Center for Risk Management Lausanne

E4S White Paper

No Surveys Required: Small Language Models Enable Scalable ESG Extraction from Annual Reports



PUBLISHED

April 2026

AUTHORS

Fabio Alessandrini

CRML, University of Lausanne · E4S Center · Banque Cantonale Vaudoise

Nathan Delacrétaz

CRML, University of Lausanne · Quanthome

Eric Jondeau

CRML, University of Lausanne · E4S Center · Swiss Finance Institute · CEPR

SUGGESTED CITATION

Center for Risk Management – Lausanne (2026). “No Surveys Required: Small Language Models Enable Scalable ESG Extraction from Annual Reports.”

All CRML real estate research is available at https://crml.ch/real-estate.

ACKNOWLEDGEMENTS

This report was commissioned by the Banque Cantonale Vaudoise and the University of Lausanne to evaluate the sustainability of the Swiss real estate market. It is part of a broader collaboration between the two institutions aimed at advancing research on the impact of ESG integration in institutional investment portfolios. We thank Rean Morinaj for valuable research assistance in building the Swiss real estate ESG database.

Enterprise for Society (E4S) is a joint venture of the University of Lausanne through its Faculty of Business and Economics (UNIL – HEC), the Institute for Management Development (IMD), and the Ecole Polytechnique Fédérale de Lausanne (EPFL) under the stewardship of its College of Management of Technology with the mission of spearheading the transition towards a more resilient, sustainable and inclusive economy.

Generative AI tools (Claude from Anthropic and GPT-5 from OpenAI) were used to assist with literature review, data analysis code development, and manuscript preparation. The authors have reviewed and edited all AI-generated content and take full responsibility for the work's content and accuracy.

Executive Summary

Climate disclosure mandates are accelerating, but ESG teams remain small. A typical team of two or three analysts cannot simultaneously cover all portfolio holdings, monitor regulatory compliance, and conduct deep analysis. Current survey-based frameworks only reach willing participants, leaving non-participating companies, recent acquisitions, and competitors unassessed.

We show that AI can extract ESG information directly from published annual reports, achieving 80% accuracy on a 36-criterion framework without requiring any input from the reporting entity. Tested on 30 Swiss real estate annual reports in three languages, compact AI models running on a standard office computer perform within one percentage point of expensive cloud services, while keeping all portfolio data on-premise.

The central finding is not which AI system scores highest, but that different configurations suit different workflows. Some minimize false alarms, making them ideal for portfolio screening where analyst time is scarce. Others prioritize comprehensive detection, better suited for regulatory compliance where missing a disclosure carries legal risk. Matching the right tool to the workflow matters more than chasing the highest accuracy number.

Off-the-shelf AI models deliver production-grade accuracy without customization or training investment. Report quality drives performance variation more than the choice of AI system, suggesting a tiered review approach: automatic acceptance for straightforward reports, targeted verification for flagged cases, and full manual review for high-stakes or ambiguous documents.

With the right tools, unstructured data can increasingly feed automated workflows, dramatically reducing the repetitive workload that currently consumes analyst capacity. This shift applies beyond ESG to any data-intensive sector built on unstructured documents, bringing qualitative analysis closer to the rigor and scalability of quantitative work.

Résumé Exécutif

Les obligations de reporting climatique se multiplient, mais les équipes ESG restent de taille modeste. Deux ou trois analystes ne peuvent pas à la fois couvrir l'ensemble d'un portefeuille, veiller à la conformité réglementaire et conduire des analyses de fond. Les dispositifs actuels reposent sur des questionnaires auxquels seules les entités volontaires répondent, excluant de fait les non-participants, les acquisitions récentes et les concurrents.

Nous montrons que l'IA peut extraire des informations ESG directement depuis les rapports annuels publiés, avec une précision de 80% sur 36 critères, sans solliciter l'entité évaluée. Testée sur 30 rapports immobiliers suisses en trois langues, une IA compacte tournant sur un ordinateur de bureau standard rivalise à moins d'un point de pourcentage avec des services cloud bien plus coûteux, tout en gardant les données du portefeuille en local.

L'enseignement principal ne porte pas sur le classement des systèmes, mais sur le fait que chaque configuration répond à un besoin distinct. Certaines limitent les fausses alertes, ce qui convient au filtrage de portefeuille. D'autres maximisent la détection, ce qui répond aux exigences de conformité réglementaire où toute omission expose à un risque juridique. Choisir le bon outil selon l'usage compte davantage que de viser le score le plus élevé.

Des modèles prêts à l'emploi atteignent une précision suffisante pour la production, sans adaptation ni coût de formation. La qualité du rapport analysé pèse plus que le choix du système, ce qui plaide pour une revue à plusieurs niveaux : validation automatique des rapports clairs, vérification ciblée des cas signalés, et examen manuel des documents ambigus.

Avec les bons outils, les données non structurées s'intègrent de plus en plus dans des chaînes de traitement automatisées, allégeant considérablement le travail répétitif des analystes. Cette dynamique dépasse l'ESG et concerne tout secteur à forte intensité documentaire, rapprochant l'analyse qualitative de la rigueur du travail quantitatif.

Zusammenfassung

Die Vorgaben zur Klimaberichterstattung werden verschärft, doch die ESG-Teams bleiben klein. Zwei bis drei Analysten können nicht gleichzeitig das gesamte Portfolio abdecken, regulatorische Vorgaben einhalten und vertiefte Analysen erstellen. Heutige Erhebungsverfahren erreichen nur teilnehmende Unternehmen.

Wir zeigen, dass KI ESG-Informationen direkt aus Geschäftsberichten gewinnen kann, mit 80% Trefferquote über 36 Nachhaltigkeitskriterien, ohne Mitwirkung der bewerteten Organisation. Getestet an 30 Schweizer Immobilienberichten in drei Sprachen liefern kompakte Modelle auf einem Bürorechner Ergebnisse, die weniger als einen Prozentpunkt von teuren Cloud-Diensten abweichen. Alle Portfoliodaten verbleiben lokal.

Die wesentliche Erkenntnis betrifft nicht die Rangfolge der Systeme, sondern dass verschiedene Konfigurationen unterschiedliche Einsatzzwecke bedienen. Manche begrenzen Fehlalarme, ideal für das Portfolio-Screening. Andere setzen auf vollständige Erfassung, entscheidend bei regulatorischen Prüfungen, wo eine übersehene Offenlegung rechtliche Folgen haben kann. Das passende Werkzeug für den Einsatz zu wählen ist wichtiger als den höchsten Wert anzustreben.

Frei verfügbare Modelle liefern praxistaugliche Genauigkeit ohne Anpassung oder Schulungsaufwand. Die Qualität der Berichte beeinflusst das Ergebnis stärker als die Wahl des Systems, was eine gestufte Prüfung nahelegt: automatische Freigabe bei klaren Berichten, gezielte Kontrolle bei auffälligen Fällen und manuelle Prüfung bei mehrdeutigen Dokumenten.

Mit den richtigen Werkzeugen lassen sich unstrukturierte Daten in automatisierte Abläufe einbinden und die wiederkehrende Arbeitslast der Analysten deutlich verringern. Diese Entwicklung betrifft jeden dokumentenintensiven Sektor und rückt qualitative Auswertungen näher an die Stringenz quantitativer Arbeit.

Contents

1 Introduction 7

2 Literature Review 11

2.1 Why AI Evaluation Is Harder Than It Looks 11

2.2 ESG Classification Presents Unique Challenges 12

2.3 Methodological Solutions We Build Upon 13

2.4 Can We Customize AI Models for ESG? 13

2.5 What This Paper Contributes 14

3 Methodology 15

3.1 Model Selection and Configuration 15

3.2 Dataset Collection and Characteristics 17

3.3 Fine-Tuning Approach 18

4 Experimental Design 18

5 Results 19

5.1 The Scaling Solution: Near-Frontier Performance at Fraction of Cost 20

5.2 Architectural Specialization Enables Workflow Matching 21

5.3 Synthetic Data Quality and Fine-Tuning Outcomes 22

5.4 Ensemble Voting Shows Minimal Gains 23

5.5 When Frontier Models Prove Necessary 25

5.6 GPT-5 Does Not Outperform GPT-5-nano 26

6 Discussion, Limitations, and Conclusions 26

6.1 Local Models Approach Cloud Accuracy with Recall Trade-offs 26

6.2 Classification Ceiling and Inherent Ambiguity 27

6.3 Model Selection Framework 28

6.4 Fine-Tuning Limitations 29

6.5 Document-Level Performance Variation 29

6.6 Limitations 30

6.7 Conclusions 31

6.8 Future Directions 31

List of Figures

List of Tables

1 Frontier vs. Best Open-Source Model Performance (n=1,080) 20
2 Zero-Shot Base Model Performance on Swiss ESG Classification (n=1,080) 21
3 Impact of Synthetic Data Fine-Tuning on 8B Parameter Models (n=1,080) 23
4 Ensemble Performance vs Individual Models 24
A1 Pairwise Model Comparisons with Bonferroni Correction 36
A2 Model Performance by Document Language 37
A3 Model Performance Summary with 95% Bootstrap Confidence Intervals 39

1 Introduction

What if ESG performance could be assessed across an entire portfolio without sending a single survey—especially for questions that are not consistently covered by public disclosures? Companies already disclose sustainability information in annual reports, and such public data have long been used in ESG analysis, but key elements can remain unavailable or difficult to capture systematically. Survey-based frameworks like GRESB require portfolio companies to complete lengthy questionnaires, creating coordination burdens and limiting coverage to willing participants.[1] This paper demonstrates that AI can now extract equivalent information directly from published documents, achieving 80% accuracy on a 36-criterion ESG framework without requiring any input from reporting entities.

