Automated Analysis of Sustainability Reports: Using Large Language Models for the Extraction and Prediction of EU Taxonomy-Compliant KPIs

Anonymous ACL submission

Abstract

The manual, resource-intensive process of complying with the EU Taxonomy presents a significant challenge for companies. While Large Language Models offer a path to automation, research is hindered by a lack of public benchmark datasets. To address this gap, we first introduce a novel, structured dataset from 190 corporate reports, containing ground-truth economic activities and quantitative Key Performance Indicators (KPIs). We then use this dataset to conduct the first systematic evaluation of LLMs on the core compliance workflow. Our results reveal a clear performance gap between qualitative and quantitative tasks. LLMs show moderate success in the qualitative task of identifying economic activities, with a multi-step agentic framework modestly enhancing precision. Conversely, the models comprehensively fail at the quantitative task of predicting financial KPIs in a zero-shot setting. We also find that concise metadata often yields superior performance to full, unstructured reports, and that model confidence scores are poorly calibrated. We conclude that while LLMs are not ready for full automation, they can serve as powerful assistive tools for human experts. Our dataset provides a public benchmark for future research.

1 Introduction

The European Union (EU) is mobilizing significant capital towards sustainable finance, guided by the European Green Deal (European Commission, 2020). A cornerstone of this strategy is the EU Taxonomy (European Parliament and Council of the European Union, 2020), a complex classification system designed to provide a standard, science-based definition of sustainable economic activities. The regulation's primary goal is to create market transparency, combat greenwashing, and direct investment toward activities essential for the green transition (European Commission, 2024a; Sneiderienе and Legenzova, 2025; Calamai et al., 2025).

This framework is operationalized through the Corporate Sustainability Reporting Directive (CSRD) (European Parliament and Council, 2022), which legally mandates that thousands of companies analyze their operations and publicly disclose their alignment with the Taxonomy. This compliance process requires undertakings to perform a complex, manual analysis of dense legal acts and unstructured internal documents to execute two core information extraction (IE) tasks: (1) identifying which of the many official economic activities the company performs, and (2) extracting the precise financial KPIs, Turnover, CapEx, and OpEx, associated with those activities (European Commission, 2021b). This manual workflow is resource-intensive, prone to misinterpretation, and scales poorly as regulations evolve (European Commission, 2021a, 2023; Schütze and Sandbaek, 2025).

The recent advancements in Large Language Models (LLMs) for complex reasoning and IE present a compelling opportunity to automate or assist these analytical tasks. However, the practical viability of general-purpose LLMs in this specialized, high-stakes domain is unproven. A significant barrier to research and development in this area is the lack of public, structured benchmark datasets; corporate disclosures are typically locked in unstructured PDF reports, making it difficult to evaluate or train models for these specific IE tasks.

To address this gap, this paper presents two primary contributions. First, we introduce and make publicly available a **novel, structured benchmark dataset** for EU Taxonomy analysis, curated from the annual reports of 190 companies. For each company, the dataset provides machine-readable reports, verified ground-truth lists of declared economic activities, and the corresponding quantitative KPIs (Turnover, CapEx, OpEx). Second, we use this dataset to conduct the **first systematic evaluation of modern LLMs** on the core compliance workflow, which we formulate as four distinct tasks:

(1) multiclass activity prediction, (2) binary activity classification, (3) quantitative KPI regression, and (4) a multi-step agentic workflow.

Our empirical results reveal a clear performance gap between qualitative and quantitative tasks. We find that LLMs demonstrate moderate success in the qualitative task of identifying relevant economic activities, with a multi-step agentic framework modestly enhancing precision. Conversely, the models comprehensively fail at the quantitative task of predicting financial KPIs in a zero-shot setting, yielding results no better than a baseline mean prediction. We also observe that concise, structured metadata often yields superior classification performance compared to the full, unstructured text of an annual report, consistent with recent findings on long-context degradation. Finally, we show that the models' self-reported confidence scores are poorly calibrated and thus unreliable for this task.

We conclude that while current-generation LLMs are not yet suitable for fully automated, end-to-end compliance, they can serve as powerful assistive tools to accelerate the discovery and analysis phases for human experts. Our dataset and baseline results provide a public benchmark for future research into developing specialized models for this complex and growing regulatory domain.

2 The EU Taxonomy Reporting Task

2.1 Regulatory Context and Reporting Obligations

The EU Taxonomy is a comprehensive classification system established by Regulation (EU) 2020/852 to define environmentally sustainable economic activities (European Parliament and Council of the European Union, 2020). It serves as a key component of the European Green Deal's sustainable finance framework, aiming to increase market transparency and combat *greenwashing* (European Commission, 2024a,b).

The regulation is operationalized through the *Corporate Sustainability Reporting Directive (CSRD)* (European Parliament and Council, 2022), which mandates that a wide range of undertakings publicly disclose their alignment with the Taxonomy. This compliance process requires companies to perform a complex internal assessment. First, an activity must make a *substantial contribution* to one of six environmental objectives (e.g., climate change mitigation). Second, it must *do no significant harm* (DNSH) to the other five objectives, while also meeting minimum social safeguards (European Parliament and Council of the European Union, 2020).

The core output of this assessment is the disclosure of three mandatory KPIs, detailed in a specific *Disclosures Delegated Act* (European Commission, 2021b). Non-financial undertakings must report the proportion of their operations that are Taxonomy-aligned, quantified across:

Turnover KPI: The proportion of net turnover derived from sustainable products or services.

CapEx KPI: The proportion of capital expenditures associated with sustainable assets or processes.

OpEx KPI: The proportion of operational expenditures related to sustainable activities.

2.2 Formulating the IE Tasks: Activity Classification and KPI Regression

For undertakings, the reporting process is a resource-intensive, manual workflow that involves interpreting dense legal texts (*Delegated Acts*) and mapping them to unstructured internal data from finance, legal, and sustainability departments (European Commission, 2021a, 2023). This process is error-prone, inconsistent, and scales poorly as the regulation evolves (Schütze and Sandbaek, 2025).

This regulatory challenge can be framed as a complex information extraction (IE) and reasoning problem. We decompose this compliance workflow into two primary computational tasks, which form the basis of our evaluation:

- **Task 1: Activity Classification.** This is a qualitative task. Before any financial data can be aggregated, a company must first determine *which* of its operations correspond to the official list of economic activities defined in the Taxonomy. We model this as a multi-label text classification problem, where the input is a company's operational description (e.g., from an annual report) and the output is a predicted set of relevant Taxonomy activities.

- **Task 2: Quantitative KPI Regression.** This is a quantitative task. Once relevant activities are identified, the associated financial metrics must be extracted and quantified. We model this as a regression problem, where the model must predict the numerical percentage for the

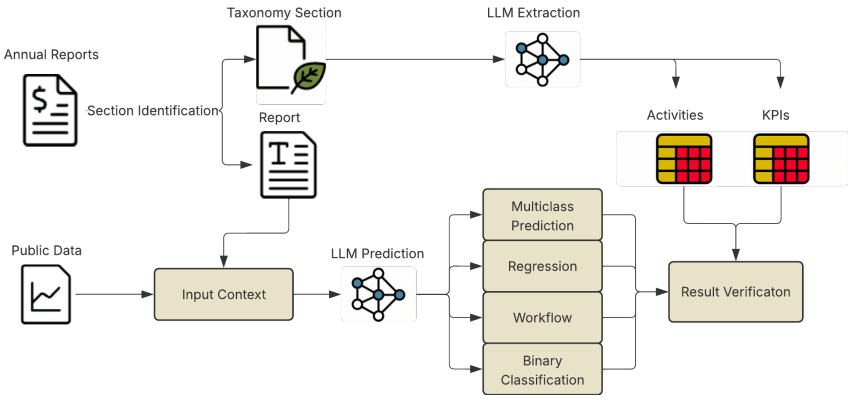

Figure 1: Overview: Extraction of source data from public reports. Multiple zero-shot experiments against the dataset to test current model capabilities.

Turnover, CapEx, and OpEx KPIs based on the textual content of a company's report.

The primary challenge lies in the model's ability to semantically bridge the gap between unstructured, narrative descriptions of a company's business and the highly structured, technical definitions of the legal framework. To ensure the evaluation tests genuine inference rather than surface-level extraction, the Taxonomy disclosure section is excluded from all report-based inputs (see Section 10.4 for a discussion of this design choice).

3 Related Work

Financial NLP has evolved from text mining to LLM-based reasoning (Li et al., 2024), with domain-specific models like FinBERT and BloombergGPT achieving strong results on financial benchmarks. In the ESG domain, RAG-based systems such as ESGReveal (Zou et al., 2023) and benchmarks like ESGenius (He et al., 2025) have advanced structured extraction and knowledge evaluation. In information extraction (IE), the field has shifted from traditional supervised methods (e.g., NER) to generative approaches using LLMs (Xu et al., 2024; Zhang et al., 2025), though applying these to specialized domains like legal and financial reporting remains challenging due to annotation costs (Oliveira et al., 2025). For EU Taxonomy compliance specifically, (Birti et al., 2025) introduced ESG-Activities, a benchmark for activity classification showing that fine-tuned smaller models can outperform larger ones, while (Ali et al., 2025) developed SustainableQA, a large-scale QA dataset for sustainability and EU Taxonomy reporting.

Our work addresses a key gap: prior studies target either general ESG analysis (Zou et al., 2023; He et al., 2025) or a single sub-task of Taxonomy compliance (Birti et al., 2025). We present the first systematic, multi-task evaluation combining qualitative activity classification with quantitative KPI regression, using a novel benchmark dataset derived from complete corporate reports. We provide an extended discussion of related work in Appendix 10.8.

4 A Benchmark Dataset for EU Taxonomy Analysis

A significant barrier to research in automated EU Taxonomy compliance is the lack of public, structured benchmark datasets. To address this, we introduce and make publicly available[1] a novel, structured dataset curated from the 2023 fiscal year reports of 190 companies subject to EU Taxonomy disclosure obligations.

4.1 Dataset Curation and Preprocessing

We collected 190 annual and sustainability reports from publicly listed companies, primarily from European stock exchanges (e.g., Frankfurter Börse, Wiener Börse) and corporate websites. All documents were originally in PDF format.

The curation process followed a multi-stage pipeline. First, raw PDFs were converted into structured Markdown using the open-source tool *Marker* (Paruchuri., 2024). We then used rule-based methods to identify and isolate the dedicated EU Taxonomy disclosure section within each report.

