. It combines conceptual research, data processing, NLP modeling, and an interactive demonstration. This guide explains how the sections relate and how new readers can navigate the document efficiently.

The document follows the logical order of the implemented workflow. It first describes the problem context and desired future outcome followed by a research overview explaining the approach.

The data sections describe the structure, theoretical grounding and preparation of the datasets used throughout the project. These sections define the GEA statement types, explain how inconsistencies and multilingual content were handled and present exploratory findings.

The discovery and web scraping section explains how candidate web pages are identified. This section clarifies the system boundary of the scraper and the technical rationale behind focusing on high-confidence links rather than full statement extraction from websites.

The PDF extraction section presents the core technical logic for extracting GEA statements from annual reports. It explains the filtering strategy, embedding-based relevance scoring, rule-based decision logic and the structured outputs used for validation and downstream analysis. This section should be read as a single unit.

The modeling and coherence sections describe how extracted statements are transformed into semantic representations, how logical consistency is evaluated and how hybrid scoring produces interpretable coherence scores. These scores are consolidated into coherence matrices that enable systematic analysis of strategic alignment across GEA dimensions.

The demo section explains how the pipeline is exposed through an interactive interface. It illustrates how non-technical users can review, correct and validate extracted statements and coherence scores using a human-in-the-loop approach.

The document concludes with a summary of completed work, identified limitations, and directions for future research. These sections define which components are stable, which remain experimental, and where future teams can extend or replace parts of the pipeline.

# Files Reading Guide

### Core Files

- `all_companies_final.xlsx` – Raw dataset containing company statements and labels
- Excel outputs – Company-level coherence matrices
- `Company_statements_200.xlsx`- Data generated from 200 company statements (stock market)
- `Data scraping Eviden.html`- Notebook data scrapper (scrapy)
- `Extracted_Statements_Final.ipynb -` Notebook with experiments on extraction
- `WebScrapper(final).ipynb`- Final working webscrapper extraction links
- `Eviden_classifier.html` - Notebook on experiments with classification of statements
- `Eviden_Clustering.html` - Notebook exploring semantic clustering of strategic statements to analyze thematic similarity and potential coherence patterns.
- `Data_Preparation_PDF_Final_v1.html`- Extracted statements (bigger chunks)
- `Training_model (1).ipynb`- training model with Sentence-BERT model.
- `PDF_extractor(final).ipynb`- Final PDF extractor (modelling)
- `Coherence Matrix(final).ipynb`- Final Coherence measurement notebook with all relationships (Final)
- `Coherence_Strategies_vs_Goals.ipynb` - Coherence measurement Notebook on strategies vs goals
- `app.py` - Demo
- `candidates.csv` – All discovered candidate URLs with scores and metadata
- `top17_live.txt` – A filtered list of the best live URLs for scraping
- `out_md/` – Folder containing cleaned Markdown versions of scraped pages
- `manifest.csv` – Log of scraping results (status, blocks, snippets)

- `best_sites_raw.csv` – All scored pages with labels and relevance scores
- `Best_sites_top.csv`– Top-ranked pages per company

# 1. Current Situation and Desired Future Outcome

The purpose of this section is to provide an overview of the different approaches that were explored to automatically extract company statements and measure their coherence under the GEA framework from publicly available web sources (webpages, open source documentation), as well as the limitations encountered with each approach. In practice such statements are rarely provided in a consistent or standardized format - they are usually scattered across various different long documents such as annual reports and websites, which makes the identification of these statements rarely done systematically, and is typically handled manually on a case by case basis.

This creates several major challenges. First, the identification and extraction process is very subjective, and requires expert input - free text chunks need to be interpreted and labeled as a mission, vision, strategy, goal or core value without explicit headings or standardized structure. Secondly, a scalable validation framework does not exist to confirm whether the extracted statements are correct, consistent or complete and aligned with what an expert would have chosen.

Our approach for this project was to start by addressing the data acquisition challenge by constructing 2 complimentary pipelines. For web source extraction, the *`WebScrapper(final)`* implements a batch scraper that collects corporate pages, cleans boilerplate content, converts pages to markdown, and produces a file with keyword-weighted rankings such as “Mission”, “Vision”, “Strategy”, “Sustainability” in order to prioritize the most relevant pages per company. For PDF sources such as annual reports, the *`PDF_extractor`* and *`Data_Preparation_PDF_Final`* notebooks focus on extracting candidate text blocks using layout-aware segmentation, cleaning and deduplication, and finally generating structured outputs (CSV/JSON tables) containing candidate statements with metadata such as PDF filename, page number, predicted label.

Only after statement extraction is complete, the project proceeds to coherence scoring as implemented in the coherence notebooks, where paired couples (pairing initially specified by GEA framework) are added into a matrix for GEA-style evaluation. This order of work ensures that coherence modeling is applied to a traceable, cleaned and structured set of statements rather than noisy raw web/PDF content.

Third, the labeled data provided by the client was very limited and not sufficient to successfully train a model.

# 2. Research Overview

## Research Overview

This project explores the automated analysis of organizational strategy through two sequential research components: **(1) identifying and classifying strategic statements**, and **(2) evaluating strategic coherence across those statements**. Rather than immediately applying supervised machine learning techniques, the project deliberately begins with **rule-based discovery**, **keyword-based scoring**, and **model-assisted reasoning using chain-of-thought (CoT)**.

The first stage of the research focuses on identifying and classifying strategic statements—such as mission, vision, goals, core values, and strategy statements—from unstructured organizational text. This task was addressed using heuristic rules and keyword patterns, allowing transparent and interpretable classification without requiring labeled training data. This approach also enabled deeper understanding of how organizations express strategy in practice and highlighted ambiguities between different statement types.

Building on this foundation, the second stage evaluates **strategic coherence** by analyzing semantic alignment and logical consistency between the identified statements. Keyword-based scoring and semantic similarity measures were used to assess whether paired couple statements (such as mission statement versus visions) meaningfully align with each other. Logical consistency rules were applied to reduce false positives and enforce structural relationships between statement types.

These design choices were motivated by the need for explainability, the lack of annotated data, and the importance of developing domain understanding before introducing statistical models. The framework was designed to be modular, allowing heuristic components to later be replaced or augmented with embedding-based methods or supervised classifiers.

## Key Findings

- Rule-based and keyword-driven methods were effective to a limited extend at identifying explicit mission, vision, goal, and value statements when conventional strategic language was used.
- Distinguishing between closely related statement types (e.g., mission vs. vision) proved challenging in cases of abstract or overlapping language.
- Semantic similarity measures improved coherence assessment compared to keyword matching alone, particularly when aligned concepts were expressed using different terminology.
- Logical consistency checks helped filter out misleading coherence scores by enforcing expected relationships between statement types.
- Heuristic-based approaches provided strong baseline performance with high interpretability, though they struggled with nuanced or implicit strategic expressions.

### Implications and Recommendations for the Next Group

- Future work should explore hybrid or embedding-based methods for classifying strategic statements, especially to handle ambiguous language.
- Starting with developing a small expert-labeled dataset would significantly improve evaluation and enable supervised learning approaches.
- The existing rule-based framework can be retained as an explainability or validation layer alongside more complex models.
- Further research should investigate how errors in statement classification propagate into coherence scoring.
- Comparing automated coherence assessments directly with expert judgments is recommended to validate accuracy.

# 3. Data

## 3.1 Data Overview and Structure

The dataset `evidendata.csv` contains the following columns:

| **Column** | **Description** |
|---|---|
| Company | Organization name |
| Statement | Strategic statement text |
| Type | Category (mission, vision, core value, goal, strategy) |

**PDF Files**
All PDF files which were used to test out the extraction notebook code were gathered by the team from various companies. For this purpose we focused mainly on obtaining only Annual reports, since this is one of the few options which are available to obtain freely from the internet. All annual reports which were obtained by us for the project can be found in the git repository.

## 3.2 Data Dictionary

The following table summarizes the conceptual definitions and relationships between key strategic statement types used in this research. These definitions were extracted from the GEA (Enterprise Coherence) framework and serve as the theoretical foundation for distinguishing between mission, vision, core values, goals, and strategy statements within organizational texts. The table was used as a structured reference framework for the rule-based identification of statement types and for evaluating strategic coherence, ensuring consistency between the underlying theory and the applied analysis.

| Statement Type | Definition | Key Characteristics | Summary |
|---|---|---|---|
| **Mission** | A mission defines the fundamental purpose of an enterprise. It represents a higher-level goal that is enduringly pursued but never fully fulfilled. The mission is derived from the vision and communicates the enterprise's reason for existence to the market and key stakeholders. | Enduring, purpose-driven, stakeholder-oriented, not time-bound | The fundamental purpose of the enterprise |
| **Vision** | A vision is a concise and aspirational statement that operationalises the mission. It expresses how the enterprise intends to be perceived by the world and presents an inspiring image of its desired future state. | Aspirational, future-oriented, externally focused | How the enterprise wants to be viewed in the future |
| **Core Values** | Core values prescribe the desired behaviour, character, and culture of an enterprise. They define the fundamental principles that guide decision-making and are not easily changed. | Normative, stable, behavioural, cultural | The guiding principles and culture of the enterprise |

| | | | |
|---|---|---|---|
| **Goals** | Goals describe desired stages of development that support the realization of the vision. They translate ambition into short-, medium-, and long-term outcomes and form a bridge between purpose and execution. | Outcome-oriented, measurable at high level, time-related | What the enterprise aims to achieve |
| **Strategy** | Strategy outlines how the enterprise intends to achieve its goals. It translates the “what” (mission, vision, values, goals) into the “how” through a coherent plan of action. | Action-oriented, adaptive, execution-focused | How the enterprise will achieve its goals |

# 4. Data Cleaning

This section documents the data cleaning and preprocessing steps applied to the company statements dataset used for classifying statements into mission, vision, strategy, goals and core values.

