A Robust Green Patent Database: A New Dataset of Green Patents Through Large Language Models

Yuan Sun, Xuan Tian and Yuanchen Yang*[1]

Abstract

Amid the escalating challenges of climate change, green technologies have become a focal point for researchers, policymakers, and industry leaders. Patents, as key indicators of technological progress, provide valuable insights into sustainable innovation. This study introduces a dynamic and comprehensive dataset of green patents issued by the United States Patent and Trademark Office (USPTO) from 1976 to June 30, 2024, leveraging a fine-tuned Large Language Model (LLM) for green patent classification.

Focusing on Y02T transportation-related green patents, our methodology refines the original USPTO dataset of 133,042 patents by identifying 107,382 high-confidence green transportation patents. This process excludes 25,660 patents (19.3%) classified as low-confidence green patents or potential misclassifications. Furthermore, our approach uncovers 17,578 previously unclassified green transportation patents, resulting in a 13.2% dataset expansion. The final dataset includes 124,960 high-confidence patents or 150,620 when incorporating all identified patents. The classification achieves an 88% accuracy rate for newly identified patents, significantly outperforming the USPTO Y02T scheme and commercial LLM-based models.

This dynamic methodology can be extended to other green technology categories, offering broader insights into sustainable innovation. By enhancing the accuracy and scope of green patent classification, this work provides a robust tool for advancing research on green innovations. All code and resources supporting this study are publicly available at https://github.com/yuanresearch/Robust-Green-Patents-Paper, fostering transparency and encouraging collaboration within the research community.

Key words: green innovation, large language model (LLM), US patents, Sustainability

---

[1]Yuan Sun (Corresponding Arthor), Assistant Professor at Shanghai University of Finance and Economics, China, sunyuan89574@gmail.com, Xuan Tian, JD Capital Chair Professor of Finance at PBC School of Finance Tsinghua University, China, tianx@pbcsf.tsinghua.edu.cn , Yuanchen Yang, Economist at the International Monetary Fund, USA, yyang6@imf.org. All errors and omissions are solely ours.

1 Introduction

Despite the intense interest in green technology (“green tech” hereafter) innovations and their impact, we know remarkably little about where or by whom these new products and services are being developed. This paper seeks to address this gap by creating a comprehensive dataset of green tech patents based on machine learning (“ML” hereafter) algorithms and existing Y02 schemes. While the definition of “green tech” has shifted over time, these patents provide a valuable window into the nature of green tech innovation.

Economists and policy makers are interested in measuring ESG innovations, as it is vital for them to assess new business opportunities and emerging markets through ESG innovations (Haščič and Migotto, 2015). There is also limited systematic evidence that firms receiving disproportional amounts of capital from ESG funds have outperformed in any measurable way (Cohen et al., 2020). Patents are not only used to measure financial innovation (Lerner et al., 2021), but can also be extended to the field of green innovation—innovation that reduces carbon footprint and facilitates climate transition. IPO firms backed by more failure-tolerant VC investors are significantly more innovative (Tian and Wang, 2014), but it is unknown whether this applied to the green technology related firms. In the realm of green technologies, the Y02 scheme is widely accepted (Angelucci et al., 2018). Y02 is a tagging scheme which enables documents related to sustainable technologies to be retrieved quickly and accurately, across classification categories. This tagging scheme is included in the European Patent Office’s CPC classification scheme, which makes it compatible with both the CPC and the IPC codes (V. Veefkind et al., 2012).

The Y02 scheme is designed with search statement algorithms defined by internal examiners with input from external peers. A quality policy has been introduced which checks the results of the algorithms written for each Y-classification code and adapts the search statements in order to reduce the number of incorrectly retrieved documents under a selected threshold. For a random sample of 150 documents, according to the adopted statistical quality control, the algorithm developers strive to reach an error rate of less than 7% (Angelucci et al., 2018). Another solid evidence for errors is that busy patent examiners exhibit significantly lower quality, which would negatively predicts the firms future stock returns (Shu et al., 2022).

For example, first, the examiners identify relevant classification entries for Y02, in this case the CO2 capture technologies. Then they look for existing classification entries and search strategies that would best locate the relevant documents for that technology. After that, they use search and classification tools to develop search algorithms which are now used to find and update all documents related to Climate Change Mitigation Technologies（CCMT）. Periodically, they use these steps to assign the Y02 tags to documents containing CCMT (Dechezleprêtre et al., 2011).

This means that tagging allows users to search across categories without affecting existing classifications.

The existing Y02 tagging schemes face several critical challenges. First, many patents that are undeniably “green” have been overlooked by current Y02 classification algorithms, leaving significant gaps in identifying cutting-edge green technologies. This issue may stem from limitations in existing methods or the varying expertise and subjective judgments of patent examiners. Second, these schemes fail to differentiate between high- and low-confidence classifications, treating all tagged patents equally. This has led to non-green patents being misclassified under Y02 labels, undermining the reliability of the dataset. In this paper, we introduce a fine-tuned machine learning architecture that addresses these challenges and significantly enhances green patent identification. Our approach uncovers overlooked innovations and redefines the boundaries of green patent classification, providing a powerful tool to advance sustainable technology research and innovation.