A semi-automated workflow was employed to extract ground-truth data. We used an

[1] Dataset available at https://github.com/anonymous-duckling/acl

LLM (Gemini-Flash-2.0) with a zero-shot prompting strategy, followed by an LLM-based self-verification step, and concluded with human verification by two PhD students (Appendix 10.6). This process was applied to extract: (1) the list of declared economic activities, and (2) the quantitative numerical values for Taxonomy-eligible and Taxonomy-aligned Turnover, CapEx, and OpEx. General firmographic metadata (e.g., sector, country, revenue) was sourced from wikipedia and also human-verified. We show all the prompts in Appendix 10.10.

4.2 Dataset Structure and Statistics

For each of the 190 companies, the final dataset contains: (1) verified company metadata (country, number of employees, sector, industry, revenue); (2) the full annual report in machine-readable format (along with a version excluding the isolated Taxonomy section); (3) the extracted text of the Taxonomy disclosure section; (4) the human-verified list of reported economic activities (ground-truth labels); and (5) the human-verified quantitative KPIs (ground-truth values).

Dataset Composition The dataset is sectorally diverse, covering 11 distinct economic sectors, with *Industrials*, *Consumer Discretionary*, and *Financials* being the most represented. It is, geographically concentrated, with companies based primarily in Germany (63%) and Austria (20%). Company size metrics (Table 1) are highly right-skewed, indicating the dataset is influenced by a number of very large, multinational corporations. This distribution is a direct result of the regulatory implementation timeline for the 2023 reporting cycle, mandatory disclosure was primarily limited to large public-interest entities under the scope of the Non-Financial Reporting Directive (NFRD) (European Parliament and Council, 2014).

Table 1: Descriptive Statistics of Company size ($N = 190$).

Statistic	Annual Revenue (in billion €)	Number of Employees
Mean	45.18	38,436
Median	2.78	8,982
Std. Dev.	187.99	101,391
Minimum	0.05	49.00
Maximum	1,837	684,025

Reporting Statistics The analysis of reported data reveals significant variance. On average, companies reported 6.22 distinct economic activities (median 5.0, max 26), as shown in Table 2. The most frequently declared activities relate to the buildings, transport, and energy sectors.

Crucially, the KPI statistics (Table 3) reveal a large gap between eligibility and alignment. While the median *eligible* CapEx is 29.75%, the median *aligned* Turnover and OpEx are both 0.00%, with aligned CapEx at only 0.42%. This highlights that while many companies operate in eligible sectors, a very small fraction currently meet the strict Technical Screening Criteria for full alignment. Additional details can be found in Appendix 10.1.

Table 2: Descriptive Statistics of Reported Activities per Company (N=190)

Statistic	Value
Mean	6.22
Median	5.00
Standard Deviation	4.63
Minimum	1.00
Maximum	26.00

Table 3: Descriptive Statistics of EU Taxonomy KPIs (%) ($N = 190$).

KPI	Cat.	Mean	Med.	Std.
Turn.	Elig.	33.06	17.05	36.24
	Alig.	8.46	0.00	19.02
CapEx	Elig.	39.44	29.75	34.93
	Alig.	13.20	0.42	24.69
OpEx	Elig.	29.84	10.37	36.33
	Alig.	9.23	0.00	20.85

5 Methodology: Evaluating LLMs for Taxonomy Reporting

5.1 Models and Experimental Setup

Our empirical evaluation utilizes five LLMs: Google's `Gemini 2.0 Flash` (Google, 2025b), `Gemini 2.5 Flash` (Google, 2025c), Meta's `Llama 4 Maverick 17B Instruct` (Meta, 2025), `gemma-3-12b-it` (Google, 2025d), and `Mistral-Small-3.1-24B-Instruct-2503` (Mistral AI, 2025). Models were accessed via Google Cloud Platform (GCP) Vertex AI or a high-performance computing (HPC) cluster. We list the prompts used in the Appendix 10.10.

5.2 Task 1: Multiclass Activity Prediction

The goal of this task is to analyze LLM performance in predicting company's EU Taxonomy-

relevant eligible economic activities, framed as a multi-label classification problem. Models must infer activities based on operational descriptions, not direct Taxonomy mentions. We tested 3 context variations to assess impact of information depth:

1. **Prediction from Company Name:** The model uses only the company name, relying on its pre-existing knowledge.

2. **Prediction from Metadata:** The model receives the company's name, industry sector, annual revenue, and employee count.

3. **Prediction from Annual Report:** The model is given the full annual report text, with the dedicated EU Taxonomy disclosure section removed.

Due to context window limitations, the 'Annual Report' variation was not run for `gemma-3-12b-it` and `Mistral-Small-3.1-24B-Instruct-2503`. Models were prompted to return a JSON list of identified activities, a justification, and a confidence score. We also test a *few-shot* setting on the best-performing model (`Gemini 2.5 Flash`) for the Name Only and Company Info conditions. We provide five example companies from different sectors (Automotive, Real Estate, Energy, Software, Steel) with their correct activities as demonstrations in the prompt. These five companies are then excluded from evaluation to avoid data leakage (see Appendix 10.2).

5.3 Task 2: Binary Activity Classification

To evaluate model precision at a more granular level, we test the ability of an LLM to validate whether a *single* economic activity applies to a given company. A comprehensive approach would require over 28,000 classifications. To manage this cost, we used a representative sampling strategy, testing a subset of 30 companies (selected via stratified random sampling by sector) against the 116 activities actively reported by at least one company in our dataset.

5.4 Task 3: Quantitative KPI Regression

In this experiment, we attempt to assess the ability of LLMs to perform regression by predicting the EU Taxonomy KPIs: Turnover, Capital Expenditure (CapEx), and Operational Expenditure (OpEx). The analysis focuses on the three *eligible* KPIs, as the corresponding *aligned* KPI data is heavily zero-inflated and unsuitable for this regression task. The task was performed by the Gemini 2.0 and 2.5 models under two contextual conditions: (1) **With Metadata** only, and (2) **With Report** (Taxonomy section removed).

5.5 Task 4: Multi-Step Agentic Framework

This experiment evaluates the effectiveness of multi-step, agentic workflows designed to simulate a more deliberative analytical process. The objective is to determine if a pipeline that combines candidate generation with a verification step can improve upon single-step prediction. We implemented and tested two distinct architectures, as illustrated in Figure 2. The implementation was done using Google's Agent Development Framework (ADK) (Google, 2025a).

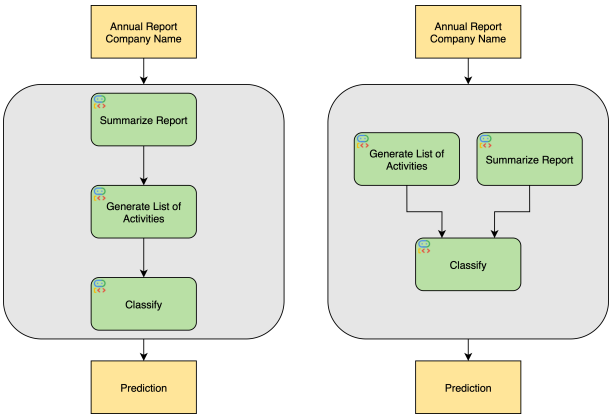

Figure 2: Architectural diagrams of the two agentic workflows. Approach 1 (left) is a sequential pipeline. Approach 2 (right) uses a parallel structure to refine predictions.

Sequential Summarization and Classification This workflow follows a linear, three-step process: (1) The full annual report is processed by an LLM to generate a comprehensive summary. (2) This summary serves as the sole input for a multiclass prediction agent to generate an initial list of activities. (3) The list of candidates is passed to a binary classification agent for verification and refinement.

Parallel Generation with Summary-Informed Classification This workflow processes the full report and its summary in parallel: (1) The highest-recall model from Task 1 (Gemini Flash 2.5) generates a broad list of candidate activities directly from the full annual report. (2) In parallel, the report is summarized. (3) Finally, a binary classification agent validates the broad candidate list, using the summary as the contextual basis for its decision.

6 Results and Analysis

6.1 Qualitative Activity Identification

We first evaluated the models' ability to perform multi-label classification of economic activities (Experiment 1). As shown in Table 4, performance was moderate. `Gemini Flash 2.5` achieved the highest $F1$-score (0.311) when provided with concise company metadata. A clear trade-off emerged: models like `Gemini Flash 2.5` and `Mistral` favored high recall (up to 0.526), while `Llama` achieved the highest precision (0.348), suggesting different models are suited for different workflow stages (e.g., discovery vs. verification). For reference, a fine-tuned `BERT-base` baseline achieves an $F1$-score of 0.097 under the same setting (Appendix 10.5). Notably, `Gemini 2.5 Pro` achieves an $F1$-score of 0.307 under the same zero-shot Company Info setting, showing no improvement over `Flash` (0.311). This is consistent with the KPI regression results and suggests that model scale alone does not improve performance on these tasks.

Table 4: Performance Metrics for Multiclass Activity Prediction (Exp.1). Best value per model is bolded; overall best per metric is underlined.

Model	Context	Precision	Recall	F1-Score
Baseline				
Fine-tuned BERT	Company Info (5-Fold CV)	0.232	0.076	0.097
Zero-Shot LLMs				
Gemini Flash 2.0	Name	**0.281**	**0.358**	**0.276**
	Company Info	0.254	0.334	0.253
	Report	0.241	0.332	0.245
Mistral	Name	0.186	**0.414**	0.213
	Company Info	**0.199**	0.375	**0.227**
Llama	Name	0.315	0.251	0.241
	Company Info	**0.348**	**0.253**	**0.251**
	Report	0.203	0.172	0.162
Gemma	Name	**0.301**	**0.222**	**0.219**
	Company Info	0.250	0.217	0.203
Gemini Flash 2.5	Name	0.268	0.452	0.297
	Company Info	**0.274**	0.493	**0.311**
	Report	0.213	**0.526**	0.275
Gemini Pro 2.5	Company Info	0.272	0.461	0.307
Gemini Flash 2.5	Name (Few-Shot)	**0.361**	0.334	**0.304**
	Co. Info (Few-Shot)	0.260	**0.389**	0.280

While Experiment 1 evaluates whether models can generate a full list of relevant activities, Experiment 2 tests the complementary question: given a single candidate activity, can the model confirm or reject it? Using only metadata, the model was highly precise (67.86%) but had very low recall (10.80%). Using the full report doubled the recall (20.45%) but nearly halved the precision (48.65%). This trade-off mirrors the pattern in Experiment 1 and motivates the study of agentic workflow, which chains generation with verification.