The objective here was to ensure that there is:label consistency, remove noise and inconsistencies in text, handle multilingual data and prepare data for exploratory data analysis (EDA) and machine learning models

This was necessary because the data contained: inconsistent category labels, typographical errors in class names, mixed languages and unstructured natural language text.

## 4.1 Initial data inspection

The dataset contained company, statement and type that was manually extracted by the group from provided documentations from the client. The initial inspection shows the column structure, sample records and distribution of statement types.

- The **Type** column had multiple inconsistencies:
    - Misspellings: `steategy` and `core Value`
    - Duplicate semantic labels: `goal` vs `goals`

These issues were normalized before any analysis or modeling.

## 4.2 Cleaning and Processing

**Standardisation:**

To ensure consistency the column names were converted to lowercase, the leading and trailing whitespaces were removed and columns were renamed to more descriptive names.This was done to improve readability and maintainability and aligns column naming with machine learning conventions.

**Label Normalization:**

Because the Type column contained the aforementioned inconsistencies they were mapped to a canonical set of 5 labels where the final labels were : mission, vision, goals, strategy and core value. This was done as machine learning models treat labels as exact strings and inconsistent labels would fragment classes and degrade performance. Doing this ensures correct class counts and reliable evaluation metrics.

**Missing Data Handling:**

Because missing text cannot be vectorized and missing labels cannot be used for supervised learning those rows were removed to ensure clean training and evaluation datasets.

**Lowercasing and Punctuation Removal:**

A basic cleaning function was applied to ensure consistent word matching, prevent duplicate tokens and remove punctuations as it adds no semantic value.

**Multilingual stopwords and tokenisation:**

Since the dataset contains English and Dutch statements, stopwords from both languages were removed and tokens of the cleaned test were made. This was done to remove high-frequency but low-information words like *the, and, de, het.* It preserves accented characters for Dutch language integrity and filters very short tokens that add little semantic value. This prepares clean tokens for: word frequency analysis, baseline keyword extraction and TF-IDF modeling.

**Language detection:**

Because English and Dutch text require different linguistic handling language detection was applied. This prevents mixing vocabularies across languages and enables language-specific EDA and modeling. This also improves model accuracy and interpretability.

# 5. Data Analysis

## 5.1 Exploratory Observations & Key Findings

The data analysis phase focused on understanding the structure, distribution, and semantic characteristics of strategy-related organizational text prior to introducing machine learning models. Exploratory Data Analysis (EDA) showed that the dataset primarily consists of short, high-level textual statements, often written in abstract and aspirational language. Many statements were compact and information-dense, which increased ambiguity when attempting to classify them into distinct strategic categories such as mission, vision, or core values.

EDA results highlighted substantial **vocabulary overlap across statement types**, particularly for high-level concepts such as innovation, sustainability, growth, and responsibility. These terms appeared consistently across mission, vision, goals, and value-related statements, confirming that keyword presence alone is insufficient for reliable classification. However, frequency and co-occurrence patterns still provided useful signals when interpreted within a rule-based and contextual framework.

Further analysis of statement length and structure indicated that **goals and investment-related statements** tended to be more concrete and action-oriented, whereas **mission and vision statements** were more abstract and descriptive. This structural distinction supported the use of linguistic cues (e.g., verbs indicating intent versus aspiration) as part of the heuristic identification process.

Clustering analysis using semantic embeddings revealed that statements grouped primarily by **thematic similarity** rather than by formal strategic role. Clusters often reflected shared focus areas—such as innovation, digital transformation, or sustainability—rather than aligning cleanly with predefined categories like mission or vision. This finding suggests that embedding-based similarity captures topical alignment effectively but lacks inherent awareness of strategic hierarchy or functional intent.

Despite this limitation, clustering proved valuable for exploratory insight. It helped identify statements that were semantically distant from others within the same organization, revealing potential inconsistencies between goals, investments, and higher-level strategic intent. These observations supported the hypothesis that **semantic similarity combined with logical consistency rules** can serve as a useful approximation of strategic coherence.

Overall, the combined EDA and clustering results reinforced the project's decision to begin with interpretable, rule-based methods. While semantic models captured nuanced meaning, they required structural guidance to distinguish between different strategic roles. This validated the staged research approach in which heuristics establish conceptual structure and embeddings are used as complementary signals rather than standalone solutions.

# 6. Discovery Phase (Candidate URL Generation)

The purpose of this initial discovery phase was to actually begin with a Web Scraping Phase. The original purpose of the web scraping component was not to scrape the entire content of each website, but to automatically extract GEA statements, specifically mission, vision, strategy, core values, and goals from each official corporate website.

The idea was that these extracted statements could then be:

- Structured into a standardized format
- Compared across companies
- Used as input for downstream NLP and embedding-based models
- Additionally, contribute to the coherence measurement between different statement components to evaluate their coherence between statements

From a technical perspective, this required a pipeline capable of discovering relevant pages, extracting clean text, and reliably classifying statement types.

However, during implementation and testing, it became clear that reliable extraction of all GEA categories was not feasible within the project constraints. As a result, the scope of the web scraping phase was **iteratively reduced** to ensure technical robustness and data quality.

### Key Design Decisions

## 6.1 Initial Design Intent

Initially, the web scraper was designed to perform the main technical steps:

- **Discover relevant pages per company:** Using keyword-based URL discovery, internal link crawling, and domain-restrictive search.
- **Scrape textual content from those pages:** By downloading HTML pages, stripping boilerplate elements, and extracting visible text.
- **Automatically extract multiple GEA statement types from each website (mission/vision/strategy/core values/goals);** Using keyword heuristics and downstream NLP-based classification.

This required:

- High-quality candidate URL discovery (to avoid irrelevant pages)
- Robust **HTML-to-text extraction** (to remove navigation, cookies, and marketing noise)
- Accurate **statement classification logic** across heterogeneous websites

In practice, this multi-step automation introduced several technical challenges:

## 6.2 Practical Challenges Encountered

During experimentation with the scraping notebooks and extracted datasets, several limitations became apparent:

### 6.2.1 Inconsistent Website Structures

Corporate websites differ heavily in:

- Navigation structure
- URL conventions
- Terminology (e.g. *“Our Purpose”*, *“Who We Are”*, *“What Drives Us”*)
- Page layout and HTML semantics

| | company_name | page_url | statement_type | statement |
|---|---|---|---|---|
| 0 | Aegon | https://www.aegon.com/about/our-purpose/commun... | mission | Deliver on our purpose by helping people in ou... |
| 1 | Aegon | https://www.aegon.com/about/our-purpose/commun... | mission | Engage as many of our colleagues as possible t... |
| 2 | Aegon | https://www.aegon.com/about/our-purpose/commun... | mission | Deliver on our purpose by helping people in ou... |
| 3 | Aegon | https://www.aegon.com/about/our-purpose/commun... | mission | Engage as many of our colleagues as possible t... |
| 4 | BE Semiconductors | https://www.besi.com/company/company-profile/s... | mission | Besi's mission is to become the world's leadin... |

Table with company_names, page_url, statement_type, and statement were printed in the end result but the statements weren’t matching the statements_type. They were not detecting correctly.

Technically, this meant that:

- Rule-based extraction(headers, sections, keywords) was brittle
- The same GEA concept appeared in different structural contexts
- Generalizable parsing rules could not be defined without site-specific.

This directly limited the scalability of automated extraction. You can check the first notebook that we made, trying to extract 200 companies  statements from different stock market companies, more precisely from the Netherlands (AEX, AMX, AScX) to have also more variation in dutch statements, US (S&P 500 & NASDAQ-100) , Belgium (BEL20) and UK (FTSE 100 & FTSE 250). The idea with this notebook was to extract the 5 GEA statements from these companies in order to later train the model to identify other statements from different website sources. You can find the whole code and explanations in: `Data scraping Eviden (1).html` and the results in `Company_statements_200.xlsx.`

### 6.2.2 Poor extraction Quality for certain GEA Statements

While Mission and Vision statements we offer short, clearly labeled and isolated on dedicated pages, other categories like (Strategy, Goals , Values ) were typically embedded in long narrative pages, mixed with sustainability, ESG, or branding content and implicit rather than explicitly labeled.

From a technical standpoint, this resulted in large text blocks with low semantic precision that polluted the dataset, difficulty defining reliable character-length or keyword thresholds and ambiguous classification signals for the model. As a result, we extracted text frequently containing more noise rather than actual statements.

Therefore, we concluded that continuing the statement filtering process required the introduction of stricter and more explicit filtering rules. These rules were designed to reduce noise and improve the quality of candidate statements before applying semantic models such as Ollama or BERT. By enforcing this structured pre-filtering step, the language models could be prompted more effectively, allowing them to operate on cleaner, more relevant input and thereby extract the most representative statement per category.

