RESEARCH ARTICLE

Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents

Sergio Pelaez[1*], Gaurav Verma[2*], Barbara Ribeiro[3,4], and Philip Shapira[1,4]

[1]School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[3]SKEMA Business School, Université Côte d'Azur, Campus Grand Paris, Paris, France
[4]Manchester Institute of Innovation Research, University of Manchester, Manchester, UK
*Co-first authors.

Keywords: AI patents, generative language models, GPT-4, large-scale classification, public value, text labeling

ABSTRACT

We put forward a novel approach using a generative language model (GPT-4) to produce labels and rationales for large-scale text analysis. The approach is used to discover public value expressions in patents. Using text (5.4 million sentences) for 154,934 US AI patent documents from the United States Patent and Trademark Office (USPTO), we design a semi-automated, human-supervised framework for identifying and labeling public value expressions in these sentences. A GPT-4 prompt is developed that includes definitions, guidelines, examples, and rationales for text classification. We evaluate the labels and rationales produced by GPT-4 using BLEU scores and topic modeling, finding that they are accurate, diverse, and faithful. GPT-4 achieved an advanced recognition of public value expressions from our framework, which it also uses to discover unseen public value expressions. The GPT-produced labels are used to train BERT-based classifiers and predict sentences on the entire database, achieving high F1 scores for the 3-class (0.85) and 2-class classification (0.91) tasks. We discuss the implications of our approach for conducting large-scale text analyses with complex and abstract concepts. With careful framework design and interactive human oversight, we suggest that generative language models can offer significant assistance in producing labels and rationales.

1. INTRODUCTION

Supervised machine learning (ML) relies on high-quality training data labeled by humans. However, the process of obtaining and validating human annotations, especially for complex and abstract concepts, is often overlooked and underemphasized in ML research and education. This can lead to costly and unreliable data that affect the performance and validity of ML models (Geiger, Yu et al., 2020). To address this gap, this paper develops a novel approach to aid the labeling of complex and abstract concepts using generative language models (GLMs). Although our method is relevant for varied disciplines and studies undertaking text mining and content analysis, we demonstrate its application in the context of science and innovation policy analysis. Specifically, we put forward a semiautomated approach to identifying public

an open access journal

Citation: Pelaez, S., Verma, G., Ribeiro, B., & Shapira, P. (2024). Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents. *Quantitative Science Studies*, *5*(1), 153–169. https://doi.org/10.1162/qss_a_00285

DOI: https://doi.org/10.1162/qss_a_00285

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00285

Supporting Information: https://doi.org/10.1162/qss_a_00285

Received: 21 May 2023
Accepted: 21 November 2023

Corresponding Author:
Philip Shapira
pshapira@manchester.ac.uk

Handling Editor:
Vincent Larivière

values associated with inventions in artificial intelligence (AI) through the analysis of text in AI patent documents.

Previous quantitative research in science and innovation policy has focused on the use of bibliometric and text-mining methods of sources that include scientific publications and patents. Trends and patterns in research and innovation performance have been discerned using a range of text processing and content analysis techniques (Antons, Grünwald et al., 2020; Porter & Cunningham, 2004). More recently, studies have used Bidirectional Encoder Representations from Transformers (BERT)–based or similar techniques to categorize research documents according to their respective subject areas based on abstract texts. Examples include tagging publications by their discipline and keywords (Färber & Ao, 2022), identifying AI patent documents (Giczy, Pairolero, & Toole, 2022) and research papers (Sachini, Sioumalas-Christodoulou et al., 2022), and obtaining multiple category label predictions for social science journal articles (Eykens, Guns, & Engels, 2021). Seed and antiseed sets of examples of labels for different categories are required to build a model using these methods. These methods have been applied to large-scale data sets. Qualitative methods have also been used to code and classify unstructured text of scientific publications, patents, and other innovation policy documents (Ribeiro & Shapira, 2020), providing nuanced analyses but with human limitations on the volume of data that can be analyzed.

In this study, we deploy an approach that uses a recently available GLM to analyze unstructured patent text. In our case, we apply the approach to discover and classify text in AI patents that conveys attention to public values. It is a semiautomated approach that involves human input and review to identify concepts that are complex and context dependent but uses a GLM and ML to accelerate and scale up classification processes. The approach addresses the shortcomings of human-based labeling, such as its high demands on time and resources, yet also supports collaborative annotation and enhances reflexivity in abductive research.

1.1. Public Value Expressions in Patent Documents

In public policy and administration, public values (PVs) have been characterized as enduring beliefs that are founded on an ideal of human society. As a result, they offer direction, meaning, and legitimacy to collective action (Rutgers, 2015). This concept has been put forward as an improvement over the traditional notion of public interest, which was viewed as ambiguous and lacking practicality as a guide. It was also proposed as an alternative theory to the market failure framework, which was deemed to offer a narrow and sometimes contradictory policy justification. In this sense, PVs are portrayed to be more concrete and practical relative to the idea of public interest while providing broader possibilities for policy deliberation and justification relative to the market failure framework (Bozeman, 2002).

Efforts to identify PVs have involved investigating, for example, scholarly literature, government documents, cultural artifacts, or opinion polls (Bozeman & Sarewitz, 2011). We frame the written articulation of a PV as a public value expression (PVE) where it indicates societal benefits that are promised to or for people, organizations, or ecosystems. A PVE is a signal—it does not mean that the PV will necessarily be realized but it does suggest that there is an idea or intent to do so. Discovering and analyzing such signals, in the context of science and innovation policy, helps understand the potential pathways and directions emphasized and promised by researchers and inventors. At the same time, identifying PVEs presents an operationalization challenge that needs to be addressed if they are to be used as an analytical tool for policy deliberation and justification. PVEs can be difficult to identify, as they are contextual, change over time, and require distinctions that are not always evident or easy to discern (Fukumoto & Bozeman, 2019).

Patent documents can be a valuable source of PVEs (Ribeiro & Shapira, 2020). Patents are manifested in written text, where inventors and legal representatives elaborate on their inventions and the context in which they operate. These descriptions provide valuable opportunities to identify and analyze PVEs. Additionally, given their radical novelty, fast growth, prominent impact, and uncertainty (Rotolo, Hicks, & Martin, 2015), emerging and general-purpose technologies, such as AI, which are often the subject of patent applications, can serve as a springboard for more comprehensive discussions on society's PVs. Because emerging technologies can change society in fundamental ways, it is important to reflect on the extent to which these changes align with our collective values and goals. This can be done by understanding how PVs are mobilized in narratives around these technologies. While doing so, we can gain insights into the intended societal benefits of emerging technologies as well as anticipate their potential negative impacts (Buhmann & Fieseler, 2021).

