The AI Innovation Compass: Constructing Semantic Networks from AI Concepts to Identify and Measure Technology Innovation

ABSTRACT

This paper explores the rapidly evolving landscape of Artificial Intelligence (AI) as a General Purpose Technology and its dual role in driving and sustaining innovation across various domains. Central to this paper is the development of an AI Concept List and it's application onto research papers to generate several semantic networks. Numerous stages are involved, including data acquisition and keyphrase extraction, extension through semantic similarities and validation using regression analysis. The AI Concept List, created through a custom unsupervised machine learning pipeline, consists of clustered keyphrases that encapsulate the broad field of AI, each annotated with an importance weight to aid in-depth analysis in various research and industry domains. The findings unveil a steady rise in the prevalence of AI concepts across certain research domains. Subsequent discussions delve into potential implications, practical applications, and inherent limitations alongside with future research directions and subsequent improvements. This work proposes a novel methodology for measuring innovation, aiming to benefit the academic and industrial communities by highlighting groundbreaking innovations and uncovering AI applications in new domains.

Keywords: artificial intelligence, machine learning, enabling technologies, concept list, semantic network, innovation, text analysis, natural language processing

The AI Innovation Compass: Constructing Semantic Networks from AI Concepts to Identify and Measure Technology Innovation

INTRODUCTION

Currently, we are witnessing a rapidly evolving landscape of technological advancements in the field of Artificial Intelligence (AI). Such advancements open up a wide range of new and innovative solutions which holds both opportunities as well as challenges for industry and research domains alike (Benbya et al., 2021; Liu et al., 2018). While the impact of AI is unquestionably powerful, it has a dual nature when it comes to innovation. On the one hand, there is lots of potential to drive radical innovation in domains like medicine, environmental science or the automotive industry (Wani et al., 2022; Lee et al., 2019; Liu et al., 2020; Badue et al., 2021), but on the other hand, it poses challenges to sustaining such innovations by raising ethical concerns, violating data privacy with huge data collections and eventually leading to technological unemployment (Du and Xie, 2021; Makridakis, 2017).

When looking at the application of novel AI frameworks and technologies in certain research domains, we find unconventional applications in subjects that traditionally relied on conventional methods or algorithms. This leads to the assumption, that AI might be utilized as a General Purpose Technology (GPT) in various fields and thus help solving problems that could not be solved so far with traditional approaches (Klinger et al., 2021; Sarker, 2021). Finding such events would require expert knowledge on the domain at hand as well as a vast amount of data in good quality and time. Clearly, an instrument (supervised or unsupervised) to track technology innovation without the need for expert knowledge would bridge the gap of detecting the use of AI in different domains and open up a new way of measuring innovation across different areas.

This paper investigates the use of AI as a general purpose technology to track technology innovation in different research domains. Therefore it utilizes a novel approach of creating an AI Concept List and applying it to build semantic networks. Fundamental to this approach is the use of concepts and conceptual spaces as introduced by Gärdenfors (2004, 2014). Due to recent

advancements in natural language processing (NLP) techniques (Yang et al., 2008; Lenz and Winker, 2020), we introduce a quantitative analysis of semantic content in texts across extensive document collections. Our methodology, detailed in subsequent sections, utilizes a custom unsupervised machine learning pipeline for data acquisition, keyphrase extraction, and semantic analysis. The findings reveal a comprehensive list of around 10k unique AI concepts, providing new insights into the interdisciplinary applications of AI and its role in driving technological innovation.

We first define the current state of research by providing an overview about AI as a general purpose technology. We then transition to the state of current methods for text-based measurements of technology innovation and the use of a suitable conceptual space for building a semantic network. Next, we lay the foundation for the AI Concept List by defining different information sources and data acquisition techniques. This step is crucial to the whole process as it is followed by a standardized pipeline consisting of preprocessing, keyphrase extension through similarity search as well as logistic regression for validation of these concepts. We then describe our findings in the Results section, where we introduce a complete AI Concept List and put it to use by building different semantic networks in various research domains. Lastly, we invoke a discussion about the usage of such an AI Concept List and semantic networks as well as its limitations and future research approaches.

ARTIFICIAL INTELLIGENCE AS A GENERAL PURPOSE TECHNOLOGY

Pioneers in fields as diverse as mathematics, psychology, and statistics initiated the genesis of Artificial Intelligence back in the 1950s; they embarked on solving concrete problems with a goal to emulate aspects of human intelligence (McCarthy et al., 2006). Since these early endeavors, AI has not only transformed significantly but also consistently pushed boundaries: it continually challenges what machines are capable of achieving. These efforts have ultimately laid our current landscape - a rich selection comprised of tools, frameworks & systems across multiple domains.

Artificial Intelligence, clearly now transcending its initial academic boundaries, has become a cornerstone in modern technological advancement and an essential element of our daily lives throughout a diverse range of domains. The development of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has been pivotal in advancing capabilities in areas like image and speech recognition (LeCun et al., 2015). More recently, advancements in transformer models, like GPT-3, have revolutionized natural language processing (NLP) (Vaswani et al., 2017). This versatility in solving complex problems across different fields is indicative of AI's role as a GPT, as it's not confined to singular or isolated applications. Evidently, AI architectures are implemented seamlessly into various sectors, revolutionizing our approaches to healthcare, transportation or environmental management. However most notably among these is consumer technology. When looking at the recent progress in generative AI, services like ChatGPT[1] or Midjourney[2] disrupt a wide range of domains by automating many processes and generating new content in various formats on the go. More applications range from subtle algorithms driving digital streaming recommendations to more recognizable forms such as intelligent assistants that manage your smart home. The pervasive adoption and profound integration of AI across diverse industries imply its classification as a fundamental, rather than merely auxiliary, technology.

But the scope extends beyond this point: Artificial Intelligence distinguishes itself in the research landscape with its dual role: as both a subject of basic research and a versatile tool utilized across various domains. This discipline encompasses a diverse array of methodologies, all carefully crafted to replicate and harness the cognitive abilities that parallel human intelligence (Goodfellow et al., 2016; Russell and Norvig, 2016). This research delves into the intricacies of machine learning, neural networks, and cognitive computing, continually advancing our understanding and capabilities in AI. Indeed, AI is not just a theoretical concept: it is used in an array of other domains as a powerful practical tool. In fields such as medicine, environmental

[1] A Service by OpenAI that provides an AI language model designed to provide information through natural language conversation (openai.com/blog/chatgpt).

[2] A tool for creating AI-generated images from textual prompts (midjourney.com).

science or engineering, we harness its power to analyze complex datasets with unparalleled accuracy and speed: for example to model intricate systems or predict outcomes more reliably than human capabilities alone could ever achieve. AI, as a tool in research, extends its application far beyond mere data analysis: because of the rapid evolution of large language models (Vaswani et al., 2017) and image diffusion (Rombach et al., 2022), a wide array of analysis tools are now possible and interactive on a natural level. A recent paper exemplifies the swift advancement of generative AI models in the domain of computational pathology. By integrating a foundational vision encoder with a large language model, this AI assistant demonstrates remarkable proficiency in diagnostic accuracy and response quality, highlighting the rapid evolution and potential of such AI systems in specialized domains (Lu et al., 2023). This is just one application and it underlines AI's ubiquity in research, from computational pathology to climate modeling, exemplifies its status as a GPT, a technology not just prevalent but foundational in various scientific domains.

By definition, a General Purpose Technology (GPT) is characterized by one or more interrelated technologies with the potential of extensive applicability across various sectors as well as technological dynamism (Bresnahan and Trajtenberg, 1995). This is evident in AI's integration into sectors as varied as healthcare, where it's used for diagnostic accuracy, and in environmental science, where it's crucial for climate modeling. Overall, a GPT plays a vital role as an "engine of growth", featuring matching innovations and novel implementations in corresponding sectors and potentially leading to a rapid deployment. But the networked nature of GPTs creates a risk of coordination failures, when rapid change makes their evolution hard to predict (Helpman and Trajtenberg, 1994). This definition can be easily applied to the current events and advancements in Artificial Intelligence, as it has unequivocally established itself as a General Purpose Technology by permeating diverse sectors and fundamentally altering operational paradigms. Its adaptability and transformative impact showcases its role in driving forward a multitude of industries. AI's unique capability to analyze, predict, and innovate has not only streamlined processes but also opened avenues for new discoveries and solutions. Being a GPT, we are able to find uses of AI as a method or tool in a diverse selection of scientific domains.

As we transition from this comprehensive understanding of AI's role as a ubiquitous driver of progress, the focus now shifts to exploring how we can measure this technological innovation. The upcoming section delves into novel methodologies, specifically examining how AI-driven tools, like an AI conceptual space, can quantitatively assess the impact and evolution of innovation in various research domains and AI's capability to bridge the gaps of different applications.