We tested this survey-free approach on 30 Swiss real estate annual reports.[2] The results establish a practical alternative to active data collection. Compact AI models that run on standard office computers[3] match the accuracy of expensive commercial cloud services while keeping our data private and eliminating survey coordination entirely. However, different AI models have distinct strengths. Some excel at minimizing false alarms while others prioritize catching every relevant disclosure. Matching the right model to our specific workflow matters more than chasing the highest accuracy number.

Climate disclosure mandates are accelerating. SEC requirements, EU taxonomy regulations, and institutional investor ESG integration demands require systematic analysis of corporate disclosures at scales that manual review cannot sustain. A typical ESG team of two or three analysts faces an impossible choice between comprehensive coverage across all portfolio holdings and deep analysis of se-

[1] The CRML conducted a survey-based data collection exercise over two years (2021–22).

[2] These criteria span environmental metrics (energy efficiency, emissions), social factors (tenant engagement, community impact), and governance practices, representative of typical ESG screening frameworks.

[3] Technically, these are 8-billion-parameter language models running locally on consumer hardware such as Apple M-series chips or NVIDIA graphics cards. For comparison, commercial cloud AI systems typically use models 10–100× larger.

lected positions. Quarterly monitoring, compliance verification, and preliminary screening all compete for limited analyst time.

Traditional approaches each carry significant drawbacks. Survey-based methodologies like GRESB (Bauer et al., 2011) provide standardized benchmarks but impose substantial completion burdens on reporting entities. Coverage is limited to companies willing and able to participate, often excluding smaller holdings, recent acquisitions, or non-cooperative portfolio companies. Manual analyst extraction from public disclosures ensures quality but scales linearly with document volume, requiring proportional headcount increases for comprehensive coverage. Commercial AI services[4] promise automation but require uploading sensitive documents to third-party servers, creating privacy concerns for proprietary frameworks and client-confidential holdings.

The survey-free alternative we propose addresses all three limitations. By extracting ESG information from documents companies already publish, we eliminate response burden and expand coverage to any entity with public disclosures. Local AI deployment preserves privacy while matching commercial accuracy. The approach generalizes beyond ESG to private equity due diligence, credit analysis, and regulatory compliance monitoring, all of which face similar needs to extract structured data from unstructured documents.

We evaluated whether AI can reliably extract ESG information that would otherwise require surveys or manual analyst review. Swiss real estate annual reports served as a demanding test case. They require understanding of multiple languages (German, French, English), technical terminology (GRESB scores, Minergie certifications, CO_2 metrics), and the ability to distinguish substantive sustainability initiatives from boilerplate regulatory language. If AI extraction works reliably in this challenging context, it should generalize well to simpler document types. We compared eight AI configurations[5] across the full 36-criterion ESG framework,

[4] Such as those offered by major cloud providers (OpenAI, Anthropic, Google) through pay-per-use API access.

[5] Three base model architectures (Ministral-8B, Llama-3.1-8B, Apertus-8B), each tested with and without domain-specific training, plus two commercial reference systems (GPT-5-nano and GPT-5).

measuring not just overall accuracy but also the specific types of errors each model makes.

Four practical findings emerged from our evaluation. First, survey-free extraction achieves production-grade accuracy. The best compact model (Ministral-8B) reaches 80.00% accuracy on our 36-criterion ESG framework, within one percentage point of commercial cloud services (GPT-5-nano at 80.65%), using only published annual reports.[6] No surveys, no questionnaires, no coordination with portfolio companies. Local deployment eliminates data exposure risks while dramatically reducing per-document costs.

Second, different models suit different workflows. Rather than a single best model, we found systematic specialization. Ministral excels at precision: when it flags something as positive, it is usually correct (74% precision versus 60% for commercial systems). This matters for screening workflows where false positives waste analyst time. Commercial cloud models excel at recall, catching more genuine positives even at the cost of more false alarms. This matters for compliance workflows where missing a disclosure creates regulatory risk. Llama provides balanced performance suitable for general monitoring.[7]

Third, domain-specific training produces mixed results. We attempted to improve model performance by training on synthetic ESG documents generated from authentic report patterns.[8] Results were disappointing. Most models became worse at catching genuine positives. Only Ministral achieved meaningful precision gains (+8 percentage points), suitable for precision-critical screening. Recent research suggests that better training data generation methods may yield improved results, but the additional complexity may not justify the effort for many practitioners.

[6] We tested on Apple M-series MacBooks and NVIDIA RTX 4090 workstations, both representative of equipment already available in many financial services IT environments.

[7] Technically, these differences reflect architectural design choices in how each model processes text and generates predictions. Appendix D provides details for readers interested in the underlying mechanisms.

[8] This fine-tuning process adapts a general-purpose AI model to perform better on specific tasks by exposing it to relevant examples. Our approach extracted style patterns and terminology from real reports to generate training data.

Fourth, error costs should drive model selection. The right model depends on the workflow's error tolerance. For preliminary screening where analyst time is precious, the chosen models should minimize false positives. For regulatory compliance where missing a disclosure creates legal exposure, the models should minimize missed detections. Section 6 provides a decision framework mapping common ESG workflows to recommended model configurations.

For ESG teams evaluating AI adoption, these findings suggest a pragmatic path forward. Off-the-shelf compact AI achieves production-viable accuracy without expensive customization or training infrastructure. Model selection should depend on whether the priority is minimizing wasted review time or ensuring comprehensive coverage. For sensitive documents, local deployment eliminates data exposure risks while matching commercial accuracy. Commercial cloud services justify their costs when comprehensive detection is critical and document sensitivity permits cloud processing.

Methodological constraints shape interpretation. Our 30-document sample provides comprehensive ESG category coverage but limits statistical power for detecting small performance differences.[9] Single-annotator classification, due to resource constraints, introduces potential bias compared to multi-rater validation. The Swiss real estate focus ensures conservative performance estimates but narrows immediate generalizability to other sectors and geographies. Despite these constraints, consistent patterns across model architectures establish reliability of core findings.

The paper proceeds as follows. Section 2 reviews relevant research on AI evaluation and ESG classification. Section 3 describes our dataset, model configurations, and evaluation protocols.[10] Section 5 presents performance comparisons across models and training configurations. Section 6 develops decision frameworks for workflow-matched model selection. Section 6 discusses broader implications and future directions.

[9] With 30 documents $\times$ 36 criteria = 1,080 individual classifications per model, we have adequate power for the large effects we report but may miss subtle differences.

[10] Readers primarily interested in practical deployment guidance may wish to skim Section 3 and proceed directly to results.

2 Literature Review

ESG data collection has historically relied on two approaches: surveys that ask companies to self-report, or manual analyst extraction from public disclosures. Survey-based frameworks like GRESB have become industry standards, but they impose completion burdens on reporting entities and limit coverage to willing participants. The emergence of capable AI models creates a third option: automated extraction from documents companies already publish. Before deploying this survey-free approach, practitioners need to understand what can go wrong and how to address it. This section reviews the reliability challenges that affect AI-based document analysis, specific issues in financial applications, and the methodological solutions we build upon.

2.1 Why AI Evaluation Is Harder Than It Looks

AI models used for document classification face reliability challenges that practitioners must understand before trusting automated decisions. Research shows these systems exhibit predictable biases: they rate familiar content higher than specialized material (Stureborg et al., 2024), and their assessments can be manipulated simply by changing the order in which documents are presented (Wang et al., 2023).[11]

A critical finding for investment applications: running an AI evaluation once produces misleadingly confident results. Schroeder and Wood-Doughty (2024) demonstrate that single-run assessments systematically underestimate uncertainty, leading to performance claims that fail when independently tested. For ESG analysts relying on AI classifications to inform portfolio decisions, this means apparent accuracy may not reflect real-world reliability.

[11] This "position bias" means that identical content may receive different scores depending on where it appears in a batch.

Perhaps most concerning, AI models sometimes fabricate supporting evidence.[12] The model may correctly classify a document but cite a paragraph that was never written, creating audit trail problems where regulators expect verifiable evidence (Ohi et al., 2024). Additional research confirms that presentation order and framing systematically influence model judgments (Koo et al., 2023), compounding reliability concerns for high-stakes applications.

2.2 ESG Classification Presents Unique Challenges

General-purpose AI models struggle with ESG-specific terminology and regulatory frameworks (Birti et al., 2025). Performance varies significantly across Environmental, Social, and Governance categories. A model that accurately identifies environmental disclosures may miss governance issues entirely (Zou et al., 2023).[13]

Three characteristics of ESG reports amplify AI reliability problems. First, relevant information is dispersed: a company's climate commitments may appear in the CEO letter, sustainability section, and financial notes, requiring synthesis across sections. Second, terminology varies widely, as different reporting frameworks (GRI, SASB, TCFD) may use different language for similar concepts. Third, companies describe identical activities in substantially different ways (Bronzini et al., 2023), making pattern recognition unreliable.

