From Critical Minerals to Circularity: Open Patent Classifiers for Material-Relevant and Circular Innovation

Fabrizio Fusillo, Martina Iori, Gianluca Orsatti, Francesco Quatraro

Extended abstract

1. Motivation and contribution

Critical minerals are increasingly central to technological change, industrial policy, the energy transition, and economic security. Yet patent data, as conventionally classified, provide no systematic information on the material inputs inventions rely upon. Standard IPC and CPC classifications identify technical fields and technological functions, but they do not detect whether an invention relies on lithium, cobalt, rare earth elements, graphite, gallium, or other strategically important materials. Keyword searches are an imperfect substitute: they miss inventions in which the relationship to a material is implicit, while at the same time generating false positives when a material is mentioned incidentally or in a technologically irrelevant context. Those errors do not occur at random, as they might depend on the specific technology and/or the patent applicant.

This measurement problem is particularly binding because technological responses to critical-mineral dependence do not operate only through substitution between materials or changes in primary extraction. They may also occur through circular strategies, namely recycling and recovery of valuable materials, extending the useful life of products and components, recirculating industrial residual streams, or reducing the requirement for virgin inputs. Measuring these responses requires going beyond conventional patent classifications in a second dimension: identifying inventions that embody circular-economy mechanisms even when neither "recycling" nor "circular economy" appears explicitly in the patent text.

We are developing an open patent data infrastructure that addresses these two related measurement problems and consists of two classification layers. The first one identifies material-specific technological relevance for Critical Raw Materials (CRMs) in U.S. patent data. The second identifies Circular-Economy (CE) technologies in U.S. and European patents[1] and further distinguishes the circularity mechanisms embodied by the invention. The two layers rely on complementary forms of machine-assisted supervision. CRM labels are initialized by matching curated material dictionaries against patent texts (i.e., titles, abstracts, and brief summaries) and then generalized by material-specific classifiers, fine-tuned on PatentSBERTa embeddings, a sentence-transformer representation pre-trained on patent texts (Bekamiri et al., 2024). CE labels are generated by a multi-agent large-language-model annotation procedure and then used as training data for a less costly patent-domain transformer that can be applied to the whole corpus.

The resulting architecture is intended as research infrastructure rather than as a stand-alone empirical application. Once released, researchers will be able to use individual layers separately or intersect them at the patent level, for example, to identify innovation related to lithium together with recycling or recovery, or rare-earth-related inventions involving material recirculation. The broader methodological contribution is a reproducible approach for constructing innovation measures in domains for which existing technological classifications are incomplete: high information but relatively costly labeling is used to train scalable open classifiers that can subsequently be applied to the universe of patents.

2. Identifying critical-mineral innovation

The first classification layer addresses material-specific technological content by considering granted USPTO patents from 1975 onward, for which titles, abstracts, and brief summaries are available through

[1] The asymmetry in coverage is a consequence of the availability in patent textual information as the CRM classification exploits the USPTO brief-summary field, which has no direct counterpart for EPO patents, whose extension is a work in progress.

EPO PATSTAT and PatentsView databases. References to 34 CRM and material groups, defined on the EU and the USGS lists of CRMs, are detected in texts through regular expressions designed to avoid spurious string matches and to recognize relevant chemical forms - for example, variants such as titania for titanium and hafnia for hafnium. Mining patents are excluded from the positive population because the objective is to identify technological uses and applications of critical materials rather than simply inventions concerning their extraction. This procedure produces a seed population of 1,426,400 patents containing at least one detectable CRM reference. The distribution across materials is strongly heterogeneous, creating a long tail in which a relatively small number of materials occur frequently, and many others appear much less often.

Rather than treating materials identification as a single multi-class problem, we estimate an independent binary classifier for each of 34 materials or material groups. Each classifier is obtained by fine-tuning a patent-domain sentence encoder using the binary classification labels from the previous step. This design allows imbalance correction, hyperparameters, and classification thresholds to vary across materials. Each classifier is trained on several examples. Positive observations are patents associated with the target material. Negative observations include year-stratified patents without detected CRM references, patents referring to other CRMs, and mining patents that mention the material but relate to extraction rather than downstream technological use. Other CRM negatives are sampled at 20 percent of the positive population. The mining cases serve as particularly useful "hard negatives" because they contain the target material vocabulary but do not represent the technological phenomenon we want the classifier to identify.

