*A*TLAS: A spend classification benchmark for estimating scope 3 carbon emissions

Andrew Dumit[*]
Watershed Technology Inc.

Krishna Rao
Watershed Technology Inc.

Travis Kwee
Watershed Technology Inc.

Jonathan Glidden
Watershed Technology Inc.

Varsha Gopalakrishnan
Watershed Technology Inc.

Katherine Tsai
Watershed Technology Inc.

Sangwon Suh[†]
Watershed Technology Inc.

Abstract

The majority (70%) of companies reporting their value chain emissions rely on financial spend ledger and emissions factors per dollar. Accurate classification of expenditures to emissions factors is critical but complex, given the sheer number of line items and the diversity of how they are categorized and described. This is an area where Large Language Models (LLMs) can play a key role. However, there is currently no benchmark dataset to evaluate the performance of LLM-based solutions. Here, we introduce the Aggregate Transaction Ledgers for Accounting Sustainability dataset or, ATLAS, and the initial evaluation results of four models using ATLAS. ATLAS is the first spend classification benchmark and is comprised of 10,000 synthetic, labeled spend items reflecting the distribution of corporate expenditures. We evaluate four baseline models, with the best model achieving a top-1 accuracy of 57.3% and a top-3 accuracy of 72.2%. ATLAS enables systematic evaluation of LLMs for spend classification. Our results provide a starting point for advancing automated carbon accounting and sustainability reporting for spend-based emissions.

1 Introduction

75% of the corporate emissions reported to CDP was from Scope 3 [1], which covers the upstream and downstream value chains [2]. Companies that report their Scope 3 emissions rely on third-party provided datasets, called emissions factors (EFs), to calculate their environmental impact [3].

The most prevalent estimation method utilizes EFs derived from Environmentally Extended Input Output (EEIO) models such as the USEEIOv2 [4]. These EFs convert expenditures on goods and services into estimated kilograms of carbon dioxide equivalents (kgCO2e), providing a crucial link between spend data and environmental impact. Applying the spend-based EFs to spend line items requires a mapping of spend line items to the EFs called “spend classification”. Beyond its role in estimating scope 3 emissions, spend classification is also essential for other corporate sustainability objectives, such as public reporting in line with the Greenhouse Gas Protocol (GHGP), developing reduction strategies through supplier engagement, and managing expenditure to meet sustainability goals, as shown in Figure 1.

[*] andrew.dumit@watershedclimate.com

[†] sangwon@watershedclimate.com

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

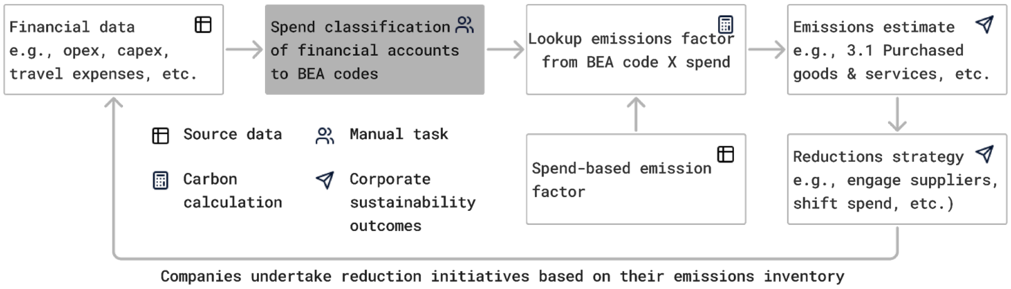
Financial data e.g., opex, capex, travel expenses, etc.
Spend classification of financial accounts to BEA codes
Lookup emissions factor from BEA code X spend
Emissions estimate e.g., 3.1 Purchased goods & services, etc.
Source data
Manual task
Carbon calculation
Corporate sustainability outcomes
Spend-based emission factor
Reductions strategy e.g., engage suppliers, shift spend, etc.)
Companies undertake reduction initiatives based on their emissions inventory

Figure 1: Processing steps involved in estimating carbon emissions from spend data. Spend classification task is highlighted for emphasis. BEA: Bureau of Economic Analysis.