Patent texts are, therefore, valuable materials for studying PVEs because, at one level, they contain detailed descriptions of inventions and the processes used to create them, which can provide insights into the current state of the art, emerging trends, commercial uses, and broader impacts of technologies. At another level, in their background section, patent texts offer clues regarding the social and economic context in which the technology was developed as well as its intended impacts in such a context. The arrival and accessibility of GLMs that can address complex and ambiguous topics, abridge them for human understanding, and organize them effectively, presents an opportunity to explore a new approach to discovering PVEs in patent texts. We describe this approach in the following sections, using the case of AI patents.

2. DATA AND METHODS

The approach used in this study proceeds through four key stages. First, we create a database of AI documents and identify potential PVEs through keyword filtering. Next, we develop a framework for sentence labeling that involves both human input and AI annotation (using a GLM). Third, we estimate BLEU scores and perform topic modeling to evaluate the faithfulness, diversity, and discovery capabilities of the AI annotator's output. Finally, we use these annotated labels to train an open-source classifier and apply it to all records. These steps are illustrated in Figure 1 and explained in detail in the following sections.

2.1. Data: Obtaining AI Patent Documents and PVEs

Our process for obtaining a set of patent documents involves two steps. First, we employ a search strategy developed by Liu, Shapira et al. (2021) that uses AI-related keywords, cooperative patent classification (CPC) codes, and international patent classification (IPC) codes to retrieve an extensive collection of U.S. patent applications and granted patents filed between 2005 and 2022. To execute the search, we entered a Boolean query into InnovationQ+®[1], a patent search tool, resulting in the retrieval of patent ID numbers and metadata at the invention level. This yielded 198,456 patent documents, which included a single record for each simple family and granted patents that overrode applications. Second, we used the bulk download option of PatentsView, provided by the United States Patent and Trademark Office (USPTO)[2], to extract background, summary (description), and abstract text. We merged the IDs generated

[1] https://ip.com/.
[2] https://patentsview.org/.

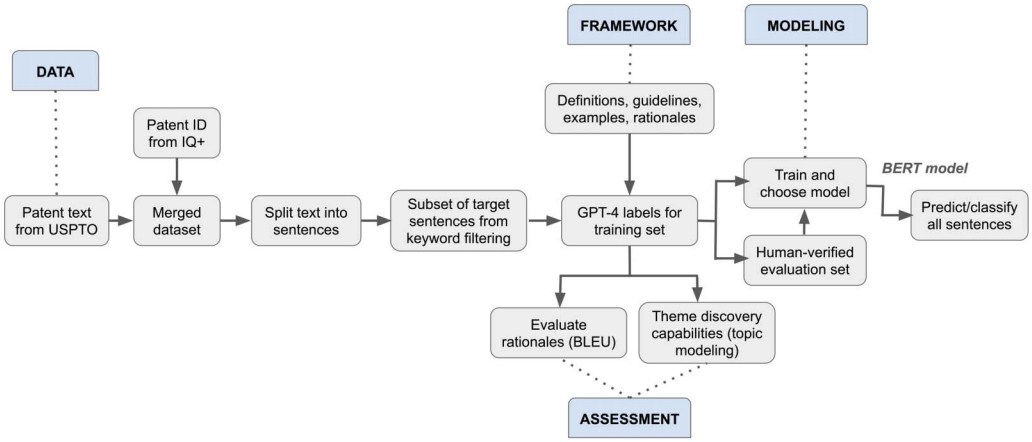

Figure 1. Schematic of main stages of the study approach.

in the first step with the text retrieved in the second step to obtain a final set of 154,934 U.S. patent documents related to AI.

The patent documents were split into individual sentences to facilitate their analysis and categorization through ML classifiers. The 154,934 patent documents yielded 5.4 million sentences. To find PVEs within such a large and sparse corpus, we required a method to obtain a manageable subset of sentences with a high density of PVEs to select a sample to annotate. To do so, we implemented a keyword filtering approach. This involved creating a list of single words, bigrams, and trigrams. Initial PV keywords were sampled from a narrative review of relevant documents, both peer-reviewed and nonpeer-reviewed, focusing on the benefits and risks of AI. An important part of this literature is reflected in publications related to the governance of this technology. Therefore, we conducted a search for guidelines published by public and private organizations, as well as journal articles analyzing and proposing frameworks for managing the impacts of AI. Our final list of terms (N = 320) represented a broad range of topics related to both the positive and adverse impacts of AI, as discussed in the available literature. By filtering sentences containing these terms, we were able to retrieve a smaller subset of sentences (N = 73,813) that potentially contained PVEs.

From this pool of sentences, reflecting practical and technical factors such as cost, time, and saturation in classification performance, we extracted a training and evaluation set of 10,000 sentences. Instead of randomly drawing sentences, we created a system to rank the importance of PV keywords, grouping them into four categories (1 to 4) based on their ability to capture relevant PVEs. On one end, keywords classified as category 1 were ranked very low in their ability to capture PVEs because they were considered ambiguous. That is, they had the potential to retrieve technical or private value sentences in addition to PVEs. Some examples of category 1 keywords include "risk management" and "emotional state." Additionally, category 1 contained short-tail keywords that retrieved a disproportionately large number of sentences, such as "health care" and "education." Keywords in category 4, on the other end, were characterized by their high ability to capture PVEs. These tended to be long-tail and to retrieve a small number of sentences. Examples of keywords in category 4 are "explainable artificial intelligence," "human safety," "privacy by design," and "societal concern." We used these categories to oversample sentences that were more relevant and undersample those that were less relevant. Specifically, we randomly sampled 4.5% of sentences from category 1, 14% from category 2, 65% from category 3, and 100% from category 4. By doing so, we ensured our training and evaluation set contained a diverse sample of sentences.

2.2. Framework: A Method for Designing Instructions for Human and AI Annotators

After obtaining the training and evaluation set, the next step was to generate labels for each sentence (i.e., classify them as PVE (or something else)). We developed a multistage process to develop a framework for labeling PVEs in patent documents. In the initial stage, we employed an abductive approach—a group of three researchers with experience in text mining of patent documents, PV theory, and emerging technologies went through the sentences and labeled them as either being PVEs or not. We achieved 66%, 56%, and 78% agreement rates in three batches of 50 sentences. The aim was to train ourselves in identifying PVEs in AI patents by observing patterns in disagreements and discussing our labeling decisions. In the second stage, we manually classified sentences on a larger scale, labeling 1,000 sentences with an agreement rate of 79%, leading to a new round of discussion.