Each material-specific dataset is divided into 70 percent training, 15 percent validation, and 15 percent test observations. We fine-tune PatentSBERTa rather than a general-language classifier, incorporating class weighting in training, automated hyperparameter optimization with Optuna, and material-specific decision-threshold calibration on the validation set. This is important because the frequency of positive cases differs across materials, making a common threshold inappropriate.

Held-out performance is stable across materials of different frequencies. Test-set F1 scores range from approximately 0.80 to 0.97: for example, 0.89 for lithium, 0.88 for rare-earth elements, 0.88 for platinum-group metals and cobalt, 0.88 for graphite and gallium, 0.90 for manganese, 0.91 for niobium, and 0.97 for phosphate rock and coking coal.

The design is weakly supervised, as the initial positive labels are derived from explicit material references. High held-out performance is therefore not necessarily evidence that the model successfully identifies implicit material relationships beyond the dictionary used to generate the labels. We are consequently implementing a second validation layer. It uses patents in which CRM references occur in the description but not the abstract, citation-based proximity to known CRM inventions, and portfolio-expansion tests among firms with established CRM-related inventive activity. This validation exercise is an important work in progress on which feedback from the I3 community would be particularly valuable.

3. From critical-mineral dependence to circular innovation

Material-specific classifiers identify where critical materials enter the technological system, but they do not provide indications on how the invention employs the material, nor whether it alters the material cycle. The same material might enter inventions that raise primary demand for it and inventions that recover it from end-of-life products.

We therefore construct a second classification infrastructure for circular-economy related technologies. Dictionary-based weak supervision performs poorly in this case because circularity is a property of the process a patent describes, rather than a readily identifiable technical vocabulary. Terms such as "waste," "efficiency," or "recycling" can occur in inventions that do not implement a circular process, while circular mechanisms can be described without using any standard circular-economy terminology.

The design combines a large language model, used as a high-information classifier on a subset of data, with a patent-domain transformer used as the scalable classifier. We first draw a stratified random sample of 260,000 U.S. and European patents, allocating observations proportionally across filing-year-by-patent-authority strata. A fixed random seed makes the sampling procedure reproducible. The LLM annotation stage labels 7,576 patents in this sample as circular-related (2.9% of documents).

The annotation procedure constructs, for each patent, a compact representation of the object of invention from the title and abstract, retrieves relevant CE examples as few-shot evidence, and asks several specialized annotator prompts to independently assess the underlying technological mechanism. A final arbiter combines their judgments. Indeed, generic environmental or sustainability language is not sufficient. A positive label requires explicit evidence of a mechanism, such as repair or life extension, reuse or remanufacturing, recycling or material recovery, take-back, use of secondary materials, or another identifiable process that preserves or recirculates resource value. The procedure also assigns circularity-strategy labels derived from the 9R framework (Potting et al., 2017), which orders circularity strategies from rethinking the need for a product, through reducing, reusing, repairing, refurbishing, remanufacturing, and repurposing, to recycling and energy recovery.

This labeling stage is not intended to be run over the entire patent universe. Instead, its output becomes training data for a supervised patent-domain transformer. The first PatentSBERTa classifier achieved an F1 score of 0.727, a precision of 0.725, and a recall of 0.728. Applying it corpus-wide identified 408,022 CE patents. Subsequent validation revealed a systematic tendency toward false positives on which the aggregate accuracy statistics were silent.

We therefore deliberately searched for difficult cases and fed disagreement information back into training. Two thousand potentially problematic BERT-positive patents were re-evaluated by the LLM procedure, identifying 307 BERT-positive/LLM-negative cases. Together with previous disagreements, these produced 327 corrective observations; 1,658 cases on which BERT and the LLM agreed positively were also added. After deduplication, the refined training dataset contains 261,966 observations, of which 3.52 percent are CE positive. The refined classifier achieves F1 = 0.759, precision = 0.719, recall = 0.804, accuracy = 0.982, and ROC-AUC = 0.985 on its evaluation sample.