However, spend classification has its challenges. This step is particularly complex because it requires company-specific context of financial ledgers and the purposes of various expenditures, domain knowledge of carbon accounting principles, and a comprehensive understanding of the possible EFs. Should two companies apply different spend classification approaches to identical expenditures, their calculated carbon footprints may vary considerably. Further complicating the process is the sheer scale of the task—large corporations often manage thousands of spend accounts and millions of line items.

LLMs have emerged as a promising solution to streamline this spend classification process. Several studies have proposed assistive or automated tools for spend classification. Balaji et al. [5] proposes a vector similarity based approach that achieves 75% EF matching precision on a dataset of 664 products. Jain et al. [6] evaluates zero-shot and supervised fine-tuning on a dataset of 2,162 financial ledger investments achieving 21% and 83% F1 industry-level classification respectively. However, to the best of our knowledge, no benchmark dataset of spend classification exists that allows systematic inter-comparison of different modeling approaches. In this study, we introduce ATLAS, the first spend classification benchmark dataset, along with the performance metrics of four baseline models.

2 ATLAS: Aggregate Transaction Ledgers for Accounting Sustainability

The ATLAS dataset comprises 10,000 synthetic spend classifications spanning 9 macro industries and 8 countries. These labels encompass 295 unique classes, with 294 corresponding to Bureau of Economic Analysis (BEA) codes and one designated for non-emissive spend. The 294 codes cover 73% of the approximately 400 codes used by the BEA to classify economic activity. Since the BEA codes form the foundation of prominent EEIO models like USEEIO that account for 48% of EEIO usage in scope 3 emissions measurement [3], the ATLAS dataset also uses BEA codes for the labels. Non-emissive spend refers to transactions recorded in a company's ledger that do not leave the company's financial boundary or are not exchanged for goods or services [2]. Of the total mappings, 1,765 (17.6%) are classified as non-emissive. The ATLAS dataset is available for academic purposes upon request to the corresponding author.

We generate the synthetic mappings based on metadata about each BEA sector A.3 using Claude 3.5 Sonnet [7] to match the label distribution of an internal private dataset of de-identified company spend accounts. We then modify the account names with rules observed in the private dataset. We review the synthesized account names for accuracy to the target label and ensure they reflect realistic inputs relative to the private dataset. Finally, we compare the performance of ATLAS against the private benchmark and observe that ATLAS is faithful to the difficulty and composition of the private dataset.

We analyze the diversity of the expenditure using an aggregation of BEA codes that measures the degree of fabrication (DoF) of produced goods. The DoF, from input-output analysis, describes where an activity lies in the value chain, between raw materials (low DoF) and finished products (high DoF) [8]. The DoF helps understand the approximate position in the value chain and highlights larger errors in reasoning. We map the BEA codes to DoF using the mapping in A.2.

As shown in Figure 2, the ATLAS dataset has wide coverage of different sectors of the economy. The distribution of labels is sourced from 9 macro industries and covers all 9 DoFs of the labels. Expenses

Table 1: Example spend account names across a sample of industries.

BEA industry code	**Industry name**	**Example spend accounts**
541800	Advertising, public relations, and related services	Video Production Expenses, Copywriting and Content Creation
230301	Nonresidential maintenance and repair	system maintenance Payment & Commerical Pressure Washer
484000	Truck transportation	Dump Truck Operations, Truck leasing Payment, long-distance auto exp.

related to finished goods and services are relatively dominant, whereas, purchases of utilities and logistics seem under-represented. We posit the under-representation is due to non-financial methods of estimating carbon emissions for such activities (e.g., using kWh of electricity consumed instead of $). Despite the skew, the ATLAS dataset is promising for benchmarking other modeling approaches because of its size and diversity of industries.

3 Analysis and baseline results

3.1 Models

We evaluate four baseline models on the ATLAS dataset, covering a spectrum of recent approaches [5] [6] utilizing simple regressions to frontier LLMs for classification.

Text embedding model: Utilizing OpenAI's text-embedding-small model [9] we compute embeddings on the test set accounts and all BEA codes. We select the top 3 examples via cosine similarity between the spend account and all BEA codes. We encode the labels via relevant metadata, see A.3.