For portfolio managers, these challenges translate into concrete risks. Systematic misclassification of ESG factors can propagate through investment screening, affecting capital allocation at scale (Chen et al., 2025). Unlike general text analysis where errors are inconvenient, ESG classification errors can trigger regulatory scrutiny or reputational damage. The stakes demand evaluation methods designed for financial applications, not adapted from general benchmarks.

[12] Technically called "hallucination." The model generates plausible-sounding justifications that do not exist in the source document.

[13] Analysis of 166 companies showed systematic performance gaps between E, S, and G dimensions.

2.3 Methodological Solutions We Build Upon

Researchers have developed practical solutions to these reliability challenges. The most important insight: run evaluations multiple times and measure consistency. Schroeder and Wood-Doughty (2024) and Ye et al. (2024) demonstrate that reliable performance assessment requires analyzing variance across trials, not trusting single runs. We adopt this approach throughout our experiments.[14]

How questions are phrased to the AI[15] significantly affects reliability. Research shows that well-designed prompts reduce output variance, but improvements must be validated for each domain (Sahoo et al., 2024). Generic optimizations do not transfer reliably to ESG applications. Structured workflows that enforce documentation and audit trails improve both reproducibility and accuracy (Zeng et al., 2025).

Citation-based quality controls address the evidence fabrication problem. Requiring AI models to quote specific passages before making classifications improves alignment with human judgments and creates verifiable audit trails (Wang et al., 2023). We implement this approach through citation validation that checks whether quoted text actually exists in source documents.[16]

2.4 Can We Customize AI Models for ESG?

We tested whether customizing AI models specifically for ESG terminology improves classification results. Modern techniques allow adapting models with minimal computational cost, modifying less than 0.1% of the model while retaining most capabilities.[17]

[14] All reported results include statistical confidence intervals from multiple independent runs with identical settings.

[15] Technically called "prompt engineering," the design and optimization of instructions given to language models.

[16] Technical details of our citation validation system appear in Section 3.

[17] Technical details on Low-Rank Adaptation (LoRA) and related methods appear in Appendix 6.8.

The short answer: not much, with current approaches. Our experiments (Section 5) show that fine-tuning on synthetic ESG examples actually degraded performance across all tested models. The likely culprit is data quality: research suggests that customization success depends on access to large quantities of authentic, labeled documents (Maini et al., 2025), which remain scarce for ESG applications. Generating synthetic training data risks creating artifacts that models memorize rather than learning transferable patterns.

2.5 What This Paper Contributes

This paper provides systematic evidence for practitioners evaluating AI deployment options for ESG classification. We address the cost-performance trade-offs that Li et al. (2023) identify as critical for financial institutions.

For methodology, we contribute: (1) metrics that separate "the model got it wrong" from "the system failed to respond," preventing misleading accuracy claims; (2) statistical protocols ensuring reported results are reproducible, not artifacts of lucky runs;[18] and (3) document handling that processes reports of varying lengths without losing context.

For deployment decisions, we provide evidence on: (1) how different small models specialize (some optimize for avoiding false positives, others for catching all relevant disclosures), enabling application-matched selection; (2) realistic expectations for model customization, documenting that synthetic training data degraded rather than improved performance; and (3) concrete cost-accuracy trade-offs showing that small models achieve near-frontier accuracy at substantially lower cost, while frontier models justify premium pricing for compliance applications requiring comprehensive coverage.[19]

[18] Following Schroeder and Wood-Doughty (2024), we use fixed settings, multiple runs, and appropriate statistical corrections.

[19] Detailed results appear in Section 5; ensemble voting analysis demonstrates that combining models provides modest gains but cannot overcome training data quality issues.

3 Methodology

We compare eight language model configurations on ESG document classification, including three compact open-source models,[20] their fine-tuned variants trained on synthetic ESG data, and two commercial models (GPT-5-nano, GPT-5). The 30 Swiss real estate annual reports generate 1,080 classification tasks (30 documents $\times$ 36 ESG categories), providing a comprehensive benchmark for ESG disclosure detection.

3.1 Model Selection and Configuration

We evaluate three compact open-source architectures representing different design priorities. Ministral-8B-Instruct-2410 (Mistral AI) employs memory-efficient processing optimized for long documents.[21] Llama-3.1-8B-Instruct (Meta) provides general-purpose baseline capabilities across diverse tasks. Apertus-8B-Instruct-2509 (Swiss AI) emphasizes European language coverage with enhanced German and French understanding, particularly relevant for Swiss multilingual contexts.

Notably, all three open-source models run on standard office hardware (a MacBook Pro or gaming PC with a mid-range graphics card) without requiring cloud subscriptions or data center access. This enables on-premise deployment where sensitive portfolio data remains within organizational boundaries.

Commercial benchmarks include GPT-5-nano-2025-08-07 as cost-optimized frontier baseline and GPT-5-2025-08-07 for full-capability comparison. Both models are accessed via OpenAI API with reproducible settings.[22]

[20] Each model contains approximately 8 billion parameters, the numerical weights that encode the model's learned patterns. This "8B" scale represents a practical sweet spot: small enough to run on consumer hardware, yet large enough to handle complex document understanding.

[21] Technically, Ministral uses interleaved sliding-window attention, which processes documents in overlapping segments to reduce memory requirements (Mistral AI, 2024).

[22] We use deterministic configuration (temperature=0.0) for GPT-5-nano, ensuring identical outputs for identical inputs. GPT-5 uses default settings as custom temperature parameters are not supported.

Fine-tuned variants train the three base models on synthetic ESG data using parameter-efficient adaptation techniques[23] on consumer hardware and High Performance Computing infrastructure.

We report five classification metrics. *Accuracy* measures overall correctness across all classifications. *Precision* captures the proportion of flagged ESG strategies that are actually present,[24] while *recall* measures the proportion of actual ESG strategies that the model detects.[25] *F1* score provides a balanced measure combining precision and recall.[26] The *Matthews Correlation Coefficient* (MCC) complements these metrics as a single summary measure of classification quality.[27] These metrics have distinct operational implications: precision-focused models reduce analyst workload by minimizing false positives, while recall-focused models ensure comprehensive coverage by minimizing missed strategies. The optimal metric emphasis depends on whether false positives or false negatives carry greater cost in a given application context.

All experiments use fixed random seeds and statistical corrections for multiple comparisons,[28] ensuring reproducible results. Complete experimental protocols appear in Appendix A and B.

[23] LoRA (Low-Rank Adaptation) modifies less than 0.1% of model parameters, enabling fine-tuning on consumer hardware rather than requiring expensive cloud GPU clusters.

[24] Precision = true positives / (true positives + false positives). High precision means fewer false alarms requiring analyst review.

[25] Recall = true positives / (true positives + false negatives). High recall means fewer missed strategies.

[26] F1 is the harmonic mean of precision and recall, penalizing models that sacrifice one metric for the other.

[27] $\text{MCC} = (\text{TP} \times \text{TN} - \text{FP} \times \text{FN}) \,/\, \sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}$, ranging from -1 (perfect inverse prediction) through 0 (no better than random or naive strategies) to $+1$ (perfect prediction). Unlike F1, which considers only true positives, false positives, and false negatives, MCC incorporates all four confusion matrix cells, making it particularly robust under class imbalance (Chicco and Jurman, 2020; Boughorbel et al., 2017).

[28] We apply Bonferroni correction to paired t-tests and use one-way ANOVA for multi-model comparisons, ensuring reported differences are statistically reliable rather than random variation.

3.2 Dataset Collection and Characteristics

We manually curate 30 annual reports from Swiss Real Estate Investment Vehicles spanning 2022-2024, using stratified sampling by legal form (fund, non-listed fund, company, foundation) and language (German, French, English in equal proportions) to ensure representative coverage of Swiss REIV diversity. A trained annotator classified each report for ESG category presence using standardized coding guidelines.

The dataset reflects real-world complexity. Documents average 8,200 words (range: 3,500–15,000), comparable to typical annual reports. The 27.5% positive / 72.5% negative label distribution reflects authentic ESG disclosure patterns, as most categories are absent from most reports. This imbalance creates a meaningful baseline: a model that blindly predicts "no ESG strategy present" for everything would achieve 72.5% accuracy without detecting anything useful.[29] Given this class imbalance, accuracy alone can be misleading. We therefore emphasize F1 score as the primary comparison metric, as it penalizes models that sacrifice recall for artificially high precision or vice versa. MCC serves as complementary validation: any constant prediction strategy—whether all-negative or all-positive—scores exactly zero by construction, providing a direct numerical test of whether a model demonstrates genuine discriminative capability beyond class distribution exploitation. For practitioners, the precision-recall trade-off remains context-dependent: regulatory compliance favors recall (missing a disclosure carries legal risk), while routine screening favors precision (minimizing analyst review burden).

Technical terminology includes GRESB scores, Minergie certification standards, and CO_2 intensity metrics embedded within multilingual corporate language, the same domain-specific vocabulary that makes manual ESG assessment time-intensive. The 30 reports generate 1,080 classification tasks (30 documents $\times$ 36 ESG categories), enabling comprehensive performance comparisons across the full spectrum of ESG disclosure patterns in Swiss real estate.

[29] This "naive baseline" establishes that models must substantially exceed 72.5% accuracy to demonstrate genuine classification capability rather than simply exploiting class imbalance.

3.3 Fine-Tuning Approach

We generate synthetic ESG training data using GPT-4o-2024-08-06 with style patterns extracted from authentic Swiss reports. Generation variables include three languages, multiple writing styles, difficulty levels, and document lengths (150–450 words).