TEXT-BASED MEASUREMENTS OF TECHNOLOGY INNOVATION: THE CASE OF CONCEPTS

Technology Innovation can be measured using different methods and techniques. Traditional quantitative metrics are for example the number of patent filings or R&D investments (de Rassenfosse and van Pottelsberghe de la Potterie, 2009). More intricate ways include big data approaches like metadata analysis of citation networks (Park et al., 2023) or economic figures (Balland et al., 2020) to measure innovativeness and link it to outside factors. But with recent advancements in the domain of natural language processing (NLP), text-based measurements are increasingly applied alongside traditional methods to scientific papers and patents (Yang et al., 2008; Lenz and Winker, 2020). They represent a significant shift from traditional quantitative metrics and sometimes surpasses simple meta-analysis by capturing more details inside the actual text.

Text-based approaches leverage the rich information embedded in text documents - such as academic papers, patents, project reports, and even social media posts - to gauge innovation trends and patterns. The key advantage of text-based analysis lies in its ability to capture the nuances and contextual subtleties of technological advancements, often missed by conventional metrics. Paired with meta information on the papers and patents like citations, venues, funding grants or even code repositories, it enables a deeper understanding of the innovation landscape, including emerging trends, technology diffusion, and the interconnectedness of different domains.

Recent literature suggests, that the development of conceptual spaces by utilizing word embeddings and semantic similarity greatly improves natural language processing applications (Mitchell and Dino, 2011; Aceves and Evans, 2022). While the groundwork for concepts is

already well defined (Gärdenfors, 2004, 2014), conceptual spaces are still evolving especially with new applications of vector representations (Hannan et al., 2019). By spanning one or more conceptual spaces around a specific technology of interest, a semantic network can be built up to analyze metrics like topic appearances or co-occurrences. A handful of papers already try to measure developments and innovation potential using state-of-the-art text-based methods. For example, Giczy et al. (2022, 2021) take patent and paper datasets and examine the use of Artificial Intelligence by employing different machine learning algorithms for classification problems. Another approach is taken by Krenn and Zeilinger (2020) - they utilize a conceptual space in quantum physics and employ a semantic network to identify recombining topics or forecast emerging fields.

Generally, it becomes evident, that new technologies and novel approaches often emerge from the combination of existing ideas and concepts. This phenomenon can be measured with a variety of existing metrics (Pelletier and Wirtz, 2022; Arts et al., 2021). Artificial Intelligence might serve as a prime example through its application in such diverse applications, combining insights from different fields to create novel solutions.

In summarizing these insights, it becomes evident that assessing technological innovation necessitates a comprehensive, multifaceted approach. Traditional metrics like patent filings and R&D investments, while foundational, are significantly enhanced by text-based methodologies and concept spaces, particularly in AI, offering a deeper, nuanced analysis. These methods exploit textual documents to uncover often-missed details and relationships, enriching our understanding of innovation trends and technology diffusion. The integration of AI as a General Purpose Technology with text-based analysis forms a vital synergy, crucial for a thorough understanding of AI's transformative impact across various sectors. This approach is further developed through the construction of an AI Concept List and subsequent semantic networks in the next sections, instrumental for organizing AI terminologies and analyzing its evolving influence.

DATA AND METHODS

In this section, we are going to explain, which data sources were used to build up an AI Concept List, capable of capturing the current landscape of AI research. Furthermore, we illustrate the pre-processing pipeline and consolidation steps towards the creation of valid concepts. Lastly, we explain the validation steps used to verify these concepts and lay out our findings towards the use and application of AI in different research fields.

Data Acquisition

With the objective to construct a strong and robust AI Concept List, the initial and perhaps most crucial phase is the acquisition of relevant data and the subsequent processing of candidate concepts. This step serves as the foundation on which the entire structure of our research is built. Data acquisition, in this context, is not merely about gathering large quantities of information in the AI domain. It is about carefully curating data that is both relevant and of high quality in the first place, ensuring that the concept list and subsequent semantic networks are built on a robust and representative substructure.

The first step constitutes a thorough collection of pre-existing AI concepts from classic scientific literature. We manually create a list of AI concepts compiled from the indices of well-known books that deal with topics of Artificial Intelligence and Machine Learning. In addition to that, we acquire suitable AI concepts from the Computer Science Ontology (Salatino et al., 2020) by selecting the high-level term `artificial intelligence` and traversing down in its reference tree to capture all phrases that are related to it. This serves as a solid foundation of well-known AI concepts.

But since this domain is exposed to an ever changing and fast paced environment, we need to further extended this list by current and state-of-the-art methods and tasks taken from research publications in the realm of AI. Simple categorizations of scientific papers into broad topics and concepts of AI are already done by OpenAlex (Priem et al., 2022) or SemanticScholar (Kinney et al., 2023) but they all rely on unsupervised classification algorithms and lack details when it

comes to tasks, methods or datasets. Furthermore, they are not providing code repositories or additional information. For this reason, we utilize the PapersWithCode[3] dataset, a project initialized by Meta AI Research[4] and run by an active community of researchers and AI enthusiasts. PapersWithCode provides an extensive collection of around 400k papers, all related to AI. In addition to that, these papers are generally tagged with tasks, areas, methods, datasets, code repositories and evaluation tables. All of this information is partly tagged by a state-of-the-art extraction algorithm[5], but mostly assigned by the community and constantly monitored. We procure the current dataset from PapersWithCode through their data dump service and subsequent API calls. Lastly, we generate keyphrases and 768-dimensional embedding vectors for abstracts and descriptions. This step is further explained in the following section, since it utilizes part of the pre-processing pipeline. All of the data acquisition scripts were written in python and designed to be run periodically to ensure an up-to-date data source for further processing. An overview about the written code and scripts can be found in the official GitHub Repository.

The results section and especially figure 7 provides a schematic illustration of the usage of different data sources and subsequent processes. To extract valid AI concepts from method and task descriptions for building the AI Concept List and creating the foundation for several semantic networks, the next section details the execution of methodical pre-processing steps. This process is crucial for refining the raw concept list into a format suitable for in-depth analysis and subsequent application.

Pre-Processing Pipeline

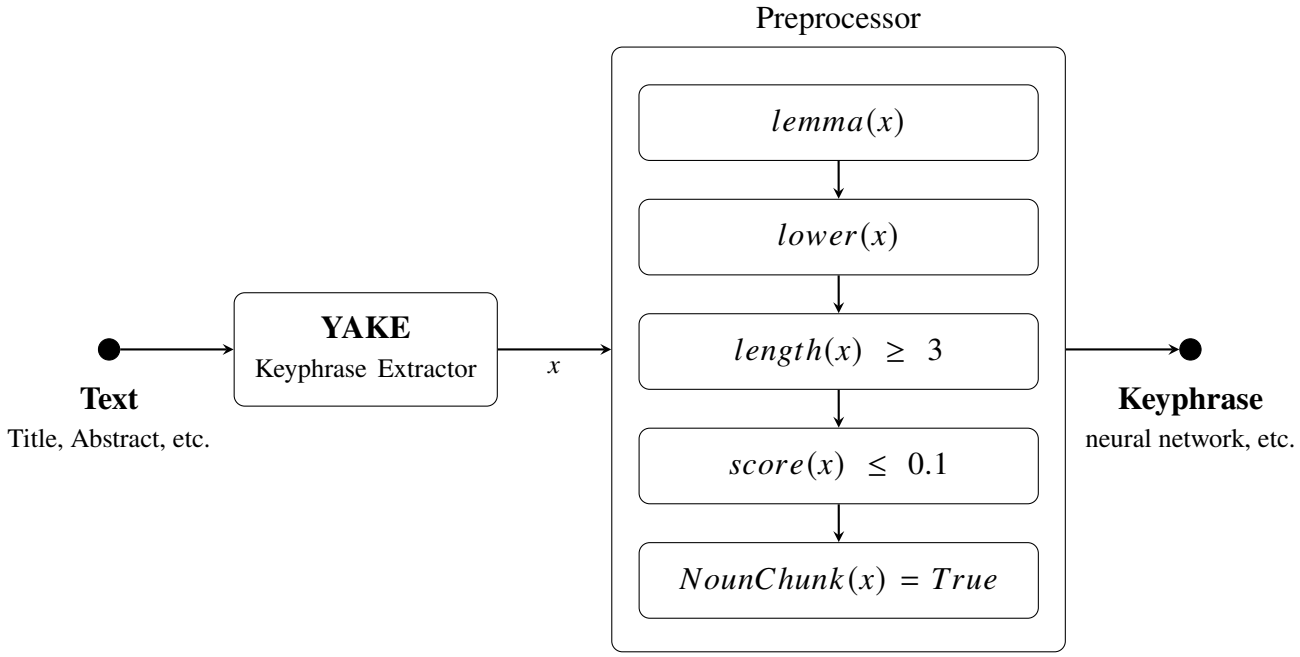
Since all relevant documents and concepts are now available, a proper pre-processing pipeline must be designed to generate high quality concept keyphrases from texts like abstracts or descriptions. A resulting keyphrase is defined as one word (*uni-gram*) or a sequence of words (*n-gram*) that appear successively in the text. Inspired by Shang et al. (2017), we define principal criteria for a candidate phrase to be accepted as a quality keyphrase that describes a valid concept:

[3]paperswithcode.com
[4]ai.meta.com
[5]github.com/paperswithcode/sota-extractor

1. **Significance**: Candidate keyphrases should be given a relevance score to assess their significance to the whole text.
2. **Semantic Network Quality**: keyphrases should appear in their root form. Therefore inflections from words must be stripped off to get a semantic network form of a word in a phrase.
3. **Descriptiveness**: A keyphrase should be able to capture the topic or concept discussed in the given document. Phrases like "paper describes" should be filtered out.