This methodology was applied consistently across the pipeline and was also implemented in the **PDF_extractor.html** component, which is described in more detail in a later section.

## 6.3. Scope Reduction Decision

Based on these findings, we made a deliberate engineering decision to reduce the scope on the Webscrapper part.
Instead of:

> *Automatically extracting all GEA statements from all discover pages*

We decided to:

> *Focus only on identifying and extracting the best Mission and Vision pages per compan*y

This decision defined a clear system boundary:

- The scraper would **stop at link discovery**
- Text extraction and classification would be deferred or handled downstream

Why Mission & Vision?

Mission and Vision were selected because they:

- Are the most consistently present across companies
- Are usually explicitly labeled in URLs, titles, or headers
- Are shorter and more self-contained
- Have a higher signal-to-noise ratio
- Are more suitable as clean, structured model input

This reduced the technical complexity while increasing reliability.

## 6.4. Final Web Scraping Objective

After scope refinement the technical objective of the web scraper became:

*For each company, identify the official website and extract the highest-confidence URL's that contain Mission and Vision statements.*

Concretely, the scraper focused on:

- Resolving the **official corporate domain**
- Detecting **Mission- and Vision-related URLs**
- Filtering and ranking links based on confidence heuristics
- Exporting the selected links in structured form

No full-page text extraction or multi-category GEA classification was performed at this stage.

## 6.5. Link Discovery & Selection Logic (Final Version)

To achieve this reduced objective, the scraper implemented the following logic:

- **Keyword-based URL detection**: URLs containing terms such as mission,vision or purpose were prioritized.

```
🏠 Homepage: https://www.ing.com
✅ Discovery complete. 82 candidates saved to candidates.csv
🟢 Live URLs exported to top17_live.txt (17 URLs)

Top 15 live candidates:
 score     reason                                                                                  final_url  status_code
  8.20 page_links                                                https://www.ing.com/About-us/Purpose-and-values.htm          200
  5.70 page_links                                                   https://www.ing.com/About-us/ING-at-a-glance.htm          200
  5.70 page_links                                                        https://www.ing.com/About-us/Strategy.htm          200
  5.70 page_links                                              https://www.ing.com/About-us/Corporate-governance.htm          200
  5.40 page_links          https://www.ing.com/About-us/Corporate-governance/Legal-structure-and-regulators.htm          200
  5.40 page_links                   https://www.ing.com/About-us/Corporate-governance/Shareholder-influence.htm          200
  5.40 page_links         https://www.ing.com/About-us/Corporate-governance/Dutch-Corporate-Governance-Code.htm          200
  5.40 page_links                     https://www.ing.com/About-us/Corporate-governance/Dutch-Banking-Code.htm          200
  5.40 page_links                  https://www.ing.com/About-us/Corporate-governance/NYSE-listing-standards.htm          200
  5.40 page_links                               https://www.ing.com/About-us/Corporate-governance/Auditors.htm          200
  5.40 page_links                           https://www.ing.com/About-us/Corporate-governance/Remuneration.htm          200
  5.40 page_links https://www.ing.com/About-us/Corporate-governance/Supervisory-Public-and-Regulatory-Affairs.htm          200
  5.40 page_links                                 https://www.ing.com/About-us/Compliance/Tax-principles.htm          200
  5.40 page_links    https://www.ing.com/Sustainability/Partnerships-and-collective-action/Equator-Principles.htm          200
  4.89 page_links                                                                 https://www.ing.com/About-us.htm          200
```

- **Semantic labeling signals:** Pages with explicit Mission/Vision indicators in the URL path or page title were favored

```
Processing 17 URL(s). Output → out_md/ and manifest.csv

[1/17] (unknown) — https://www.ing.com/About-us/Purpose-and-values.htm
✅ Saved: out_md/content_purpose-and-values-htm-b75a9602.md (blocks=19)

[2/17] (unknown) — https://www.ing.com/About-us/ING-at-a-glance.htm
✅ Saved: out_md/content_ing-at-a-glance-htm-6fba7b37.md (blocks=13)

[3/17] (unknown) — https://www.ing.com/About-us/Strategy.htm
✅ Saved: out_md/content_strategy-htm-1713d8e3.md (blocks=8)
```

- **Noise filtering:** Pages were excluded if they: Were excessively long and generic or did not clearly isolate Mission or Vision content.
- **Domain validation:** All selected URLs had to belong to the official corporate domain to avoid third-party sources.

The output was one or a small number of high-confidence links per company, rather than large scraped text corpora as you saw before.

### 6.6. Output of the Webscraping Phase

The final output of the web scraping phase consisted of a structured dataset containing:

- Company name
- Official website
- Selected Mission link(s)
- Selected Vision link(s)

These outputs were stored in **Excel format** and intended to be:

- Manually reviewable
- Used as clean, high-quality input for downstream NLP or LLM-based extraction

You can find the exact code in Webscrapper (3). Later in the demo explanation. I'll show interactively how we code it in a way that is simpler for the user or the experts to collect additional information about Vision and Mission statements only.

## 7. PDF Statement Extraction

In this section, we will explain the technical approach used for extracting GEA-related statements from PDF documents. Instead of describing the system as a linear pipeline, the extraction logic is presented through four core components: filtering, training and model selection, structuring and modelling. Each component was developed and refined across the

three notebooks (`Data_Preparation_PDF_Final`, `Training_model`, and the final `PDF_extractor`).

## 7.1. Statement Filtering

The first major challenge in PDF-based extraction is that company reports contain large amounts of non-strategic content, including legal disclaimers, financial tables, operational descriptions, and repeated boilerplate text. To address this, a dedicated filtering component was implemented to reduce noise before any semantic modelling was applied aggressively.

Text extraction from PDFs was performed using **PyMuPDF (`fitz`)**, which provides page-level access to raw text. After extraction, multiple filtering techniques were applied using **Python**, **regular expressions (`re`)**, and **pandas** to retain only meaningful candidate statements.

Filtering was implemented using a combination of rule-based structural constraints and lightweight semantic heuristics:

- **Length-based filtering** was applied to remove text fragments below a minimum character threshold, eliminating headings, captions, bullet points, and incomplete sentences.
- **Maximum-length constraints** were introduced to exclude very long text blocks, which typically represent narrative sections or contextual explanations rather than concise mission or strategy statements.
- **Regex-based cleaning** was used to remove common sources of noise such as page numbers, section numbering, copyright notices, headers, footers, and repeated document metadata.
- **Deduplication logic** was applied across pages to remove identical or near-identical fragments, preventing the same statement from appearing multiple times due to repeated headers or layout artifacts.

After structural filtering, the remaining text blocks were tokenized at sentence or paragraph level and prepared for semantic analysis. At this stage, only linguistically coherent and self-contained text fragments were retained.

This early filtering step significantly reduced the search space and ensured that downstream models only processed high-quality candidate statements, improving both efficiency and precision in later classification and relevance scoring stages. At this stage, we encountered two main issues: the extracted text chunks were often too long, and this caused the classification results to become noisy and inconsistent. Because the model was receiving overly broad and unfocused input, it struggled to clearly distinguish between the different categories. To address this, we needed to rethink our approach and find a way to both reduce the chunk size and improve the overall quality of the classification.

## 7.2. Training & Model Selection

Due to the limited availability of labeled training data and the highly abstract nature of mission, vision, strategy, and values statements, training a fully supervised classification model was not feasible within the scope of this project. Instead, the focus was placed on **leveraging pretrained transformer-based language models** and empirically evaluating their suitability for semantic statement identification.

All experiments were conducted in Python using **Jupyter Notebooks**, primarily relying on the **SentenceTransformers** library for embedding generation and **scikit-learn** utilities for similarity computation.

#### Embedding-Based Representation

Text fragments produced during the filtering phase were converted into fixed-length vector representations using **Sentence-BERT–based models**. These embeddings capture semantic similarity beyond surface-level keywords, which is essential for strategic language that is often implicit and abstract. Embeddings were also normalized to ensure stable cosine similarity measurements, and pretrained models were used without fine-tuning to avoid overfitting on the small dataset.

#### Category Anchoring & Similarity Scoring

Instead of training a classifier, each GEA category (Mission, Vision, Strategy, Core Values, Goals) was represented using semantic anchor definitions, consisting of: short conceptual descriptions and example phrases derived from validated statements

Candidate statements were compared against these anchors using **cosine similarity**, implemented via `sklearn.metrics.pairwise.cosine_similarity`.

#### Threshold Experiments

Multiple similarity thresholds were tested to evaluate trade-offs between recall and precision:

- Lower thresholds increased recall but introduced generic corporate statements.
- Higher thresholds improved precision but excluded abstract yet valid strategic statements.

Experiments demonstrated that **pure similarity scoring was insufficient** on its own. Statements with high semantic similarity were often thematically related but not structurally suitable as GEA statements.

**Outcome of the Training Phase**

Rather than producing a deployable trained model, this notebook served as a **controlled experimentation and validation environment**. Its main outputs were:

- Selection of the most stable embedding approach
- Empirically validated similarity thresholds
- Identification of failure cases requiring rule-based correction

These findings directly informed the final modelling logic implemented in the `PDF_extractor`, where semantic scores are combined with explicit decision rules to maintain interpretability and control.