Fine-tuning uses parameter-efficient techniques that modify less than 0.1% of the model,[30] enabling training on consumer hardware rather than requiring cloud GPU clusters. This substantially reduces infrastructure requirements compared to full model training.

Synthetic data undergoes 80/20 train/validation splits stratified by ESG category. All procedures use fixed seeds for reproducibility. Complete synthetic data generation protocols and platform-specific configurations appear in Appendix B.

4 Experimental Design

We evaluate models through two experiments: zero-shot baseline assessment and fine-tuning with synthetic data.

Baseline evaluation provides identical prompts to all models with balanced examples and structured JSON output requirements. Temperature fixed at 0.0 ensures deterministic outputs. Each model classifies all 1,080 tasks (30 documents $\times$ 36 ESG categories) without domain training.

Fine-tuning experiments train the three base models on synthetic ESG data generated by GPT-4o-2024-08-06. Training employs LoRA (rank=8, α=20) with AdamW optimizer (learning rate=5e-5) across 5 epochs. Data undergoes 80/20 train/validation splits stratified by ESG category. Fixed seeds ensure reproducibility.

[30] Specifically, LoRA (Low-Rank Adaptation) with rank=8 and scaling factor α=20. Training uses AdamW optimizer with learning rate 5e-5 and weight decay 0.01 across 5 epochs with early stopping.

Binary classification (present/absent) applies to 36 ESG categories spanning environmental, social, and governance dimensions across all documents.

Document-level analysis computes precision, recall, and F1 for each document independently, identifying three failure patterns: zero recall (when ground truth contains no positive labels), negative prediction bias (>80% negative predictions), and precision-recall imbalance ($|P_d - R_d| > 0.3$).

Complete experimental protocols, failure pattern definitions, and quality control procedures appear in Appendix C.

5 Results

Each evaluation task presents a model with the full text of a Swiss real estate annual report and a binary classification prompt for one of 36 ESG categories. For instance, the *green certifications* prompt instructs the model: "Analyze the document and determine if it discusses green certifications," defined as "official sustainability certifications awarded to buildings, specifically Swiss labels such as Minergie, SNBS, LEED, DGNB, or BREEAM." The model returns a yes/no classification and a supporting text citation. This process is repeated for all 36 categories across 30 reports, yielding 1,080 classification tasks.

Many ESG strategies are not stated explicitly but embedded in broader operational descriptions. For instance, the ground truth labels a passage about "involvement of local businesses during construction creating local value and regional anchoring" as evidence for the *circular economy* category, and a description of "real-time monitoring of energy consumption through technological tools" as evidence for *smart building use*—neither passage mentions its category by name. Both were correctly classified by all five evaluated models, confirming that the approach captures conceptual relevance beyond surface-level keyword matching.

Compact open-source models resolve the ESG assessment scaling challenge. However, no model achieves both high precision and high recall simultaneously,

creating workflow-specific deployment niches rather than universal "best model" recommendations.

These findings demonstrate that frontier-level accuracy is achievable using freely available, locally-deployed models. The optimal model choice depends on whether an application prioritizes comprehensive detection (recall) or minimizing false positives (precision).

5.1 The Scaling Solution: Near-Frontier Performance at Fraction of Cost

Compact models achieve 63–80% accuracy on 1,080 ESG classification tasks across 30 Swiss real estate annual reports (Table 1). Ministral-8B (80.00%) approaches GPT-5-nano (80.65%) within 0.65 percentage points, enabling local deployment without sacrificing classification reliability.

Model	Accuracy	Precision	Recall	F1 Score	MCC
Ministral-8B (Base)	80.00%	73.68%	42.42%	53.85%	0.449
GPT-5-nano	80.65%	60.05%	88.55%	71.56%	0.602
GPT-5	80.65%	60.73%	83.84%	70.44%	0.582
Gap: GPT-5-nano vs. Ministral-8B					
Absolute Difference	+0.65pp	-13.63pp	+46.13pp	+17.71pp	+0.153
Relative Improvement	+0.8%	-18.5%	+108.8%	+32.9%	+34.1%

Table 1. Frontier vs. Best Open-Source Model Performance (n=1,080)

The minimal accuracy gap (0.65pp) establishes that compact open-source models can match commercial API performance for ESG classification. Local deployment on consumer hardware enables privacy-preserving analysis without third-party API dependencies, addressing data sensitivity concerns for organizations with proprietary ESG frameworks or confidential portfolio holdings.

However, accuracy alone does not determine optimal model selection. The following subsections detail how different models specialize in precision versus recall, enabling application-specific deployment decisions.

5.2 Architectural Specialization Enables Workflow Matching

Base models show substantial performance variation reflecting architectural specialization (Table 2). Ministral-8B leads in accuracy (80.00%) and precision (73.68%, meaning when it flags an ESG strategy, it is correct nearly three-quarters of the time). Llama-3.1-8B balances metrics with 77.96% accuracy and the highest F1 score (57.65%)[31] among open-source models. Apertus-8B maximizes recall (75.76%, catching over three-quarters of all ESG strategies) at precision cost (40.54%, yielding only 62.78% accuracy).[32]

Model	Accuracy	Precision	Recall	F1 Score	MCC
Ministral-8B	80.00%	73.68%	42.42%	53.85%	0.449
Llama-3.1-8B	77.96%	61.13%	54.55%	57.65%	0.429
Apertus-8B	62.78%	40.54%	75.76%	52.82%	0.300
Naive Baseline (All-Negative)	72.50%	—	0.00%	—	0.000

Table 2. Zero-Shot Base Model Performance on Swiss ESG Classification (n=1,080)

Ministral-8B and Llama-3.1-8B substantially exceed the naive baseline of 72.50% accuracy,[33] while Apertus-8B (62.78%) trades accuracy for the highest recall among compact models (75.76%). MCC scores confirm genuine classification capability across all three: the naive baseline scores exactly zero while all base models achieve positive MCC (0.30-0.45), demonstrating discriminative ability

[31] F1 is the harmonic mean of precision and recall.

[32] Apertus performs significantly worse than all other models after multiple comparison correction ($p < 0.001$); see Appendix A.

[33] A model predicting "no ESG strategy" for every classification would achieve 72.50% accuracy with 0% recall, detecting nothing while appearing moderately accurate due to class imbalance.

independent of class distribution. However, no compact model achieves both high precision ($> 70\%$) and high recall ($> 70\%$) simultaneously, a fundamental trade-off at this model scale.

These architectural specializations have distinct operational implications. Ministral-8B suits applications where false positives create costly review burdens, as its 73.68% precision minimizes unnecessary analyst effort; however, its 42.42% recall requires downstream verification for comprehensive coverage. Llama-3.1-8B provides balanced performance (57.65% F1) suitable for general-purpose classification where neither error type dominates costs. Apertus-8B suits recall-critical applications such as regulatory compliance screening, catching 75.76% of strategies, though its 40.54% precision generates substantial false positives requiring verification overhead. Model selection should therefore depend on whether false positives or false negatives carry greater cost in a given application context.

5.3 Synthetic Data Quality and Fine-Tuning Outcomes

Fine-tuning with synthetic ESG data reveals the importance of data quality over quantity. Our synthetic generation extracted style patterns and terminology from authentic Swiss reports but did not employ proper data augmentation strategies.[34] Results show systematic recall degradation across all three models (Ministral: -11.44pp, Llama: -10.11pp, Apertus: -57.91pp), with no model improving its F1 score or recall (Table 3).

Fine-tuning outcomes vary substantially by architecture. Ministral-8B fine-tuned shows the sole positive outcome: precision improves from 73.68% to 82.14% (+8.46pp), reaching the highest precision across all configurations, though recall drops to 30.98% (-11.44pp). This configuration suits applications where false positives carry high costs. Llama-3.1-8B fine-tuned shows consistent degradation without compensating gains (-1.66pp accuracy, -10.11pp recall, -6.88pp

[34] Data augmentation involves systematically varying authentic examples (rephrasing, paraphrasing, or restructuring real documents) rather than generating entirely new synthetic text.

Model	Accuracy	Precision	Recall	F1	MCC
Ministral-8B	80.00% (-0.83)	73.68% (+8.46)	42.42% (-11.44)	53.85% (-8.86)	0.449 (-0.03)
Llama-3.1-8B	77.96% (-1.66)	61.13% (-1.94)	54.55% (-10.11)	57.65% (-6.88)	0.429 (-0.07)
Apertus-8B	62.78% (+5.37)	40.54% (-5.90)	75.76% (-57.91)	52.82% (-29.26)	0.300 (-0.24)

Note: Base model values shown; parentheses indicate change after fine-tuning (percentage points for Accuracy, Precision, Recall, F1; absolute change for MCC).

Table 3. Impact of Synthetic Data Fine-Tuning on 8B Parameter Models (n=1,080)

F1), indicating fine-tuning provides no benefit for this architecture. Apertus-8B fine-tuned exhibits severe recall degradation despite accuracy improvements (+5.37pp accuracy); recall drops from 75.76% to 17.85% (-57.91pp), eliminating the model's recall advantage and suggesting overfitting to synthetic patterns that do not transfer to authentic reports.

These results suggest that base models provide immediate viability (77–80% accuracy) without training investment, representing efficient resource allocation for most applications. Organizations with extensive authentic ESG document collections might explore grounded augmentation approaches, though this requires significant data curation effort and rigorous validation.