A key learning from our discussions was that Although the three coders had many areas of agreement about PVEs, in other cases there were differences in the interpretation of PV concepts. This resulted in the rise of differences in labeling sentences. Our individual understanding of PV theories was often implicit and challenging to articulate to one another. Therefore, our labeling decisions were nonstandard and sometimes contradictory. Additionally, in the process of repeated coding, we disagreed with our past selves, failed to be consistent across sentences, and provided varied justifications for our choices. As explained below, we imputed that one of the key reasons for this behavior was the lack of a consolidated paradigm in PV theory, which required us to develop our own framework to support the labeling.

There may be other reasons that generally explain the challenges of labeling complex and abstract topics, such as PVs. In such cases, human labeling capabilities can be diminished by cognitive limitations, such as working memory constraints, mental fatigue from repetitive and lengthy coding tasks, attentional biases, and inflexibility (i.e., once we learn a heuristic to label, we get stuck with it, even if it is wrong). Cognitive biases can also play a role. For instance, a tendency to rely too heavily on initial information (anchoring bias) could result in misclassifying later sentences. The impact of labeling discrepancies could be significant given a small set of labelers. Earlier research in fields such as political science, for example, has relied on crowdsourcing to deploy large-scale, qualitative analyses of text to overcome the challenges of bias and inconsistent coding. In this context, the deployment of an increased number of labelers was justified as a way to compensate for individual errors (Benoit, Conway et al., 2016).

To overcome the variability of individual coding, we progressed to a third stage, where we provided written justifications for decisions regarding 100 labeled sentences. Justifications made our disagreements and the limitations of PV theory more visible. We discussed how PV theory differed from similar frameworks, such as public interest. We also debated whether PVEs and private value statements were mutually exclusive, the assumptions that were acceptable to infer a PVE from an ambiguous sentence, the role of the target beneficiary in determining whether a sentence was a PVE, the unequal standards to annotate sentences across PV themes, and the consequences of assigning hierarchies to PVEs. Notably, some of these debates are at the center of current discussions in PV theory, as highlighted, for example, in Fukumoto and Bozeman (2019). Our efforts to provide justifications and address contested points led us to create a set of guidelines that encompassed positive heuristics for identifying a sentence as a PVE and negative heuristics for determining when not to label a sentence as a PVE.

These guidelines were initially conceived as a written reference to aid researchers' labeling decisions. The aim was to help us incorporate a deductive approach, where the guidelines

were seen as principles that generalized consistently to observable examples. However, with the introduction in March 2023 of GPT-4 (Generative Pre-Trained Transformer 4; https://openai.com/research/gpt-4), we realized there was an opportunity to use this multimodal large language model (LLM) to interpret, contextualize, and organize text. We anticipated that our guidelines would be potentially helpful, not only to human annotators but also to AI annotators. The guidelines were both a framework to support human deductive reasoning and a prompt with instructions for a GLM to understand PV theory and classify patent text accordingly. This led to a fourth stage where, using GPT-4's Application Programming Interface (API), we designed a framework where GPT-4 could label the training and evaluation set. This framework included:

- a concrete and comprehensive definition of PV;
- a categorization of sentences as Direct-PVE (i.e., the sentence demonstrates that the invention addresses a PV), Contextual-PVE (i.e., the sentence demonstrates awareness of a PV, but does not provide enough information to link it with the invention), or No-PVE;
- a reduced number of two positive heuristics and three negative heuristics;
- four examples for each heuristic; and
- rationales for each example.

This framework is included in the Supplementary material.

The high-level form of the framework, which most effectively communicates with humans, must be adapted to function as a prompt with instructions for the AI annotator. We provided system instructions to the model to clarify its role and supply necessary definitions to aid categorization. In Figure 2, we show the instructions provided to the GPT-4 model and the role that different components play. Before specifying the definitions crafted as part of the developed framework, we specify the role of the GPT-4 agent (see "Task specification" in the figure). Subsequently, we use the definitions developed as part of our framework (Supplementary material). The definitions are concise and actionable, and can include both inclusionary as well as exclusionary identifiers for each of the categories. Furthermore, based on our early experiments, we added statements to correct the behavior of the GLM. In the illustrated example, it is to stop the model from making assumptions and inferences based on its inherent

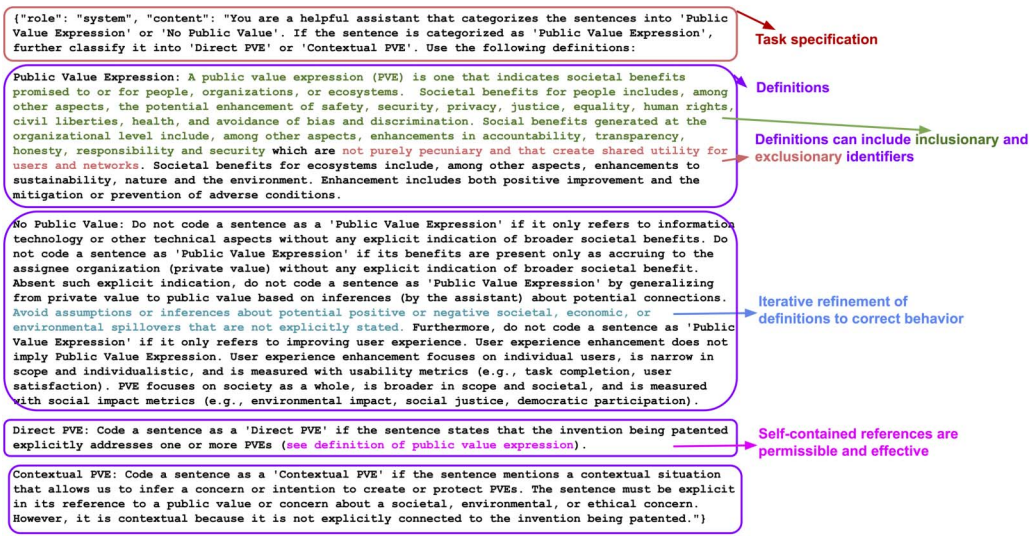

Figure 2. Illustration of the instructions provided to GPT-4 and the key components involved.