Logistic regression: We train a logistic regression on the training set after encoding them via normalized word frequency. We apply the same encoding to the test set during evaluation.

LLM with prompt engineering: We employ Anthropic's Claude 3.5 Sonnet [7], a state-of-the-art language model, to predict BEA codes from account names. The prompt incorporates chain-of-thought prompting [10] and meta-prompting [11], to enhance the model's reasoning capabilities.

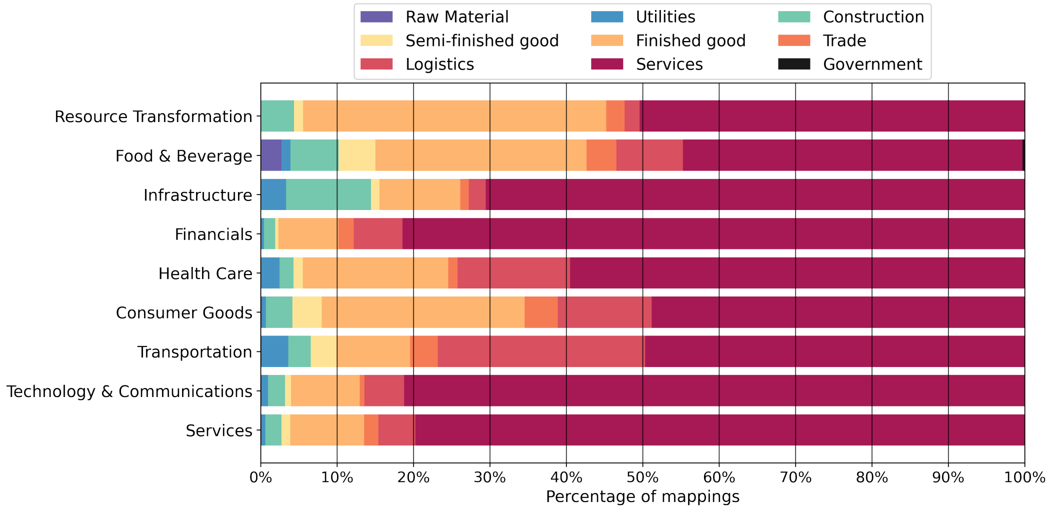

Figure 2: Distribution of the degree of fabrication amongst mappings within each macro-industry used to source the labels in the ATLAS dataset.

Table 2: Performance comparison of different model types across various metrics. Higher is better for metrics marked with (↑), lower is better for metrics marked with (↓).

Model	**Private benchmark Top 3 acc (↑)**	**Top 1 acc (↑)**	**Top 3 acc (↑)**	**Top-1 DoF Accuracy (↑)**	**Top-1 EF RMSE (↓)**
Text-embedding	37.6%	37.1%	57.3%	68.7%	0.331
Logistic regression	53.5%	48.5%	60.2%	73.8%	0.376
LLM	56.6%	40.6%	57.3%	75.9%	0.326
Fewshot LLM	**61.4%**	**57.3%**	**72.2%**	**82.8%**	**0.293**

LLM enhanced with few shot learning: We build on the previous approach by integrating examples for improved in-context learning [12]. We dynamically retrieve 50 examples from the training dataset which are incorporated into the prompt.

3.2 Results and discussion

We evaluate our four baseline models against five metrics. These metrics are defined in A.1. The results of this evaluation are presented in Table 2, which offers several key insights.

First, our results demonstrate that frontier LLMs exhibit strong performance on this task. Notably, their effectiveness is enhanced when utilized as an assistive tool (when top 3 accuracy is more important) rather than an autonomous process (when top 1 accuracy is relevant). In the former, a human analyst with limited domain knowledge could reasonably adjudicate among the predictions.

Second, there is a marked difference between the accuracies on the private benchmark vs. ATLAS for all models except the LLM model. We posit the difference is due to at least two reasons. First, the account names in ATLAS are generated from the industry metadata which leads to a greater similarity between the synthetic data and the data used to train or source fewshot examples. Second, the divergent gap in performance between the base LLM and Fewshot LLM on the private benchmark compared to the same metrics on ATLAS indicates that the train and test data may be more similar in ATLAS than the private benchmark. Despite the imperfections, ATLAS remains a good benchmark because improvements on ATLAS reflect improvements on the private benchmark. If a practitioner optimizes on ATLAS, we expect the model to do well on the spend classification task broadly.