6.2 Quantitative KPI Prediction

In contrast to the moderate success in classification, both `Gemini 2.5 Flash` and `Gemini 2.5 Pro` failed at the quantitative KPI regression task (Experiment 3). As shown in Table 5, all R^2 values are negative for both models, indicating that their predictions are less accurate than a naive mean baseline. Notably, the larger `Pro` model does not outperform `Flash`, suggesting that increased model capacity alone is insufficient for this task (see Appendix 10.3 for a detailed analysis).

Table 5: KPI Regression results (Exp. 3). Negative R^2 values indicate predictions worse than a mean baseline. Both models fail across all KPIs.

Model	KPI Target	MAE	RMSE	R^2
Gemini 2.5 Flash	Turnover Eligible %	30.04	40.68	-0.21
	CapEx Eligible %	30.33	39.16	-0.29
	OpEx Eligible %	26.91	37.19	-0.002
Gemini 2.5 Pro	Turnover Eligible %	34.71	48.05	-0.77
	CapEx Eligible %	38.27	48.76	-0.96
	OpEx Eligible %	39.01	50.37	-0.94

Error analysis confirms this failure is systematic. Figure 3 shows that the predicted distributions do not capture the skewed, multi-modal shape of the ground truth and instead collapse into a narrow, overconfident range. Both models consistently over-predict eligibility and fail to ground their numerical estimates in the report content, indicating that this task cannot be addressed through prompting alone without domain-specific fine-tuning.

6.3 Efficacy of Agentic Workflows

Given the precision/recall trade-off in single-step classification, we evaluated a multi-step agentic framework (Experiment 4) to combine a high-recall generation step with a high-precision verification step. We tested two architectures on a sample of 50 companies using `Gemini Flash 2.5`.

Sequential Workflow first summarized the report, then generated candidates from the summary, and finally verified them. This workflow underperformed the baseline, with a final $F1$-score of 0.2957 (Table 6), likely due to information loss during the initial summarization.

Summarization Workflow first generated a high-recall list of candidates from the *full report* ($Recall = 0.5404$) and, in parallel, generated a summary. This summary was then used as the context for a verification agent to filter the candidate list. As shown in Table 7, this approach was more

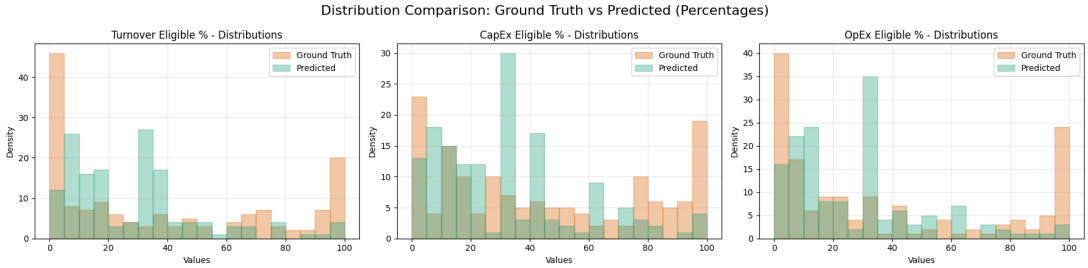

Figure 3: Comparison of ground-truth (orange) vs. predicted (green) KPI distributions. The model (gemini-flash-2.5) fails to capture the true data distributions.

effective. The verification step significantly improved precision (from 0.211 to 0.3436), resulting in a final $F1$-score of 0.3285. This modestly outperforms the best single-step model (0.311) and demonstrates the value of a 'generate-then-verify' pipeline for improving reliability.

Table 6: Performance of Sequential Workflow.

Pipeline Stage	Precision	Recall	F1-Score
Step 1: Candidate Generation	0.3487	0.3842	0.3250
End-to-End (Full Pipeline)	**0.3987**	**0.2961**	**0.2957**

Table 7: Performance of Parallel Workflow.

Pipeline Stage	Precision	Recall	F1-Score
Step 1: Candidate Generation	0.2110	0.5404	0.2796
End-to-End (Full Pipeline)	**0.3436**	**0.4016**	**0.3285**

6.4 Analysis of Influencing Factors

Our analysis of factors influencing multiclass prediction (Exp. 1) revealed that providing more data is not always better. The average $F1$-score when using the full, unstructured annual report was 0.2264, which was lower than using concise, structured company metadata (0.2493) or just the company name (0.2492). This is consistent with recent findings that long input sequences can degrade zero-shot performance (Wang et al., 2024), and suggests the extensive, noisy text of a full report impedes inference for this task.

Other factors were also highly predictive. Company sector was a key determinant, with *Utilities* and *Real Estate* yielding the highest $F1$-scores (0.4226 and 0.3972, respectively), while *Financials* (0.1551) was the most challenging. A Random Forest analysis identified company size (specifically `log_employees` and `log_revenue`) as the most important predictive features of model performance, accounting for over 47% of feature importance.

Table 8: Top 5 Most Important Features from Random Forest Analysis

Feature	Importance Score
log_employees	0.2387
log_revenue	0.2377
sector_Utilities	0.0779
sector_Real Estate	0.0687
context_with_report	0.0337

6.5 Analysis of Model Reliability

Finally, we assessed whether a model's self-reported confidence is a reliable indicator of its accuracy. For the multiclass classification task (Exp. 1), we found all models to be significantly miscalibrated. Table 9 shows high Expected Calibration Error (ECE) (Kadavath et al., 2022; Tian et al., 2023) scores across the board.

Notably, `Gemini Flash 2.5`, the model with the highest recall, was also the most poorly calibrated, with its Expected Calibration Error (ECE) peaking at 0.684 when using the full report. This indicates a strong tendency toward overconfidence; the model is not "aware" of its high rate of false positives. `Llama-4` was more stable, but still miscalibrated (ECE $\approx$ 0.40). These findings show that in a zero-shot setting, verbalized confidence scores for this task are not dependable.

7 Discussion

Our evaluation reveals a significant performance disparity between qualitative and quantitative tasks. For qualitative activity classification, the best single-step model achieved a moderate $F1$-score of 0.311, with a multi-step agentic workflow modestly improving it to 0.3285. This suggests current LLMs can act as "analyst assistants" for an initial discovery phase, especially high-recall models like `Gemini Flash 2.5`, but are not reliable

Table 9: Expected Calibration Error (ECE) for Multiclass Prediction (Exp. 1) (Bin size $m = 15$). Best (lowest) ECE per model is bolded; overall best is underlined.

Model	Context Level	ECE Score	N
Gemini Flash 2.0	Name Only	0.392	1161
	With Company Info	0.361	1210
	With Report	**0.353**	1449
Gemini Flash 2.5	Name Only	**0.600**	1860
	With Company Info	0.611	2031
	With Report	0.684	2644
Gemma-3	Name Only	**0.430**	193
	With Company Info	0.483	602
Llama-4	Name Only	0.407	664
	With Company Info	**0.390**	622
	With Report	0.393	885
Mistral	Name Only	0.606	3344
	With Company Info	**0.546**	2139

for full automation. An error analysis reveals that models tend to over-predict broad, cross-sector activities while missing incidental activities related to general corporate operations (Appendix 10.7), confirming that a human-in-the-loop is required.

Adding few-shot examples to `Gemini Flash 2.5` shows mixed results depending on the input context. With only the company name, few-shot examples clearly help: precision increases from 0.268 to 0.361, and $F1$ improves slightly ($0.297 \rightarrow 0.304$). However, when company metadata is already provided, few-shot examples hurt performance ($F1$: $0.311 \rightarrow 0.280$). The reason is over-prediction: the model predicts 7.55 activities per company on average, compared to 6.23 in the ground truth. We find that 34.9% of false positives are activities copied from the examples, meaning the model tends to reuse example activities even when they do not apply (see Appendix 10.2).

Replacing `Flash` with the larger `Gemini 2.5 Pro` does not improve performance on either task: $F1$ remains comparable on activity prediction (0.307 vs. 0.311) and all R^2 values remain negative on KPI regression (Appendix 10.3). This confirms that the challenge lies in the task structure rather than in model capacity, and motivates the need for task-specific approaches in future work.

We also observe that providing the full annual report does not always improve performance over concise metadata. This is consistent with recent findings that long input sequences can reduce the effectiveness of zero-shot inference (Wang et al., 2024), and highlights the value of structured input preprocessing in practical applications.

Finally, the models' self-reported confidence scores were found to be poorly calibrated, with ECE scores as high as 0.684. We attribute this not just to model overconfidence, but to a fundamental information gap. The LLM acts as an external auditor, trying to replicate a complex internal assessment (the "ground truth") using only public data. It is being asked to make a context-rich decision with incomplete information, making its inability to produce reliable confidence scores an expected outcome. This reinforces the intuition that domain-specific specialization is crucial (Birti et al., 2025).

8 Conclusion

We presented a systematic evaluation of LLMs on the core analytical tasks required for EU Taxonomy reporting, using a novel, curated dataset of 190 corporate reports which we have made public.

We make two primary contributions: (1) We release the first structured dataset for this multi-task (qualitative and quantitative) regulatory challenge. (2) We establish a comprehensive performance benchmark for prompting-based LLMs, demonstrating their current capabilities and limitations.

Our findings demonstrate that LLMs show moderate promise as assistive tools for the qualitative task of activity identification. However, their outputs are not yet reliable for full automation. We conclude that the path toward robust automation will require specialized, fine-tuned models and sophisticated, human-in-the-loop agentic systems.

9 Limitations

Our findings are subject to the following considerations. First, our dataset draws primarily from German (63%) and Austrian (20%) companies, as detailed in Section 4. Future work can extend coverage as the CSRD broadens the reporting scope. Second, not all models were evaluated on every task due to practical constraints such as context window limits and API costs. Third, our study focuses on prompting-based strategies; fine-tuned models may achieve different results. Finally, the moderate performance levels observed highlight the importance of responsible deployment: errors in automated assessments could lead to incorrect regulatory conclusions, and our benchmark provides clear performance boundaries that can guide practitioners in determining where human expertise remains essential. We discuss the broader societal

implications of automating EU Taxonomy compliance in Appendix 10.9.

References

Mohammed Ali, Abdelrahman Abdallah, Amit Agarwal, Hitesh Laxmichand Patel, and Adam Jatowt. 2026. Recor: Reasoning-focused multi-turn conversational retrieval benchmark. *arXiv preprint arXiv:2601.05461*.