In accordance to these criteria, the pre-processing pipeline was designed as shown in figure 1.

FIGURE 1
Pre-Processing Pipeline



Notes: Standardized Pre-Processing Pipeline to create quality keyphrases from texts

First, the text is extracted from a certain document in the corpus (e.g. an abstract from a paper or a description of a method). This text is already slightly processed as we need to convert it into UTF-8 encoding and strip elements like hyperlinks and exotic characters. Next, we chose to integrate the YAKE keyphrase extractor as it comes with a useful relevance score (Campos et al., 2018). This relevance score captures the criterion of significance, since it assigns a numerical

value to each candidate keyphrase extracted from the given text. It does so by multiplying the bi-gram probability scores for each word of the calculated candidate keyphrase divided by the sum of all bi-gram probability scores weighted by the candidate keyphrase frequency. In short, it captures the distance of a given candidate keyphrase to the whole text and thus provides a local, text-wide measurement of significance to a given candidate keyphrase. In our experiments, we found that a YAKE score of $S(x) \leq 0.1$, with x as the candidate keyphrase, delivers quality keyphrases that are not generic and can be used for further processing.

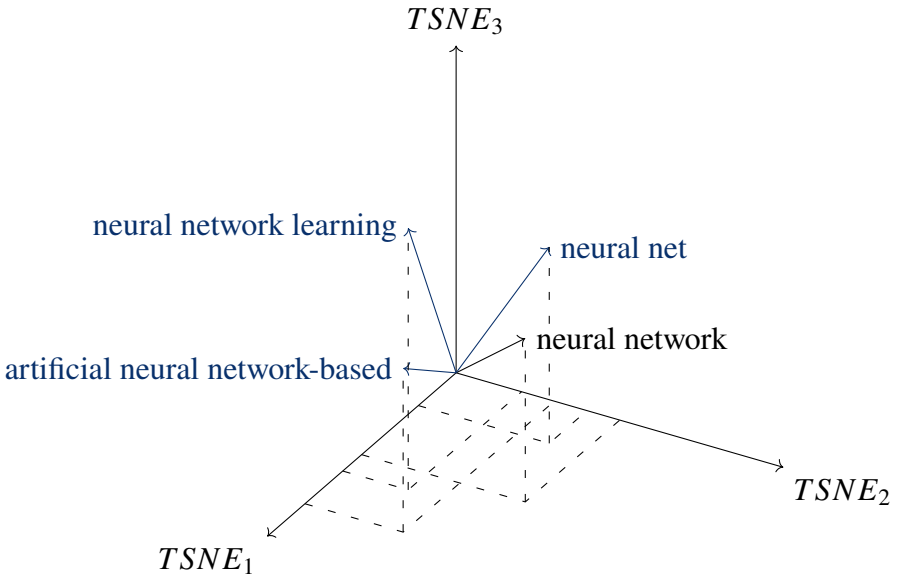
Through YAKE, we are generating up to 15 quality keyphrases for each given text. To further satisfy criteria two (Semantic Network Quality) and three (Descriptiveness), we employ a custom-built preprocessor. Starting with the semantic network quality, candidate keyphrases are brought into their canonical form through lemmatization. This process considers the context and morphological analysis of words inside the candidate keyphrases, ensuring the root word (lemma) is a valid linguistic entity. For example, the words "networks", "networking", "networked" will be converted to their lemmatized form: "network". To further enhance the semantic network quality, all keyphrases are converted into lowercase and very short terms (with less than 4 characters) are dropped. After this process, the aforementioned YAKE relevance score is examined and candidate keyphrases (x) with a score $S(x) > 0.1$ are dropped as well. Lastly, to comply with quality criteria three (Descriptiveness), we utilize a noun chunk check. This function takes in the whole text and generates noun chunks. Noun chunks contain at least one noun, may include adjectives, determiners, or pronouns associated with the noun and do not extend beyond a simple or compound noun phrase. Therefore, these noun chunks do not include verbs or clauses that aren't part of the noun phrase itself. The candidate keyphrases are checked against these noun chunks (of course in their lemmatized form) and candidate phrases that are not also noun chunks are dropped.

All in all, this leaves us with a clean set of quality keyphrases for a given text and constitutes the pipeline for all keyphrase extraction activities in this work.

Extending the Concept List Through Similarity Search

As we are aiming to create a comprehensive AI Concept List and enrich this list with more similar phrases taken from a large variety of scientific papers, a method has to be applied to find semantic similarities. Several techniques are known to produce high quality results (Mihalcea et al., 2006; Ali et al., 2018), we chose to opt for a rather new method that utilizes state of the art word embeddings from large language models to represent the given keyphrases in a 768-dimensional vector space and calculate nearest neighbors through cosine similarities between vectors to capture semantic similarity. An example can be seen in figure 2. Similar terms to the original term "neural network" are positioned in proximity and exhibit a high degree of similarity through the calculation of $cos(\theta)$ between its vector and all other vectors from other candidate keyphrases. This depiction is of course representative and does not capture the whole 768 dimensions.

FIGURE 2
Exemplary Similarity Search Mapping



Notes: An exemplary gold-standard keyphrase "neural network" with its closest semantic similar phrases reduced to three dimensions through TSNE (Maaten and Hinton, 2008)

Similar to approaches used by Mitchell and Dino (2011) as well as Liu et al. (2023), we

first take a set of concepts and embed these using the SentenceTransformers library (Reimers and Gurevych, 2019) with a carefully chosen model named "SciNCL" (Ostendorff et al., 2022). This embedding model was meticulously tuned to scientific language and allows for a good clustering performance on scientific topics. It performed best as a transformers baseline in the *SciRepEval* Benchmark (Singh et al., 2023). Next, we also embed all candidate keyphrases from a secondary source (like PapersWithCode abstracts, see the results section for details). By utilizing t-distributed stochastic neighbor embedding (TSNE) (Maaten and Hinton, 2008), we reduce the 768-dimensional embedding vectors of each keyphrase into two dimensional vectors for visualization and clustering purposes. This is done using the Python library `OpenTSNE`[6].

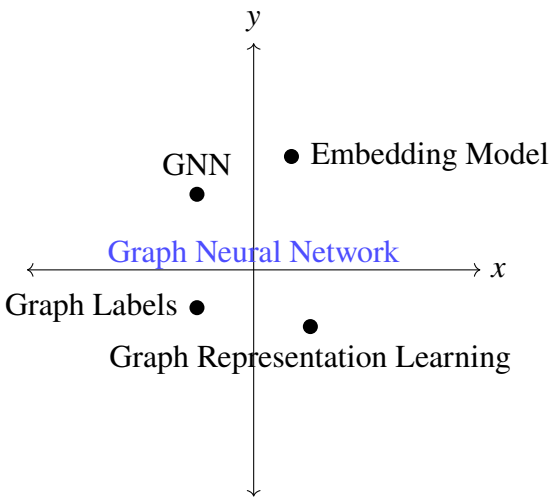
After obtaining two-dimensional vectors for each concept, we group them into various topics, followed by the determination of centroids through the computation of mean embedding vectors for all concepts within each cluster. This strategy is adept at mitigating the influence of outlier cases during the neighbor identification process for each keyphrase, thereby enhancing the resilience of the topic generation process. Regarding the selection of a clustering technique, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) method is employed (McInnes and Healy, 2017; McInnes et al., 2017). HDBSCAN surpasses conventional clustering algorithms in accuracy and consistency, facilitating the identification of appropriate clustering centroids. Further, topic representations were generated using the `transformers`[7] library from Huggingface and as a model, we chose a reasonable `flan-t5-xl` (Wei et al., 2022). Keyphrases of a given cluster were fed in alongside with a prompt: *"Given the following phrases, come up with a topic name that is specific and precise: [KEYWORDS]"*. With $t = 0.1$, a very deterministic temperature for the language model, we generate precise and descriptive topic representations.

The resulting topic centroids are taken as a starting point to extend the given concept list with candidate keyphrases (an example of which can be seen in in figure 3. We employ a simple k-nearest neighbor search algorithm onto the generated keyphrases obtained from the

[6]opentsne.readthedocs.io
[7]github.com/huggingface/transformers

PapersWithCode abstracts to match a fixed number of suitable concepts from current literature to the given topic centroids through semantic similarity by utilizing the `NearestNeighbors` function from Scikit-learn (Pedregosa et al., 2011).

FIGURE 3
Similarity in Cluster Centroids



Notes: An exemplary AI concept topic "Graph Neural Network" with its closest semantic similar phrases reduced to two dimensions through TSNE.

Finally, we merge the initial AI Concept List with semantically similar nearest neighbors to cluster centroids. This collection of quality phrases for AI concepts will now be put into a validation stage which additionally provides valuable weights for future use cases.