## 7.3. Statement Structuring

After filtering and semantic relevance scoring, the remaining candidate statements were transformed into a structured representation suitable for validation and downstream analysis. This structuring step was essential to ensure traceability, consistency, and compatibility with later coherence modelling.

Structuring was implemented using **pandas DataFrames**, where each extracted statement was treated as a single atomic record. The following design principles were applied:

- Each statement corresponds to exactly one row in the table.
- Statements are assigned to **one GEA category only** (Mission, Vision, Strategy, Core Values, or Goals).
- Statements that could not be confidently assigned to a single category were labelled as ‘ Other’.

For each statement, the following metadata fields were retained:

- Cleaned statement text
- Assigned GEA category
- Semantic similarity score
- Source document identifier
- Page number within the PDF

This structured format was deliberately chosen to align with **Excel-based review workflows**, which are commonly used by enterprise architects. The table representation enables:

- Manual validation of extracted statements
- Transparent inspection of category assignments
- Direct reuse as input for coherence and alignment analysis

By separating statement extraction from coherence modelling, the system ensures that domain experts can validate the extracted data independently before any higher-level analysis is performed.

## 7.4. Modelling & Decision Logic

The final modelling logic combines semantic similarity scoring with explicit rule-based decision constraints. Rather than relying on end-to-end classification, the system uses a controlled hybrid approach to ensure interpretability and robustness.

**Semantic Scoring**

For each candidate statement:

- Sentence embeddings are generated using a pretrained **Sentence-BERT** model.
- Cosine similarity is computed between the statement embedding and each GEA category anchor.
- The highest similarity score determines the most likely category candidate.

**Rule-Based Constraints**

To prevent false positives and unstable classifications, semantic scores are evaluated within a rule-based decision framework:

- Minimum similarity thresholds ensure that only semantically meaningful matches are accepted.
- Length constraints guarantee that statements are sufficiently complete but not overly verbose.
- Deduplication checks prevent repeated statements from being included multiple times.
- Single-label enforcement ensures that each statement maps to only one GEA category.

Statements that fail to meet all criteria are excluded from the final output. This conservative strategy prioritizes precision and explainability over full automation.

## 7.5. Role Within the Overall GEA Framework

The PDF-based statement extraction component functions as a data preparation and validation layer within the broader GEA framework. Its primary role is not to generate final judgments, but to supply high-quality, structured inputs for coherence and alignment analysis.

Specifically, the output of the PDF extractor serves as the validated input for company-level coherence matrices and supports expert-driven interpretation rather than automated decision-making.

# 8. Keyword Extraction

The keyword extraction was meant to the most semantically meaningful words within the statements, this is because these statements can be long, descriptive and multitopic. This approach is the current practice of the client's coherence determination.

Keyword extraction reduces noise and dilution of semantic relevance by isolating core concepts that best represent each statement's intent.

Keyword extraction was applied selectively for abstract or long statements like: mission vs vision. However it was disabled where full contextual meaning is required like values vs mission where the statements are short.

## 8.1 Method Chosen: YAKE

The pipeline uses YAKE (Yet Another Keyword Extractor), an unsupervised keyword extraction algorithm.

YAKE was selected because:

- Unsupervised (no training data required)
- Fast and domain-agnostic
- Performs well on short texts

YAKE determines keyword relevance based on:

- Term frequency
- Position in the document
- Contextual dispersion
- Casing and word form features

Configuration Parameters

- Language: English
- Maximum n-gram length: 3
- Number of keywords per statement: 5 (configurable)

The fallback logic in place if YAKE fails to extract keywords, the full statement text is used as a fallback so that no statement is dropped and the pipeline remains robust across all inputs.

The keyword extraction stage produces a list of extracted keywords per statement and aggregated semantic embeddings which will be used as inputs for classification, similarity scoring and coherence evaluation.

# 9. Classification and Modeling

## 9.1 Purpose

Classification and modeling is responsible for transforming the textual company data into structured and a comparable semantic representation and determining the relationships between pairs of the statements. The keyword extraction reduces noise and the classification and modeling is what provides the analysis of the coherence scoring framework.

This stage combines two complementary modeling approaches:

1. **Semantic similarity modeling**, which captures topical and conceptual alignment.
2. **Natural Language Inference (NLI)** modeling, which identifies logical consistency or contradiction.

These two models enable a better evaluation of coherence rather than the two approaches independently.

## 9.2 Semantic Representation via Sentence Embeddings

To model the semantic content of company statements, a sentence embedding approach is taken. Sentence embeddings encode entire statements into dense numerical vectors such that semantically similar statements occupy nearby positions in vector space.

The sentence embeddings are generated using Sentence-BERT, specifically the all-mpnet-base-v2 variant. This model was chosen because:

- Strong performance on semantic similarity
- Sentence-level representations
- Widely validated in academic literature

Unlike traditional bag-of-words or TF–IDF representations, sentence embeddings preserve contextual meaning and word order, which is critical when analysing mission, vision, goal or strategy statements.

#### Normalisation and Vector Space Properties

Before the comparison, the embeddings are normalized to a unit length so the vector magnitude does not influence similarity and the comparison is based on the vector orientation. Because of this, cosine similarity can be better calculated using the dot product of the two embedding vectors.

## 9.3 Natural Language Inference (NLI)

The semantic alignment captures topical alignment however it does not detect logical incompatibilities. Statements may make use of similar wordings but express contradictory ideas.

This is why natural language inference modelling is used to address the limits of semantic alignment.

NLI determines whether a hypothesis:

Entailed
Contradicted
Neutral

This allows the system to distinguish between surface level similarity and intent.

**NLI model selection**

The NLI component is implemented using a transformer-based entailment model trained on the Multi-Genre Natural Language Inference (MNLI) dataset. This model has been shown to generalise effectively across diverse textual domains, including formal and abstract language which is common in corporate documentation.

The model outputs probability scores for each inference class enabling probabilistic reasoning rather than binary decisions.

**Contradiction detection**

It is possible that there are statements that are similar on a semantic level but contradictory logically, this can mean misalignment. Such cases are treated as stronger signals of incoherence than statements that are merely unrelated. Therefore, contradiction probabilities are explicitly incorporated into the final scoring.

## 9.4 Hybrid scoring model

Rather than relying on a single metric, the framework employs a hybrid rule-based scoring model that integrates semantic similarity and NLI outputs.

**Semantic Similarity Thresholding**

Semantic similarity scores are discretised into ordinal coherence levels. Higher similarity values correspond to stronger alignment between statements, while lower values indicate weak or absent alignment. This discretisation improves interpretability, allowing coherence to be expressed in meaningful categories rather than continuous values.

**Logical Penalty Mechanism**

When high semantic similarity co-occurs with elevated contradiction probability, a penalty is applied. This reflects the insight that:

- Conceptually similar but logically conflicting statements represent strategic inconsistency
- Such inconsistencies are more critical than simple thematic divergence

The penalty mechanism ensures that logical conflicts override superficial semantic alignment.

**Interpretability of the Scoring Model**

The final scoring model produces a small set of discrete values representing:

- Strong alignment
- Moderate alignment
- Weak alignment
- Neutral relationship
- Logical contradiction

This design prioritises interpretability and transparency, making the results accessible to non-technical stakeholders such as managers or policy analysts.

# 10. Coherence Matrix

The coherence matrix is the analytical representation of strategic alignment within an organisation. Its purpose is to consolidate the outputs of keyword extraction, semantic modeling, and logical inference into a structured, interpretable format that enables systematic comparison across different categories of corporate statements.

The statements were arranged into a matrix, and this allows coherence to be assessed:

- Across multiple strategic dimensions
- At both the individual statement level and the organisational level
- In a manner that supports both quantitative analysis and qualitative interpretation

## 10.1 Structure and dimensions

Each coherence matrix is constructed as a two-dimensional grid in which:

- Rows (premises) represent one category of strategic statements
- Columns (hypotheses) represent another category

Each cell in the matrix corresponds to a pair comparison between a premise statement and a hypothesis statement.

## 10.2 Pairwise evaluation

For each cell in the matrix, a pairwise evaluation is performed using the hybrid classification and modeling framework described previously. The evaluation integrates:

1. Semantic similarity, which measures conceptual alignment
2. Logical inference, which identifies entailment, neutrality, or contradiction

The result of this evaluation is a single coherence score that summarises the relationship between the two statements.

## 10.3 Scoring

Rather than using raw similarity or probability values, the framework maps each pairwise comparison to a discrete coherence score. This scoring scheme improves interpretability

**Positive Alignment Scores**

Positive scores indicate increasing levels of strategic coherence:

- High scores represent strong semantic alignment and logical consistency
- Moderate scores indicate partial or thematic alignment
- Lower positive scores reflect weak but non-random alignment

These scores signal that operational or strategic statements are meaningfully connected.

**Neutral Scores**

A neutral score is assigned when statements exhibit neither meaningful alignment nor explicit contradiction. This outcome suggests:

- Conceptual independence between the statements
- Potential gaps in strategic integration

Neutral relationships are particularly informative when they occur systematically across multiple statement pairs.

**Negative Scores**

Negative scores are for cases where there is logical contradiction and semantic overlap. This means that there is  a strongest incoherence, indicating statements addressing similar themes promote incompatible strategic directions. This incoherence enables the framework to identify not only absence of alignment, but active misalignment.