Mohammed Ali, Abdelrahman Abdallah, and Adam Jatowt. 2025. Sustainableqa: A comprehensive question answering dataset for corporate sustainability and eu taxonomy reporting. *arXiv preprint arXiv:2508.03000*.

Mattia Birti, Francesco Osborne, and Andrea Maurino. 2025. Optimizing large language models for esg activity detection in financial texts. *Preprint*, arXiv:2502.21112.

Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M. Suchanek. 2025. Corporate greenwashing detection in text – a survey. *Preprint*, arXiv:2502.07541.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

European Commission. 2020. Financing the green transition: The european green deal investment plan and just transition mechanism. Press Corner. Accessed: 2025-08-06.

European Commission. 2021a. Commission Delegated Regulation (EU) 2021/2139. Official Journal of the European Union, L 442, pp. 1-633.

European Commission. 2021b. Commission Delegated Regulation (EU) 2021/2178. Official Journal of the European Union, L 443, pp. 9-67.

European Commission. 2023. Commission Delegated Regulation (EU) 2023/2486 . Official Journal of the European Union, L 2486, pp. 1-215.

European Commission. 2024a. Eu taxonomy for sustainable activities. Finance. Accessed: 2025-08-06.

European Commission. 2024b. The european green deal. Strategy and policy. Accessed: 2025-08-06.

European Parliament and Council. 2014. Directive 2014/95/EU of the European Parliament and of the Council . Official Journal of the European Union, L 330, 15.11.2014, p. 1-7. Directive 2014/95/EU (NFRD).

European Parliament and Council. 2022. Directive (EU) 2022/2464 of the European Parliament and of the Council . Official Journal of the European Union, L 322, 16.12.2022, p. 15-80. Directive (EU) 2022/2464 (CSRD).

European Parliament and Council of the European Union. 2020. Regulation (EU) 2020/852 of the European Parliament and of the Council . Official Journal of the European Union, L 198, pp. 13-43.

Google. 2025a. Agent Development Kit (ADK). https://google.github.io/adk-docs/.

Google. 2025b. Gemini 2.0 flash model card. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash. Accessed: 2025-07-28.

Google. 2025c. Gemini 2.5 flash model card. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash. Accessed: 2025-07-28.

Google. 2025d. gemma-3-12b-it model card. https://huggingface.co/google/gemma-3-12b-it. Accessed: 2025-07-28.

Chaoyue He, Xin Zhou, Yi Wu, Xinjia Yu, Yan Zhang, Lei Zhang, Di Wang, Shengfei Lyu, Hong Xu, Xiaoqiao Wang, Wei Liu, and Chunyan Miao. 2025. Esgenius: Benchmarking llms on environmental, social, and governance (esg) and sustainability knowledge. *Preprint*, arXiv:2506.01646.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2024. Large language models in finance: A survey. *Preprint*, arXiv:2311.10723.

Meta. 2025. Llama 4 maverick 17b instruct on vertex ai. https://console.cloud.google.com/vertex-ai/publishers/meta/model-garden/llama-4-maverick-17b-128e-instruct-maas. Accessed: 2025-07-28.

Mistral AI. 2025. Mistral-small-3.1-24b-instruct-2503 model card. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503. Accessed: 2025-07-28.

Vitor Oliveira, Gabriel Nogueira, Thiago Faleiros, and Ricardo Marcacini. 2025. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law*, 33(2):361–381.

Vik Paruchuri. 2024. Marker. https://github.com/datalab-to/marker.

Franziska Schütze and Benedikte Sandbaek. 2025. Taxonomy disclosure in the eu–a useful framework, despite current challenges. *The Economists' Voice*, 22(1):161–172.

Agne Sneideriene and Renata Legenzova. 2025. Greenwashing prevention in environmental, social, and governance (esg) disclosures: A bibliometric analysis. *Research in International Business and Finance*, 74:102720.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024. Beyond the limits: A survey of techniques to extend the context length in large language models. *Preprint*, arXiv:2402.02244.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Preprint*, arXiv:2312.17617.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025. A survey of generative information extraction. In *Proceedings of the 31st International conference on computational linguistics*, pages 4840–4870.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, HongXiang Tong, Lei Xiao, and Wenwen Zhou. 2023. Esgreveal: An llm-based approach for extracting structured data from esg reports. *Preprint*, arXiv:2312.17264.

10 Appendix

This appendix contains supplementary visualizations that provide additional detail on the dataset's characteristics.

10.1 Dataset characteristics

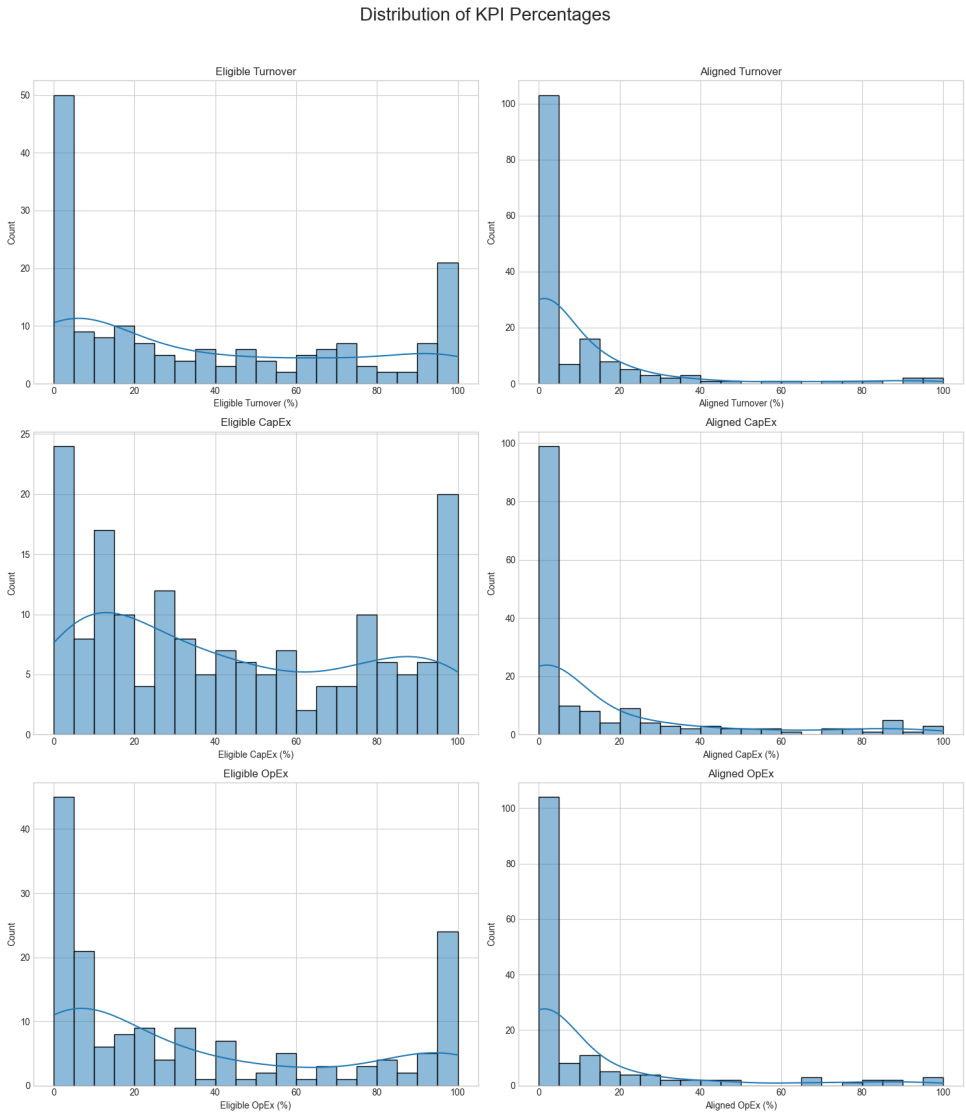

Figure 4: Detailed distributions of the six EU Taxonomy KPI percentages. The histograms for aligned KPIs (right column) show a strong concentration at or near zero, while eligible KPIs (left column) exhibit more varied distributions.

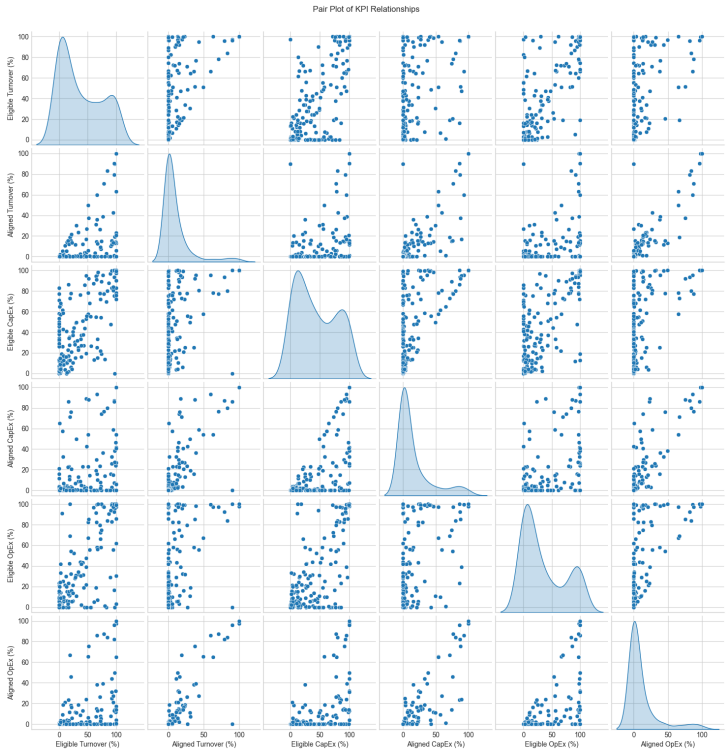

Figure 5: Pair plot matrix of the six EU Taxonomy KPI percentages. The diagonal shows the kernel density estimate (KDE) for each variable, while the off-diagonal plots show the scatter relationship between each pair of variables.