2. Dataset Construction

In this study, we constructed a comprehensive dataset of green patents issued by the United States Patent and Trademark Office (USPTO). The dataset encompasses patents classified under various categories of green technologies, aiming to provide a holistic view of innovations in environmentally sustainable solutions. By expanding upon the existing Y02 classification scheme—which is specifically designed for tagging climate change mitigation technologies—we enhance the understanding of the evolution and breadth of green technologies over time.

2.1 Downloading Green Patents from USPTO

We began by downloading the complete set of patents from the PatentsView (Toole et al., 2021), a platform provided by the USPTO that offers comprehensive and accessible patent data. From this extensive collection, we extracted patents that the USPTO defines as green patents, focusing on those classified under the Cooperative Patent Classification (CPC) codes associated with green technologies.

For illustration purposes, we use the Y02T category as an example. The Y02T classification pertains to climate change mitigation technologies related to transportation, including innovations in electric vehicles, hybrid propulsion, and energy-efficient transportation systems.

2.2 Identifying Potentially Underrepresented CPC Classes

To enhance the comprehensiveness of our green patent dataset, we identified a list of CPC classes that are potentially underrepresented in the traditional Y02T classification scheme. Based on expert input, we compiled the following list of CPC classes:

C10L - Fuels Not Otherwise Provided For; Natural Gas; Synthetic Natural Gas; Liquefied Petroleum Gas; Additives to Fuels or Fires

Why It Might Be Overlooked: This category primarily deals with fuel types and additives rather than direct emission-reduction technologies.

Relevance: Additives in fuels and alternative gas sources, such as liquefied petroleum gas (LPG) and synthetic natural gas, can improve combustion efficiency and reduce harmful emissions. Including this category in climate change mitigation could enhance cleaner fuel options for transportation and industrial processes.

B29D - Producing Particular Articles from Plastics or Substances in a Plastic State

Why It Might Be Overlooked: This category is often associated with manufacturing processes for plastic products rather than sustainability efforts.

Relevance: The plastic manufacturing process can contribute to emissions if not managed efficiently. Incorporating methods to reduce waste and energy consumption in the production of plastic products can align with sustainability goals, particularly in reducing plastic waste and improving recycling technologies.

F25B - Refrigeration Machines, Plants, or Systems; Combined Heating and Refrigeration Systems; Heat Pump Systems

Why It Might Be Overlooked: Refrigeration systems may not be considered in the same light as other technologies like electric vehicles or renewable energy sources when addressing emissions.

Relevance: Energy-efficient refrigeration and heat pump systems are key technologies for reducing energy consumption and emissions in both residential and industrial applications. They can play a significant role in mitigating climate change by enhancing energy efficiency in heating and cooling processes.

G08G - Traffic Control Systems for Road Vehicles

Why It Might Be Overlooked: Y02T often focuses on technologies that reduce vehicle emissions, rather than optimizing the broader transportation system.

Relevance: Intelligent traffic management systems can help optimize traffic flow, reduce congestion, and cut fuel consumption, leading to lower emissions. Including this category could foster a more holistic approach to climate change mitigation by integrating traffic control with vehicle emission reduction strategies.

H05B - Electric Heating; Electric Lighting Not Otherwise Provided For

Why It Might Be Overlooked: This category is often associated with heating and lighting technologies that are not typically viewed as emission-reduction solutions.

Relevance: The development of energy-efficient electric heating and lighting systems can reduce overall energy consumption and greenhouse gas emissions, particularly in

industrial and commercial sectors. Incorporating these technologies into green innovations could further support sustainable energy use in a variety of sectors.

2.3 Preliminary Assessment Using GPT-4o-Mini

To efficiently evaluate whether patents from the identified CPC classes qualify as green patents, we used GPT-4o-Mini, an advanced AI language model developed by OpenAI. This model was utilized to quickly assess the environmental sustainability of patents, focusing on those that contribute to green innovations such as energy efficiency, emission reduction, and clean technologies. The purpose of this initial assessment was to filter patents and create a reliable training dataset for further analysis.

A specific prompt was crafted to guide the AI in performing the task of patent evaluation. This prompt instructed GPT-4o-Mini to act as an expert green patent examiner. The key task was to determine whether a given patent qualifies as a "green patent," defined as a patent related to environmentally sustainable innovations.

The prompt provided to the AI was as follows:

"You are an expert green patent examiner. Your task is to determine whether the patent in question qualifies as a 'green patent.' related to transportation. A 'green patent' refers to an innovation focused on environmental sustainability, including technologies that reduce environmental impact. Please provide your answer in the following format:

Decision: Yes or No (Indicate 'Yes' if the patent qualifies as a green patent, or 'No' if it does not.)

Confidence Score: A numerical value between 1 and 100 that reflects your confidence in the decision."

Using GPT-4o-Mini in this way enables a rapid screening of patents, allowing us to quickly assess large patent datasets while gauging the AI's confidence in its classifications. This early-stage filtering is crucial for constructing a high-quality training dataset and enables efficient identification of patents relevant to green technologies. Furthermore, the confidence score provides a measure of the reliability of the AI's decisions, allowing for adjustments or further verification as necessary.

This preliminary assessment process is essential for creating a scalable and reliable patent classification system, ensuring that subsequent analyses are based on the most relevant and accurate patent data. By automating the initial evaluation with measurable confidence, we can handle large volumes of patent data without the need for extensive manual review.