## 10.4 Multi matrix

The framework constructs multiple coherence matrices, each corresponding to a different strategic relationship, such as:

- Mission versus Vision
- Core Values versus Mission
- Goals versus Vision
- Goals versus Strategy

This allows coherence to be examined from multiple perspectives and prevents overgeneralisation from a single relational view.

The matrix-based representation offers several practical advantages:

- Results are visually interpretable and can be easily communicated
- Patterns of alignment and misalignment become immediately apparent
- Decision makers can identify specific statements responsible for incoherence

As a result, the coherence matrix functions not only as an analytical tool, but also as a diagnostic instrument for strategic evaluation.

# 11. Demo (Interactive Interface)

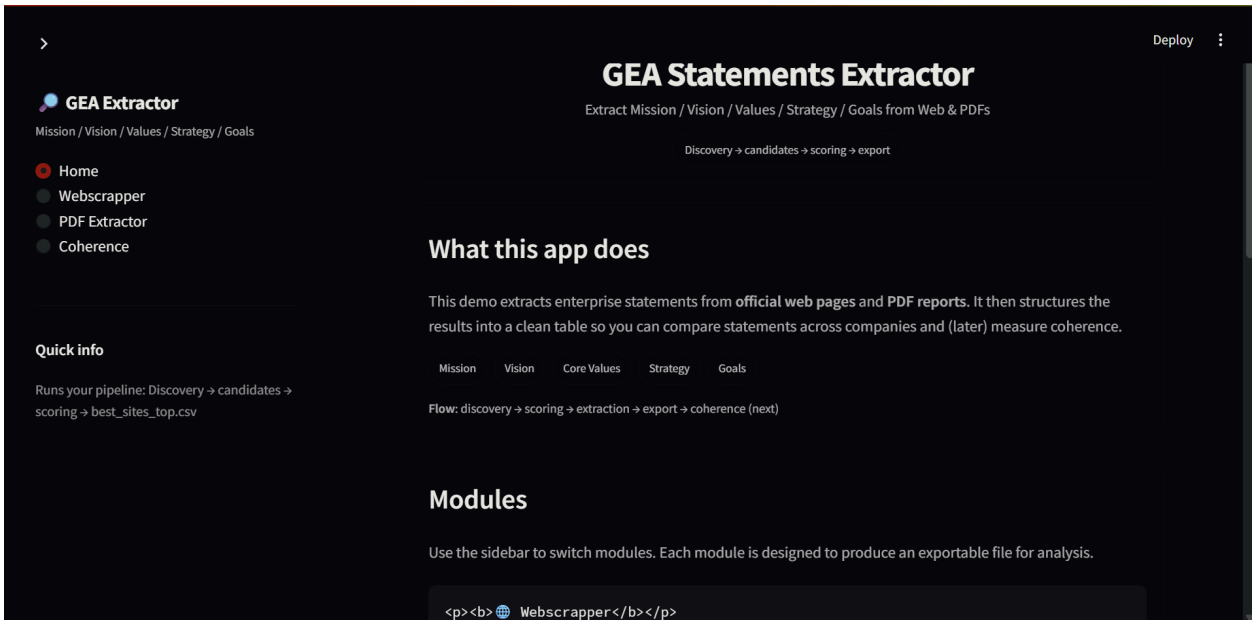
## Purpose of the Demo

The interactive demo was developed to present the Extraction-to-Coherence pipeline in a clear, structured, and accessible way for non-technical stakeholders. The purpose of the demo is not to showcase a fully automated AI system, but to demonstrate how the implemented methods work in practice and how the outputs can support expert analysis and validation.

The interface allows users to interact with the system step by step, inspect intermediate results, and understand how design choices affect the final output. This ensures transparency and aligns with the project's goal of supporting enterprise architects rather than replacing their judgment.

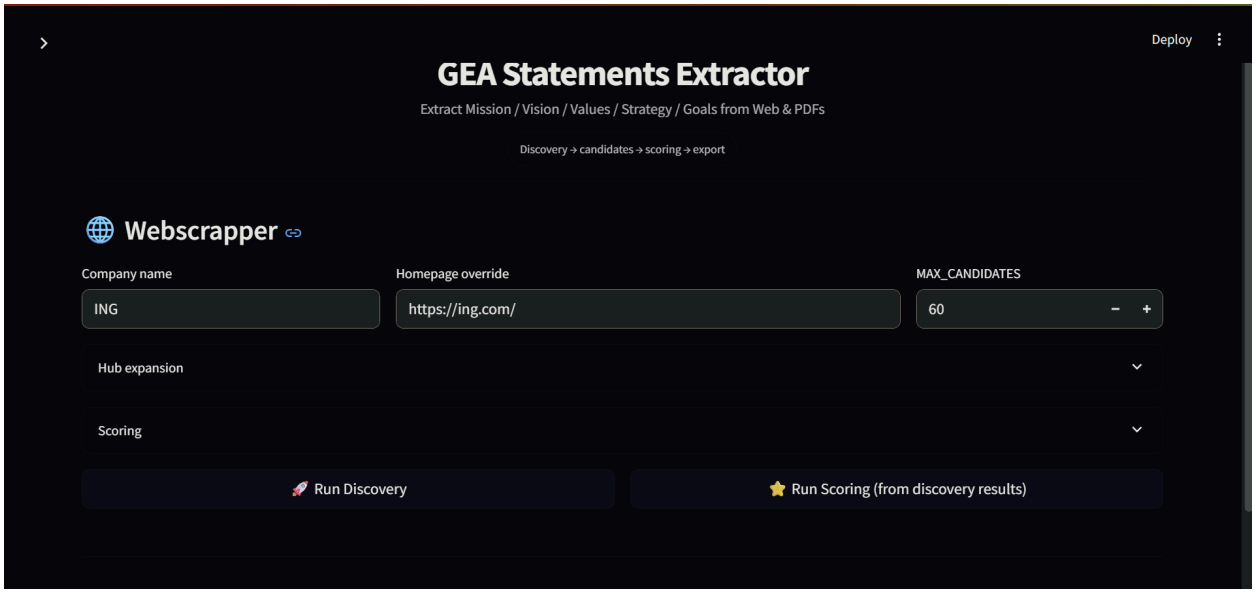
The demo is intended for enterprise architects, consultants, researchers, and future project teams. No programming knowledge is required to use the interface.

The demo consists of 1 main section and 3 subsections:



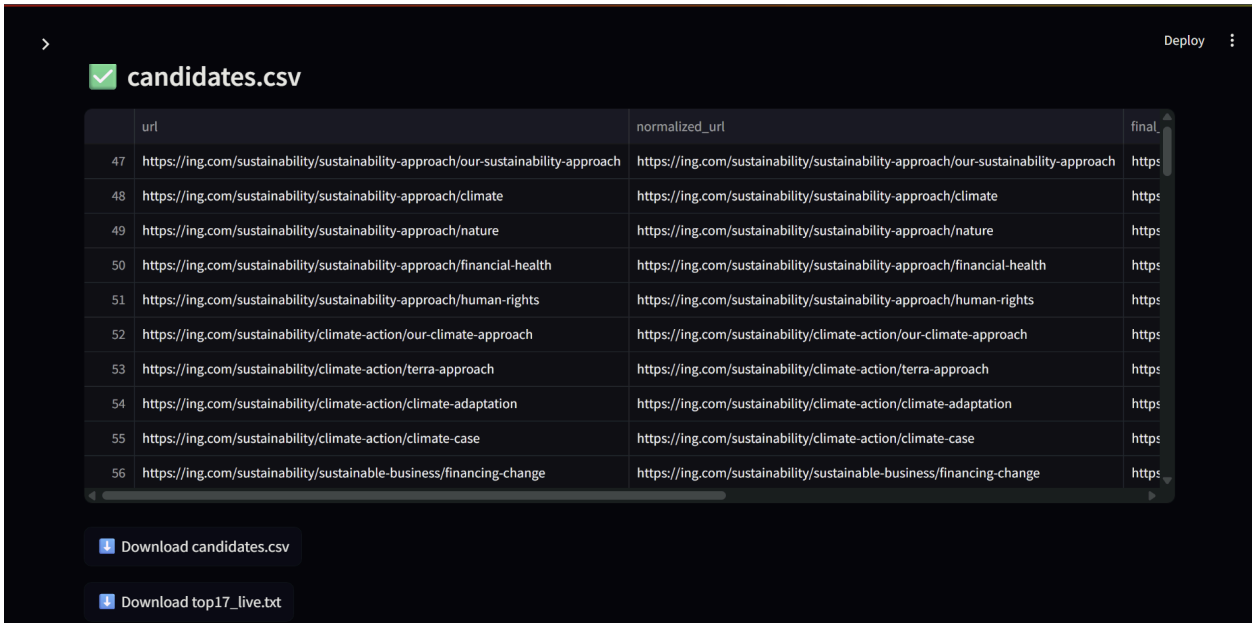

The first main section is a small description of how the demo works. What sections we work on and how the app is used in each subsection.