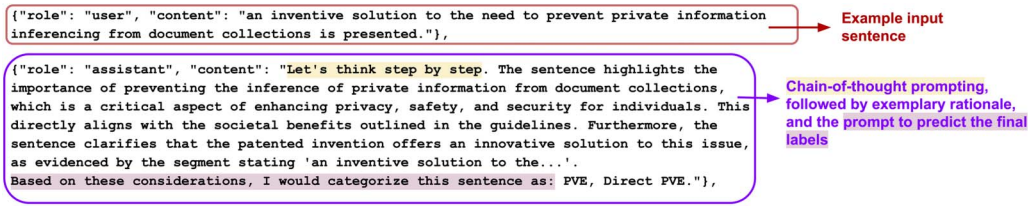

Figure 3. Illustration of example sentence and rationale provided to GPT-4, with key components indicated.

knowledge of a language (see "Iterative refinement of definitions to correct behavior" in the figure). Finally, we note that some definitions can refer to the properties discussed as part of other definitions, as long as the entire set of instructions provided to the model is kept self-contained. As GLMs demonstrate long-form context modeling abilities, they can use the considerations discussed under the definition of "Public Value Expression" to infer "Direct-PVE. "

Following the instructions, we provided 14 examples to GPT-4, spanning the categories under consideration. For each example, we start the rationale with the phrase "Let's think step by step" to trigger chain-of-thought reasoning by GPT-4, a known technique to help the agent arrive at a prediction that follows an exemplified reasoning scheme. Each rationale follows: "Based on these considerations, I would categorize this sentence as: <final prediction>." The 14 examples were chosen to illustrate a diverse and nonredundant reasoning scheme and were chosen iteratively while analyzing performance on a small set of unseen examples (N = 100). We show one such example in Figure 3.

GPT-4 produced a label (i.e., Direct-PVE, Contextual-PVE, or No-PVE) and a rationale for its decision. Our agreement with GPT's labels went up to 95% in a conservative estimation in a separate evaluation set of 1,000 randomly sampled sentences (from the subset of 10,000 sentences).

Out of the 1,000 sentences in the validation set, we found only one case where the three human annotators disagreed with GPT-4. This sentence mentioned privacy concerns, which GPT-4 classified as No-PVE, when it should be a Contextual-PVE. GPT-4 wrongly suggested that privacy was not a broad societal concern. However, this was an isolated error, and other privacy sentences were correctly labeled by GPT-4.

We also found 42 cases where one annotator disagreed with GPT-4 and the other two agreed with it. We counted these as disagreements so as to be conservative in our evaluation. These cases involved nuances in the use of words such as "environment," "smart cities," "education," and "government programs," which could imply value or technical meanings, depending on the context. Although further refining the prompts for GPT-4 could reduce some disagreements, it could also increase others. Overall, we concluded that GPT-4 was generally accurate in labeling PVEs and that achieving small marginal improvements beyond this point would be difficult. As we demonstrate in the analysis section, not only did GPT-4 produce sensitive labels, but its rationales demonstrated a consistent understanding of our framework and provided logical reasons to support its classifications.

2.3. Assessment: Methods for Measuring the Faithfulness and Diversity of the AI Annotator's Outputs

In contrast to discriminative models such as BERT and its variants, GLMs such as GPT-4 can be used for generating coherent long-form rationales along with the final predictions. These rationales serve three purposes. First, chain-of-thought prompting (i.e., encouraging the GLM to

adopt a reasoning scheme by using phrases such as "Let's think step by step" before predicting the final label) has demonstrated effectiveness in eliciting an accurate final prediction by the model (Wei, Wang et al., 2022). Second, these rationales are material that allows human researchers to assess the labels, making the decision process of the GLM more transparent and subject to human review. Third, given the high quality of the rationales, these function as persuasive and standardization mechanisms—they help humans in their own labeling decisions and focus their justifications on a smaller set of standard and organized possibilities, thus enabling them to overcome cognitive limitations and biases.

To authenticate this line of argument, we use the rationales that GPT-4 generates while labeling the 10,000 examples to answer the following question: *How diverse and faithful are the rationales (or reasoning mechanisms) and can the generative language model strike a balance between the two measures?*

In prior work, diversity has been measured as the complement of similarity between items (Ma, Lyu, & King, 2010; Verma, Vinay et al., 2022; Zhu, Lu et al., 2018). In this work, we adopt the same approach and measure the similarity between two rationales using the BLEU score (Papineni, Roukos et al., 2002). The BLEU score measures the similarity between two sentences by comparing *n*-grams (sequences of *n* words) in a given sentence with the reference sentence. Mathematically,

$$\text{BLEU} = BP \cdot e^{\sum_{n=1}^{N} w_n \cdot \log(p_n)}$$

where w_n are weights for each *n*-gram, p_n is the precision of an *n*-gram, and N is the maximum order of *n*-grams used (we compute up to four *n*-grams).

The BLEU score ranges from 0 to 1, where 1 denotes the maximum similarity between two sentences. However, as our goal is to measure the diversity of rationales, we interpret the scores as low values demonstrating higher diversity. We are interested in the following forms of diversity: (a) How diverse are the generated rationales with respect to the rationales provided to the model, and (b) how diverse are the generated rationales among each other. For (a), we quantify the category-wise average of maximum similarity that a given generated rationale has with the supplied rationales of the same category. For (b), we take the average pairwise similarity scores of all the within-category generated rationales. The same similarity scores can be used to comment on how "faithful" the generated rationales are with respect to the provided rationales. High faithfulness is desired to ensure that the model's reasoning mechanism abides by the same set of reasoning principles as depicted by the supplied rationales.

In addition to using the BLEU similarity scores to assess the quality of the rationales generated by GPT-4, especially regarding striking a balance between diversity and rationales, we assess GPT-4's ability to discover PV themes with topic modeling. To ensure cost-effectiveness, the number of illustrative examples that can be provided to GPT-4 is limited. Therefore, we carefully curate four examples that cover nonredundant themes demonstrating Direct-PVEs. However, as these four themes do not cover all possible themes related to PVs, we aim to evaluate whether GPT-4 can produce rationales belonging to different themes not included in the examples. We perform topic modeling on the generated rationales of the subset of sentences labeled "Direct-PVE" by GPT-4. More specifically, we use Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to discover 10 topics among all the aforementioned rationales. We chose the number 10 after qualitatively analyzing the coherence and overlap of produced topics. We then qualitatively inspect the 10 topics and assess whether they relate

to the themes covered in the provided examples or have been revealed because of the reasoning capabilities of GPT-4.