2.4 Extracting the Training Dataset

Using the results from the preliminary assessments, we constructed our training dataset with a clear focus on balancing and ensuring the quality of the data. The construction process is outlined below:

1.Positive Training Data: This subset includes patents where the primary CPC class belongs to the underrepresented categories identified earlier (['C10L', 'B29D', 'F25B', 'G08G', 'H05B']). We considered these patents as positive examples because their classification within these specific CPC classes suggests that the USPTO examiners have recognized them as green patents. These patents are related to innovations in fields such as alternative fuels, energy-efficient technologies, refrigeration, traffic control, and electrical heating, all of which contribute to environmental sustainability. Despite their relevance to green technologies, these categories have been underrepresented in the Y02T scheme, which primarily targets patent classifications explicitly related to climate change mitigation in transportation. As a result, including these patents helps ensure that our model captures a wider range of green innovations that might not have been considered under the more traditional green patent classification schemes.

2.Negative Training Data: This subset includes patents for which the AI model's decision was "No" with a high confidence score (greater than or equal to 85). The confidence score reflects the model's certainty that the patent does not qualify as a green patent, ensuring that the negative examples are clearly non-green and easily distinguishable from positive examples. These patents typically pertain to technologies or innovations that do not contribute to environmental sustainability, such as those focused on conventional fossil fuels, general manufacturing processes, or other areas not directly linked to reducing environmental impact. By incorporating these patents, we ensure that the model is capable of distinguishing between green and non-green technologies with high accuracy.

To ensure that our training dataset was both balanced and effective for model training, we carefully sampled the negative patents to be twice the number of positive patents. This ratio was chosen to prevent an imbalance that could lead to biased predictions. By providing twice as many negative examples as positive ones, we ensure that the model has sufficient exposure to a variety of non-green patents without overwhelming the positive examples, which could reduce the effectiveness of the model in correctly identifying green patents. This approach allows for the optimal training of the AI model, helping it to learn to accurately classify patents into green and non-green categories.

The creation of this training dataset was crucial for ensuring that our model would not only be accurate but also generalizable. By considering underrepresented categories as positive examples, we ensure that the model is trained to recognize a wider range of green patents, including those that might otherwise be overlooked. Additionally,

the balanced ratio of positive and negative data points allows the model to learn the nuances of classification, ultimately improving its performance in subsequent evaluations.

2.5 Obtaining Detailed Explanations from the Advanced Model

To enhance our dataset with qualitative insights and improve transparency in the decision-making process, we employed GPT-4o-latest, the most advanced AI language model available. This model was used to generate detailed explanations for each patent's classification, helping us to better understand the rationale behind each decision. These explanations offer valuable context that improves the interpretability and reliability of our model, ensuring that the decision-making process is both clear and transparent.

The core objective was to instruct the AI to provide a comprehensive and detailed analysis of each patent, focusing on its environmental benefits or shortcomings. Rather than simply providing a classification, the AI was asked to elaborate on the sustainability aspects of each patent. This required the model to identify the key environmental features that contributed to the patent's classification and to explain how these features aligned with recognized green patent standards.

The GPT-4o-latest was asked to provide a concise summary of 200 to 300 words for each patent. The explanation should:

1. Identify the core environmental benefits or shortcomings of the patent.

2. Highlight unique or innovative features that contribute to its sustainability.

3. Discuss how the patent aligns with established green patent standards and contributes to environmental sustainability goals.

The purpose of obtaining these detailed explanations was to achieve two main objectives. First, the explanations provided insight into why a patent was classified as green or non-green, allowing us to better understand the reasoning behind the AI's decisions. This transparency was crucial for validating the accuracy of the classifications and ensuring that the AI's decision-making process aligned with environmental sustainability standards. Second, the explanations themselves serve as valuable training data for further refining the AI model, enhancing its future predictions and decision-making capabilities.

By generating detailed, context-rich explanations, we were able to thoroughly evaluate the performance of the best available Large Language Models (LLMs) and their capacity to classify patents based on sustainability criteria. This process was critical for enhancing the interpretability of the model, ensuring that patent classifications were grounded in a consistent and evidence-based understanding of

green technologies. Such interpretability is essential for building trust in the model's outputs and for advancing research on the intersection of artificial intelligence and sustainable innovation.

2.6 Fine-Tuning the Open-Source Language Model

With the enriched dataset in hand, we proceeded to fine-tune an open-source language model—specifically, the LLM3.1 model developed by Facebook—using the **Unsloth** (Zheng et al., 2024) FastLanguageModel framework. Unsloth is an open-source framework designed for the rapid fine-tuning of large language models (LLMs). By utilizing this framework, we could tailor the model to specialize in the task of patent classification related to green technologies, providing several key advantages that were crucial for the success of our project.

Unsloth simplifies and accelerates the fine-tuning process by enabling efficient model customization. The FastLanguageModel tool within unsloth is optimized for handling large amounts of data and complex tasks, such as processing patent descriptions in a specialized domain like green technology. The framework’s advantages include:

1.Efficiency: Traditional fine-tuning of large language models, particularly closed-source models, can take several days or even weeks to complete. In contrast, the fine-tuning process with unsloth’s FastLanguageModel takes only a few minutes. This drastically reduces the time required for model training and testing, allowing for rapid iteration and adjustments. This speed was vital to keeping the project on track and ensuring we could quickly integrate feedback and improve model performance.