10.1.1 Prediction Plots

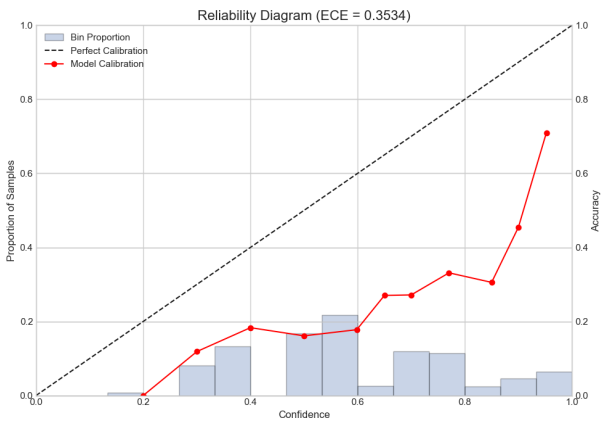

Figure 6: ECE Reliability Diagram Gemini Flash 2.0 Multiclass Task

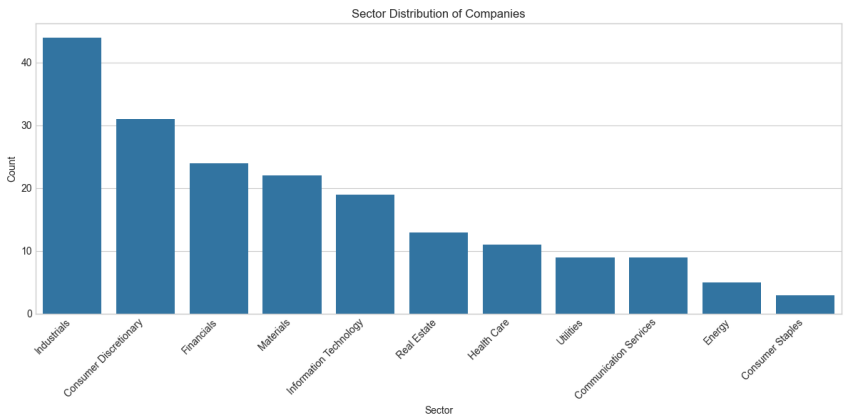

Figure 7: Sectoral distribution of companies in the dataset.

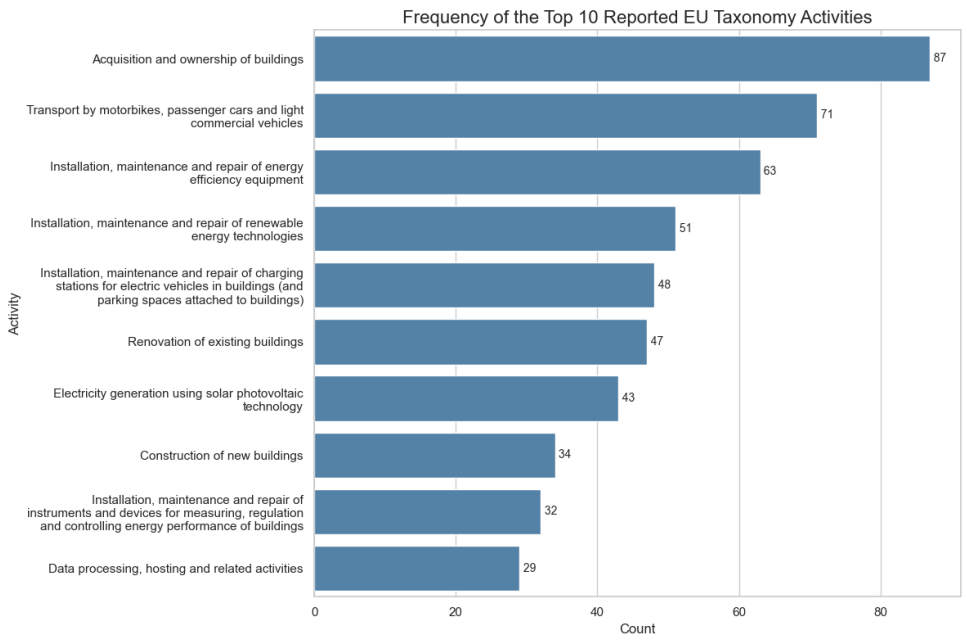

Figure 8: Frequency of the Top 10 Reported EU Taxonomy Activities

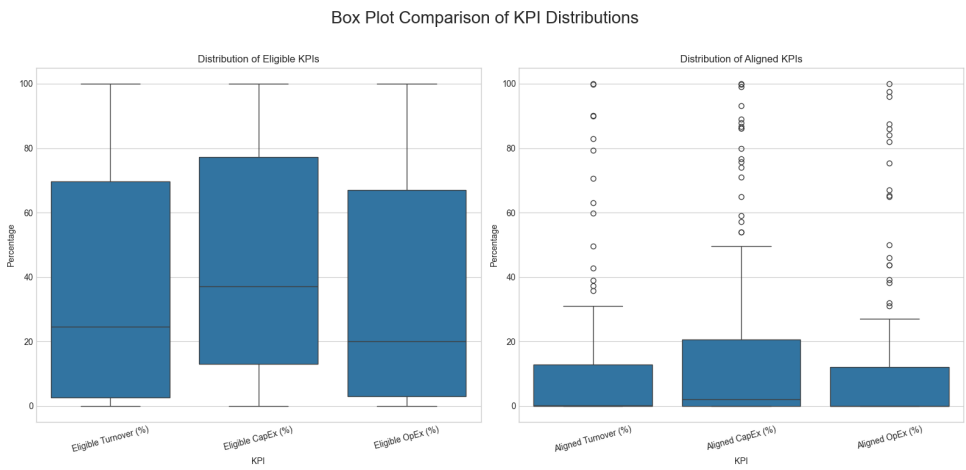

Figure 9: Box Plot Comparison of KPI Distributions.

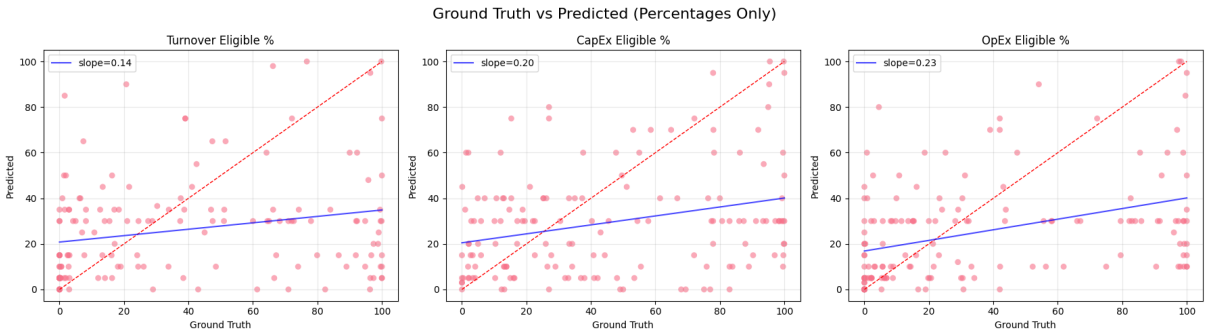

Figure 10: Comparison of ground-truth values versus model predictions for EU Taxonomy eligible KPIs.

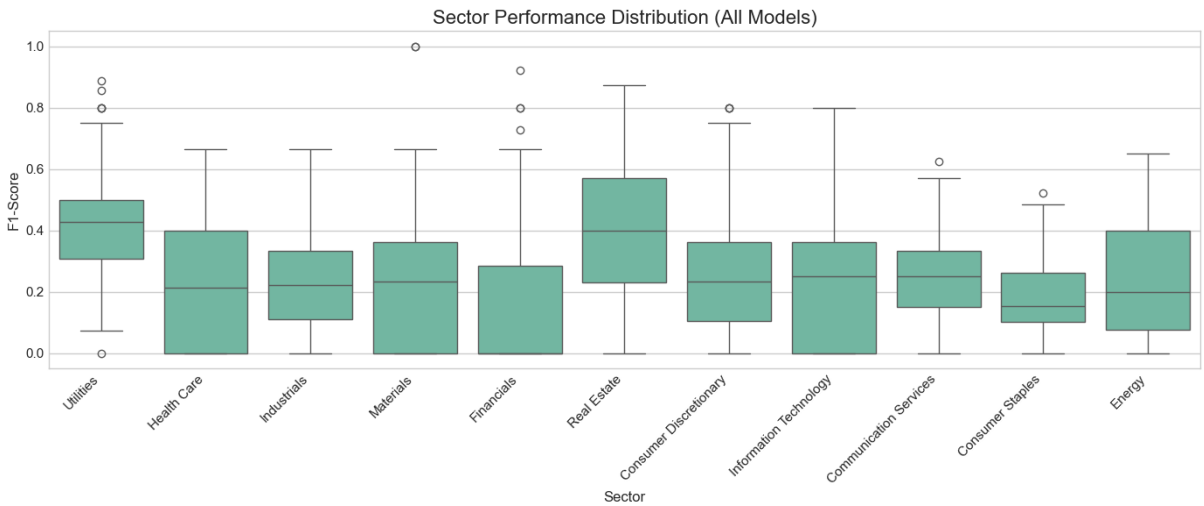

Figure 11: Distribution of F1-Scores by Company Sector across all models.