Regression and Validation

As a crucial step in the creation of the AI Concept List and subsequent semantic networks, all given concepts have to be validated against a test sample. This step will provide us with two very valuable insights: Firstly, we are able to filter out concepts that are still too generic and thus not beneficial in describing the domain of AI. Secondly, when choosing the right validation method, we can derive certain coefficients from the model to assign numeric values to the given concepts in order to mark their importance in the overall AI domain.

Initially, a suitable dataset has to be determined which acts as a negative sample to test the concept list against. In a perfect setting, this would be a large selection of scientific papers that are

not dealing with the topics of Artificial Intelligence or Machine Learning at all and are free to access. Therefore, we again utilize the extensive OpenAlex Database, which contains around 240 million scientific papers. To obtain a large sample of abstracts from papers with no ties to topics in the field of AI, a locally hosted version of this extensive database is built up since the API service from OpenAlex does not provide such large requests. We create a local version of the OpenAlex database in a dockerized PostgreSQL instance. This allows us to execute several queries and select specific papers with available abstracts that exclude concepts (like "Artificial Intelligence" or "Machine Learning"). Naturally, after acquisition of this large text corpus, we apply the pre-processing pipeline to generate quality keyphrases for that negative sample.

Next, we proceed to construct a phrase-document matrix. This process is facilitated by a custom Python package we developed, which leverages a trie data structure for efficient keyphrase search within lists or documents The underlying concept was outlined by (Brass, 2008) but the package provides additional features like creating whole semantic networks given a valid list of keyphrases and a selection of documents.

Initially, we categorize our document samples as either AI-related or Non-AI-related. Each document is represented by a list of keyphrases extracted from it. Let D be the set of all documents and P the set of all concept phrases in the AI Concept List. The Phrase Document Matrix M is defined as:

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1|P|} \\ M_{21} & M_{22} & \cdots & M_{2|P|} \\ \vdots & \vdots & \ddots & \vdots \\ M_{|D|1} & M_{|D|2} & \cdots & M_{|D||P|} \end{bmatrix} \tag{1}$$

where M_{ij} is the frequency of the j-th concept phrase in the i-th document. Each document is labeled as AI-related or Non-AI-related, forming the label vector $\mathbf{y}$:

$$\mathbf{y} = [y_1, y_2, \ldots, y_{|D|}]^\top \tag{2}$$

where $y_i = 1$ indicates an AI-related document and $y_i = 0$ otherwise.

As a validation method, it should be able to fit the data efficiently and provide interpretable coefficients to each concept phrase so that we can derive weights to the concepts. We used a logistic regression model in this setting since it provides a very fast and parallel runtime and can output regression coefficients to each given AI concept. This is rather difficult when dealing with machine learning classifiers. After careful consideration for the best hyper parameters to run the logistic regression, the resulting coefficients as well as the whole regression model can be used to validate the AI Concept List. The probability for the i-th document being AI-related is given by the sigmoid function:

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{|P|} x_{i|P|})}} \tag{3}$$

Here, $\mathbf{x}_i$ represents the feature vector (extracted from the phrase-document matrix M) for the i-th document, and β_j are the coefficients learned during the model training.

The logistic regression model is trained to minimize the negative log-likelihood, defined as:

$$\text{Cost}(\beta) = -\sum_{i=1}^{|D|} [y_i \log(P(y_i = 1|\mathbf{x}_i)) + (1 - y_i) \log(1 - P(y_i = 1|\mathbf{x}_i))] \tag{4}$$

The learned coefficients β_j indicate the importance of each AI concept in predicting the classification of a given document. Therefore, we can normalize each positive coefficient and treat it as a weighing factor.

To evaluate the applied logistic regression model, we employ k-Fold Cross-Validation as a robust statistical technique, to assess performance and stability. Specifically, we utilize a 10-fold cross-validation approach, effectively partitioning the dataset into ten distinct subsets. For each fold, we train the model on nine subsets and validate it on the remaining subset. This method aids

in mitigating overfitting and provides a more generalizable performance metric. The Cross-Validation Accuracy (CV Accuracy) is given by the formula:

$$\text{CV Accuracy} = \frac{1}{k} \sum_{i=1}^{k} \text{Accuracy}_i \tag{5}$$

where k is the number of folds (in our case, 10), and Accuracy_i is the accuracy score for the i^{th} fold. This formula calculates the average accuracy across all folds, providing a comprehensive picture of the model's performance.

The next section will now apply the given method on real-world data and lay out the creation of the AI Semantic Network and its findings as well as its implications and limitations.

RESULTS

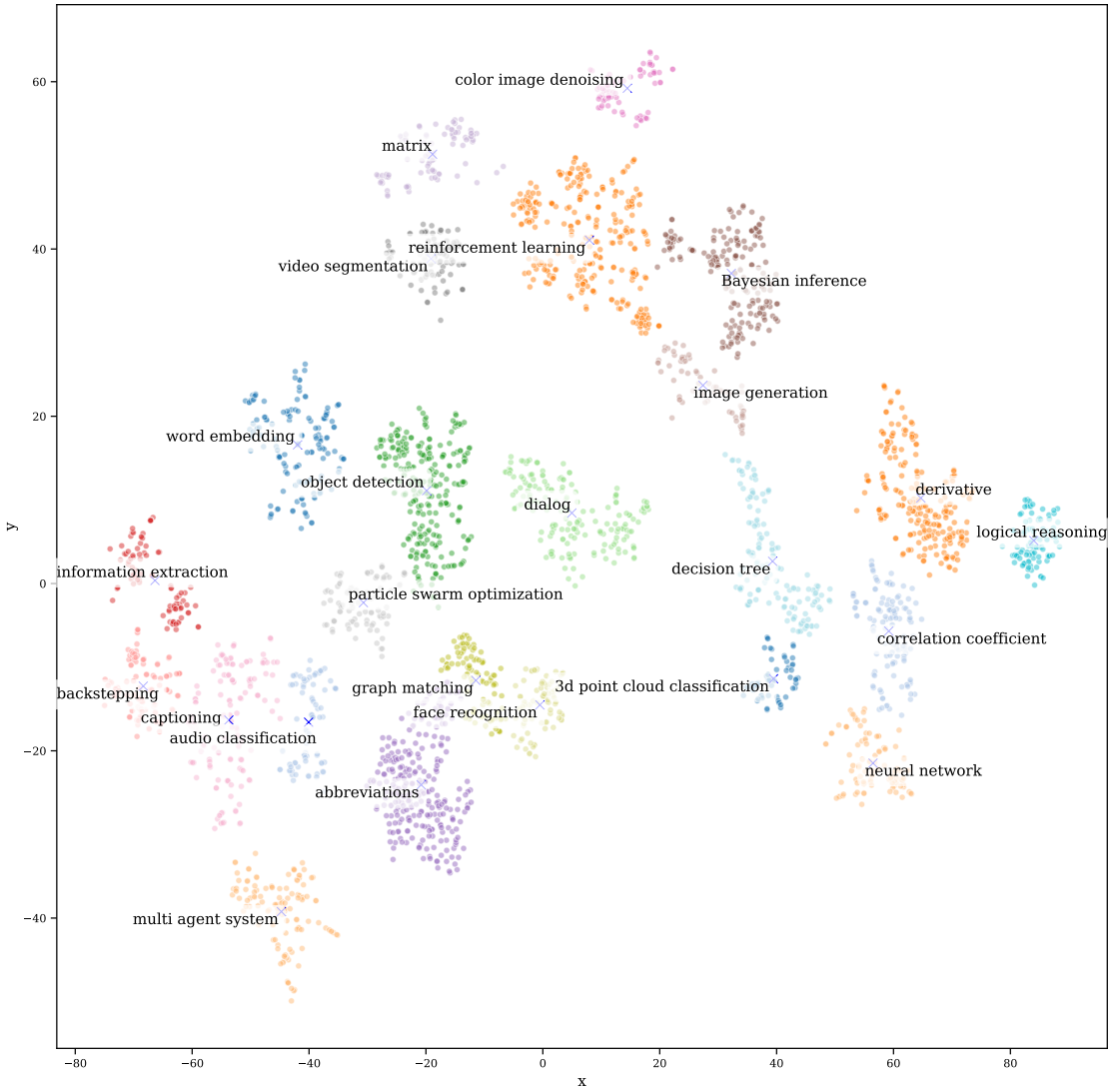
With the data acquisition, pre-processing pipeline, concept list extension and validation steps laid out, we will now apply those steps onto real-world data and present an exemplary AI Concept List as well as various Semantic Networks.

Initial Data Acquisition

As earlier mentioned, the AI Concept List consists of a mixture of different sources. First, we collect concepts from book indices out of three well known scientific books in the domain of Artificial Intelligence (Goodfellow et al., 2016; Murphy, 2022; Prince, 2023). To enrich these rather theoretical concepts, we also gather concepts from the Computer Science Ontology as well as from method and task names obtained from the PapersWithCode dataset. From these sources, we derive `14.655` raw candidate concepts. After processing these concepts (lowercase, removal of abbreviations, lemmatizing, etc.) and deduplicating the dataset, we end up with `9.070` unique AI concepts. These concepts are then embedded and clustered as described in the Data and Methods section. The resulting `25` clusters represent current tasks and topics of AI research in a wide variety of applications ranging from computer vision to language processing. Figure 3 depicts a 2-D scatter plot of the clusters with their calculated centroid, which will become

important in the next section.