2.Large Input Window: One of the standout features of the LLM3.1 model, facilitated by unsloth, is its input window of 128,000 tokens. This extended input window is particularly important for processing patent texts, which are often long and detailed. The model’s ability to handle such large inputs enables it to capture all the relevant information contained in patent descriptions, ensuring that no crucial data is lost during classification. By processing 99% of the content without truncation, the model can analyze the nuances of patent documents and make more accurate classifications as to whether they are related to green technology.

3.Customization: Fine-tuning the model on our specific dataset, which focused on green patents, allowed us to customize it for this particular task. By training the model with patent descriptions that have already been classified, we enabled it to specialize in identifying the features that define green patents. This level of customization improved the model’s ability to accurately identify patents that align with environmental sustainability goals and green technology standards.

In addition to LLM3.1, we evaluated a range of four-bit models provided by unsloth. These models are optimized for high efficiency, offering faster processing speeds without compromising accuracy. Notable models evaluated include:

- unsloth/Meta-Llama-3.1-8B-bnb-4bit – Selected for its superior balance of speed and accuracy, with 2x faster processing compared to other Llama-3.1 models.
- unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit – Optimized for instruction-based tasks.
- unsloth/Meta-Llama-3.1-70B-bnb-4bit – A larger variant designed for processing expansive datasets efficiently.
- unsloth/Llama-3.2 Series – Including instruction-optimized and scaled versions like Llama-3.2-3B-bnb-4bit.

After comprehensive evaluation, the Meta-Llama-3.1-8B-bnb-4bit model was selected as the primary choice for fine-tuning. Its 8 billion parameters provided an ideal tradeoff between computational efficiency and modeling capability, ensuring robust handling of complex patent texts. Moreover, its twofold speed advantage over comparable models facilitated rapid processing of our large dataset while maintaining high accuracy.

The integration of unsloth’s FastLanguageModel framework and four-bit optimization significantly enhanced the efficiency and precision of our fine-tuning process. The Meta-Llama-3.1-8B-bnb-4bit model excelled in processing intricate patent descriptions, achieving superior performance in green patent classification. The reduced computational overhead and faster training cycles allowed us to iteratively refine the model, ensuring it was well-aligned with the specific demands of identifying green technologies.

In conclusion, the combination of unsloth innovative tools and the Meta-Llama-3.1-8B-bnb-4bit model enabled the rapid deployment of an efficient and accurate green patent classification system. This approach not only streamlined the development process but also set a new standard for fine-tuning language models in specialized domains, paving the way for further advancements in AI-driven patent analysis.

2.7 Application of the Fine-Tuned Model

With the fine-tuned Y02T-enhanced model in place, we applied it to assess the green patent status of patents from our candidate datasets. This model, which has been specifically trained to recognize underrepresented green technologies, is now capable of identifying patents that may have been overlooked by traditional classification schemes. By fine-tuning the model to specialize in recognizing green technologies, it became adept at identifying patents that contribute to environmental sustainability.

The application of this fine-tuned model allowed us to significantly expand our green patent dataset. The model is now capable of capturing a more comprehensive picture of innovations that contribute to environmental sustainability across various industries. This is particularly important because many potentially impactful green technologies might not have been classified under traditional green patent categories.

By utilizing this model, we were able to identify these overlooked patents, thereby improving the accuracy and breadth of our dataset.

To deploy the model for practical use, we first ensured that the fine-tuned model and the tokenizer were correctly loaded into the system. The tokenizer is essential for converting patent descriptions into a format that the model can understand, while the fine-tuned model processes these inputs and generates a classification.

We chose to load the model using a configuration optimized for large-scale input, capable of handling the lengthy and detailed descriptions typical of patent documents. This is crucial, as patent descriptions often contain complex language and intricate technical details that need to be fully captured in order to make accurate classifications. Additionally, we configured the model to use 4-bit quantization, which reduces memory usage while maintaining high performance, making the model more efficient for inference tasks.

The system was then set to use a predefined prompt, designed to guide the model in assessing whether a patent qualifies as a "green patent." The prompt instructed the model to evaluate the patent based on its focus on environmentally sustainable innovations. The output from the model was standardized, consisting of two key elements:

1. **Decision:** The model outputs a simple "Yes" or "No" decision, indicating whether the patent qualifies as a green patent.
2. **Confidence Score:** The model provides a confidence score between 1 and 100, indicating the level of certainty in its decision. A higher confidence score reflects a stronger belief in the correctness of the classification.

This structured output ensures consistency in the model's assessments and provides additional insight into the reliability of the classification decisions. For example, a patent classified with a high confidence score provides a greater degree of certainty, while lower confidence scores may suggest the need for further review or refinement.

By applying the fine-tuned model, we were able to assess a larger number of patents and expand our green patent dataset. This automated approach significantly increased the efficiency of the classification process, enabling us to quickly analyze and classify patents on a much larger scale than manual methods would have allowed. Moreover, the confidence scores provided an additional layer of transparency, allowing us to evaluate the reliability of the model's classifications.

This step was vital in creating a more comprehensive and robust green patent dataset. It allowed us to capture a broader range of innovations, ensuring that our dataset reflected the latest advances in green technologies across various sectors. The fine-tuned model not only improved the speed of the classification process but also

enhanced the accuracy of identifying patents that align with environmental sustainability goals.

In conclusion, the application of the fine-tuned model was a crucial step in our process of expanding and refining the green patent dataset. By leveraging the model's specialized capabilities, we were able to classify patents with greater efficiency and accuracy, contributing to a deeper understanding of the innovations driving environmental sustainability.