2.4. Modeling: Training and Prediction Methods

The goal of obtaining labels for 10,000 sentences is to train open-source classifiers such as BERT (Devlin, Chang et al., 2018), which could then be used for inferring the labels of millions of sentences. Given the remarkable performance of GPT-4, which was validated to align with human experts more than 95% of the time, a desirable setting would use GPT-4 to label all the millions of examples in the target corpus. However, using closed-source GLM for large-scale inference would incur infeasible costs for us (and many other academic research groups). Despite open-source counterparts of proprietary generative language models, such as LLaMA (Touvron, Lavril et al., 2023), OPT (Zhang, Roller et al., 2022), and Flan-T5 (Chung, Hou et al., 2022), hosting them requires a large GPU infrastructure that may not be accessible to many scholars. Furthermore, current open-source language models are also limited in their maximum context length, which limits the information that can be provided as part of the instructions and examples.

To this end, we deploy a hybrid approach where we use the 10,000 examples labeled using GPT-4 to train open-source pretrained BERT-like models. BERT (Devlin et al., 2018), is a language representation model that uses transformers and bidirectional training to understand the context of words within a sentence. It is pretrained using a combination of masked language modeling objective and next-sentence prediction on a large corpus of text. Such models can be hosted on a single GPU (NVIDIA A100 80GB, available via platforms such as Google Colab) and can be trained for the specific use case at hand with relative ease. For training the models, we use the 9,000 examples labeled using GPT-4 as the training corpus and the 1,000 GPT-4-identified, human-validated examples as the evaluation corpus. We use two formulations of the classification task: 3-class classification, under which the models are tasked to distinguish between Direct-PVE (D-PVE), Contextual-PVE (C-PVE), and No-PVE; and 2-class classification, under which the models are tasked to distinguish between PVE and No-PVE. By design, the former setting is more challenging, as it involves distinguishing between the subcategories of PVEs. To quantify the performance of the models, we use macro averages of class-wise F1 scores, precision, recall, and accuracy.

We evaluate a range of pretrained language models that are similar to the BERT model. We describe these models and the key differences between them. BERT-base-uncased is a version of BERT with 110 million parameters. It deals with lowercase English text, allowing it to be smaller and faster while maintaining robust performance on many tasks. BERT-large-uncased, an expanded version of BERT-base, contains 340 million parameters. Despite being computationally heavier, its larger size grants it improved language understanding capabilities. DistilBERT (Sanh, Debut et al., 2019) is a distilled, lighter version of BERT, maintaining comparable performance while being smaller (66 million parameters), faster, and cheaper to run. It uses a teacher-student training paradigm, learning from a larger "teacher" BERT model. RoBERTa-large (354 million parameters), or Robustly Optimized BERT, is a variant of BERT that modifies BERT's training process for improved performance, using larger batches of data, more data, and removing the next-sentence prediction task (Liu, Ott et al., 2019). ALBERT-xxl-v2 (223 million parameters), or A Lite BERT, is a version of BERT that introduces parameter-reduction techniques to lower memory consumption and increase training speed. It achieves this by sharing parameters across layers and factorizing the embedding layer (Lan, Chen et al.,

2019). Finally, DeBERTa-xxl-v2 (1.5 billion parameters), or Decoding-enhanced BERT with disentangled attention, improves upon BERT and RoBERTa with a disentangled attention mechanism (He, Liu et al., 2020). This allows the model to dynamically integrate contextual information, improving its understanding of complex language phenomena[3].

To provide context for the performance of the classifiers obtained from training these pretrained language models, we also include the performance that random classifiers would achieve on the evaluation set. These random classifiers, uniform and biased based on class ratio, serve as a baseline comparison for evaluating the effectiveness of our trained classifiers.

3. RESULTS

3.1. Evaluating GPT-4 Rationales

Table 1 shows three examples of sentences from patent texts, their labels, and the sentences generated by GPT-4. These examples illustrate how GPT-4 interprets the patent text and provides rationales for its decisions. The rationales reveal the following steps that GPT-4 takes:

- it checks for an explicit reference to a PV in the sentence;
- it identifies possible PV themes that are related to the reference; and
- it looks for evidence that connects the patented invention and the PV (e.g., "an objective of the present invention ...").

In addition, GPT-4 considers the negative heuristics that we instructed it to use, as seen in the rationale of the "No-PVE" sentence, where it explains that the sentence focuses on the technical aspects of the invention rather than on a PV. In other "No-PVE" sentences, GPT-4 states that enhancing user experience or creating private value does not necessarily imply contributing to PVs.

To evaluate the quality of GPT-4's rationales, we can use an additional method based on BLEU scores. Our goal is to ensure that these rationales are both faithful and diverse. This means that our framework should serve as a reliable guide for generating rationales for unseen examples. Additionally, the GLM should be able to deduce the label of any PV theme—even if it is not included in the examples—by leveraging the first principles outlined in the framework.

We analyze the diversity of the GPT-4 rationales with respect to the illustrative examples supplied to GPT-4. Recall that a lower BLEU score indicates more diversity. We use BLEU scores (see Table 2) to quantify the pairwise similarity among within-category rationales provided to GPT-4 (first column after categories), the average maximum similarity of generated rationales with within-category provided rationales (second column), and the average pairwise similarity among within-category generated rationales (third column).

[3] We also experimented with BERT for patents (Lee & Hsiang, 2019), which was trained on 100 million patents and has demonstrated notable performance in tasks such as searching for prior art while drafting patents, autocomplete, and automatically generating classification codes. Therefore, one might expect BERT for patents to be among the models being used for this task. However, we observed that the performance of BERT for patents is lacking (F1 score of 0.78 for the 3-class classification task) in comparison to the other models mentioned here (F1 score > 0.80 for the 3-class classification task). We believe that this is because identification of public values is not a conventional patent-related task and requires the understanding of broader context, which domain-specific models such as BERT for patents may lose because of restrictions to the training corpus.

Table 1. GPT-4 rationales for PVE labeling with three examples

Patent sentence	Human-verified GPT-4 label	GPT-4 rationale
It is further an objective of the present invention to assess the risk of a service provider engaging in collusion with one or more partners in the ecosystem of service providers.	Direct-PVE	"The sentence explicitly states that the invention aims to assess the risk of collusion among service providers within an ecosystem. This aligns with the societal benefits for organizations, as it enhances transparency, honesty, and responsibility, which are not purely pecuniary and create shared utility for users and networks. The invention is explicitly identified as addressing this issue, as indicated by the segment of the sentence that reads 'it is further an objective of the present invention to ...'".
In all cases, data security is a priority due to the sensitive nature of medical information.	Contextual-PVE	"The sentence explicitly highlights the importance of data security in the context of medical information, which is a critical aspect of enhancing privacy, safety, and security for individuals. This directly aligns with the societal benefits outlined in the guidelines. However, the sentence does not explicitly connect this priority to a specific invention or patent."
In marked contrast are artificial neural networks, which have been disparagingly referred to as "black boxes" for their inability to model a problem in an intuitive fashion.	No-PVE	"The sentence discusses the limitations of artificial neural networks, referring to them as "black boxes" due to their lack of intuitive problem modeling. However, it does not explicitly or implicitly mention any broader societal benefits or public values. The focus is on the technical aspects of artificial neural networks rather than their potential impact on society."