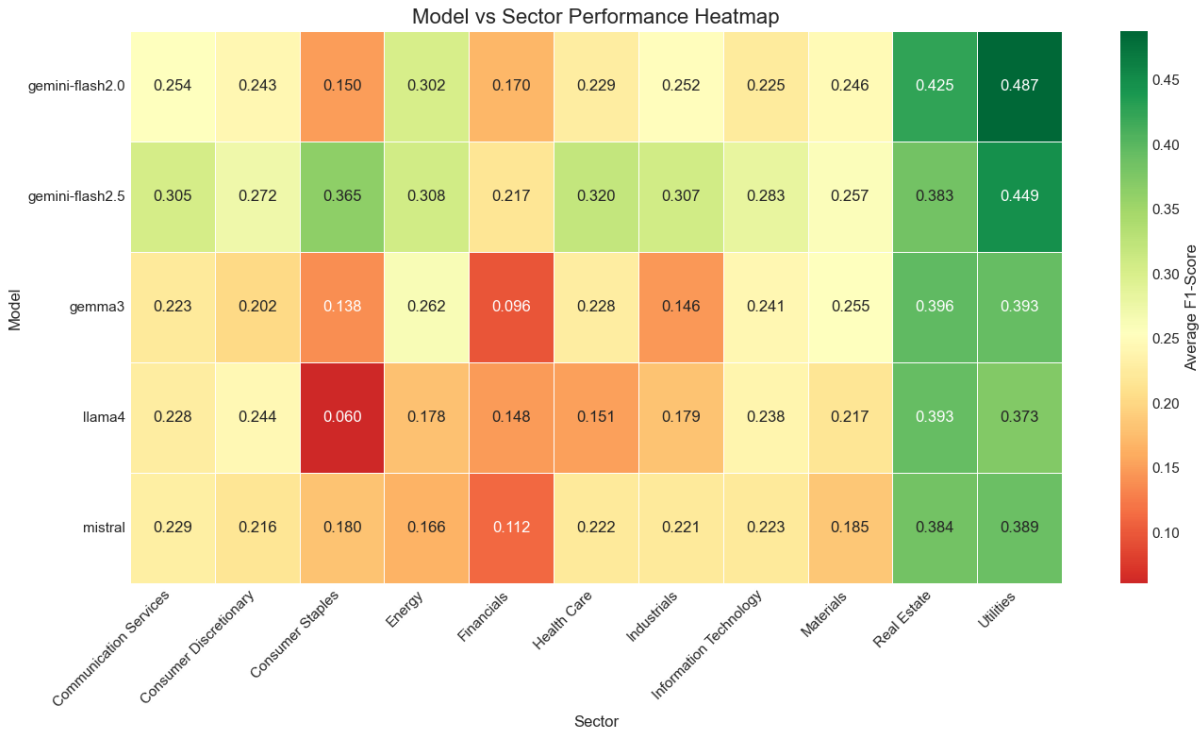

Figure 12: Heatmap of Average F1-Score by Model and Sector.

10.2 Few-Shot Experiment Details

For the few-shot setting, we selected five example companies from different sectors: **BMW** (Automotive), **LEG Immobilien** (Real Estate), **E.ON** (Energy/Utilities), **Nemetschek** (Software/Technology), and **Salzgitter** (Steel/Materials). For each example company, we include its correct EU Taxonomy activities in the same JSON format the model is expected to output. These five companies are removed from the test set, leaving 184 companies for evaluation.

The prompt is the same as the zero-shot version (Appendix 10.10.1), with the only change being the added examples before the target company. This way, any difference in results comes only from the examples.

Over-Prediction Analysis. When using company metadata with few-shot examples, the model predicts too many activities: 7.55 per company on average, while the ground truth has 6.23. The most common wrong predictions were "Close to market research, development and innovation" (89 times) and "Data-driven solutions for GHG emissions reductions" (65 times). We found that 34.9% of all wrong predictions were activities that appeared in the five examples. This shows the model copies activities from the examples and applies them to companies where they do not belong. This problem gets worse when the model already has rich input (company metadata), because the extra context combined with the examples leads the model to predict even more activities.

10.3 KPI Regression: Frontier Model Analysis

To investigate whether increased model capacity can address the failure observed in Experiment 3, we evaluate `Gemini 2.5 Pro` on all 190 companies using the identical prompt (Appendix 10.10.3) and evaluation pipeline used for `Flash`. Table 10 compares the two models.

As shown in Table 10, `Pro` performs consistently worse than `Flash` across all three KPIs, with higher MAE and more strongly negative R^2 values. This result is counter-intuitive, as one would expect a larger model to at least match a smaller one on the same task. We identify two patterns that explain this behavior.

Over-Prediction Bias. The ground-truth eligibility percentages are heavily right-skewed: the median is 17.1% for Turnover, 29.8% for CapEx, and 11.2% for OpEx, with a large number of companies reporting zero eligibility (48 for Turnover, 24 for CapEx, 55 for OpEx). In contrast, `Pro` predicts much higher values, with medians of 49.5%, 72.0%, and 65.0% respectively. The model rarely predicts low values: out of 24 companies with zero CapEx eligibility in the ground truth, `Pro` never predicts zero for this KPI. For Turnover, 48 companies have zero eligibility, but `Pro` predicts zero for only 15 of them. Overall, `Pro` predicts values above 50% for 93 companies on Turnover, 129 on CapEx, and 113 on OpEx, while the actual data is concentrated at the low end. This systematic upward bias is the main driver of the poor R^2 scores.

Table 10: KPI Regression: `Gemini 2.5 Pro` vs. `Flash` on all 190 companies (eligible percentage KPIs).

Model	KPI	MAE	RMSE	R^2
Flash	Turnover	30.04	40.68	−0.21
	CapEx	30.33	39.16	−0.29
	OpEx	26.91	37.19	−0.002
Pro	Turnover	34.71	48.05	−0.77
	CapEx	38.27	48.76	−0.96
	OpEx	39.01	50.37	−0.94

10.4 Task Design: Excluding the Taxonomy Section

In our dataset, each annual report is provided in two versions: the complete report and a version from which the dedicated EU Taxonomy disclosure section has been removed. All report-based experiments in this paper use the version without the Taxonomy section.

The Taxonomy disclosure section contains the final output of the compliance process: the list of eligible activities and the corresponding KPI values, typically presented in structured tables. If this section were included in the input, the activity classification task would reduce to extracting a pre-existing list, and the KPI regression task would reduce to reading values from a table. Neither would test the model's ability to reason about a company's operations.

By removing this section, we evaluate whether LLMs can perform the analytical steps that precede the final disclosure: identifying which business operations correspond to Taxonomy-defined activities, and estimating the associated financial proportions from operational and financial descriptions throughout the report. This reflects the practical setting in which a company or auditor must perform this analysis before the disclosure section has been drafted.

10.5 Baseline: Fine-Tuned BERT for Activity Prediction

To establish a supervised baseline for the activity prediction task, we fine-tuned `BERT-base-uncased` (Devlin et al., 2019) for multi-label classification using 5-fold cross-validation. The model receives the same company metadata used in the LLM experiments (name, industry, country, revenue, and employee count) and predicts across all 156 activity classes. Training was conducted for 10 epochs per fold with a learning rate of 2×10^{-5}.

Table 11: Fine-tuned BERT baseline vs. best zero-shot LLM.

Model	Precision	Recall	F1
Fine-tuned BERT (5-Fold CV)	0.232	0.076	0.097
Gemini 2.5 Flash (Zero-Shot)	0.274	0.493	0.311

The fine-tuned model averages only 1.6 predicted activities per company compared to 6.2 in the ground truth, resulting in low recall (0.076). The zero-shot LLMs achieve over three times the $F1$-score, with substantially higher recall (0.493), indicating that pre-trained world knowledge about companies and industries enables LLMs to identify relevant activities that a supervised model trained on this dataset alone cannot recover.

10.6 Annotation and Verification Protocol

The ground-truth labels in our dataset are not subjective annotations but factual data drawn from official company disclosures: each company explicitly declares its eligible activities and KPI values in its EU Taxonomy reporting section. The extraction followed a three-stage pipeline: (1) automated extraction using Gemini-Flash-2.0 with zero-shot prompting, (2) an LLM-based self-verification step in which the model cross-checked its outputs against the source text, and (3) human verification by two PhD students, who split the 190 companies between them and independently checked each extraction against the original report, correcting any errors.

10.7 Error Analysis: Activity Prediction

To understand how the models fail, we analyze the false positives and false negatives of the best-performing zero-shot configuration across all evaluated companies. Two clear patterns emerge. On the false-positive side, the models frequently assign activities that are broadly applicable across industries, such as *Data-driven solutions for GHG emissions reductions* and *Close to market research, development and innovation*, even when companies do not report them. These activities appear as false positives for over 60 companies each, indicating that the models treat them as safe defaults rather than company-specific predictions. On the false-negative side, the most commonly missed activities are those that arise from ancillary corporate operations: *Transport by motorbikes, passenger cars and light commercial vehicles* (missed 63 times) and *Acquisition and ownership of buildings* (missed 44 times) are reported by companies across all sectors through their vehicle fleets and property holdings, yet the models cannot infer them from metadata that describes only the primary line of business. This asymmetry highlights a key challenge for metadata-based approaches: while they capture core operations effectively, the secondary activities that appear in real-world Taxonomy disclosures require richer contextual input.

10.8 Extended Related Work

NLP for Financial and ESG Reporting. The application of NLP in the financial domain has rapidly evolved from text mining to complex reasoning tasks using LLMs (Li et al., 2024). A key driver has been the development of domain-specific models. FinBERT, an adaptation of BERT pre-trained on financial documents, demonstrated high performance on financial sentiment analysis. More recently, models like BloombergGPT have shown state-of-the-art performance on financial NLP benchmarks by training on a mix of general and domain-specific corpora (Li et al., 2024).

This trend extends to the field of Environmental, Social, and Governance (ESG) analysis. ESGReveal employs a RAG approach to extract structured numerical and qualitative data from corporate ESG reports, with GPT-4 achieving high accuracy on disclosure analysis tasks (Zou et al., 2023). Concurrently, benchmarks like ESGenius evaluate the foundational ESG knowledge of LLMs using expert-validated multiple-choice questions (He et al., 2025). More recently, SustainableQA (Ali et al., 2025) introduced a large-scale dataset of over 195,000 question-answer pairs extracted from corporate sustainability reports, demonstrating that a compact fine-tuned model can outperform larger general-purpose systems on sustainability compliance tasks.

Generative Information Extraction. The core compliance tasks mandated by the EU Taxonomy, identifying relevant economic activities and extracting quantitative KPIs from corporate reports, are applications of Information Extraction (IE). LLMs have catalyzed a shift in IE from traditional supervised, token-level classification (e.g., Named Entity Recognition) to generative IE (Xu et al., 2024; Zhang et al., 2025). In this paradigm, extraction is reformulated as a text-to-text task, where a model is prompted to generate the desired structured output directly, often demonstrating strong zero-shot capabilities (Xu et al., 2024).

However, applying IE to specialized domains like legal and financial reporting presents distinct challenges, particularly the high cost of manual annotation (Oliveira et al., 2025). Research has shown that combining LLM-based prompting with weak supervision and a small fraction of human-labeled data can be effective for creating training data for NER in complex legal documents (Oliveira et al., 2025). More specific to our focus, (Birti et al., 2025) introduced ESG-Activities, a benchmark for classifying financial texts against EU Taxonomy activities. Their work highlights that fine-tuning smaller models on domain-specific data (both manual and synthetic) can significantly improve performance, even surpassing larger proprietary models. RECOR (Ali et al., 2026) introduced a reasoning-focused multi-turn conversational retrieval benchmark spanning 11 domains including sustainable living, highlighting the broader challenge of combining complex reasoning with domain-specific information retrieval.