3. Results and Validation

3.1 Results: Dataset Composition and Key Insights

This study introduces a comprehensive dataset of green patents developed to improve the classification and understanding of green transportation technologies. By leveraging fine-tuned Large Language Models (LLMs), the methodology enhances the existing Y02T classification system while uncovering previously overlooked patents. Key results include:

- High-confidence green patents: From the USPTO-defined Y02T dataset of 133,042 green patents, the model identified 107,382 high-confidence green patents, excluding 25,660 lower-confidence patents (19.3%), which were likely misclassifications.
- Previously unclassified green patents: The methodology discovered an additional 17,578 patents related to green transportation technologies that were not classified under the original Y02T scheme, representing a 13.2% increase in the dataset size.
- Expanded dataset size: Combining high-confidence green patents and newly identified patents resulted in an enhanced dataset of 124,960 patents. When including the original dataset for reference, the total expanded dataset encompasses 150,620 patents.
- Model accuracy: The fine-tuned LLM achieved an 88% accuracy rate when classifying newly identified green patents, significantly outperforming the USPTO Y02T classification and other commercial LLM benchmarks.

These results underscore the robustness of the methodology in refining existing classifications, uncovering underrepresented patents, and creating a more complete dataset for green technology research.

3.2 Validation: Dataset and Ground Truth

The dataset used in this study was carefully curated to ensure comprehensive evaluation of green and non-green patent classification models. It integrates the following components:

- USPTO-defined classifications: The baseline categorization provided by the United States Patent and Trademark Office, representing the standard Y02 classification system.

- Human annotations: Serving as the ground truth, these classifications were meticulously curated to provide a high-confidence reference for model evaluation.
- Predictions from GPT-4o-mini: Outputs from a smaller, state-of-the-art language model.
- Predictions from GPT-4o: Outputs from a larger, more advanced language model.
- Predictions from FT-LLM: Results from the fine-tuned large language model developed specifically for this study.

The dataset includes patents classified as either green or non-green, with significant challenges arising from overlapping features between categories and the nuanced language often present in patent descriptions. Human annotations provided a reliable benchmark for assessing the adaptability and precision of the models, making them crucial for validating performance metrics.

3.3 Model Performance and Comparative Analysis

To evaluate the models, key metrics such as accuracy, precision, recall, F1 score, and confusion matrices were employed. The table below provides a summary of the performance metrics for each model:

Table1 Model Performance

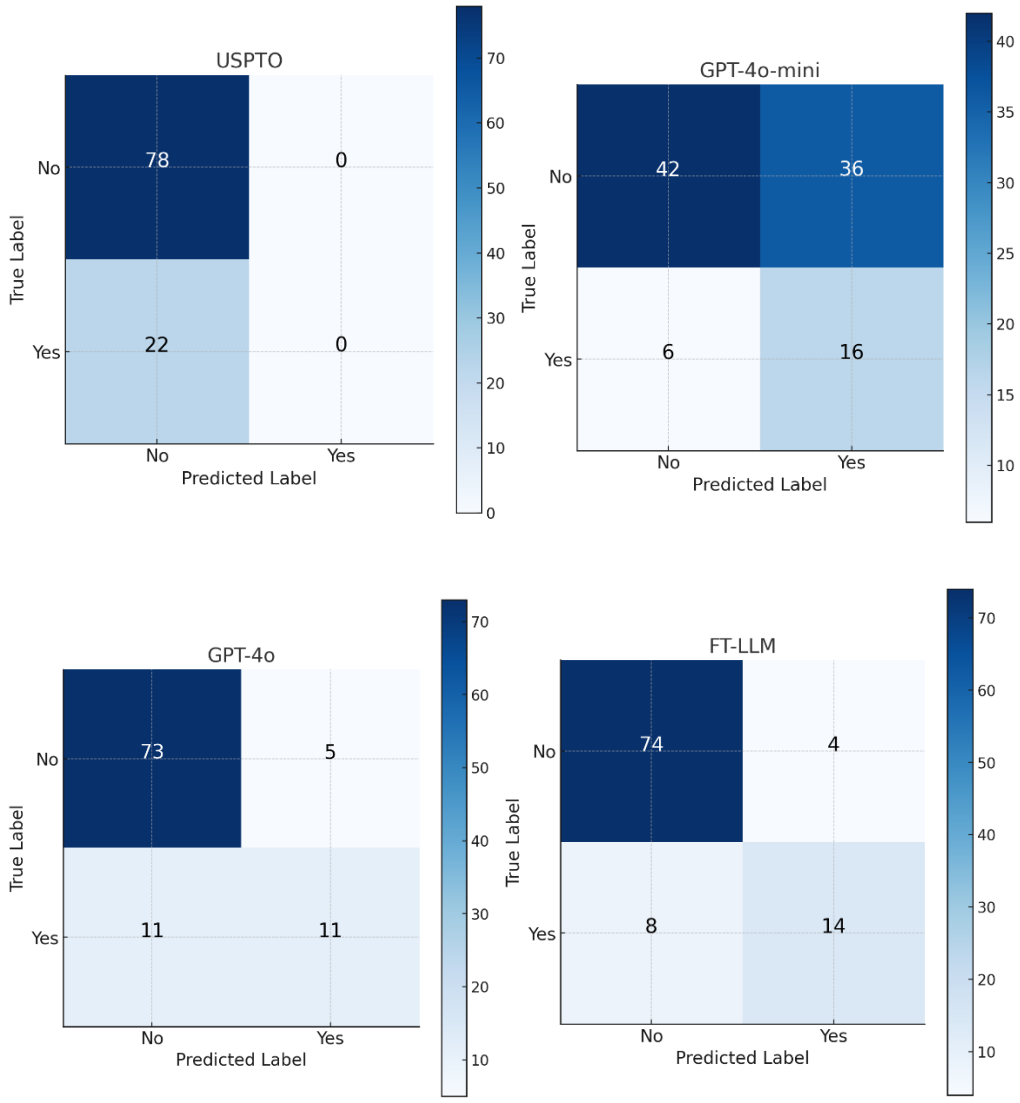
Model	Accuracy	Precision	Recall	F1 Score
USPTO	0.78	0.00	0.00	0.00
GPT-4o-mini	0.58	0.31	0.73	0.43
GPT-4o	0.84	0.69	0.50	0.58
FT-LLM	0.88	0.78	0.64	0.70