Lastly, despite the promising results, there remains considerable scope for enhancement, particularly in error reduction. While the Root Mean Square Error (RMSE) of the top-1 emissions factor shows meaningful improvement from the worst to the best-performing method, further refinement of this metric could facilitate more confident autonomous mappings.

4 Climate impact and conclusions

Consistent spend classification is crucial for evaluating progress towards climate goals and developing effective reduction strategies. Misclassified expenditures can lead to ineffective initiatives, highlighting the importance of accurate measurement and management in this context. To address these challenges, we introduce ATLAS, the first benchmark dataset for spend classification in carbon accounting. With 10,000 synthetic mappings across diverse industries, ATLAS enables the evaluation of automated methods.

Future work could explore improved prompt techniques, fine-tuning on spend data, and enhanced retrieval strategies for in-context learning[13]. Additionally, expanding the contextual information in the dataset could further improve model performance.

Semi-automated or fully automated spend classification could reduce resources required for carbon accounting, enabling more comprehensive and reliable sustainability reporting, particularly for scope 3 emissions calculations. The ATLAS dataset and baseline models provide a foundation for accelerating global corporate sustainability efforts.

References

[1] CDP. Aug. 2024. URL: `https://cdn.cdp.net/cdp-production/cms/guidance_docs/pdfs/000/003/504/original/CDP-technical-note-scope-3-relevance-by-sector.pdf`.

[2] World Resources Institute and World Business Council for Sustainable Development. *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard (Revised Edition)*. Tech. rep. Accessed on 2024-08-28. Washington, DC and Geneva, Switzerland: World Resources Institute and World Business Council for Sustainable Development, 2004. URL: `https://ghgprotocol.org/corporate-standard`.

[3] CDP. 2023. URL: `https://www.cdp.net/en/companies/cdp-2023-disclosure-data-factsheet`.

[4] Wesley W Ingwersen et al. "USEEIO v2. 0, the US environmentally-extended input-output model v2. 0". In: *Scientific Data* 9.1 (2022), p. 194.

[5] Bharathan Balaji et al. "Flamingo: Environmental impact factor matching for life cycle assessment with zero-shot machine learning". In: *ACM Journal on Computing and Sustainable Societies* 1.2 (2023), pp. 1–23.

[6] Ayush Jain et al. "Empowering Sustainable Finance: Leveraging Large Language Models for Climate-Aware Investments". In: *International Conference on Learning Representations*. 2024.

[7] Anthropic. June 2024. URL: `https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf`.

[8] Shinichiro Nakamura et al. "The waste input-output approach to materials flow analysis". In: *Journal of Industrial Ecology* 11.4 (2007), pp. 50–63.

[9] OpenAI. *New embedding models and API updates*. Jan. 2024. URL: `https://openai.com/index/new-embedding-models-and-api-updates`.

[10] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[11] Yifan Zhang. "Meta prompting for agi systems". In: *arXiv preprint arXiv:2311.11482* (2023).

[12] Qingxiu Dong et al. "A survey on in-context learning". In: *arXiv preprint arXiv:2301.00234* (2022).

[13] Alexander Scarlatos and Andrew Lan. "Reticl: Sequential retrieval of in-context examples with reinforcement learning". In: *arXiv preprint arXiv:2305.14502* (2023).