5.4 Ensemble Voting Shows Minimal Gains

Given the systematic precision-recall trade-offs observed across individual models, we investigate whether combining multiple models through majority voting can achieve balanced performance.

Majority voting across three compact models (Ministral + Llama + Apertus) provides minimal improvement (Table 4). The base model ensemble achieves 57.97% F1, only 0.32pp above the best individual model (Llama-3.1-8B at 57.65%).

Running three models triples computational costs (processing time, hardware requirements, energy consumption) while delivering negligible performance improvement. The fine-tuned ensemble performs even worse: F1 drops to 41.95%

Model/Ensemble	Accuracy	Precision	Recall	F1	MCC
Individual Base Models					
Ministral-8B	80.00%	73.68%	42.42%	53.85%	0.449
Llama-3.1-8B	77.96%	61.13%	54.55%	57.65%	0.429
Apertus-8B	62.78%	40.54%	75.76%	52.82%	0.300
Base Ensemble	**78.52%**	**62.75%**	**53.87%**	**57.97%**	**0.439**
Δ from best individual				+0.32pp	
Individual Fine-tuned Models					
Ministral-8B-FT	79.17%	82.14%	30.98%	44.99%	0.416
Llama-3.1-8B-FT	76.30%	59.19%	44.44%	50.77%	0.362
Apertus-8B-FT	68.15%	34.64%	17.85%	23.56%	0.065
Fine-tuned Ensemble	**77.96%**	**76.11%**	**28.96%**	**41.95%**	**0.372**
Δ from base ensemble				-16.02pp	
Hybrid Ensemble (Base + Fine-tuned)					
Hybrid Ensemble	**78.33%**	**62.65%**	**52.53%**	**57.14%**	**0.431**
Δ from best individual				-0.51pp	
Frontier Reference					
GPT-5-nano	80.65%	60.05%	88.55%	71.56%	0.602
Δ from best ensemble				+13.59pp	

Table 4. Ensemble Performance vs Individual Models

(-16.02pp from base ensemble) because all models overfit to the same synthetic training patterns, amplifying rather than correcting systematic errors. MCC confirms this pattern: the base ensemble (0.44) remains well below GPT-5-nano (0.60), and fine-tuning further reduces ensemble MCC to 0.37. Base ensembles remain 13.59pp below frontier performance (GPT-5-nano: 71.56% F1) while requiring triple computational resources, suggesting that single-model deployment represents more efficient resource allocation than ensemble approaches for this task.

We also tested cross-tier ensembles combining the best local model (Ministral-8B) with GPT-5-nano. With only two models, majority voting does not apply; instead we evaluate OR-logic (predict positive if either model does) and AND-logic (predict positive only if both agree). The OR-logic ensemble slightly improves recall over GPT-5-nano alone (90.24% vs. 88.55%) but does not improve F1 (71.37% vs. 71.56%). The AND-logic variant yields high precision (78.06%) but low recall (40.74%, F1: 53.54%), essentially replicating Ministral's conservative profile (F1: 53.85%). Neither variant outperforms GPT-5-nano on F1, indicating that simple combination rules do not bridge the precision-recall trade-off between local and frontier models.

5.5 When Frontier Models Prove Necessary

GPT-5-nano's minimal accuracy advantage (0.65pp, not statistically significant) contrasts with a substantial recall advantage: 88.55% versus 42.42%, detecting more than twice as many genuine ESG strategies as the best compact model.

The recall advantage suggests contexts where frontier models may justify their higher costs. Regulatory compliance screening (SEC climate disclosure rules, EU taxonomy reporting) requires comprehensive ESG detection where missing material initiatives carries significant penalties. High-stakes investment decisions benefit from superior balanced accuracy (71.56% vs. 53.85%), reducing error-driven capital misallocations. More generally, recall-critical applications where missing ESG initiatives creates material financial or reputational risk may warrant frontier model deployment.

However, higher recall (88.55%) comes with lower precision (60.05%), increasing false positives. GPT-5-nano's overall error rate (19.35%) is comparable to Ministral's (20.00%).[35] The frontier model catches more genuine strategies but generates additional false positives requiring verification. This trade-off favors GPT-5-nano when missing strategies carries greater cost than reviewing false alarms, typically in compliance and audit contexts rather than routine screening.

5.6 GPT-5 Does Not Outperform GPT-5-nano

GPT-5 (the largest frontier model) achieves 80.65% accuracy and 70.44% F1. Despite matching GPT-5-nano on overall accuracy, the two models make different individual predictions (GPT-5: 249 true positives; GPT-5-nano: 263), with GPT-5 underperforming on F1 (70.44% vs. 71.56%) despite substantially higher API costs.

This counterintuitive result suggests capability saturation: Swiss ESG classification complexity does not require capabilities beyond GPT-5-nano's scale. Larger models provide no additional benefit for this specific task domain, indicating that premium pricing for the largest available model does not guarantee better classification performance. Task-specific benchmarking is therefore advisable before model selection, as general capability rankings may not transfer to domain-specific applications.

6 Discussion, Limitations, and Conclusions

6.1 Local Models Approach Cloud Accuracy with Recall Trade-offs

The central finding is that 8-billion parameter models achieve comparable accuracy to frontier cloud APIs (80.00% vs 80.65%), but with recall trade-off (42.42%

[35] Error rates computed as 100% minus accuracy from Table 1.

vs 88.55%). Local models match frontier accuracy while detecting less than half the positive cases. Three factors explain the accuracy convergence.

First, ESG classification represents a well-defined task. The binary decision of whether a document contains a substantive sustainability strategy across 36 ESG categories does not require the complex reasoning that differentiates expensive cloud models from smaller alternatives. Second, smaller models achieve higher precision for this task: Ministral-8B correctly identifies 74% of flagged strategies as genuine, compared to 60% for GPT-5-nano.[36] Third, all models have encountered ESG content during training. Swiss real estate disclosures use standard financial terminology, sustainability frameworks, and regulatory language that both 8-billion and frontier models have seen, reducing the capability gap for this domain.

6.2 Classification Ceiling and Inherent Ambiguity

The clustering of all models between 63% and 81% accuracy suggests inherent classification difficulty rather than purely technological limitations. Four judgment calls that challenge human analysts also challenge AI systems. (1) The distinction between aspiration and action ("We value sustainability" versus "We achieved 15% energy reduction through solar installation") requires subjective boundary-setting. (2) Implicit strategies, such as "comprehensive employee training programs" described without ESG framing, may or may not constitute social commitment depending on interpretation. (3) The treatment of future commitments versus completed actions remains ambiguous across reporting frameworks. (4) Finally, materiality thresholds vary: a single sentence mentioning environmental certification may represent either substantive strategy or passing reference.

These ambiguities likely contribute to the observed performance ceiling, though other factors such as training data composition and model architecture may also play a role. The remaining accuracy gap appears to reflect, at least in part, gen-

[36] This counterintuitive result reflects architectural design choices. Ministral's sliding-window attention mechanism is particularly effective for document classification where evidence appears in localized passages rather than requiring synthesis across the entire document.

uine classification difficulty where reasonable experts would disagree. If this interpretation holds, improvements beyond 80% may require clearer ESG category definitions and standardized disclosure templates rather than more capable models alone.

6.3 Model Selection Framework

Model selection depends on the relative costs of false positives versus false negatives in a given application context. MCC scores provide a single summary of overall classification quality: GPT-5-nano leads (0.60), followed by GPT-5 (0.58), while base compact models cluster at 0.30–0.45, confirming frontier superiority on balanced evaluation while compact models still demonstrate substantial capability above the zero-baseline. For initial portfolio screening where false alarms create review burden, precision-focused models like Ministral-8B (74% precision) minimize unnecessary analyst effort. For regulatory compliance applications (SEC climate disclosure, EU Taxonomy reporting) where missing disclosures carries legal risk, recall-focused models like GPT-5-nano (89% recall) provide more comprehensive coverage. For routine ESG monitoring where neither error type dominates, balanced models like Llama-3.1-8B (58% F1) offer reasonable performance across both dimensions.

Frontier models justify their higher costs in specific contexts: applications requiring recall above 85%, situations where missing ESG initiatives creates material financial or legal consequences, or environments lacking hardware for local deployment. For most screening workflows, local models achieve comparable accuracy while preserving data privacy.

6.4 Fine-Tuning Limitations

Fine-tuning on synthetic ESG data produced uniformly negative results: all three models performed worse after training.[37] The synthetic training data mimicked Swiss report style (terminology, sentence structure) but lacked the substance of authentic disclosures. Swiss annual reports contain company-specific programs, careful hedging language ("committed to" versus "achieved"), and cross-references that synthetic data cannot replicate. Models appear to have memorized surface patterns rather than learning transferable ESG classification skills.

These findings suggest that base models provide immediate viability (77–80% accuracy) without training investment or degradation risk. Fine-tuning may benefit organizations with hundreds of labeled authentic ESG documents, but this exceeds typical resource constraints. Ensemble voting similarly failed to improve outcomes: when all models trained on the same flawed synthetic data, voting amplified shared mistakes rather than correcting them.