10.9 Broader Impact

The automation of EU Taxonomy compliance has significant societal implications. First, professional advisory services for Taxonomy reporting are costly, creating a barrier for small and medium-sized enterprises (SMEs) that lack dedicated sustainability departments. LLMs offer a substantially cheaper alternative that could democratize access to regulatory compliance, enabling financially constrained companies to participate in the green transition on more equal footing.

Second, the EU Taxonomy is not static. The regulatory framework evolves continuously through new and amended Delegated Acts, expanded environmental objectives, and a broadening scope under the CSRD. Human experts require significant time and training to stay current with these changes. LLM-based systems, by contrast, can be rapidly updated with new regulatory texts, making them well-suited to keep pace with a fast-moving legal landscape.

However, our results also highlight important risks. The moderate performance levels we observe mean that uncritical reliance on LLM outputs could lead to incorrect regulatory conclusions, potentially resulting in misallocation of sustainable finance or unintended greenwashing. Our benchmark provides clear performance boundaries that can guide practitioners in determining where human expertise remains essential. We therefore advocate for human-in-the-loop deployment, where LLMs accelerate the initial analysis but final decisions are validated by domain experts.

10.10 Prompts

10.10.1 Multiclass Prediction Prompt

Prompt designed for the multiclass prediction of EU Taxonomy activities (Experiment 1).

```
**Task:** Your task is to act as a Sustainable Finance Analyst. Your goal is
to thoroughly analyze the provided company ‘annual_report‘ and predict all EU
Taxonomy activities the company is likely involved in.
**Important Constraint:** The annual report will likely not mention the EU
Taxonomy directly.  Your task is to **infer** the activities based on the
company’s described operations, projects, products, services, and environmental
initiatives.
**Context: The EU Taxonomy Framework** The EU Taxonomy Regulation aims to help
channel capital towards activities that substantially contribute to reaching the
objectives of the European Green Deal. By providing clear, common definitions
of environmentally sustainable economic activities, it helps to increase
transparency and mitigate the risk of ’greenwashing’.
An economic activity must meet four conditions to qualify as environmentally
sustainable: 1. It must contribute substantially to one or more of the six
environmental objectives. 2. It must ’do no significant harm’ (DNSH) to any of
the other environmental objectives. 3. It must be carried out in compliance
with minimum social safeguards. 4. It must comply with detailed technical
screening criteria (TSC) established by the Commission.
The six environmental objectives are: climate change mitigation; climate change
adaptation; sustainable use and protection of water and marine resources;
transition to a circular economy; pollution prevention and control; and the
protection and restoration of biodiversity and ecosystems.
**Core Directive:** Your analysis must be comprehensive and evidence-based. It
is better to include a plausible activity supported by textual evidence than
to miss one.
**Inputs:** * **Company Name:** {company_name} * **Company Information (JSON
Object):** Company Information: {company_info}, * **Annual Report:** The full
text of the report will be at the End of the Instuctions * **Authoritative List
of EU Taxonomy Activities:** [ "Afforestation", "Rehabilitation and restoration
of forests, including reforestation and natural forest regeneration after an
extreme event", "Forest management", "Conservation forestry", "Restoration of
wetlands", "Conservation, including restoration, of habitats[1], ecosystems[2]
and species", "Manufacture of renewable energy technologies", "Manufacture
of equipment for the production and use of hydrogen", "Manufacture of low
carbon technologies for transport", "Manufacture of batteries", "Manufacture
of energy efficiency equipment for buildings", "Manufacture of other low
carbon technologies", "Manufacture of cement", "Manufacture of aluminium",
"Manufacture of iron and steel", "Manufacture of hydrogen", "Manufacture
of carbon black", "Manufacture of soda ash", "Manufacture of chlorine",
"Manufacture of organic basic chemicals", "Manufacture of anhydrous ammonia",
"Manufacture of nitric acid", "Manufacture of plastics in primary form",
"Manufacture of automotive and mobility components", "Manufacture of rail
rolling stock constituents", "Manufacture, installation, and servicing of high,
medium and low voltage electrical equipment for electrical transmission and
distribution that result in or enable a substantial contribution to climate
change mitigation", "Manufacturing of aircraft", "Manufacture, installation and
```

associated services for leakage control technologies enabling leakage reduction and prevention in water supply systems", "Manufacture of plastic packaging goods", "Manufacture of electrical and electronic equipment", "Manufacture of active pharmaceutical ingredients (API) or active substances", "Manufacture of medicinal products", "Electricity generation using solar photovoltaic technology", "Electricity generation using concentrated solar power (CSP) technology", "Electricity generation from wind power", "Electricity generation from ocean energy technologies", "Electricity generation from hydropower", "Electricity generation from geothermal energy", "Electricity generation from renewable non-fossil gaseous and liquid fuels", "Electricity generation from bioenergy", "Transmission and distribution of electricity", "Storage of electricity", "Storage of thermal energy", "Storage of hydrogen", "Manufacture of biogas and biofuels for use in transport and of bioliquids", "Transmission and distribution networks for renewable and low-carbon gases", "District heating/cooling distribution", "Installation and operation of electric heat pumps", "Cogeneration of heat/cool and power from solar energy", "Cogeneration of heat/cool and power from geothermal energy", "Cogeneration of heat/cool and power from renewable non-fossil gaseous and liquid fuels", "Cogeneration of heat/cool and power from bioenergy", "Production of heat/cool from solar thermal heating", "Production of heat/cool from geothermal energy", "Production of heat/cool from renewable non-fossil gaseous and liquid fuels", "Production of heat/cool from bioenergy", "Production of heat/cool using waste heat", "Pre-commercial stages of advanced technologies to produce energy from nuclear processes with minimal waste from the fuel cycle", "Electricity generation from nuclear energy in existing installations", "Electricity generation from fossil gaseous fuels", "High-efficiency co-generation of heat/cool and power from fossil gaseous fuels", "Production of heat/cool from fossil gaseous fuels in an efficient district heating and cooling system", "Construction and safe operation of new nuclear power plants, for the generation of electricity and/or heat, including for hydrogen production, using best-available technologies", "Construction, extension and operation of water collection, treatment and supply systems", "Renewal of water collection, treatment and supply systems", "Construction, extension and operation of waste water collection and treatment", "Renewal of waste water collection and treatment", "Collection and transport of non-hazardous waste in source segregated fractions", "Anaerobic digestion of sewage sludge", "Anaerobic digestion of bio-waste", "Composting of bio-waste", "Material recovery from non-hazardous waste", "Landfill gas capture and utilisation", "Transport of CO2", "Underground permanent geological storage of CO2", "Desalination", "Water supply", "Urban Waste Water Treatment", "Sustainable urban drainage systems (SUDS)", "Phosphorus recovery from waste water", "Production of alternative water resources for purposes other than human consumption", "Collection and transport of non-hazardous and hazardous waste", "Treatment of hazardous waste", "Recovery of bio-waste by anaerobic digestion or composting", "Depollution and dismantling of end-of-life products", "Sorting and material recovery of non-hazardous waste", "Collection and transport of hazardous waste", "Remediation of legally non-conforming landfills and abandoned or illegal waste dumps", "Remediation of contaminated sites and areas", "Passenger interurban rail transport", "Freight rail transport", "Urban and suburban transport, road passenger transport", "Operation of personal mobility devices, cycle logistics", "Transport by motorbikes, passenger cars

and light commercial vehicles", "Freight transport services by road", "Inland passenger water transport", "Inland freight water transport", "Retrofitting of inland water passenger and freight transport", "Sea and coastal freight water transport, vessels for port operations and auxiliary activities", "Sea and coastal passenger water transport", "Retrofitting of sea and coastal freight and passenger water transport", "Infrastructure for personal mobility, cycle logistics", "Infrastructure for rail transport", "Infrastructure enabling low-carbon road transport and public transport", "Infrastructure enabling low carbon water transport", "Low carbon airport infrastructure", "Leasing of aircraft", "Passenger and freight air transport", "Air transport ground handling operations", "Infrastructure enabling road transport and public transport", "Infrastructure for water transport", "Airport infrastructure", "Data processing, hosting and related activities", "Data-driven solutions for GHG emissions reductions", "Computer programming, consultancy and related activities", "Programming and broadcasting activities", "Software enabling physical climate risk management and adaptation", "Provision of IT/OT data-driven solutions for leakage reduction", "Provision of IT/OT data-driven solutions", "Close to market research, development and innovation", "Research, development and innovation for direct air capture of CO2", "Professional services related to energy performance of buildings", "Engineering activities and related technical consultancy dedicated to adaptation to climate change", "Consultancy for physical climate risk management and adaptation", "Non-life insurance: underwriting of climate-related perils", "Reinsurance", "Education", "Residential care activities", "Creative, arts and entertainment activities", "Libraries, archives, museums and cultural activities", "Motion picture, video and television programme production, sound recording and music publishing activities", "Construction of new buildings", "Renovation of existing buildings", "Installation, maintenance and repair of energy efficiency equipment", "Installation, maintenance and repair of charging stations for electric vehicles in buildings (and parking spaces attached to buildings)", "Installation, maintenance and repair of instruments and devices for measuring, regulation and controlling energy performance of buildings", "Installation, maintenance and repair of renewable energy technologies", "Acquisition and ownership of buildings", "Demolition and wrecking of buildings and other structures", "Maintenance of roads and motorways", "Use of concrete in civil engineering", "Emergency Services", "Flood risk prevention and protection infrastructure", "Nature-based solutions for flood and drought risk prevention and protection", "Repair, refurbishment and remanufacturing", "Sale of spare parts", "Preparation for re-use of end-of-life products and product components", "Sale of second-hand goods", "Product-as-a-service and other circular use- and result-oriented service models", "Marketplace for the trade of second-hand goods for reuse", "Hotels, holiday, camping grounds and similar accommodation"]

Instructions for Analysis and Prediction:

1. **Strategic Analysis of the Report:** Do not just read the report linearly. Analyze it strategically to find the most relevant information. Focus on sections such as: * 'Business Description', 'Our Operations', 'Products & Services', or 'Business Segments'. * 'Sustainability Report' or 'Corporate Social Responsibility (CSR) Report'. * 'Management Discussion and Analysis (MD&A)' for details on projects and investments.