A Appendix

A.1 Metric definitions

We evaluate the performance of our model using multiple metrics that capture different aspects of the mapping task. First, we use top-1 classification accuracy to measure whether the model correctly identifies the BEA code for each expense item as chosen by the experts. To account for some uncertainty in classification and the possibility of human-to-human differences in the codes chosen by the experts, we also assess top-3 classification accuracy, where the model predicts three possible BEA codes, and we check if one of them matches the correct code.

$$\text{Top3 Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\text{True Label} \in \{\text{Top 3 Predictions}\})$$

$$\begin{aligned} &\text{Where:} \\ N &: \text{is the total number of labels} \\ \mathbf{1}(\cdot) &: \text{is the indicator function, which is 1 if the true label} \\ &\quad \text{is within the top 3 predictions and 0 otherwise.} \end{aligned}$$

Additionally, we map the top prediction and label to its DoF by the applying te mapping described in A.2 and compare equality between them:

To further evaluate the model's impact on emissions calculations, we calculate the top-1 root mean squared error (RMSE) between the USEEIO emissions factors of the predicted and correct BEA code.

A.2 Degree of Fabrication Mapping

Here, we provide the mapping for all codes in the dataset to their respective degree of fabrication. Most codes can be mapped at the two-digit level. Manufacturing sectors are split into finished or semi-finished goods. Where an industry contains both, expert judgement was used to map them to the DoF that is most prevalent.

A.2.1 Two-digit code mappings

Code	Category	Code	Category	Code	Category
11	Raw Material	21	Raw Material	22	Utilities
23	Construction	42	Trade	44	Trade
4B	Trade	45	Trade	48	Logistics
49	Logistics	51	Services	52	Services
53	Services	54	Services	55	Services
56	Services	61	Services	62	Services
71	Services	72	Services	81	Services
GS	Government	S0	Government		

A.2.2 Four-Digit Codes

Code	Category	Code	Category	Code	Category
3111	Finished good	3112	Semi-finished good	3113	Finished good
3114	Finished good	3115	Finished good	3116	Semi-finished good

Code	Category	Code	Category	Code	Category
3117	Semi-finished good	3118	Finished good	3119	Finished good
3121	Finished good	3122	Finished good	3131	Semi-finished good
3132	Semi-finished good	3133	Semi-finished good	3141	Finished good
3149	Finished good	3151	Semi-finished good	3152	Semi-finished good
3159	Semi-finished good	3161	Semi-finished good	3169	Semi-finished good
3211	Semi-finished good	3212	Semi-finished good	3219	Semi-finished good
3221	Semi-finished good	3222	Finished good	3231	Finished good
3241	Semi-finished good	3251	Semi-finished good	3252	Semi-finished good
3253	Finished good	3254	Finished good	3255	Finished good
3256	Finished good	3259	Finished good	3261	Finished good
3262	Finished good	3271	Finished good	3272	Finished good
3273	Finished good	3274	Finished good	3279	Finished good
3311	Semi-finished good	3312	Semi-finished good	3313	Semi-finished good
3314	Semi-finished good	3315	Semi-finished good	3321	Semi-finished good
3322	Finished good	3323	Finished good	3324	Finished good
3325	Finished good	3326	Finished good	3327	Finished good
3328	Finished good	3329	Finished good	3331	Finished good
3332	Finished good	3333	Finished good	3334	Finished good
3335	Finished good	3336	Finished good	3339	Finished good
3341	Finished good	3342	Finished good	3343	Finished good
3344	Finished good	3345	Finished good	3346	Finished good
3351	Finished good	3352	Finished good	3353	Finished good
3359	Finished good	3361	Finished good	3362	Semi-finished good
3363	Semi-finished good	3364	Semi-finished good	3365	Finished good
3366	Finished good	3369	Finished good	3371	Finished good
3372	Finished good	3391	Finished good	3399	Finished good
3150	Finished good	3379	Finished good		

A.3 Example industry metadata

For the industry embedding model, we use metadata about the industry to create the vectorized representation for that industry. One example of that metadata is below:

1111A0, Oilseed Farming: This industry comprises establishments primarily engaged in: growing soybeans and/or producing soybean seeds; growing fibrous oilseed producing plants and/or producing oilseed seeds, such as sunflower, safflower, flax, rape, canola, and sesame; growing dry peas, beans, and/or lentils.

A.4 Additional model specifics

Logistic regression: For the output, we add one additional class for predicting non-emissive.

Text embedding model: We encode the “non-emissive” code as a concatenation of nouns describing non-emissive business activities that we collected from in-house experts. Specifically, we use: "Salaries, Benefits, Taxes, Donations, and Amortization".