Applied to the full USPTO/EPO population for 1975-2022, the refined model identifies 464,160 CE patents. External coherence checks show that it captures all 34,536 patents with a Y02W waste-management/environmental-treatment CPC code, while also extending beyond that relatively narrow taxonomy. The leading technological classes of the resulting population include separation and filtration, wastewater treatment, environmentally oriented production technologies, catalytic processes, and energy-related environmental technologies.

Additional LLM validation on two independent samples of 1,000 patents yields an F1 of 0.836 for 2012-2022 patents and 0.770 for the full 1975-2022 period, with recall above 93 percent in both exercises. The remaining error remains asymmetric: the BERT classifier tends to classify more patents as circular than the LLM benchmark does. The released documentation will report threshold-specific precision and recall, as well as the confusion matrices for both validation samples.

4. Measuring the type of circularity

A binary CE indicator is insufficient for most research questions concerning material dependence. Recycling and material recovery are quite different technological responses, from prolonging the life of a product to reducing material requirements at the design stage.

Therefore, the annotation prompt also identifies the 9R strategies the invention implements for each patent labeled as circular. Because individual R categories are unevenly represented and some distinctions are difficult to infer reliably from abstracts, we aggregate them into three economically meaningful groups: G1 - smarter product use and manufacture (Refuse, Rethink, Reduce); G2 - extension of product and component lifetimes (Reuse, Repair, Refurbish, Remanufacture, Repurpose); and G3 - useful application of materials (Recycle, Recover). The grouping corresponds approximately to

narrowing, slowing, and closing resource loops and avoids distinctions that patent abstracts do not support.

Separate PatentSBERTa classifiers are trained for each group, preserving the multi-label nature of circularity as each patent can implement more than one strategy. Held-out F1 is 0.65 for G1, 0.78 for G2, and 0.82 for G3. Applying the classifiers to the 464,160 CE patents assigns 169,120 patents to G1, 214,339 to G2, and 252,324 to G3; only 1.2 percent of CE patents remain below all three decision thresholds.

This layer is particularly important for the connection to critical minerals. Intersecting material classifications with G3, for example, can identify inventions combining specific critical materials with recycling or recovery mechanisms, while G2 can capture technologies that may reduce material demand by prolonging the service life of material-intensive equipment.

5. Research opportunities and value to the innovation community

The combined data architecture opens new opportunities to study how material constraints shape the direction, organization, and geography of inventive activity. Its contribution is to connect material-specific technological relevance with the circular mechanisms embodied in inventions. Researchers can use these dimensions separately or jointly and link them to assignees, inventors, citations, technology classes, and locations. The resulting measures complement existing environmental patent classifications, including CPC Y02 (Angelucci et al., 2018) and OECD ENV-TECH (Haščič and Migotto, 2015), by providing a systematic material-specific dimension and distinguishing different circularity strategies.

A first application concerns directed technological change. Do firms exposed to critical-mineral supply disruptions or price increases expand inventive activity within existing technological trajectories, or redirect it toward resource efficiency, product-life extension, and recycling or recovery? For example, within lithium-related patent portfolios, researchers could examine whether inventive activity shifts toward technologies that prolong battery service life or recover materials from spent batteries. Linking the classifications to firms' pre-existing technological portfolios and external measures of material exposure would help distinguish changes in the volume of inventive activity from changes in its composition. Such analyses could inform research on induced innovation and adjustment to resource constraints, with causal interpretation requiring an appropriate identification strategy.

A second application concerns the capabilities underlying circular innovation. Do material-related circular inventions originate among firms already specialized in the relevant material, established recycling businesses, or entrants from other technological domains? The material classifiers can characterize firms' prior technological specialization, while the circularity classifiers identify diversification into different circular mechanisms. Combining these measures with inventor histories and citation networks would allow researchers to examine whether such diversification builds on internal capabilities, recruitment, or connections to external knowledge. Comparing G1, G2, and G3 could further reveal whether different circularity strategies draw on different knowledge bases and involve different patterns of entry.