6.5 Document-Level Performance Variation

The 80% accuracy figure represents an average across substantial document-level variation: some documents achieve 99% accuracy while others reach only 54%.[38]

Three error patterns emerge most frequently. Hallucinated strategies (5 of 30 documents) occur when models flag 1–6 ESG strategies that do not exist in the source document, typically when boilerplate language resembles genuine ESG disclosure. Overly conservative predictions (11 of 30 documents) occur when models miss genuine strategies, defaulting to "no ESG content" despite

[37] Fine-tuning degraded recall across all models: Ministral dropped 11 percentage points, Llama dropped 10 points, and Apertus dropped 58 points. Only Ministral showed compensating precision gains (+8 points to 82%), which may suit precision-critical applications but represents overall degradation.

[38] Error rates ranged 76-fold across our 30-document test set (1% to 46%, standard deviation 11 percentage points). Roughly 65% of this variation comes from document characteristics (legal form, language, topic mix) rather than model choice.

documentary evidence. Inconsistent thresholds (2 of 30 documents) manifest as 30+ point divergence between precision and recall for the same document, suggesting decision boundaries shift based on document style.

This variation suggests value in tiered review approaches. Approximately 83% of documents show no quality flags and may warrant automatic acceptance. Roughly 12% flagged for single concerning patterns may benefit from spot-check verification. The remaining 5% with multiple quality flags or high-stakes implications may require full manual review. Such tiered approaches can substantially reduce analyst workload while maintaining quality control for high-risk classifications.

6.6 Limitations

Several methodological constraints shape interpretation of these findings. The 30-document Swiss real estate sample (1,080 individual ESG classifications) provides sufficient statistical power for the large effects reported but may not generalize to different report styles, industries, or geographies. Swiss real estate ESG disclosures may be more standardized than other domains, potentially underestimating performance variation in more heterogeneous portfolios.

Ground truth labeling by a single domain expert introduces potential bias compared to multi-rater validation. Different analysts may draw different boundaries for what constitutes a "substantive strategy" versus boilerplate, suggesting that deployment contexts should include calibration against organization-specific standards.

Fine-tuning experiments used a single parameter configuration rather than exhaustive search.[39] Different parameters might yield better results, though we prioritized conservative findings over potentially cherry-picked successes.

Finally, these results reflect model capabilities as of September–October 2025. Model capabilities evolve continuously, suggesting periodic reassessment of deployment decisions as new models emerge.

[39] We tested LoRA rank=8, learning rate 5×10^{-5}, 5 training epochs. Exhaustive parameter search would cost $10,000+ and risk overfitting to our small test set.

6.7 Conclusions

This study demonstrates that AI-based extraction of ESG information from published annual reports achieves 80% accuracy on a 36-criterion framework, matching commercial services without requiring survey responses or company participation. This survey-free approach enables coverage of any entity with public disclosures, including non-participating companies, recent acquisitions, and competitors.

The findings suggest several deployment considerations. Base models (Ministral-8B, Llama-3.1-8B) provide immediate viability at 77–80% accuracy without training investment, as fine-tuning with synthetic data did not improve outcomes in our experiments. Model selection should depend on application-specific error costs: precision-focused models suit screening applications where false positives create review burden, while recall-focused models suit compliance applications where missed disclosures carry greater risk. Document-level performance variation (76-fold range in our sample) suggests value in tiered review approaches that concentrate human oversight on flagged or high-stakes cases.

More broadly, these results illustrate how the right tools can turn unstructured documents into structured, actionable data. As this capability matures, it has the potential to dramatically reduce the repetitive workload that currently consumes analyst capacity in ESG and other data-intensive sectors built on qualitative sources. The long-term implication is a shift toward quantitatively backed analysis in domains that have traditionally relied on manual interpretation, enabling systematic coverage at scales previously impractical through manual methods.

6.8 Future Directions

Three areas warrant further investigation. Multi-expert validation would quantify inter-rater reliability among ESG analysts, establishing realistic accuracy ceilings and calibrating model expectations against human disagreement rates. While simple cross-tier ensembles did not outperform GPT-5-nano on F1 (Section 5), more sophisticated combination strategies—weighted voting, confidence-based

routing, or ensembles combining precision-focused and recall-focused models—might achieve balanced performance without the training complexity encountered in our fine-tuning experiments. Active learning approaches that focus annotation effort on cases where models disagree most might improve both ground truth quality and model performance simultaneously.

The survey-free approach demonstrated here expands ESG assessment coverage beyond survey participants while reducing coordination burden. Further validation across diverse industries, geographies, and reporting frameworks would strengthen generalizability claims and refine model selection guidelines for specific application contexts.

References

Bauer, R., Eichholtz, P.M.A., Kok, N., Quigley, J.M., 2011. How green is your property portfolio? The Global Real Estate Sustainability Benchmark. Rotman International Journal of Pension Management 4(1), 34–43. URL: https://ssrn.com/abstract=1829276.

Birti, M., Osborne, F., Maurino, A., 2025. Optimizing Large Language Models for ESG Activity Detection in Financial Texts. arXiv 2502.21112. doi:10.48550/arXiv.2502.21112.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE 12. doi:https://doi.org/10.1371/journal.pone.0177678.

Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., Staiano, J., 2023. Glitter or gold? Deriving structured insights from sustainability reports via large language models. EPJ Data Science 13, 1–41. doi:10.48550/arXiv.2310.05628.

Chen, Z., Chen, J., Chen, J., Sra, M., 2025. Standard Benchmarks Fail – Auditing LLM Agents in Finance Must Prioritize Risk. arXiv 2502.15865. doi:10.48550/arXiv.2502.15865.

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6. doi:https://doi.org/10.1186/s12864-019-6413-7.

Koo, R., Lee, M., Raheja, V., Park, J.I., Kim, Z.M., Kang, D., 2023. Benchmarking Cognitive Biases in Large Language Models as Evaluators. arXiv 2309.17012. doi:10.48550/arXiv.2309.17012.

Li, Y., Wang, S., Ding, H., Chen, H., 2023. Large Language Models in Finance: A Survey. Proceedings of the Fourth ACM International Conference on AI in Finance. doi:10.1145/3604237.3626869.

Maini, P., Dorna, V., Doshi, P., Carranza, A., Pan, F., Urbanek, J., Burstein, P., Fang, A., Deng, A., Abbas, A., Larsen, B., Blakeney, C., Bannur, C., Baek, C., Teh, D., Schwab, D., Mongstad, H., Yin, H., Wills, J., Mentzer, K., Merrick, L., Monti, R., Adiga, R., Joshi, S., Das, S., Wang, Z., Gaza, B., Morcos, A., Leavitt, M., 2025.

Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining. arXiv 2508.10975. doi:10.48550/arXiv.2508.10975.

Mistral AI, 2024. Ministral-8b-instruct-2410. Hugging Face Model Repository. URL: https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.

Ohi, M., Kaneko, M., Koike, R., Loem, M., Okazaki, N., 2024. Likelihood-based Mitigation of Evaluation Bias in Large Language Models. arXiv 2402.15987. doi:10.48550/arXiv.2402.15987.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A., 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. arXiv 2402.07927. doi:10.48550/arXiv.2402.07927.

Schroeder, K., Wood-Doughty, Z., 2024. Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. arXiv 2412.12509. doi:10.48550/arXiv.2412.12509.

Stureborg, R., Alikaniotis, D., Suhara, Y., 2024. Large Language Models are Inconsistent and Biased Evaluators. arXiv 2405.01724. doi:10.48550/arXiv.2405.01724.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z., 2023. Large Language Models are not Fair Evaluators. arXiv 2305.17926. doi:10.48550/arXiv.2305.17926.

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D., Yilmaz, E., Shi, S., Tu, Z., 2024. Benchmarking LLMs via Uncertainty Quantification. arXiv 2401.12794. doi:10.48550/arXiv.2401.12794.

Zeng, Q., Jin, C., Wang, X., Zheng, Y., Li, Q., 2025. An Analyst-Inspector Framework for Evaluating Reproducibility of LLMs in Data Science. arXiv 2502.16395. doi:10.48550/arXiv.2502.16395.

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., Zhou, W., 2023. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. arXiv 2312.17264. doi:10.48550/arXiv.2312.17264.

Appendix

Appendix A: Statistical Validation

This section provides statistical validation for claims in the main paper. We report multiple comparison corrections, language impact analysis, and effect size interpretations.

Multiple Comparison Correction

We test differences between eight model configurations. Testing multiple hypotheses simultaneously inflates false positive rates.

Without correction, testing 10 pairs at $\alpha = 0.05$ yields 34% probability of at least one spurious significance claim. We apply Bonferroni correction to control family-wise error rate.

Correction method: Adjusted significance threshold $\alpha_{adj} = \alpha/k$ where k is the number of comparisons. For 10 pairwise tests, $\alpha_{adj} = 0.05/10 = 0.005$.

Pairwise comparisons tested:

1. **Base vs Frontier** (6 comparisons): Each base model (Ministral-8B, Llama-3.1-8B, Apertus-8B) compared against each frontier model (GPT-5-nano, GPT-5). These tests validate cost-benefit claims.
2. **Among Base Models** (3 comparisons): Ministral-8B vs Llama-3.1-8B, Ministral-8B vs Apertus-8B, Llama-3.1-8B vs Apertus-8B. These tests establish precision-recall specialization patterns.
3. **Frontier Models** (1 comparison): GPT-5-nano vs GPT-5. Tests whether 4 $\times$ cost premium justifies performance differences.

Statistical method: Paired t-tests on per-document accuracy. Each model evaluates the same 1,080 tasks (30 documents $\times$ 36 ESG categories). Paired design controls for document-level difficulty variation.