- USPTO Baseline: While achieving a reasonable accuracy of 0.78, the USPTO baseline failed entirely in precision, recall, and F1 score, each registering 0. This result highlights the rigidity of rule-based systems, which rely heavily on true negatives and lack the flexibility to handle the complex and nuanced features of patent datasets.
- GPT-4o-mini: This model emphasized recall (0.73), reflecting its ability to identify green patents effectively. However, its low precision (0.31) resulted in a high rate of false positives, where non-green patents were misclassified as green. This imbalance led to an F1 score of 0.43, limiting its utility for tasks requiring both precision and recall.
- GPT-4o: Achieved strong precision (0.69) and accuracy (0.84), but its recall (0.50) was comparatively lower. This reflects a conservative classification approach that reduced false positives but missed true positives. The moderate F1 score of 0.58 illustrates this trade-off, making GPT-4o suitable for applications

requiring moderate precision and recall but less effective for nuanced classifications.

- FT-LLM: The fine-tuned model outperformed all others, achieving the highest accuracy (0.88), precision (0.78), recall (0.64), and F1 score (0.70). These results highlight the effectiveness of domain-specific fine-tuning, enabling the model to balance precision and recall while adapting to the complexities of green patent data. The FT-LLM demonstrated the highest number of true positives with low false positives and negatives, underscoring its adaptability and superior performance.

Figure 1: Confusion Matrix Analysis



The confusion matrices reveal critical classification behaviors across the models:

- USPTO Baseline: Relied heavily on true negatives and failed to classify any true positives, indicating an inability to adapt to complex patent data.
- GPT-4o-mini: Showed a high number of false positives, boosting recall at the expense of precision.
- GPT-4o: Achieved a more balanced matrix with fewer false positives but introduced false negatives, reflecting a cautious classification approach.
- FT-LLM: Outperformed other models by achieving the highest number of true positives while maintaining low false positive and false negative rates. This result highlights its adaptability to nuanced and overlapping patent features.

The FT-LLM demonstrated remarkable strengths, making it the most effective model in the study. By fine-tuning on green patent data, the FT-LLM exhibited exceptional domain adaptability, capturing intricate linguistic patterns unique to green technologies that generic models failed to recognize. Its ability to balance precision and recall resulted in the highest F1 score (0.70), reflecting its capacity to minimize classification errors while maintaining robustness. Additionally, the scalability of its architecture and methodology allows for broader application across diverse patent domains, underscoring its versatility and value as a classification tool.

However, despite its strong performance, the FT-LLM has areas for improvement. Its recall, at 0.64, indicates a need to enhance sensitivity to true positives, particularly for green patents with more ambiguous descriptions. Furthermore, the computational demands of fine-tuning large models like the FT-LLM pose challenges for smaller-scale implementations, as the resource intensity may limit accessibility to users with constrained computational infrastructure. Addressing these limitations could further enhance the model’s utility and broaden its applicability.

3.4 Implications and Conclusion

The superior performance of the FT-LLM has far-reaching implications for green innovation. By accurately identifying green patents, the model provides a powerful tool for researchers, investors, and policymakers to track technological advancements, promote sustainable development, and drive informed decision-making in green finance. Its ability to uncover underrepresented technologies further enhances its value, offering insights that could otherwise be overlooked by traditional methods.

Despite its strengths, optimizing recall and reducing the computational intensity of fine-tuning would further broaden the model’s applicability. Nevertheless, the FT-LLM sets a new benchmark for AI-driven patent classification, achieving unmatched accuracy and reliability. These results validate the efficacy of domain-specific fine-tuning and highlight its transformative potential for advancing sustainability-focused research and policy development.

In conclusion, the FT-LLM represents a significant step forward in leveraging AI for green patent classification. Its robust performance and adaptability underscore its critical role in ensuring green technologies are identified and utilized effectively, contributing to global sustainability efforts.

4. A Robust Framework for Transformative Green Patent Discovery

The methodology presented in this study extends far beyond being a mere classification tool; it is a robust framework capable of transforming how green patents are discovered, understood, and applied. By integrating adaptability, scalability, and cutting-edge technology, this approach establishes itself as a cornerstone for future exploration of green innovation. This chapter delves deeper into the multifaceted ways this framework demonstrates its robustness, from expanding into diverse green domains to adapting to advancements in AI and hardware, and even empowering global accessibility through portable applications.