FIGURE 4
2D Scatter Plot with clusters of initial AI Concepts



Notes: This figure depicts a 2-D scatter plot of the initial AI Concepts taken from book indices, PwC methods and tasks as well as CSO topics. These concepts are clustered by their semantic similarity and represented by a topic name.

Extending the AI Concept List

To enrich this list and therefore capture the current state of research as well as nuances in specific sub-domains, we gather keyphrases from all abstracts of papers within the PapersWithCode dataset utilizing our well-defined pre-processing pipeline. This results in

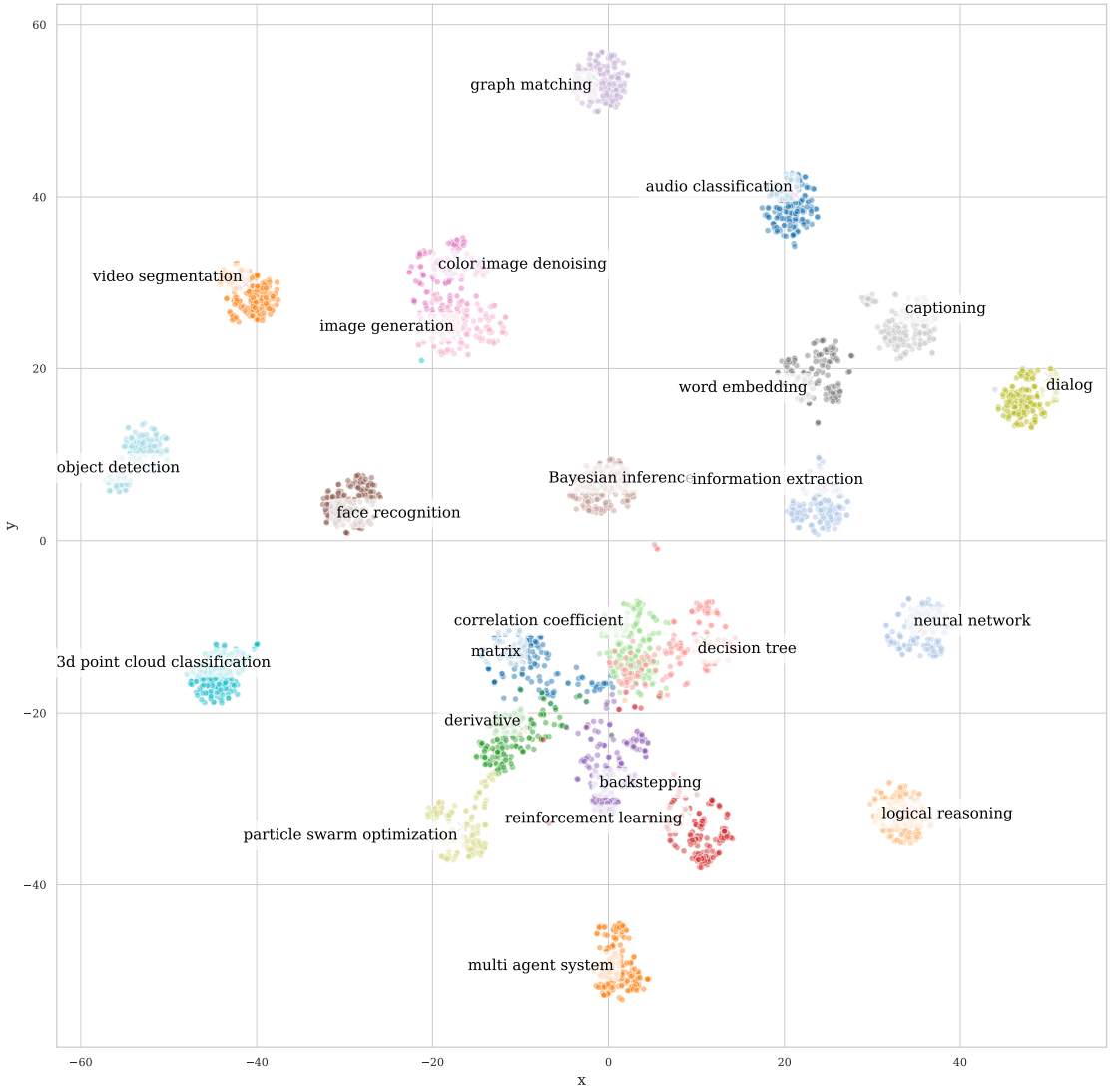
562.066 unique keyphrases after preprocessing and deduplicating. Our implementation of similarity search, as outlined in the Data and Methods section, incorporates abstract keyphrases as a secondary source, while the initial AI concepts and their resulting clusters are the primary source. We conduct a series of k-Nearest Neighbor calculations to AI concept cluster centroids by gathering the 100 nearest neighboring candidate AI concept phrases to each cluster. We also exclude the cluster of abbreviations (like gnn, rnn, etc.) since semantic similarity is not interpretable here. This results in an extension to the initial AI Concept List of 2.507 new keyphrases and leaves us with a comprehensive AI Concept List consisting of 11.577 phrases. Figure 5 depicts the 2-D scatter plot of the given AI Concept centroids and the resulting nearest neighbors around those points. Only a few of them are actually chosen, since they overlap strongly with the initial AI concept list.

Regression and Validation

Finally, we employ our validation method with 1.219.378 negative keyphrase sample documents obtained from a large sample of Non-AI related OpenAlex Publications as well as 415.941 positive keyphrase sample documents from the PapersWithCode dataset. As described before, this results in a stacked phrase document matrix, where the negative and positive keyphrases are grouped to their original document as D (totalling 1.635.319 entries) and the AI concept phrases as P (totalling 11.577 entries). Because of the slightly imbalanced nature of this dataset, we set the class weight to balanced in the scikit learn logistic regression model. We calculate the frequency of each AI concept in each document keyphrase collection to populate the matrix M. The resulting dimensions of M are $1.635.319 \times 11.577$, while the label vector y has a length of 1.635.319.

Fitting the logistic regression model and applying our 10-fold cross validation step, the resulting accuracy converges at 87%. While this accuracy level indicates a reasonable degree of correctness in predictions, it is important to note that the primary objective of employing logistic regression in our study is not to achieve optimal classification accuracy per se. Instead, our focus

FIGURE 5
2D Scatter Plot with AI Concept Centroids and k-Nearest Neighbors

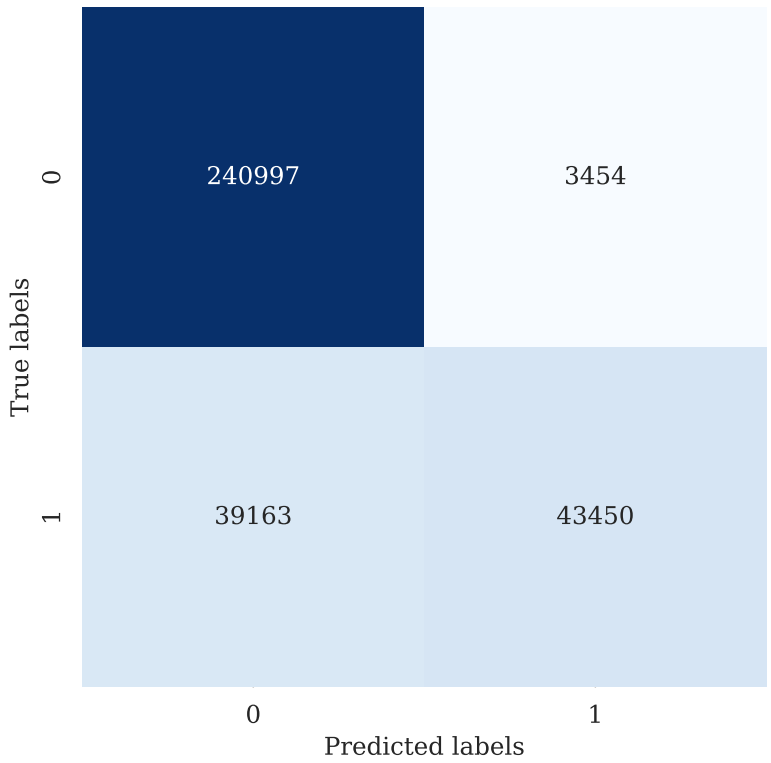


Notes: This figure depicts a 2-D scatter plot of the initial cluster centroids taken from the Data Acquisition section. These AI clusters are extended by their nearest neighbor candidate AI concept phrases taken from PwC abstracts. The position of cluster centroids is different to figure 4 due to the TSNE algorithm.

is on leveraging the regression coefficients to assign importance weights to the keyphrases. These weights are instrumental in evaluating the significance and relevance of each phrase within the broader context of AI research and innovation. The confusion matrix, which provides insight into the model's performance in terms of true positives, true negatives, false positives, and false negatives, is presented in figure 6. This matrix reveals that the model correctly identified a

significant number of true negatives as well as true positives, while the number of false positives remained considerably lower. The number of false negatives indicates a specific area for improvement.