Key findings after correction:

- Ministral-8B vs GPT-5-nano: NOT significant ($p = 0.686$). Validates "near-frontier accuracy" claim. The 0.65pp gap does not represent systematic performance difference.
- Ministral-8B vs Llama-3.1-8B: NOT significant after correction ($p = 0.045$, above $\alpha_{adj} = 0.005$). Raw p-value significant, but does not survive multiple comparison correction.

Model 1	Model 2	p-value	Adj. Sig.	Cohen's d
Base vs Frontier Models				
Ministral-8B	GPT-5-nano	0.6858	No	-0.016 (negligible)
Ministral-8B	GPT-5	0.6817	No	-0.016 (negligible)
Llama-3.1-8B	GPT-5-nano	0.0759	No	-0.066 (negligible)
Llama-3.1-8B	GPT-5	0.0715	No	-0.066 (negligible)
Apertus-8B	GPT-5-nano	0.0000***	Yes	-0.405 (small)
Apertus-8B	GPT-5	0.0000***	Yes	-0.405 (small)
Among Base Models				
Ministral-8B	Llama-3.1-8B	0.0446*	No	0.050 (negligible)
Ministral-8B	Apertus-8B	0.0000***	Yes	0.388 (small)
Llama-3.1-8B	Apertus-8B	0.0000***	Yes	0.337 (small)
Frontier Models				
GPT-5-nano	GPT-5	1.0000	No	0.000 (negligible)

Note: Bonferroni-adjusted α = 0.005 for 10 comparisons. *** p<0.001, ** p<0.01, * p<0.05.

Table A1. Pairwise Model Comparisons with Bonferroni Correction

- Apertus-8B vs all others: SIGNIFICANT ($p < 0.001$). Apertus performs consistently worse across comparisons.
- GPT-5-nano vs GPT-5: NOT significant ($p = 1.0$). Both models achieve identical overall accuracy (80.65%) but differ in individual predictions (GPT-5-nano: 263 TP, 175 FP; GPT-5: 249 TP, 161 FP).

Effect sizes: We report Cohen's d for practical significance assessment. Cohen's d is a standardized effect size used to quantify practical significance by expressing the difference between two group means in standard deviation units. Small models vs frontier models show negligible effects (d $<$ 0.1) for Ministral and Llama, confirming performance parity despite massive parameter count differences.

Language Impact Analysis

Document language significantly affects model performance. We test whether German, French, and English documents present different difficulty levels.

Statistical method: One-way ANOVA tests whether mean accuracy differs across languages. Pairwise follow-up tests use Bonferroni correction for three comparisons ($\alpha_{adj} = 0.017$).

ANOVA results: $F(2, 8637) = 20.94$, $p < 0.001$. Language has significant effect on accuracy.

Language	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
German (n=2880)	78.09	57.11	61.02	59.00
	[76.53, 79.62]	[53.40, 60.74]	[57.29, 64.47]	[55.91, 61.96]
English (n=2880)	77.71	50.98	55.34	53.07
	[75.97, 79.13]	[47.27, 54.46]	[51.49, 59.18]	[49.67, 56.10]
French (n=2880)	71.60	60.03	48.46	53.63
	[70.00, 73.16]	[56.94, 63.45]	[45.60, 51.45]	[51.04, 56.41]

Note: ANOVA shows significant language effect (F=20.94, p<0.001). German vs French and English vs French differences significant after Bonferroni correction.

Table A2. Model Performance by Document Language

Pairwise comparisons:

- German vs English: NOT significant ($p = 0.727$, Cohen's d = -0.009). Performance nearly identical.
- German vs French: SIGNIFICANT ($p < 0.001$, Cohen's d = -0.150). German documents 6.49pp easier than French.
- English vs French: SIGNIFICANT ($p < 0.001$, Cohen's d = -0.141). English documents 6.11pp easier than French.

Interpretation: French documents present systematically higher difficulty. This could reflect:

1. More complex sentence structures in French corporate reporting
2. Domain terminology differences across language communities
3. Training data imbalances (models pretrained predominantly on English)

The 6-7pp performance gap across languages exceeds the 0.65pp gap between Ministral and GPT-5-nano. Language effects dominate model choice effects for multilingual applications.

Effect Size Interpretation

We use Cohen's d to measure practical significance. Statistical significance (p-values) indicates whether differences are real. Effect sizes indicate whether differences matter.

Cohen's d interpretation guidelines:

- $|d| < 0.2$: Negligible effect
- $0.2 \leq |d| < 0.5$: Small effect
- $0.5 \leq |d| < 0.8$: Medium effect
- $|d| \geq 0.8$: Large effect

Model comparison effect sizes:

All base vs frontier comparisons show negligible to small effects. The largest effect is Apertus vs frontier models (d = -0.405, small effect). Even this represents less than half a standard deviation difference.

Ministral vs GPT-5-nano: d = -0.016 (negligible). This confirms practical equivalence despite statistical testing framework.

Among base models, Ministral vs Apertus shows d = 0.388 (small effect). These models target different use cases rather than representing strict quality hierarchy.

Bootstrap Confidence Intervals

All performance metrics include 95% bootstrap confidence intervals. We use 1,000 bootstrap samples with replacement, seeded for reproducibility (seed=42).

Bootstrap procedure:

1. Sample 1,080 predictions with replacement from model's evaluation set
2. Calculate accuracy, precision, recall, and F1 for sample
3. Repeat 1,000 times
4. Report 2.5th and 97.5th percentiles as confidence bounds

Confidence interval interpretation: If confidence intervals for two models overlap substantially, performance differences may not be meaningful. Ministral (80%,

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	MCC
GPT-5	80.65	60.73	83.84	70.44	0.582
	[78.24, 83.06]	[56.04, 65.37]	[79.42, 87.84]	[66.76, 73.99]	[0.533, 0.631]
GPT-5-nano	80.65	60.05	88.55	71.56	0.602
	[78.33, 82.97]	[55.48, 64.80]	[84.72, 91.90]	[67.79, 75.10]	[0.556, 0.648]
Ministral-8B	80.00	73.68	42.42	53.85	0.449
	[77.69, 82.22]	[65.66, 80.21]	[36.91, 48.10]	[47.84, 59.37]	[0.386, 0.508]
Llama-3.1-8B	77.96	61.13	54.55	57.65	0.429
	[75.37, 80.37]	[54.90, 66.93]	[48.77, 59.93]	[52.25, 62.19]	[0.369, 0.489]
Apertus-8B	62.78	40.54	75.76	52.82	0.300
	[59.91, 65.56]	[36.45, 44.69]	[71.11, 80.34]	[48.80, 56.66]	[0.246, 0.354]

Note: Confidence intervals shown in brackets below each metric. MCC bootstrap CIs computed with 10,000 resamples (seed=42).

Table A3. Model Performance Summary with 95% Bootstrap Confidence Intervals

[77.69%, 82.22%]) and GPT-5-nano (80.65%, [78.33%, 82.97%]) show complete CI overlap, supporting non-significance finding.

Narrow confidence intervals (typically $\pm$2-3pp) indicate stable performance estimates. The 1,080-sample evaluation provides sufficient statistical power for detecting meaningful differences.

Sample Size and Statistical Power

We evaluate 30 documents $\times$ 36 ESG categories = 1,080 classification tasks per model. This sample size provides adequate power for detecting meaningful differences.

Power analysis framework: Statistical power (1-β) is the probability of detecting a real effect when it exists. Standard practice targets 80% power.

Achieved power for observed effects:

For the 0.65pp accuracy difference between Ministral and GPT-5-nano, we achieve >99% power to detect this difference as significant at $\alpha = 0.05$. The non-significant result therefore reflects true absence of meaningful difference, not insufficient sample size.

For medium effect sizes (Cohen's d = 0.5), our sample provides 95% power. For small effects (d = 0.3), power exceeds 75%. The evaluation design provides adequate sensitivity.

Document-level sample size: The 30 documents represent stratified sample of Swiss real estate annual reports. Documents span legal forms (funds, companies, foundations) and languages (German, French, English). Ground truth labels were assigned by a single trained annotator following standardized coding guidelines, with binary classification criteria (present/absent) designed to minimize subjective interpretation.

Limitations: The sample focuses on Swiss real estate ESG reporting. Generalization to other sectors or regions requires validation. The 36 ESG categories provide comprehensive coverage but specific category performance varies.

Reproducibility Notes

All statistical analyses use fixed random seeds (seed=42) for reproducibility. Bootstrap confidence intervals, model training, and data sampling use consistent seeding.

Software and methods:

- Statistical tests: scipy.stats (Python)
- Bootstrap sampling: numpy.random with fixed seed
- Performance metrics: scikit-learn
- Data processing: Polars

Significance thresholds: We use conventional $\alpha = 0.05$ before Bonferroni correction. Adjusted thresholds depend on number of comparisons in each analysis family.

Data availability: Model predictions, ground truth labels, and analysis scripts enable full replication. Contact authors for access to evaluation datasets and experimental configurations.

Appendix B: Detailed Methodology

B.1 Platform-Specific Implementation

All models receive identical inputs: document text (max 4,096 tokens), task specification with balanced examples, structured JSON output requirements. Documents exceeding token limits undergo semantic chunking with majority-vote aggregation.