4.1 Unlocking New Frontiers in Green Innovation

The robust nature of this framework lies in its ability to transcend a single focus area, such as green transportation, and seamlessly expand into other domains of green technology. Green energy, with its wide range of subfields like renewable energy sources, energy storage solutions, and grid integration technologies, is a prime candidate for this methodology. Similarly, the framework can be applied to green building innovations, identifying patents related to energy-efficient designs, sustainable construction materials, and smart building technologies.

Green biology represents another frontier, encompassing patents in biofuels, sustainable agricultural practices, and biodegradable materials. This versatility ensures that the framework keeps pace with the rapidly evolving landscape of sustainability-focused innovations, enabling stakeholders to uncover transformative ideas across diverse industries. With each expansion, the framework’s capacity to identify untapped opportunities and generate actionable insights only grows, reinforcing its role as an indispensable tool for green innovation.

4.2 Customizable Frameworks for Deeper Insights

At its core, this framework is designed to evolve and respond to specific research or industry needs. By modifying the input data and training parameters, it can be tailored to capture increasingly nuanced aspects of green technologies. For example, even within green transportation, additional CPC classifications—such as those focusing on autonomous electric vehicles, hydrogen-powered engines, or smart logistics systems—can be incorporated to broaden the scope and precision of the analysis.

This customization extends beyond expanding categories to refining the granularity of analysis. Researchers investigating emerging materials in renewable energy or investors targeting circular economy innovations can adjust the model to prioritize these areas. This modularity ensures that the framework remains relevant and robust,

capable of addressing both broad trends and highly specific objectives in green technology research.

4.3 Staying at the Cutting Edge of AI Advancements

The framework’s robustness is further demonstrated by its ability to integrate seamlessly with advanced closed-source LLMs for specific tasks, such as identifying high-confidence green patents and generating detailed explanations. As more sophisticated proprietary models emerge, this methodology can incorporate these tools to further enhance accuracy, precision, and interpretability. This ensures that the framework remains at the forefront of AI-driven research, setting new benchmarks in green patent classification.

Moreover, the ability to integrate state-of-the-art models facilitates continuous improvement. As closed-source LLMs evolve to handle more complex linguistic structures or larger datasets, the framework can capitalize on these advancements, significantly improving its ability to identify, classify, and contextualize patents with unparalleled accuracy. This forward-looking design ensures that it evolves alongside technological advancements, maintaining its position as a leader in green patent discovery.

4.4 Future-Proofed with Open-Source Innovations

The framework's reliance on open-source LLMs for inference adds another layer of robustness, enabling it to leverage advancements in both model development and hardware capabilities. As new open-source models with improved efficiency and accuracy are released, they can be seamlessly integrated into the system. For example, transitioning to models with larger token windows or improved understanding of domain-specific language can significantly enhance performance.

Likewise, advancements in hardware—such as next-generation GPUs, TPUs, or energy-efficient computing platforms—further optimize the framework’s efficiency and scalability. These hardware improvements not only reduce computational costs but also enable the system to process larger datasets and deliver faster results. This ensures the framework remains future-proof, capable of adapting to technological advancements and continuously pushing the boundaries of green patent classification.

4.5 Green Innovation in Your Pocket: Portable and Accessible

One of the most transformative aspects of this robust framework is its potential to become portable, making it accessible to a global audience. By deploying the developed LLM on mobile devices, the framework can empower researchers, policymakers, and industry professionals with real-time green patent discovery capabilities. Imagine a policymaker at an international conference or an entrepreneur

in a remote region having access to state-of-the-art green patent analysis directly from their smartphone.

This portability democratizes access to advanced AI-driven tools and ensures that the methodology can be utilized in diverse environments, from urban research centers to rural innovation hubs. Furthermore, the open-source nature of the framework guarantees that it remains free of charge, breaking down barriers to entry and encouraging widespread adoption. This level of accessibility transforms the framework into a global resource for advancing green innovation and fostering sustainable development.

5: Contributions and Future Pathways for Green Innovation

This chapter highlights the significant contributions of this research to the academic world, business and investment communities, and policymakers. Furthermore, it explores how the integration of Large Language Models (LLMs) is shaping the future of innovation research and paving new pathways for advancements in green technology.

5.1 Contribution of the Research

The methodology and findings presented in this study represent a major leap forward in the classification and analysis of green patents. By combining advanced LLMs with dynamic and customizable frameworks, this research offers a more accurate, adaptable, and comprehensive tool for identifying and analyzing green technologies. Unlike traditional classification systems, this approach uncovers underrepresented patents, highlights gaps in innovation, and ensures that emerging technologies receive the recognition they deserve.

This study not only addresses the limitations of existing systems but also introduces a replicable and scalable framework for future green innovation research. The public availability of the dataset and tools further underscores its contribution, fostering collaboration and enabling stakeholders across academia, industry, and government to benefit from its insights.

5.2 Advancing Academic Research

For the academic community, this research serves as a valuable resource for studying the evolution of green technologies and their broader implications. The dataset provides a detailed foundation for analyzing technological trends, exploring interdisciplinary connections, and understanding the role of green patents in addressing environmental challenges.