As indicated, we first looked at the diversity of the rationales that were provided, and the average of pairwise similarity among within-category rationales was calculated. These low values indicate that the provided rationales are very diverse with respect to each other, which is expected as the examples were chosen to be nonredundant. Next, we analyzed the diversity of the generated rationales with respect to the provided rationales. The average maximum similarity between each generated rationale and the provided rationales for the same label was calculated. The increase in BLEU scores indicates that the provided rationales are being used as guidelines for generating the rationales of unseen examples, a behavior often referred to as faithfulness. However, it is worth noting that the BLEU scores are not exceptionally high, which demonstrates diversity in the sampled data and the rationales required to categorize

Table 2. Diversity and faithfulness of rationales

Category	Pairwise similarity among within-category rationales	Average maximum similarity with provided rationales	Pairwise similarity among generated rationales
No-PVE	0.02	0.25	0.23
Direct-PVE	0.04	0.16	0.07
Contextual-PVE	0.08	0.20	0.13

the examples. Finally, we looked at the diversity of the generated rationales with respect to each other. The average of pairwise similarity among within-category rationales was calculated. The observed results are expected, as the generated rationales are anchored in the provided rationales and show some similarity but not too much. The diversity among Direct-PVE is notably higher in comparison to Contextual-PVE and No-PVE. One possible explanation for why Direct-PVEs are more diverse than Contextual-PVEs or No-PVEs is that Direct-PVEs are more likely to reflect the specificity and novelty of the AI invention and how it addresses a societal challenge or benefit. In contrast, Contextual-PVEs or No-PVEs are more likely to reflect the generality and commonality of AI technologies and their potential applications.

We also assess GPT-4's ability to discover PV themes through topic modeling. We provided four examples to GPT-4 for the category "Direct-PVE," which spanned the following themes: security and privacy, disease risk prediction, occupational health, and exiting poverty. To assess whether GPT-4 rationales cover themes beyond the ones provided in the examples, we use topic modeling to identify 10 themes in the rationales produced for 1,258 unseen sentences labeled as "Direct-PVE." With qualitative analysis of frequently occurring words within each topic, we found the following 10 topics. We also indicate whether these topics are "covered" in the provided rationales or are "revealed." We note that we adopt a very conservative approach while considering a rationale as "revealed" only considering topics completely unrelated to the topics provided in examples.

- Topic 1: Security and safety enhancement (*covered*)
- Topic 2: Mental health, well-being, patient, and medical monitoring (*covered*)
- Topic 3: Privacy, data security, and personal information protection (*covered*)
- Topic 4: Clean energy, sustainability, and renewable resources (*revealed*)
- Topic 5: Economic and financial development (*revealed*)
- Topic 6: Transparency in decision-making, machine bias, and participation consent (*revealed*)
- Topic 7: Developing organizational trust among people (*revealed*)
- Topics 8 and 9: Modeling and assessment of disease risk (*covered*)
- Topic 10: Applications for assessing human behavior (*revealed*)

We note that GPT-4 rationales for categorizing content as Direct-PVE cover some of the new themes that are unrelated to themes covered in the examples (five out of 10), whereas the other half are related to the ones that have been included in the examples (five out of 10). These results further advance the argument that GPT-4 can strike a balance between faithfulness with the provided rationales while exhibiting reasoning based on "revealed" themes.

Based on the above two tests, using BLEU scores and topic modeling, the results indicate that GPT-4 has balanced diversity and faithfulness in justifying its PV classification decisions, and GPT-4 internalized our framework and used it as a deductive device to classify the various types of PVEs it encountered.

3.2. Training Open-Source Classifiers

In the final major step of our approach, we use the labels produced by GPT-4 to train open-source discriminative classifiers. We report the results for the two settings—three-way classification and two-way classification in Tables 3 and 4. First, we note that the performance for both the classification settings is satisfactory—the best models achieve an F1 score of

Table 3. Classification performance of models trained on data labeled using GPT-4: 3-class classification (No-PVE, D-PVE, and C-PVE)

Model	# Parameters	F1 (Macro)	Precision (Macro)	Recall (Macro)	Accuracy
Random (uniform)	–	0.3159	0.3357	0.3454	0.3370
Random (biased)	–	0.3336	0.3337	0.3336	0.4260
DistilBERT-base-uncased	66 million	0.8107	0.8124	0.8091	0.8370
BERT-base-uncased	110 million	0.8210	0.8207	0.8221	0.8480
BERT-large-uncased	340 million	0.8149	0.8236	0.8072	0.8510
RoBERTa-large	354 million	0.8373	0.8464	0.8291	0.8630
ALBERT-xxl-v2	223 million	0.8343	0.8435	0.8269	0.8570
DeBERTa-xxl-v2	1.5 billion	0.8414	0.8518	0.8351	0.8617

almost 0.85 for the 3-class classification task and an F1 score of above 0.90 for the 2-class classification task. This means that the models were able to learn with the labels obtained from GPT-4 and predict accurately on a held-out evaluation set. Second, we note that the fine-grained classification task is relatively more difficult for the BERT-based classifiers, as it involves distinguishing between Contextual-PVEs and Direct-PVEs. Finally, we observe a clear trend across the two tasks. The number of parameters in the pretrained language model positively affects the classification performance. We can note that larger models (DeBERTa-xxl-v2 with 1.5 billion parameters) perform better than models with fewer parameters. This pattern aligns with the scaling law, suggesting that substantial performance improvements are achieved by scaling up LLMs (Bowman, 2023). We investigated the use of Latent Dirichlet Allocation (LDA) to model topics from the entire data set, but this did not prove to be viable. However, our framework was not designed to identify specific topics but rather to categorize sentences into three broad categories. The rationales presented well-defined and distinct topics, but these were only available for 10,000 sentences. The rest of the 5.4 million sentences did not have enough semantic content for the LDA approach to generate meaningful and distinct topics.