FIGURE 6
Resulting confusion matrix of the logistic regression



Notes: This figure depicts the resulting confusion matrix from fitting label vector y to matrix M in the logistic regression. Here, class 0 stands for Non-AI samples, while 1 are AI samples.

Furthermore, the classification report in table 1 provides detailed insights into the model's performance. The precision, recall, and F1-score for each class highlight the model's strengths and weaknesses in classifying each category. The high precision in class 1 indicates a low false positive rate, while the recall and F1-score suggest areas for potential improvement in model sensitivity and the harmonic mean of precision and recall, respectively.

When mapping the regression coefficients to the AI keyphrases, we found that `4.551` out of the `11.577` keyphrases had no regression coefficient and were therefore perfectly aligned with

TABLE 1
Logistic Regression Classification Report for the AI Concept List

Class	Precision	Recall	F-score	Support
0	0.87	0.99	0.92	244451
1	0.93	0.53	0.67	82613
accuracy			0.87	327064
macro avg	0.89	0.76	0.79	327064
weighted avg	0.88	0.87	0.86	327064

Notes: Classification Report generated on the test sample, which is 20% of the original sample.

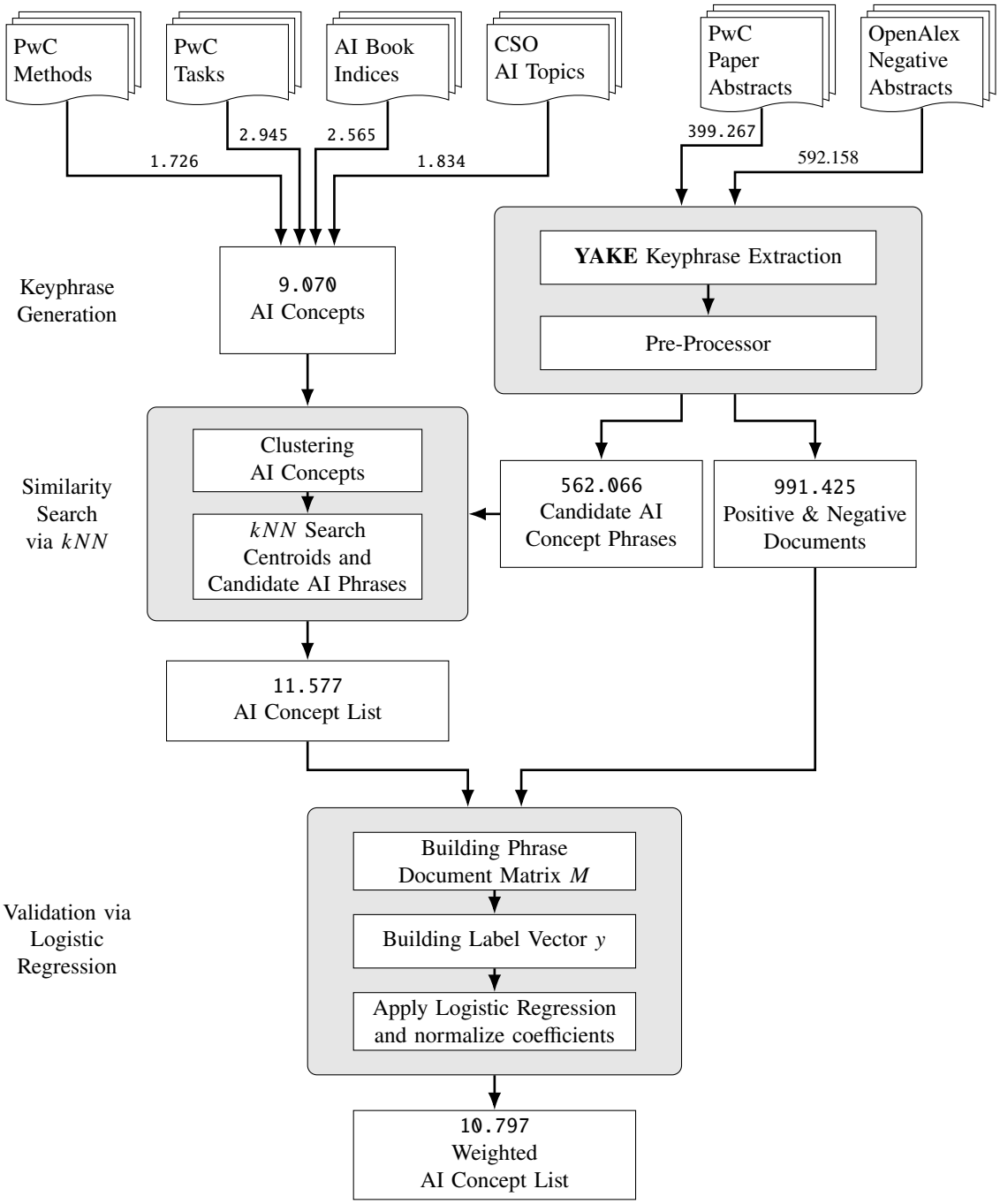
the classification set (only appeared in the positive sample). Further, we found that `780` keyphrases were associated with a negative regression coefficient and thus not contributing to the overall performance of the regression model. To improve our AI Concept List, we removed these negative concepts, leaving us with `10.797` high-quality AI Concept Phrases. Next, we used a k-Nearest Neighbor approach to find phrases lacking regression coefficients and assigned them coefficients from their closest semantic matches for consistency. Lastly, we normalize these regression coefficients to create a set of importance weights for each of the `10.797` AI Concepts.

Each step can be seen in an overview figure 7. A sample of the AI Concept List with importance weights, regression coefficients as well as some graphical representations are provided in Appendix A. The complete AI Concept List can be found on Huggingface for further examination.

Creating Semantic Networks

With the refined AI Concept List now comprising numerous concept phrases with importance weights, an interesting pattern emerges from the histogram in figure 8. It reveals that only few keyphrases are assigned with very high weights, while the majority are assigned medium to lower weights. This could suggest that the AI Concept List is focusing on a select group of highly relevant phrases, possibly due to their frequent occurrence or strong association with specific topics. The distribution of weights might also reflect the natural language use, where

FIGURE 7
AI Concept List Pipeline

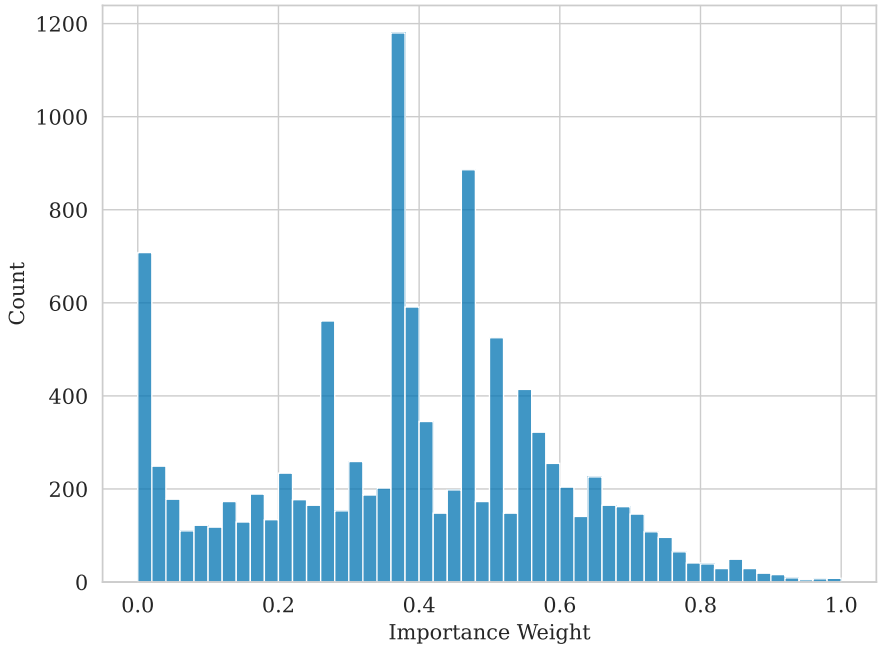


Notes: Standardized Pipeline to create a weighted AI Concept List out of a diverse set of sources. See the results section for details.

certain phrases are more central or pivotal to discussions than others. The concept list, therefore, appears to be not uniformly distributed but rather concentrated around a few significant phrases,

indicating a skewed importance towards certain terms within the dataset.

FIGURE 8
Histogram of AI keyphrase Importance Weights



Notes: This figure depicts a histogram showing the counts of AI concepts to their respective importance weights.

With the finalized AI Concept List, we are now able to create individual semantic networks on different scientific domains. We take the aforementioned OpenAlex Concepts, since they provide a broad categorization of scientific papers into multi-level abstractions and gather a sample of 100k scientific paper abstracts for each existing level 0 concepts (19 samples in total). Each sample only contains english abstracts which are also fed through our Pre-Processing Pipeline to gather high-quality texts. An overview can be seen in table 2.