The methodology itself offers a replicable model for other domains, encouraging the academic world to adopt advanced AI techniques for patent analysis. Scholars can use this framework to investigate green innovation within specific sectors, such as renewable energy, sustainable agriculture, or energy-efficient transportation. Furthermore, the ability to customize the methodology ensures its relevance to diverse research questions, from evaluating the economic impact of green technologies to assessing their scalability and adoption.

By bridging the gap between AI and sustainability research, this study fosters interdisciplinary collaboration, paving the way for new insights into the intersections of technology, policy, and environmental science.

5.3 Empowering Businesses and Investors

For businesses and investors, this research offers a powerful tool for making strategic decisions in the rapidly evolving green technology market. The dataset provides detailed insights into emerging trends, allowing stakeholders to identify promising technologies and assess their potential impact. For example, venture capitalists and private equity firms can leverage this tool to evaluate startups or established companies innovating in green technologies, such as energy storage, waste-to-energy solutions, or carbon capture systems.

The dynamic nature of the methodology ensures that investors stay ahead of market trends, enabling them to allocate resources to areas with the highest potential for growth and sustainability impact. Businesses, on the other hand, can use the insights to align their innovation strategies with market needs, ensuring their products and services remain competitive and relevant. The combination of accurate data, customizable models, and actionable insights positions this research as a critical resource for driving green investment and innovation.

5.4 Informing Policymaking for Sustainable Development

Policymakers play a pivotal role in shaping the regulatory and financial landscape for green technologies. This research equips them with a robust tool to evaluate the effectiveness of existing policies, identify areas requiring additional support, and direct funding toward impactful projects. By analyzing the dataset, policymakers can assess technological progress in specific sectors, such as renewable energy or sustainable transportation, and determine which innovations align with national or global sustainability goals.

The ability to identify underrepresented technologies further enhances the utility of this research. Policymakers can use these insights to design targeted incentives for emerging green innovations, ensuring that critical but overlooked areas receive the attention and funding they need. Moreover, the contextual explanations provided by LLMs make it easier for policymakers to understand the significance of specific patents, enabling data-driven decisions that promote long-term environmental and economic benefits.

5.5 LLMs Shaping the Future of Innovation Research

The integration of Large Language Models (LLMs) into this study represents a transformative shift in how innovation research is conducted. LLMs, with their ability to process vast amounts of textual data and extract nuanced insights, have opened new frontiers in patent analysis. By fine-tuning these models for green patent classification, this research demonstrates their potential to uncover hidden patterns, provide detailed explanations, and enhance interpretability.

Looking ahead, LLMs are poised to play an even greater role in shaping the landscape of innovation research. Their adaptability makes them ideal for exploring new domains, integrating interdisciplinary data, and generating predictive insights about future technological trends. As LLMs continue to evolve, their applications in green innovation will expand, enabling researchers to tackle increasingly complex questions about sustainability, scalability, and impact.

The future path for this research lies in further enhancing the synergy between AI and innovation studies. By integrating LLM-based patent analysis with other data sources, such as market adoption rates and environmental impact assessments, researchers can create a more holistic understanding of green technology ecosystems. This multidisciplinary approach will not only advance academic inquiry but also support practical applications, from designing better policies to accelerating the adoption of transformative green technologies.

Acknowledgments

The authors gratefully acknowledge the financial support provided by assistant professor research starting funding from Shanghai University of Finance and Economics. This research was also supported by resources provided by PBC School of Finance, Tsinghua University at its early stage. The authors declare that they have no financial relationships or other conflicts of interest that could influence the interpretation or outcome of this study. All opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the supporting organizations.

References:

Angelucci, S., Hurtado-Albir, F.J., Volpe, A., 2018. Supporting global initiatives on climate change: The EPO's "Y02-Y04S" tagging scheme. World Pat. Inf. 54, S85–S92. https://doi.org/10.1016/j.wpi.2017.04.006

Cohen, L., Gurun, U.G., Nguyen, Q.H., 2020. The ESG-innovation disconnect: Evidence from green patenting. National Bureau of Economic Research.

Dechezleprêtre, A., Glachant, M., Haščič, I., Johnstone, N., Ménière, Y., 2011. Invention and transfer of climate change–mitigation technologies: A global analysis. Rev. Environ. Econ. Policy 5, 109–130. https://doi.org/10.1093/reep/req023

Haščič, I., Migotto, M., 2015. Measuring environmental innovation using patent data.

Lerner, J., Seru, A., Short, N., Sun, Y., 2021. Financial innovation in the 21st century: Evidence from US patents. National Bureau of Economic Research.

Shu, T., Tian, X., Zhan, X., 2022. Patent quality, firm value, and investor underreaction: Evidence from patent examiner busyness. J. Financ. Econ. 143, 1043–1069. https://doi.org/10.1016/j.jfineco.2021.10.013

Tian, X., Wang, T.Y., 2014. Tolerance for failure and corporate innovation. Rev. Financ. Stud. 27, 211–255. https://doi.org/10.1093/rfs/hhr130

Toole, A., Jones, C., Madhavan, S., 2021. Patentsview: An open data platform to advance science and technology policy. https://doi.org/10.2139/ssrn.3874213

V. Veefkind, J. Hurtado-Albir, S. Angelucci, K. Karachalios, N. Thumm, 2012. A new EPO classification scheme for climate change mitigation technologies. World Pat. Inf. 34, 106–111. https://doi.org/10.1016/J.WPI.2011.12.004

Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y., 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. https://doi.org/10.48550/arXiv.2403.13372