Table 4. Classification performance of models trained on data labeled using GPT-4: 2-class classification (No-PVE and PVE)

Model	# Parameters	F1 (Macro)	Precision (Macro)	Recall (Macro)	Accuracy
Random (uniform)	–	0.4983	0.5001	0.5001	0.4990
Random (biased)	–	0.5397	0.5397	0.5398	0.5430
DistilBERT-base-uncased	66 million	0.8107	0.8124	0.8091	0.8370
BERT-base-uncased	110 million	0.8716	0.8861	0.8701	0.9042
BERT-large-uncased	340 million	0.8901	0.8993	0.8867	0.9114
RoBERTa-large	354 million	0.9013	0.9136	0.9004	0.9218
ALBERT-xxl-v2	223 million	0.9046	0.9172	0.9081	0.9278
DeBERTa-xxl-v2	1.5 billion	0.9103	0.9256	0.9129	0.9310

4. DISCUSSION AND CONCLUSIONS

The availability of labeled data is an important requirement for training language classifiers for subsequent large-scale machine-based text analysis. However, the manual labeling of text is a resource-intensive and time-consuming process, limiting the volume of data that can reasonably be coded by humans. Additionally, when dealing with complex and abstract concepts such as PVs, manual labeling also carries cognitive limitations and biases that can reduce the quality of the labels. To address this problem, we put forward an alternative semiautomated approach where a GLM serves as the main annotator, with human validators providing verification on smaller samples. This approach requires crafting a comprehensive and intelligible framework with guidelines on what to do (i.e., positive heuristics), what not to do (i.e., negative heuristics), precise definitions, examples, and rationales. Such a framework acts as a prompt with instructions for the AI annotator.

The quality of the framework's output depends on both the design of the framework and prompt and the capabilities of the GLM. We refined our definitions, guidelines, and examples through multiple iterations to reduce the error rate of GPT-4's results. We note that Although our framework design worked well with GPT-4, it had limitations with GPT-3.5. When we tested the same framework on both models, GPT-4 produced outstanding results, but GPT-3.5 showed only a slight improvement over human labeling. Other text-based experiments have found a similar contrast between these two versions of GPT (Nori, King et al., 2023). GPT-4 appears to be already capable of addressing complex classification tasks, and future versions will likely require less framework design work and provide more accurate labels and sophisticated rationales for abstract concepts (Bubeck, Chandrasekaran et al., 2023).

In our framework, we used GPT-4 to label a training set (at cost) which was then used to predict within a much larger set using open-source discriminative classifiers. We found that the labels generated using this approach were consistent and accurate, with convincing rationales that aided humans in organizing their ideas and differences around intricate topics. The GLM produced rationales that followed the instructed guidelines (i.e., it is faithful) while generalizing to label unseen topics (i.e., it is diverse). Our analysis of the rationales offers insights into the reasoning capabilities of GLMs, demonstrating that they can strike a balance between faithfulness and diversity when prompted with carefully crafted instructions and examples.

To be clear, the approach presented in this paper was not costless. The cost for labeling the 10,000 sentences in March 2023 via the Open-AI GPT-4 API was about US$ 1,200, with just under 40 million tokens used. However, the cost in terms of time and money was far less than the fully manual coding of 10,000 sentences. GLMs may become less expensive to use or increasingly available for large-scale language processing on an open-access basis. With the support of GLMs, the semiautomated framework that we put forward in this study is likely to be useful in many use cases, especially when analyzing complex concepts. In particular, the approach has the potential to be widely useful in the social sciences, including in science and innovation policy analysis.

A key element of our approach is human–machine interaction. The capabilities of the human side are important. We recognize that the rise of GLMs raises concerns about the risk of bias, even in classifying text. Complex and abstract concepts such as PVs require training in and understanding of public administration, policy, responsible innovation and ethics, and related fields, meaning that, unlike labeling examples where it is easy to discriminate between categories (e.g., an apple is a fruit and an iris is a plant), complex and abstract concepts cannot be left for outsourced labeling (e.g., Mechanical Turk). On the machine side, even with more capable GLMs, it would be a mistake to assume that human labor is no longer necessary for

large-scale text analysis. On the contrary, it is more critical than ever. Human researchers are indispensable in designing frameworks, engineering prompts, and validating results. Crucially, the careful development of training instructions depends on a series of human-led steps to ensure the quality of the analytical lens being constructed. Further, human supervision of intermediary outputs is equally necessary. In their supervised ML approach to management analysis, Harrison, Josefy et al. (2022), show how different phases of the process, from construct identification to data scaling, depend on human manual labor. This is in line with the findings of recent studies looking at the impact of digitalization on scientific work (Ribeiro, Meckin et al., 2023). As this research demonstrates, these activities remain a time- and intellectually intensive task that requires the expertise and creativity of skilled professionals.

Although we deploy a framework for AI-assisted analysis of text, the depth of this part of the analysis remains largely descriptive, in content analysis terms (Krippendorff, 2019). It is through human iterations with the text, with each other and, crucially, with AI through its labeling and justifications behind its choices, that the analysis becomes more interpretive and subjective. This movement between human and AI labeling resembles hybrid approaches adopted elsewhere through collaboration between computational and qualitative researchers (Li, Dohan, & Abramson, 2021). We argue that using generative language and machine learning models can both enhance and restrict the subjectivity component of content analyses and qualitative research. Ultimately, the descriptive or interpretive tendencies of the analysis will depend on the context of each study, including its research questions, disciplinary orientations, the characteristics of the theoretical frameworks with which researchers are engaging, and the shortcomings of AI models in dealing with more subjective interpretations of the data.

The approach used in this study builds on the strengths of GLMs in classification and pattern identification. There are limitations that should be kept in mind when interpreting the results. Our empirical data is restricted to text found in US AI patent documents. The expression of PVs in that text depends on the motivations of the inventors, assignees, and agents involved in preparing these documents. Our interpretation of PVs is collated from a comprehensive review of PV discussions in available, English-language, peer-reviewed publications and other documents. Broader or narrower interpretations of what is a PV could be posited. The semiautomated approach to using GLMs has a cost and requires coding skills, as well as research methodology skills.

In work already under way, and aware of these limitations, we will use our approach to label, organize, and analyze trends and patterns of PVE expressions in AI patent documents. We will focus attention on the drivers behind the inclusion of PVEs and the potential societal and policy implications of the presence (and absence) of PVEs in emerging AI technologies. We anticipate that approaches similar to those described in this study to use GLMs for labeling and organizing complex and abstract concepts, as well as classifying text at large scales, will be taken up by other researchers. Although we found substantial improvements over manual coding, there is a need for further research to compare and validate the use of GLMs in classifying complex social science concepts in texts. However, our initial experience with embedding a semiautomated GLM approach is promising; the findings suggest significant potential to assist researchers—in science and innovation policy analysis as well as other domains—in text analysis.