## 11.1. Webscrapper



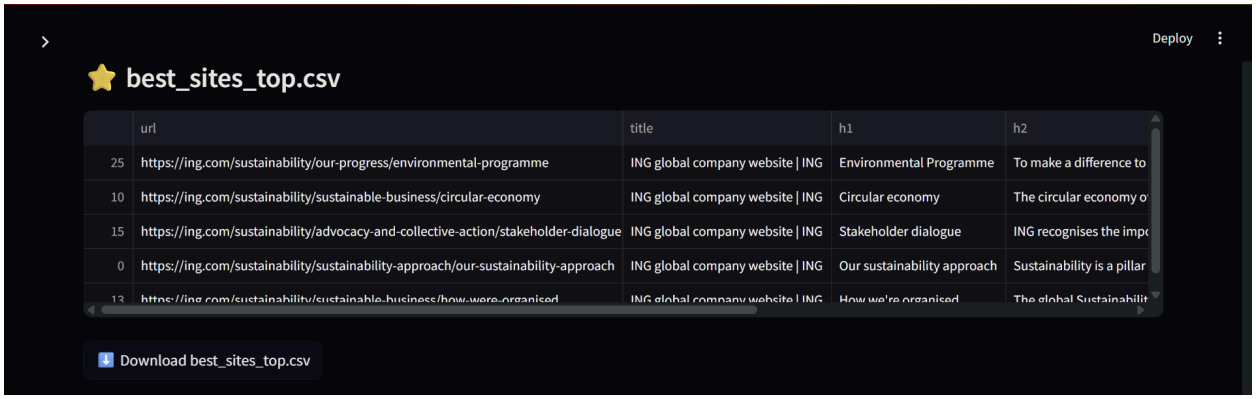

In this subsection, the Web Scraper is presented as a simple interactive demo that can be used by the client and domain experts. The interface was intentionally designed to be as minimal as possible. The required inputs are the company name, the official company website, and the maximum number of candidate links to extract from that website.In addition, optional filtering parameters are available, such as the maximum number of hubs to expand, which determines how many key sections of the website are explored, and the maximum number of links per hub, which controls how deeply each section is crawled. These optional filters allow users to perform

a more targeted or deeper search when needed.After configuring the inputs, the user can run the discovery process. The system then returns a table containing the highest-ranked links based on the applied filtering logic. This table can be exported as a CSV or TXT file. The final output consists exclusively of the most relevant links for Mission and Vision statements.



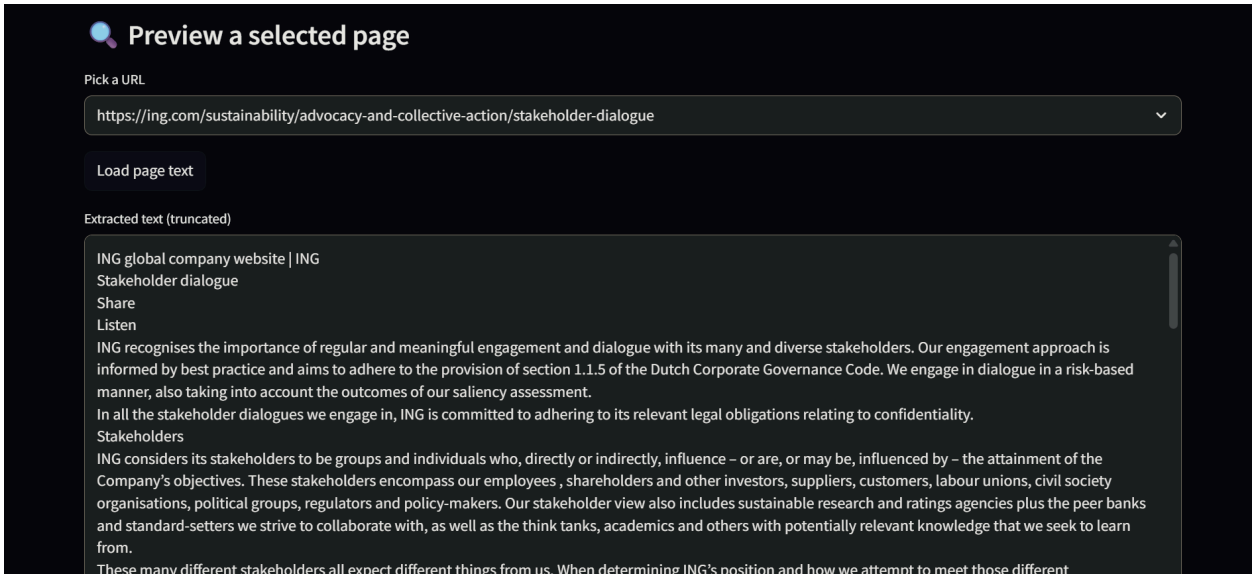

| | url | normalized_url | final_ |
|---|---|---|---|
| 47 | https://ing.com/sustainability/sustainability-approach/our-sustainability-approach | https://ing.com/sustainability/sustainability-approach/our-sustainability-approach | https |
| 48 | https://ing.com/sustainability/sustainability-approach/climate | https://ing.com/sustainability/sustainability-approach/climate | https |
| 49 | https://ing.com/sustainability/sustainability-approach/nature | https://ing.com/sustainability/sustainability-approach/nature | https |
| 50 | https://ing.com/sustainability/sustainability-approach/financial-health | https://ing.com/sustainability/sustainability-approach/financial-health | https |
| 51 | https://ing.com/sustainability/sustainability-approach/human-rights | https://ing.com/sustainability/sustainability-approach/human-rights | https |
| 52 | https://ing.com/sustainability/climate-action/our-climate-approach | https://ing.com/sustainability/climate-action/our-climate-approach | https |
| 53 | https://ing.com/sustainability/climate-action/terra-approach | https://ing.com/sustainability/climate-action/terra-approach | https |
| 54 | https://ing.com/sustainability/climate-action/climate-adaptation | https://ing.com/sustainability/climate-action/climate-adaptation | https |
| 55 | https://ing.com/sustainability/climate-action/climate-case | https://ing.com/sustainability/climate-action/climate-case | https |
| 56 | https://ing.com/sustainability/sustainable-business/financing-change | https://ing.com/sustainability/sustainable-business/financing-change | https |

If you want better results you can click now the Run Best Score. And then you will get the best top 5 link results from the best of the best candidates:




| | url | title | h1 | h2 |
|---|---|---|---|---|
| 25 | https://ing.com/sustainability/our-progress/environmental-programme | ING global company website \| ING | Environmental Programme | To make a difference to |
| 10 | https://ing.com/sustainability/sustainable-business/circular-economy | ING global company website \| ING | Circular economy | The circular economy o |
| 15 | https://ing.com/sustainability/advocacy-and-collective-action/stakeholder-dialogue | ING global company website \| ING | Stakeholder dialogue | ING recognises the impo |
| 0 | https://ing.com/sustainability/sustainability-approach/our-sustainability-approach | ING global company website \| ING | Our sustainability approach | Sustainability is a pillar |
| 13 | https://ing.com/sustainability/sustainable-business/how-were-organised | ING global company website \| ING | How we're organised | The global Sustainabilit |

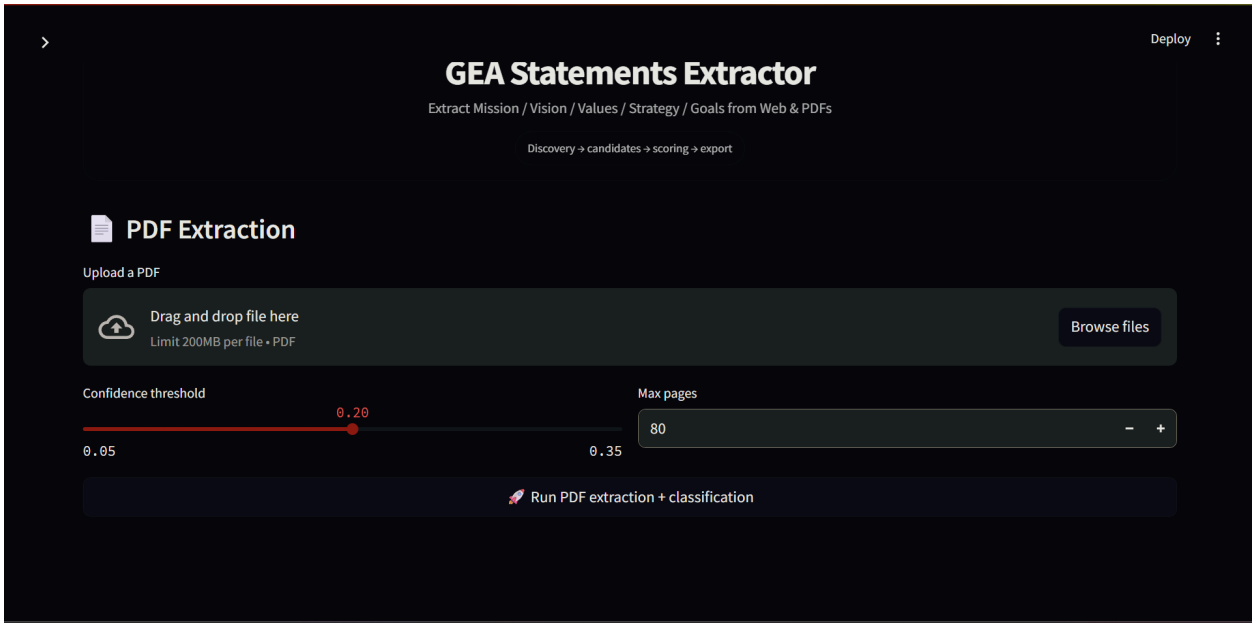
After that you can preview the page of the best top candidates to actually get additional information:



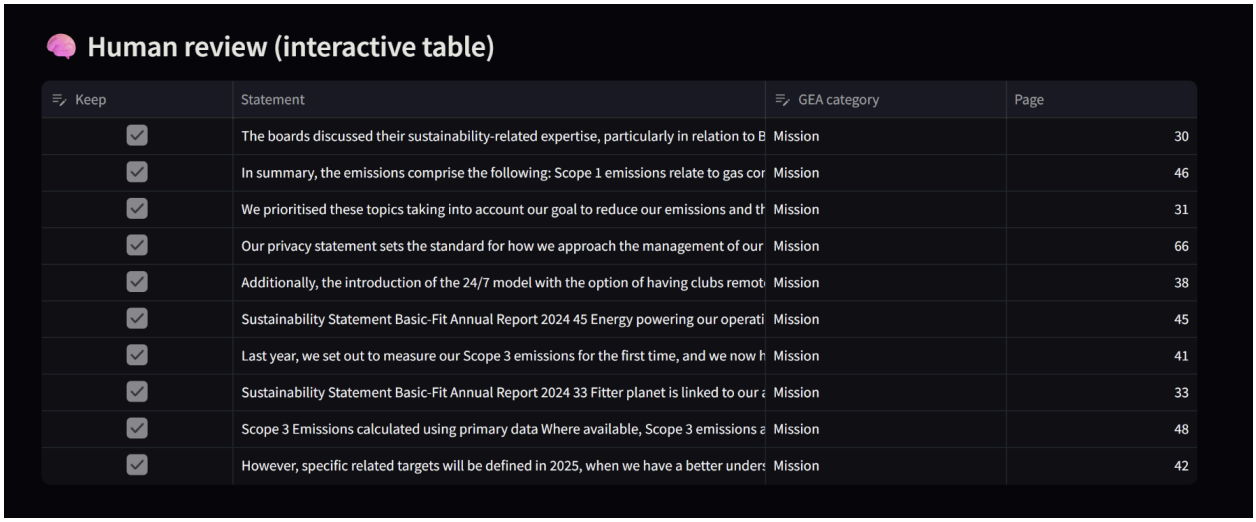

This helps the user to get a better perspective of the link instead of clicking the link on its own.

## 11.2. PDF extraction

In this subsection, we have the PDF extractor, that it will extract all the statements from PDF's only to the GEA Framework (mission/vision/core values/ strategy and goals):

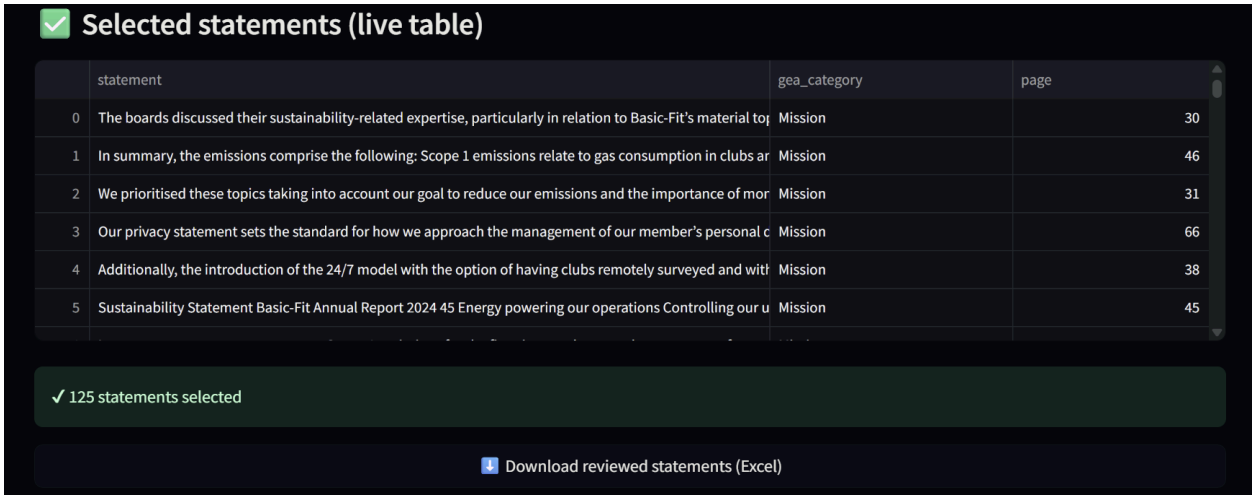



Here you can upload a pdf document and then set the confidence between these parameters to get a better result and the pages that you want to extract. We incorporate these extra features because there can be bigger pdf's that can contain all the information that we want to extract in the middle.
After we upload the document, we will obtain a table with all the (possible) statements with the GEA category:



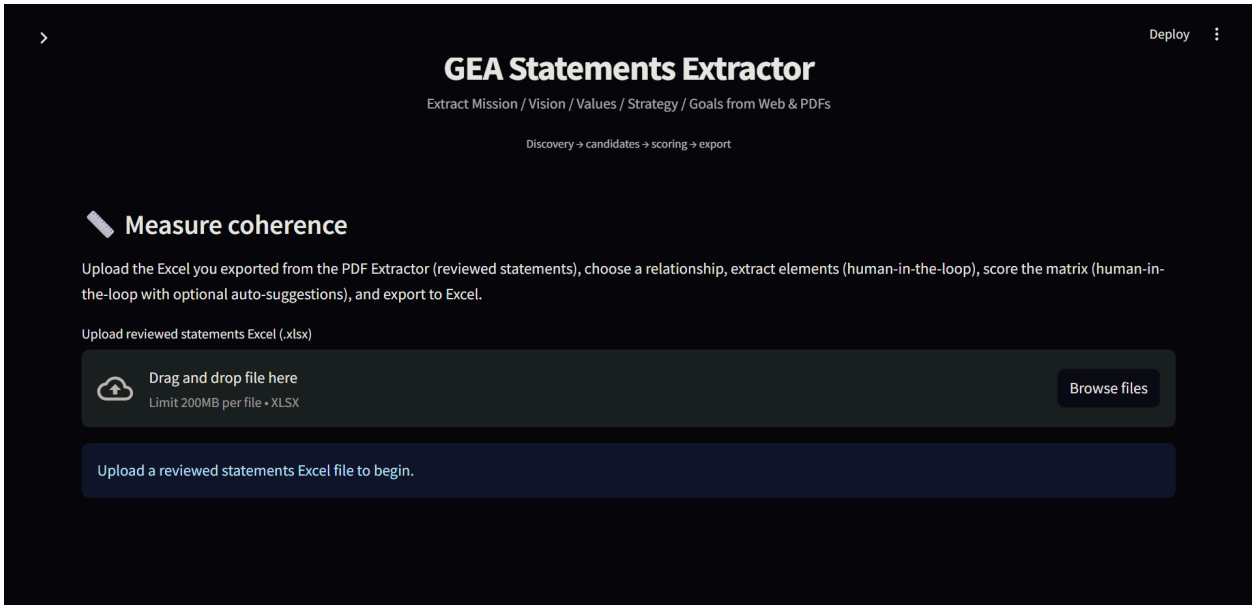

In this step, a human-in-the-loop approach is integrated to support the model in distinguishing between correct and incorrect statements. If a statement is assigned to an incorrect GEA category, for example, if a Vision statement is mistakenly labeled as a Mission, the user can manually adjust the category.

Within the web application, these category corrections can be applied directly, following the same validation principle as in the previous steps. This feedback helps improve the overall quality of the extracted data. After the suggested changes are applied, a final table containing the corrected statements is generated.

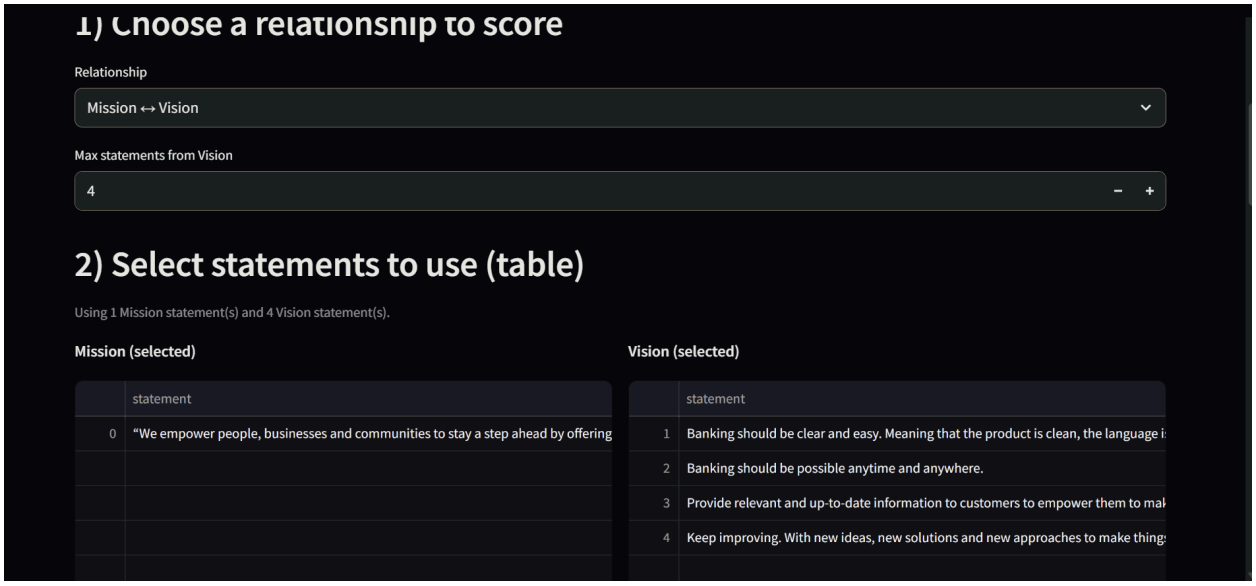



At the end, when you are satisfied with your result, you can download this table in an Excel file. This way you can use it for the next subsection. Coherence measurement:

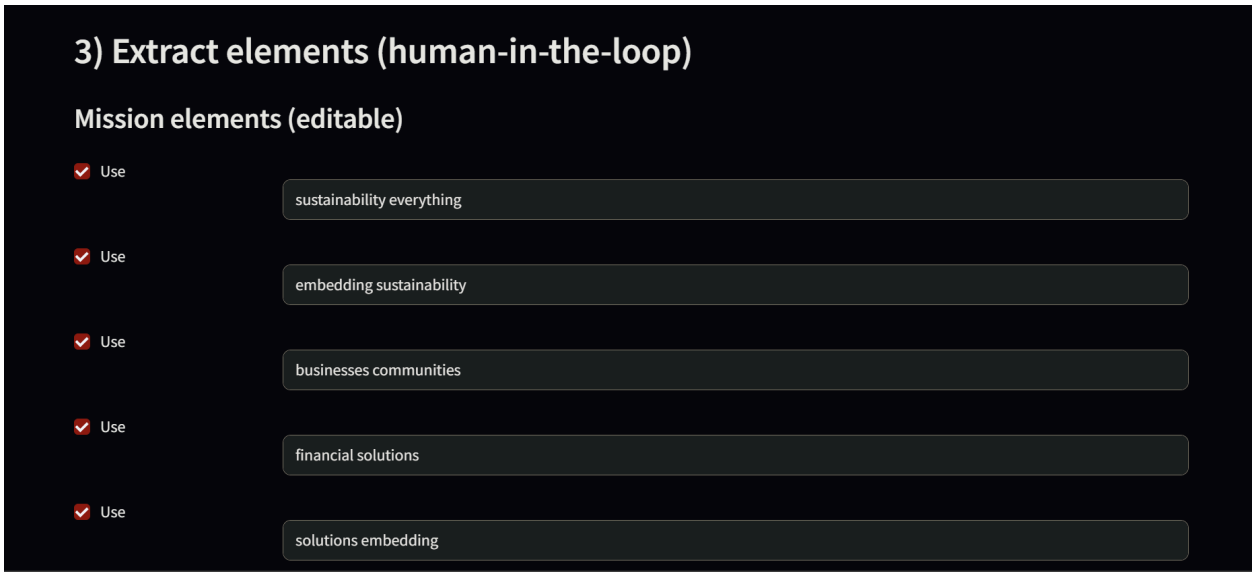
## 11.3. Measure coherence



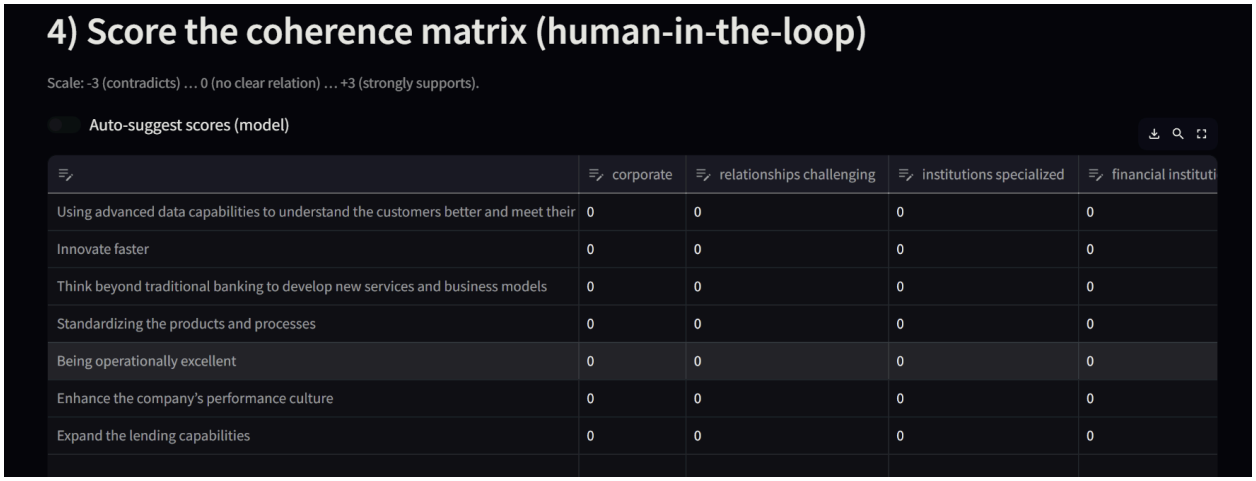

As described, the previously generated Excel file can be uploaded to measure the coherence between statements. In this demo, the relationships included for coherence evaluation are Mission–Vision and Strategy–Goals:



In this step, the user can select how many statements to compare and measure their coherence. The interface also supports a human-in-the-loop approach for selecting the relevant elements (such as Mission and Goals). Users may keep the automatically extracted elements or manually add additional elements if they believe more input is needed to enable a more accurate comparison between Mission and Strategy.




At the final stage, a table is generated in which each element is organized and can be individually scored. Users can assign a coherence score based on their assessment, where strong coherence is scored as 3 and weak or conflicting coherence as −3. An auto-suggested scoring option is available to provide an initial indication of coherence; however, users can always adjust the scores manually if they disagree with the suggested values.




| | corporate | relationships challenging | institutions specialized | financial instituti |
|---|---|---|---|---|
| Using advanced data capabilities to understand the customers better and meet their | 0 | 0 | 0 | 0 |
| Innovate faster | 0 | 0 | 0 | 0 |
| Think beyond traditional banking to develop new services and business models | 0 | 0 | 0 | 0 |
| Standardizing the products and processes | 0 | 0 | 0 | 0 |
| Being operationally excellent | 0 | 0 | 0 | 0 |
| Enhance the company's performance culture | 0 | 0 | 0 | 0 |
| Expand the lending capabilities | 0 | 0 | 0 | 0 |

At the end of the process, the final table containing the coherence matrix can be downloaded as an Excel file. This application was developed to demonstrate, in an interactive way, how the GEA extractor works and which statements can be extracted from different data sources. The extracted statements can then be compared to produce a structured coherence matrix.

The application will be deployed in the ATOS GitHub repository, where the running version will be available. The source code for the application can be found in the `app.py` file.

# 12. Work Done

During this project, a complete Extraction-to-Coherence was designed, implemented, and validated to support enterprise architects in assessing strategic alignment under the GEA framework:

The following work was successfully completed:

- Implemented an **end-to-end NLP pipeline** to extract and analyze GEA statements from PDF's and websites
- Automatically extracted and structured **Mission, Vision, Strategy, Goals and Core Values**
- Built **company-level coherence matrices** to preserve strategic content
- Designed and explainable **hybrid scoring system** combining keywords, embeddings, and contradiction detection.
- Delivered **Excel-based outputs** for expert validation and review.
- Developed an **interactive demo** to showcase extraction and coherence analysis

# 13. Limitations

- **Dataset size is limited**, preventing full statistical validation
- Coherence scores are **indicative, not final judgments**
- Results depend on **clarity and structure of company statements**
- Web scraping was **intentionally limited** to ensure data quality
- **Expert review** remains necessary for final interpretation

# 14. Future Project Concept

The current Extraction-to-Coherence pipeline provides a robust foundation, but several enhancements could improve precision, usability, and scalability:

- Improved PDF extraction: Implement layout aware models or fine tuned language models to better identify GEA statements in documents.
- Improved candidate canking: Combine semantic relevance, structural cues and past expert choices to present a smaller, better ordered set of candidate statements.
- Separation of retrieval and extraction: Clearly distinguish web scraping from statement extraction to reduce noise.
- Expanded expert labeled data: Collect and store corrections to build a growing dataset for semi supervised or supervised learning.

- Refined coherence scoring: Incorporate weighting, ambiguity handling and multi-label assignments to better reflect real-world statements.
- Interactive visualization: Extend Excel outputs with dashboards or heatmaps for easier interpretation.

# Final Note

This project demonstrates that strategic coherence within the GEA framework can be **systematically supported using AI**, without replacing expert judgment. By combining explainable NLP techniques with structured outputs, the solution transforms unstructured corporate information into **actionable and reviewable insights**.

The delivered pipeline is not a finished product, but a **robust foundation** that future teams can extend by expanding the dataset, refining scoring logic, and integrating deeper expert validation.