The aforementioned custom-made Python Package "Wordtrie" is also designed to analyze and map a corpus of documents to various AI Concepts. These documents are mapped to the AI Concept List by constructing an occurrence matrix. This matrix, denoted as **O**, forms the basis of each AI Semantic Network and is defined such that each row corresponds to a unique AI concept,

TABLE 2
Overview of Domain Samples for the AI Semantic Networks

Domain Concept Name	$N_{Documents}$	$N_{AvgWords}$
Art	19.073	156, 96
Biology	100.000	231, 03
Business	100.000	144, 17
Chemistry	100.000	130, 21
Computer Science	100.000	149, 28
Economics	100.000	141, 94
Engineering	17.979	98, 83
Environmental Science	100.000	223, 92
Geology	582.167	218, 61
History	4.487	53, 42
Materials Sciences	100.000	150, 04
Mathematics	100.000	102, 31
Medicine	100.000	224, 94
Philosophy	39.817	188, 17
Physics	100.000	141, 92
Political Science	17.506	121, 40
Psychology	100.000	170, 92
Sociology	15.970	163, 68

Notes: Overview of all Samples taken from the OpenAlex Database with their average abstract word count. Some domains come with a smaller sample size because of a lack of coverage in the OpenAlex database.

and each column represents a document within the corpus. The entries of **O**, denoted as o_{ij}, are the counts of occurrences of the i^{th} AI concept within the j^{th} document. The dimensions of **O** are thus $n \times m$, where n is the total number of AI concepts considered (10.797), and m is the number of documents in the corpus.

The occurrence matrix **O** constitutes an AI Semantic Network for each domain, which is instrumental in visualizing and analyzing the interconnections between documents and AI concepts. This network allows for the identification of prevalent themes and trends within the domain of AI research. By treating AI concepts as nodes and their occurrences within documents as edges, we can construct a graph that represents the semantic relationships inherent in the

corpus. One of the primary applications of this semantic network is the temporal analysis of AI research trends. By aggregating the occurrences of AI concepts across documents over time, we can plot the number of documents containing AI concepts as a function of time. This yields valuable insights into the evolution of interest and research focus within the field of AI.

In summary, we see an active prevalence of AI concepts across various research domains. Figure 9 depicts the percentage of documents that have at least 3 occurrences of AI concepts inside their titles and abstracts. Not surprisingly, domains like mathematics or computer science show high engagements while other domains like art or history are not engaging with AI tools or methods. Figure 10 highlights the expanding footprint of AI in certain research domains, suggesting its application is becoming more widespread across disciplines such as Computer Science, Environmental Sciences and Medicine, sometimes as a fundamental pillar, other times as an important tool in specific contexts.

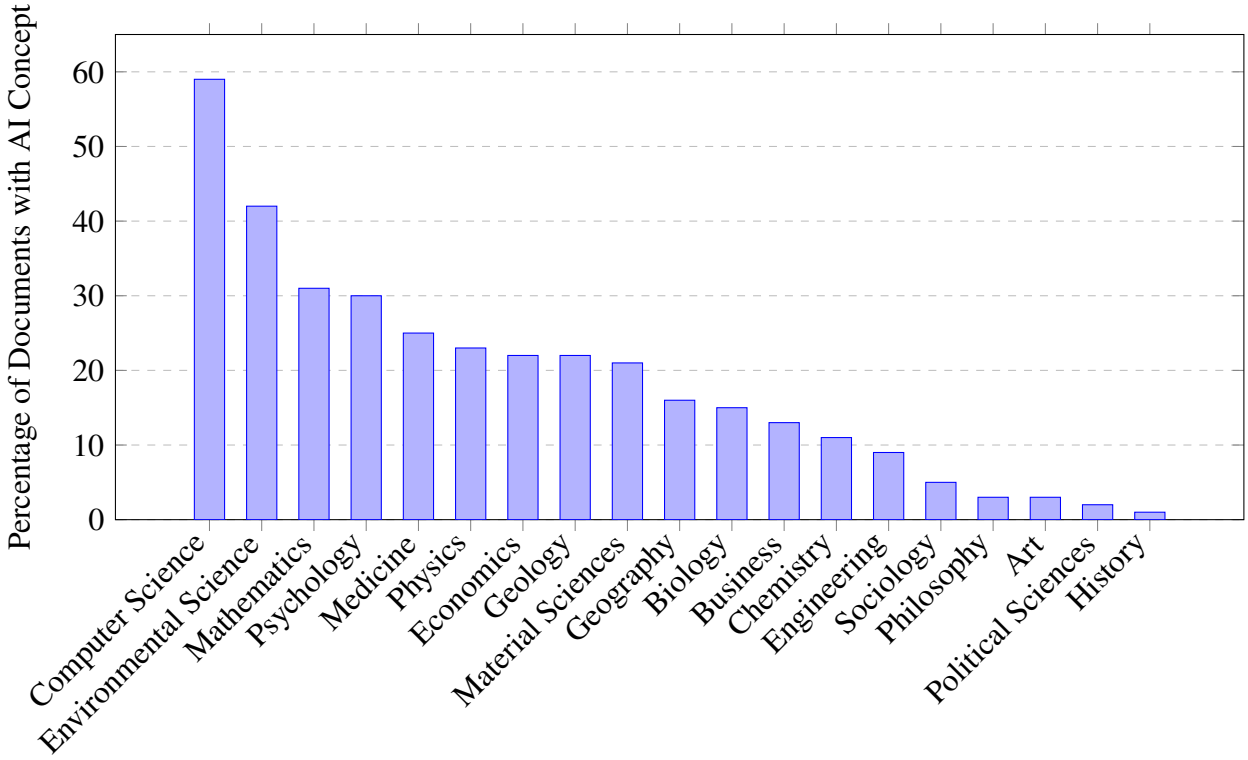
To put these results into perspective, we now transition into the discussion and future research. We will explore, how these patterns reflect current trends and gaps in AI research and what they reveal about the field's evolution. Further, we will look at possible applications and implications for both research and industry as well as some limitations to this approach.

DISCUSSION AND IMPLICATIONS

In this final section, we will engage in a comprehensive analysis of our findings, particularly the distribution and usage of AI in research. We further shed light on possible applications of various semantic networks as a measurement of innovation in certain domains and a foundation for decision support systems in critical situations.

When looking at the AI Concept List and its resulting topics (figure 4), we can clearly see, that the initial concept phrases capture a wide bandwidth of domains and tasks. Unsurprisingly, we are able to find multiple topics that are covered by the PapersWithCode Dataset (e.g. "face recognition") as well as foundational topics (e.g. "reinforcement learning") discussed in recent books and papers. When tuning the clustering algorithm, more intricate topics can be found as

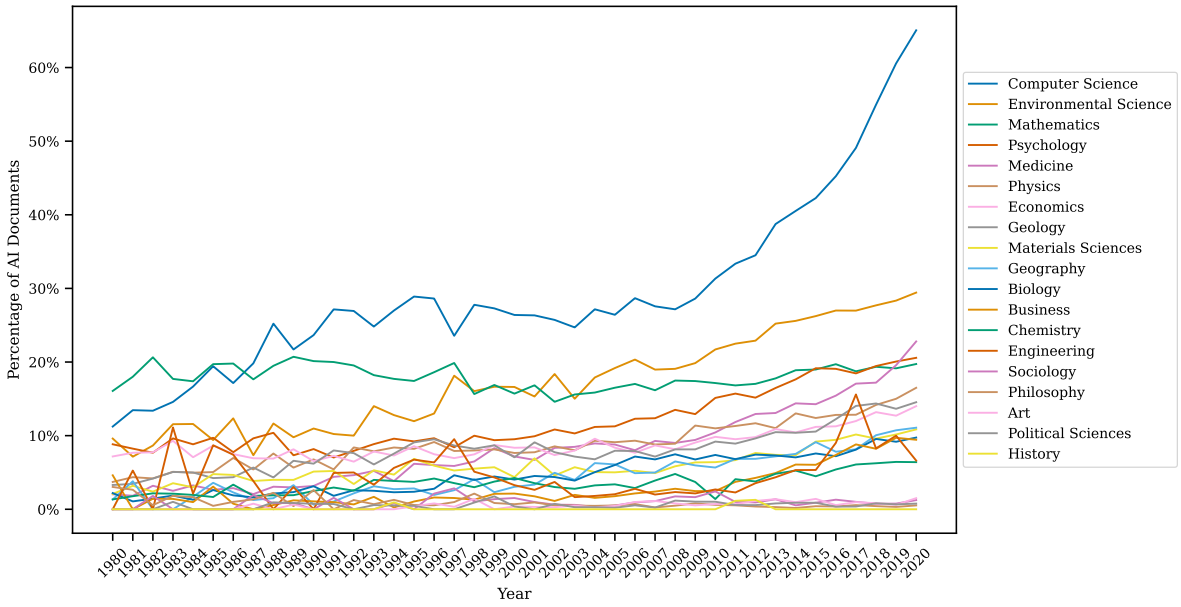
FIGURE 9
Percentage of Documents with AI Concept per Domain



Notes: Accumulated percentage of the occurrence of at least 3 AI concepts in different research domains.

well. The superiority of our approach is evident in the AI Concept Lists ability to accurately and dynamically reflect the rapidly evolving landscape of AI research. Unlike earlier attempts, which often relied on narrowly focused datasets (Baek et al., 2021; Giczy et al., 2022), meta information (Park et al., 2023) or simpler text-mining methods like named entity recognition (Fleuren and Alkema, 2015) or TF-IDF (Tseng et al., 2007), our methodology integrates a wider array of sources and employs advanced text-based analytics. Further, thanks to a modular implementation, new sources can be added easily. This ensures a more robust and holistic representation of the AI field, encompassing emerging trends and niche areas that were previously underrepresented or overlooked in academic discourse. In addition to that, we are also planning on incorporating more features into our pipeline, as we transition to an extensive knowledge graph that extends beyond the scope of the current AI Concept List (with code repositories, fulltext analysis, authors,

FIGURE 10
Level 0 AI Concept occurrences in different research domains over time



Notes: This figure depicts the occurrence of AI concepts in different scientific papers associated with various domains over time.

institutions, funding patterns, etc.).