AUTHOR CONTRIBUTIONS

Sergio Pelaez: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Gaurav

Verma: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Barbara Ribeiro: Conceptualization, Data curation, Investigation, Methodology, Supervision, Validation, Writing—review & editing. Philip Shapira: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work was supported in part by the Partnership for the Organization of Innovation and New Technologies, SSHRC [grant number 895-2018-1006] (SP, PS); and the Biotechnology and Biological Sciences Research Council [grant number BB/W013770/1] (PS). GV is partially supported by the Snap Research Fellowship.

DATA AND CODE AVAILABILITY

For legal reasons, data from InnovationQ+® cannot be made openly available. The Liu et al. (2021) AI patent search approach that was applied to InnovationQ+® is openly available at https://doi.org/10.1371/journal.pone.0262050. USPTO patent records and text are openly available at https://patentsview.org. The training guidelines used in the study are provided in the Supplementary material. Examples of code and instructions provided to GPT-4 via API are provided in Figures 3 and 4 in the Supplementary material. The code for obtaining GPT-4 predictions, output labels and returned rationales are available at https://github.com/pshapira/pve.

REFERENCES

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020), The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management*, *50*(3), 329–351. https://doi.org/10.1111/radm.12408

Benoit, K., Conway, D., Lauderlade, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295. https://doi.org/10.1017/S0003055416000058

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bowman, S. R. (2023). Eight things to know about large language models. *arXiv*. https://doi.org/10.48550/arXiv.2304.00612

Bozeman, B. (2002). Public-value failure: When efficient markets may not do. *Public Administration Review*, *62*(2), 145–161. https://doi.org/10.1111/0033-3352.00165

Bozeman, B., & Sarewitz, D. (2011). Public value mapping and science policy evaluation. *Minerva*, *49*(1), 1–23. https://doi.org/10.1007/s11024-011-9161-7

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., ... Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. https://doi.org/10.48550/arXiv.2303.12712

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., ... Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv*. https://doi.org/10.48550/arXiv.2210.11416

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Eykens, J., Guns, R., & Engels, T. C. (2021). Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. *Quantitative Science Studies*, *2*(1), 89–110. https://doi.org/10.1162/qss_a_00106

Färber, M., & Ao, L. (2022). The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings. *Quantitative Science Studies*, *3*(1), 51–98. https://doi.org/10.1162/qss_a_00183

Fukumoto, E., & Bozeman, B. (2019). Public values theory: What is missing? *The American Review of Public Administration*, *49*(6), 635–648. https://doi.org/10.1177/0275074018814244

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., ... Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325–336). https://doi.org/10.1145/3351095.3372862

Giczy, A. V., Pairolero, N. A., & Toole, A. A. (2022). Identifying artificial intelligence (AI) invention: A novel AI patent dataset. *Journal of Technology Transfer*, *47*(2), 476–505. https://doi.org/10.1007/s10961-021-09900-2

Harrison, J. S., Josefy, M. A., Kalm, M., & Krause, R. (2022). Using supervised machine learning to scale human-coded data: A method and dataset in the board leadership context. *Strategic Management Journal*, *44*(7), 1780–1802. https://doi.org/10.1002/smj.3480

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv*. https://doi.org/10.48550/arXiv.2006.03654

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781071878781

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*. https://doi.org/10.48550/arXiv.1909.11942

Lee, J.-S., & Hsiang, J. (2019). PatentBERT: Patent classification with fine-tuning a pre-trained BERT model. *arXiv*. https://doi.org/10.48550/arXiv.1906.02124

Li, Z., Dohan, D., & Abramson, C. M. (2021). Qualitative coding in the computational era: A hybrid approach to improve reliability and reduce effort for coding ethnographic interviews. *Socius*, *7*. https://doi.org/10.1177/23780231211062345, PubMed: 37091692

Liu, N., Shapira, P., Yue, X., & Guan, J. (2021). Mapping technological innovation dynamics in artificial intelligence domains: Evidence from a global patent analysis. *PLOS ONE*, *16*(12), e0262050. https://doi.org/10.1371/journal.pone.0262050, PubMed: 34972173

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

Ma, H., Lyu, M. R., & King, I. (2010). Diversifying query suggestion results. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence* (pp. 1399–1404). New York: AAAI Press. https://doi.org/10.1609/aaai.v24i1.7514

NIST. (2023). *AI risk management framework*. Gaithersburg, MD: National Institute for Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv*. https://doi.org/10.48550/arXiv.2303.13375

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). https://aclanthology.org/P02-1040. https://doi.org/10.3115/1073083.1073135

Porter, A. L., & Cunningham, S. W. (2004). *Tech mining: Exploiting new technologies for competitive advantage*. Hoboken, NJ: Wiley. https://doi.org/10.1002/0471698466

Ribeiro, B., Meckin, R., Balmer, A. & Shapira, P. (2023). The digitalisation paradox of everyday scientific labour: How mundane knowledge work is amplified and diversified in the biosciences, *Research Policy*, *52*(1), 104607. https://doi.org/10.1016/j.respol.2022.104607

Ribeiro, B., & Shapira, P. (2020). Private and public values of innovation: A patent analysis of synthetic biology. *Research Policy*, *49*(1), 103875. https://doi.org/10.1016/j.respol.2019.103875, PubMed: 32015589

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Rutgers, M. R. (2015). As good as it gets? On the meaning of public value in the study of policy and management. *American Review of Public Administration*, *45*(1), 29–45. https://doi.org/10.1177/0275074014525833

Sachini, E., Sioumalas-Christodoulou, K., Christopoulos, S., & Karampekios, N. (2022). AI for AI: Using AI methods for classifying AI science documents. *Quantitative Science Studies*, *3*(4), 1119–1132. https://doi.org/10.1162/qss_a_00223

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*. https://doi.org/10.48550/arXiv.1910.01108

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., ... Lample, G. (2023). LLaMa: Open and efficient foundation language models. *arXiv*. https://doi.org/10.48550/arXiv.2302.13971

Verma, G., Vinay, V., Rossi, R. A., & Kumar, S. (2022). Robustness of fusion-based multimodal classifiers to cross-modal content dilutions. *arXiv*. https://doi.org/10.48550/arXiv.2211.02646

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2201.11903

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., ... Zettlemoyer, L. (2022). OPT: Open pre-trained transformer language models. *arXiv*. https://doi.org/10.48550/arXiv.2205.01068

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., ... Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *SIGIR'18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1097–1100). https://doi.org/10.1145/3209978.3210080