2. **Extract Core Business Activities:** From these sections, identify the company's fundamental business model. What does it physically produce, build, manage, or service?
3. **Search for Thematic Evidence:** Look for descriptions of projects or processes that align with the themes of the six environmental objectives, even if the terminology is different. For example: * For **Circular Economy**, search for terms like "recycling," "waste reduction," "re-use," "circular design," "remanufacturing," "recycled content." * For **Pollution Prevention**, search for "emissions control," "waste treatment," "pollution reduction," "scrubbers," "hazardous substance management." * For **Water Protection**, look for "wastewater treatment," "water efficiency," "leakage reduction."
4. **Map Evidence to Taxonomy Activities:** Based on the evidence you've gathered, match the company's operations to the specific activities in the 'Authoritative List'.
5. **Generate Structured Output:** Return your prediction as a JSON list of objects.
* 'activity_name': The exact name of the activity from the provided list. * 'justification': **This is critical.** You must justify your prediction by citing the specific evidence from the 'annual_report'. Briefly quote the relevant phrase or sentence that supports your conclusion. * 'estimated_relevance': Assess how central the activity appears to be to the company's business model as described in the report. Use "High," "Medium," or "Low." * 'confidence_score': A numerical score (0-1) reflecting your confidence that the described operation maps to the chosen Taxonomy activity.
Example Output:
[{{ "activity_name": "Manufacture of batteries", "justification": "The report states in the 'Strategic Investments' section: 'This year we began construction of our new giga-factory for the production of lithium-ion batteries for electric vehicles.'", "estimated_relevance": "High", "confidence_score": 0.98 }}
]
Input: Annual Report: {annual_report} Begin your analysis now.

10.10.2 Binary Prediction Prompt

Prompt designed for the binary classification of EU Taxonomy activities (Experiment 2).

Task: Your task is to act as a precision-focused Business Analyst. Your goal is to determine if a company performs a single, specific economic activity based on the provided company information and, when available, its annual report.
Core Directive: Your analysis must be based **strictly** on the evidence within the provided texts. You must make a definitive binary decision ('true' or 'false'). If there is no clear evidence that the company performs the activity, you must conclude 'false'.
Context: This classification is crucial for accurately categorizing a company's operations for regulatory, investment, and supply-chain analysis. The decision must be precise and directly supported by the provided information.
–
Inputs:
* **Company Information:** A JSON object or text describing the company's business, products, and services. * {company_info}
* **Activity to Classify:** The specific activity and its official description

that you need to check for. * **Activity Name:** {activity_name} * **Activity Description:** {activity_description} * **Activity Sector:** {activity_sector} * **Annual Report (Optional):** The text of the company's annual report. This input is optional and may not always be provided. It is provided at the end of this prompt.
—
Instructions for Analysis and Decision:
1. **Understand the Target Activity:** First, carefully analyze the provided ‘activity_name‘ and ‘activity_description‘. You must have a precise understanding of what the activity entails before searching for it.
2. **Scrutinize Company Data:** Thoroughly examine the ‘company_info‘ and the ‘annual_report‘ (if available). Search for explicit mentions or unambiguous descriptions of operations, products, services, projects, or revenue streams that directly match the ‘activity_description‘.
3. **Make a Binary Decision:** Based on your findings, decide if the company performs this activity. * **‘true‘**: Only if you find clear, direct evidence. * **‘false‘**: If there is no supporting evidence, if the evidence is ambiguous, or if it only relates to a minor or indirect part of the business (e.g., a supplier's activity).
4. **Determine Confidence:** Assign a confidence score between 0.0 and 1.0. This score reflects your certainty in the ‘applies‘ decision based on the quality of the evidence. * A direct quote mentioning the activity warrants a high score (e.g., 0.95-1.0). * A decision based on very strong inference from a product description would have a slightly lower score (e.g., 0.8-0.9). * A ‘false‘ decision due to a complete lack of evidence should have very high confidence (e.g., 0.99).
5. **Generate Structured Output:** Format your response as a **single JSON object**. Do not include any other text, explanation, or justification in your response. The JSON object must contain only two keys: ‘applies‘ (boolean), ‘confidence_score‘ (float), and ‘justification‘ (string).
—
Example Scenario:
* **Company Info:** ‘{{"name": "Global Auto Corp", "description": "Global Auto Corp is a leading manufacturer of passenger cars, including a new line of fully electric vehicles (EVs)."}}‘ * **Activity Name:** ‘"Manufacture of low carbon technologies for transport"‘ * **Activity Description:** ‘"Manufacturing of transport equipment with zero or low tail-pipe CO2 emissions, such as electric vehicles."‘
Correct Output for Example:
“‘json {{ "applies": true, "confidence_score": 0.99, "justification": "The company explicitly states it manufactures fully electric vehicles, which directly aligns with the activity description." }} “‘ * **Annual Report (Optional):** The text of the company's annual report. This input is optional and may not always be provided. If it is provided, it should be used as the primary source of evidence. * {annual_report}
[Begin the Analysis based on the inputs]

10.10.3 KPI Prediction Prompt

Prompt designed for prediction of EU Taxonomy KPIs (Experiment 3).

```
You are an expert financial analyst with deep knowledge of the EU Taxonomy framework.
The Challenge The provided annual report contains no direct mentions of the EU Taxonomy or its specific criteria. Your task is to analyze the company's general operations and financials and then use your internal, expert knowledge of the EU Taxonomy to predict the company's eligibility KPIs.
Your Reasoning Process (Internal Monologue) 1. First, read the report to form a high-level understanding of the company's business model: What does it produce? What services does it offer? Where do its revenues and capital expenditures come from?
2. Next, compare this understanding of the company's activities against your internal knowledge of the broad economic sectors covered by the EU Taxonomy (e.g., energy, manufacturing, transport, buildings, etc.).
3. Based on this comparison, estimate the percentage of the company's Turnover, CapEx, and OpEx that corresponds to these potentially eligible sectors.
Your Final Task Provide your final quantitative prediction in the specified JSON format. Do not include your reasoning in the output; provide only the clean JSON object.
Context:
"company_sector": {company_industry}, "financials": {{ "currency": "EUR", "units": {units}, "total_turnover": {total_revenue}, "total_capex": {total_capex}, "total_opex": {total_opex} }}
The annual report:
{annual_report}
Required Output JSON Format:
{{ "predictedTurnoverKPI": {{ "eligiblePercentage": "number", "eligibleValue": "number", "alignedPercentage": "number", "alignedValue": "number" }}, "predictedCapexKPI": {{ "eligiblePercentage": "number", "eligibleValue": "number", "alignedPercentage": "number", "alignedValue": "number" }}, "predictedOpexKPI": {{ "eligiblePercentage": "number", "eligibleValue": "number", "alignedPercentage": "number", "alignedValue": "number" }} }}
```

10.10.4 Summarization Prompt

Prompt designed for summarization of the annual report.

```
**Role:** You are an expert financial and sustainability analyst. Your specialization is in dissecting corporate annual reports to perform a detailed operational analysis and understand the core business of a company.
**Context:** You will be provided with the full text of a company's annual report. This report may not have a specific, pre-formatted section detailing all its economic activities in a standardized way.
**Primary Goal:** Your task is to perform a deep analysis of the annual report to produce an extensive and detailed summary. The summary must be structured into two main components: a comprehensive company profile and a detailed breakdown of the company's economic activities, which you will infer and construct from the report's content.
**Detailed Instructions:**
Carefully read and analyze the entire provided annual report. Pay close
```

attention to sections such as "Business Description," "Strategy," "Management's Discussion and Analysis (MD&A)," "Description of Operations," "Products and Services," "Segment Information," and any sustainability or ESG sections.
Based on your analysis, generate the output in two distinct parts as follows:
Part 1: Comprehensive Company Profile
Synthesize information from across the report to build a detailed company profile. This profile should be extensive and cover the following points in a clear, narrative format:
1. **Company Overview:** State the company's full name, headquarters location, and the primary stock exchange it is listed on (if mentioned). 2. **Mission and Vision:** Summarize the company's stated mission, vision, or core purpose. 3. **Core Business & Industry:** Describe the main industry the company operates in. What is its fundamental business model? How does it create value? 4. **Products & Services:** Detail the key products, services, or solutions the company offers. Group them logically if the company has multiple lines of business. 5. **Operational Scale & Geographic Footprint:** Describe the scale of the company's operations. Include details on key figures like revenue, number of employees, number of countries of operation, and key markets. 6. **Major Business Segments:** If the company reports in segments, describe each segment's function, its contribution to the overall business (e.g., percentage of revenue), and the specific activities it undertakes. 7. **Strategic Direction & Outlook:** Summarize the company's forward-looking strategy, including stated goals for growth, innovation, market expansion, or strategic shifts as mentioned in the report.
Part 2: Identification and Description of Economic Activities
This is the most critical part of your task. Based on your holistic understanding from Part 1, you will identify, describe, and categorize the distinct economic activities the company engages in.
An "economic activity" is a process that uses a combination of resources (like labor, capital, goods, and services) to produce a specific set of products or services. A company may have one primary activity or several distinct ones, often corresponding to its business segments or major product lines.
For *each* distinct economic activity you identify, create a dedicated subsection with the following structure:
1. **Economic Activity Title:** Create a concise, descriptive title for the activity. This title should be derived from the language and context of the report. Be as specific as the report allows (e.g., instead of a generic "Manufacturing," identify "Manufacturing of Consumer Electronics" or "Manufacturing of Heavy Industrial Machinery"). 2. **Detailed Description of the Activity:** Provide a comprehensive paragraph describing what this economic activity entails for the company. Explain the process, its purpose within the company's value chain, and its key characteristics. 3. **Associated Business Segments:** List the official business segment(s) from the report that this activity falls under. 4. **Key Products/Services Generated:** List the specific products or services that are the output of this economic activity. 5. **Evidence-Based Justification:** This is essential. Provide a detailed justification for identifying this activity by quoting or directly paraphrasing specific sentences or passages from the annual report. Cite the section or page number if available. Explicitly connect the evidence from the report to your description of the activity.

```
If the company has only one primary economic activity, create a single, highly
detailed section for it.  If it has multiple, create a separate, structured
section for each one.
**Output Format and Tone:**
* The final output should be extensive, well-organized, and professional. * Use
Markdown for clear headings (‘#‘, ‘##‘, ‘###‘) and bullet points to structure
your response. * Maintain an objective, analytical tone throughout the document.
* Do not invent information.  Your entire analysis must be grounded in the
provided annual report. * A long and detailed response is highly encouraged to
ensure a thorough summary.
—
**[START OF INPUT]**
**Annual Report:** “‘ report “‘
**[END OF INPUT]**
```