A third application concerns the geography and diffusion of these capabilities. Researchers could investigate whether circular inventive activity develops in regions with established material-related expertise or emerges in new locations, and trace the knowledge connections between them. These analyses would measure the location and diffusion of inventive activity; studying subsequent adoption, production, or changes in material dependence would require additional data.

The project also creates a useful setting for methodological work on innovation measurement. We plan to release model scores alongside binary labels so that users can choose thresholds appropriate to their application, conduct sensitivity analysis, and explicitly study the consequences of classification uncertainty. The availability of keyword seeds, LLM labels, transformer outputs, and external validation signals within the same infrastructure can also support comparisons of alternative approaches to text-based technology measurement. In this sense, the project is intended not only to produce two new patent

indicators but also to provide a transparent benchmark for studying how weak supervision, LLM-assisted labeling, and domain-specific language models can be combined in economic research.

Finally, the modular design is meant to support extension by other researchers. Additional material categories, alternative circularity definitions, or new patent jurisdictions can be incorporated without rebuilding the entire system. We see the I3 community as particularly valuable for determining which components should be standardized as shared infrastructure, which uncertainty measures are most useful to downstream empirical researchers, and which validation exercises would make the released data credible across different applications in innovation, productivity, trade, and regional economics.

6. Planned open data and dissemination

The central output of the project will be a versioned, openly documented research resource. Subject to the redistribution restrictions attached to the underlying patent text, we plan to release patent identifiers and derived variables, rather than proprietary full-text data, as patent-level CSV and Parquet tables that can be merged with PatentsView, EPO PATSTAT, or Google Patents. The public release will include material-specific CRM classifications and model scores; the binary CE classification and scores; the G1-G3 circularity annotations; training and validation samples where licensing permits; material dictionaries and taxonomy documentation; LLM prompts and aggregation rules; scripts for sampling, training, threshold calibration, inference, and validation; and model cards documenting known error modes and population coverage.

We also plan to make the fine-tuned transformer checkpoints available where the underlying model's license permits. The release will be versioned so that subsequent improvements do not change historical classifications. Because U.S. and European policy taxonomies do not coincide perfectly, an additional priority for the public version is a transparent crosswalk between the material categories used in the current CRM classifiers and U.S. critical-mineral definitions, together with procedures for adding or retraining material classifiers as official lists change.

The presentation at the I3 Technical Working Group would focus on data construction and validation, and in particular on the three steps that can be scaled up to other measurement or classification problems, i.e., the conversion of dictionary-based weak supervision into material-specific classifiers, the use of an LLM as an annotator whose labels train a scalable patent-domain model, and the disagreement analysis that exposed and reduced over-classification in the CE model.

We would particularly welcome feedback on three issues before the public release: external validation strategies for implicit critical-mineral relevance; harmonization between European CRM and U.S. critical-mineral taxonomies; and the appropriate balance between precision and recall when releasing general-purpose patent classifications intended to support heterogeneous research applications. The project is deliberately at a stage at which community feedback can influence both the final validation design and the structure of the open data release.

Acknowledgements. This research was supported by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (D13C22002160001). The authors declare that they have no financial relationships or other potential conflicts of interest relevant to this work.

References:

Angelucci, S., Hurtado-Albir, F. J., & Volpe, A. (2018). Supporting global initiatives on climate change: The EPO's "Y02-Y04S" tagging scheme. *World Patent Information*, *54*, S85-S92.

Bekamiri, H., Hain, D. S., & Jurowetzki, R. (2024). PatentsBERTa: A deep nlp based hybrid model for patent distance and classification using augmented sbert. *Technological Forecasting and Social Change*, *206*, 123536.

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

Potting, J., Hekkert, M. P., Worrell, E., & Hanemaaijer, A. (2017). Circular economy: measuring innovation in the product chain.