Local inference uses MLX (Apple Silicon) or PyTorch (Linux/Windows) in 16-bit precision. Commercial models accessed via official APIs with deterministic

settings (temperature=0.0 for GPT-5-nano and GPT-4o; GPT-5 uses default configuration), max_tokens=500.

B.2 Synthetic Data Generation Protocols

Synthetic data generation employs GPT-4o-2024-08-06 via OpenAI API with temperature=0.7 and max_tokens=500. The system loads ESG categories from configuration files and extracts linguistic style patterns from authentic Swiss annual reports.

Generation variables include three languages (English, German, French), multiple writing styles extracted from real reports, difficulty levels (easy, medium, hard), and length ranges (150-450 words). Each synthetic example receives style-specific prompts combining category definitions, language instructions, and difficulty modifiers to ensure realistic text generation.

Quality control applies JSON parsing validation and tracks generation metadata through a dedicated QualityTracker class. All synthetic examples undergo structured output formatting with boolean classification labels and supporting evidence text, maintaining consistency with real dataset formats.

B.3 Fine-Tuning Hyperparameters

Fine-tuning employs parameter-efficient adaptation techniques, primarily Low-Rank Adaptation (LoRA) with rank=8 and scaling factor=20, and Quantized LoRA (QLoRA) for memory-constrained scenarios. These methods modify <0.1% of model parameters while achieving >95% of full fine-tuning performance.

Synthetic training data undergoes 80/20 train/validation splits stratified by ESG category. AdamW optimizer employs learning rate=5e-5 and weight decay=0.01 across 5 epochs with early stopping based on validation loss.

Platform-specific optimization includes MLX for local testing and DCSR cluster configurations for large-scale experiments. Post-training, LoRA/QLoRA adapters undergo fusion with base models using the cross-platform adapter fusion system for deployment efficiency. All training procedures maintain reproducibility through fixed random seeds and comprehensive logging.

Appendix C: Experimental Design Procedures

C.1 Baseline Establishment Details

All models undergo zero-shot evaluation without domain-specific training. Commercial models serve as performance benchmarks with comprehensive resource

tracking. GPT-4o-2024-08-06 provides the primary baseline given its established performance in document classification tasks. API interactions include systematic logging of token consumption, response latencies measured from request initiation to completion, and retry logic with exponential backoff to handle rate limiting. Cost tracking captures both input and output token charges at current API pricing tiers.

C.2 Fine-Tuning Configuration Details

Training protocol employs LoRA (rank=8, α=20) and QLoRA configurations depending on memory constraints. AdamW optimizer operates with learning rate=5e-5, weight decay=0.01, and β coefficients (0.9, 0.999). Training spans 5 epochs with early stopping triggered by validation loss plateaus (patience=2 epochs). Validation monitoring occurs every 100 optimization steps with automatic checkpoint saving.

C.3 Document-Level Analysis Procedures

We computed classification metrics for each document independently. For document d, we calculated precision P_d, recall R_d, and F1-score $F_{1,d}$ with bootstrap confidence intervals (1000 iterations, $\alpha = 0.05$). Cross-model consistency was measured via Spearman rank correlation of document-level error rates across all model pairs. Documents with $R_d = 0$ across all models were flagged as universally challenging.

C.4 Technical Implementation

The evaluation system employs deterministic settings to ensure reproducibility. Documents are segmented as needed to minimize position bias and preserve contextual integrity. Standardized prompts provide clear definitions and necessary context, including balanced positive and negative examples to enable effective few-shot learning. Models generate structured JSON outputs with binary classification decisions for each ESG category applied systematically across all documents. Quality controls encompass automated output parsing, fallback logic for incomplete outputs, systematic error logging, and detailed experimental records, supporting complete verification and auditability.

Appendix D: Parameter-Efficient Fine-Tuning Methods

This appendix provides technical background on fine-tuning approaches referenced in the literature review.

D.1 Low-Rank Adaptation (LoRA)

Recent advances in parameter-efficient fine-tuning enable domain adaptation of large language models with minimal computational overhead. Low-Rank Adaptation (LoRA) decomposes weight updates into low-rank matrices, modifying $< 0.1\%$ of model parameters while achieving $> 95\%$ of full fine-tuning performance on benchmark tasks. For a pretrained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA represents the update as $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with rank $r \ll \min(d, k)$.

Quantized LoRA (QLoRA) extends this approach to memory-constrained environments through 4-bit NormalFloat quantization of base weights, enabling fine-tuning of 7B+ parameter models on consumer GPUs with 16GB VRAM.

D.2 Domain Adaptation Challenges

Domain adaptation introduces catastrophic forgetting risks when task distributions differ substantially from pretraining data. Financial text classification presents particular challenges: ESG disclosures combine technical terminology, regulatory frameworks, and narrative strategies requiring nuanced interpretation (Birti et al., 2025). Small models' limited parameter budgets may prove insufficient to encode both general language understanding and domain-specific knowledge simultaneously.

D.3 Synthetic Data Quality Considerations

Research on trillion-scale pretraining (Maini et al., 2025) demonstrates that data quality matters more than quantity for fine-tuning success. Grounded augmentation—diverse rephrasing of authentic documents—dramatically outperforms pure generation approaches. Augmentation strategies preserve distribution properties while scaling training data, whereas generation risks creating artifacts that models memorize rather than learning transferable patterns.

For ESG applications, this suggests fine-tuning success depends on access to extensive authentic document corpora for augmentation rather than relying on pattern extraction or synthetic generation. Zou et al. (2023) show performance differences across Environmental, Social, and Governance dimensions, suggesting domain adaptation must account for category-specific complexity rather than treating ESG as a monolithic domain.

Annual Reports Sampling Design

The sampling strategy uses stratified simple random sampling of annual reports by legal form—fund, non-listed fund, company, and foundation—and by language, including English, French, and German. Within each stratum, a fixed number of reports is drawn randomly without replacement. To prevent duplication, a unique identifier is applied; once selected, all other entries with the same identifier are removed from the sampling frame for all years. This ensures each real estate investment vehicle appears only once. Empty strata are excluded and their quota redistributed across the remaining strata, keeping the total sample at thirty reports. Sampling is implemented in Python 3.12 using random sampling with a fixed seed for reproducibility. The final sampling frame and seed state are archived with the replication materials. The analysis focuses on within-stratum comparisons, omitting sampling weights to maintain statistical validity for subgroup inferences while avoiding increased variance from disproportionate allocation. Inclusion probabilities are uniform within each stratum.

Ensemble Voting Methodology and Agreement Analysis

Voting Algorithm. Three ensembles combine model predictions through simple majority rule. Each model produces binary classification (0 or 1) for each document-prompt pair. The ensemble prediction equals 1 if at least two models predict 1, otherwise 0. Formally: ensemble_pred $= \mathbf{1}[\sum_{i=1}^{3} \text{pred}_i \geq 2]$ where $\mathbf{1}[\cdot]$ is the indicator function. No weighting by model confidence or F1 score is applied.

Data Alignment. Predictions from three experiments are aligned on (document name, prompt name) keys. Each experiment produces 1,080 predictions (30 documents $\times$ 36 ESG categories). Inner join ensures exact alignment across all three models. Misaligned predictions are excluded. All ensembles validated to contain exactly 1,080 predictions.

Ensemble Configurations. Base model ensemble combines Ministral-8B-Instruct, Llama-3.1-8B-Instruct, and Apertus-8B-Instruct (all zero-shot base models). Fine-tuned ensemble combines the same three architectures after LoRA fine-tuning on 4,500 synthetic examples (rank=8, α=20, 5 epochs). Hybrid ensemble combines Ministral-8B fine-tuned (highest precision: 82.14%), Apertus-8B base (highest recall: 75.76%), and Llama-3.1-8B base (highest F1: 57.65%).

Agreement Statistics. Unanimous agreement occurs when all three models predict the same class (either all 0 or all 1). Two-one splits occur when exactly two models agree. Error correction counts predictions where at least one individual model is wrong but the ensemble is correct. Error introduction counts predictions where at least two individual models are correct but the ensemble is wrong.

Detailed Results. Base ensemble: 654/1,080 (60.6%) unanimous, 426/1,080 (39.4%) two-one splits, 293/1,080 (27.1%) error corrections, 0/1,080 (0.0%) error introductions. Fine-tuned ensemble: 763/1,080 (70.6%) unanimous, 317/1,080 (29.4%) two-one splits, 214/1,080 (19.8%) error corrections, 0/1,080 (0.0%) error introductions. Hybrid ensemble: 601/1,080 (55.6%) unanimous, 479/1,080 (44.4%) two-one splits, 321/1,080 (29.7%) error corrections, 0/1,080 (0.0%) error introductions.

Zero Error Introduction. All ensembles show 0.0% error introduction rate. Majority voting never creates errors when two or more individual models are correct. This validates the voting mechanism as reliable. Error corrections (19.8%-29.7%) represent genuine ensemble benefits from architectural diversity.

Implementation. Ensemble voting implementation uses Polars DataFrames for efficient data alignment and vectorized operations. All ensemble outputs match existing experiment format for reproducibility.

Prompts Format

You are a precise document classifier. Analyze the document and determine if it discusses {category}.

Definition: {category_definition}

Examples of documents that ARE about this topic:
{good_examples}

Examples of documents that are NOT about this topic:
{bad_examples}

Respond with a JSON object in exactly this format:
{{
"answer_bool": true/false
}}

Requirements:
- answer_bool: true if the document discusses {category}, false otherwise

Document text:
{document_text}