When applying the AI Concept List to form semantic networks in certain research domains, we identify a steady rise in the use of AI methods, tasks and theorems. The steady increase in AI-related publications as seen in figure 10 mirrors significant technological advancements, including the rise of deep learning and enhanced computational capacities, which have made AI a more accessible and valuable resource for tackling complex research questions. Moreover, the temporal growth pattern underscores periods of intensified activity, likely influenced by technological breakthroughs and shifts in research funding towards AI. While AI's integration is varied and its impact differs across domains, its role in facilitating research and offering novel solutions is increasingly recognized. This trend points to an increased interest in AI's applications and its potential to contribute to diverse fields of study, reflecting its role as a valuable, albeit not universal, tool in the scientific research toolkit.

These AI Semantic Networks serve as a starting point for many use cases. By tracking the frequency and evolution of the usage of given concepts over time, researchers can gain insights into emerging trends, shifts in focus areas and the evolution of thought within specific domains. This tool could therefore help to identify emerging technologies or theories, providing valuable quantitative measurements of how certain concepts gain traction in the academic community either in a certain domain or in a broader perspective. It can also assist in literature review processes, helping researchers to quickly find relevant works based on the prevalence of concept phrases. Such a tool can also be employed to uncover new connections between different research domains. Through the unconventional and fast paced evolution of AI methods and frameworks, more and more researchers apply these methods to their subject and stumble across unexpected results. With the use of semantic networks and an efficient keyphrase extraction tool, the utilization of AI systems could be easily identified and connected to other domains. This could be applied to virtually any area, as long as there is an access to high-quality textual information like scientific papers, patents, press releases or websites.

In the industrial realm, the AI Concept List and subsequent semantic networks could be pivotal as a component of decision support systems. When looking at recent radical innovations in the domain of AI, decision-makers face ever increasing dilemmas as they have to navigate a complex web of choices under growing pressure and accelerating timelines (Eisenhardt, 1989; Kengpol and O'Brien, 2001; Duan et al., 2019). By utilizing the AI Concept List and analyzing the prevalence as well as the importance weights of AI concepts, companies can gauge market trends, technological advancements, and emerging consumer needs. This could be particularly useful for product development, marketing strategies, and competitive analysis. For instance, a company in the tech sector could use this instrument to stay informed of the latest developments in AI, ensuring their products align with current trends and consumer expectations. Moreover, the applications could extend to predictive analytics. By examining the trajectory of certain concepts over time, organizations could predict future trends in technology and consumer behavior. This predictive capability would be invaluable for strategic planning and long-term decision-making.

LIMITATIONS AND FUTURE RESEARCH

The applications of such an AI Concept List and semantic network must be carefully considered. Each step of the creation pipeline could be extended to incorporate more data, produce more fine-grained results or generate higher-quality outputs. The Data Acquisition part is a pivotal factor in digesting novel and recent developments in the realm of AI. The already quite extensive selection of sources could be expanded with more papers, books or other documents to capture even more aspects and gain deeper expert knowledge on certain parts or whole domains. Key phase generation could be done on several levels (inspired by Shang et al. (2017)) to better adhere to the keyphrase criteria. Additionally, the phase of Regression & Validation could be optimized by implementing a streaming service that constantly updates and re-assesses the regression coefficients as well as resulting weights for the given concepts. This would also benefit the accuracy of our validation method. Lastly, when turning to the results, the selection of domain samples can be improved. OpenAlex provides a comprehensive categorization into different topics, but the selection of papers could be further filtered by venue, institution or other factors. It is possible, that publications are mislabeled in individual cases.

On a more qualitative note, the AI Concept List and semantic network might be instrumentalized in driving radical innovation. It is essential to consider how certain concepts might highlight or overshadow emerging and disruptive technologies. There is no guarantee, that each and every development in the domain of AI is reflected in these concepts or subsequently in the semantic networks. The risk lies in the potential to reinforce existing knowledge and paradigms, possibly at the cost of novel or radical ideas. Therefore, it is critical to explore ways to calibrate the concept list to recognize and elevate groundbreaking concepts, ensuring it becomes a tool that not only tracks but also fosters innovation. Another area of exploration could be the development of filters or lenses within the semantic network that focus on identifying and highlighting potentially disruptive technologies or theories, thereby aligning more closely with the goal of driving radical innovation.

In conclusion, this paper presents a comprehensive AI Concept List and its application to

form semantic networks in different research domains. It provides a novel tool consisting of concept phrases that describe the domain of Artificial Intelligence. It surpasses traditional methods of measuring technology innovation by incorporating a wide range of sources and offering insights into emerging trends across various domains. Its utility is significant in both academic research and industry, particularly when it comes to identifying and measuring radical innovation. While this tool is promising, future enhancements are necessary to address its limitations, such as refining the pipeline and ensuring it highlights disruptive innovations without reinforcing existing paradigms. All in all, the AI Concept List's role as a foundation for an instrument to measure technological innovation is essential for further research in this field.

REFERENCES

Aceves, P. and Evans, J. 2022. Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory within Organization Science. preprint, SocArXiv.

Ali, A., Alfayez, F., and Alquhayz, H. 2018. Semantic Similarity Measures Between Words: A Brief Survey. ***Science International***, 30.

Arts, S., Hou, J., and Gomez, J. C. 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. ***Research Policy***, 50(2): 104144.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., and De Souza, A. F. 2021. Self-driving cars: A survey. ***Expert Systems with Applications***, 165: 113816.

Baek, S., Jung, W., and Han, S. H. 2021. A critical review of text-based research in construction: Data source, analysis method, and implications. ***Automation in Construction***, 132: 103915.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., and Hidalgo, C. A. 2020. Complex economic activities concentrate in large cities. ***Nature Human Behaviour***, 4(3): 248–254. Number: 3 Publisher: Nature Publishing Group.

Benbya, H., Pachidi, S., and Jarvenpaa, S. 2021. Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. ***Journal of the Association for Information Systems***, 22(2).

Brass, P. 2008. ***Advanced Data Structures***. Cambridge: Cambridge University Press.

Bresnahan, T. F. and Trajtenberg, M. 1995. General purpose technologies 'Engines of growth'? ***Journal of Econometrics***, 65(1): 83–108.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., and Jatowt, A. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A.(Eds.), ***Advances in Information Retrieval***, volume 10772, 806–810. Cham: Springer International Publishing. Series Title: Lecture Notes in Computer Science.

de Rassenfosse, G. and van Pottelsberghe de la Potterie, B. 2009. A policy insight into the R&D–patent relationship. ***Research Policy***, 38(5): 779–792.

Du, S. and Xie, C. 2021. Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. ***Journal of Business Research***, 129: 961–974.

Duan, Y., Edwards, J., and Dwivedi, Y. K. 2019. Artificial intelligence for decision making in the era of Big Data - evolution, challenges and research agenda. ***Int. J. Inf. Manag.***, 48: 63–71.

Eisenhardt, K. M. 1989. Making Fast Strategic Decisions in High-Velocity Environments. ***The Academy of Management Journal***, 32(3): 543–576. Publisher: Academy of Management.

Fleuren, W. W. M. and Alkema, W. 2015. Application of text mining in the biomedical domain. ***Methods***, 74: 97–106.

Giczy, A. V., Pairolero, N. A., and Toole, A. 2021. Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset.

Giczy, A. V., Pairolero, N. A., and Toole, A. A. 2022. Identifying artificial intelligence (AI) invention: a novel AI patent dataset. ***The Journal of Technology Transfer***, 47(2): 476–505.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. ***Deep Learning***. MIT Press. Google-Books-ID: omivDQAAQBAJ.

Gärdenfors, P. 2004. ***Conceptual Spaces: The Geometry of Thought***. MIT Press. Google-Books-ID: FSLFjw1EcBwC.

Gärdenfors, P. 2014. ***The Geometry of Meaning: Semantics Based on Conceptual Spaces***. The MIT Press.

Hannan, M. T., Le Mens, G., Hsu, G., Kovács, B., Negro, G., Pólos, L., Pontikes, E. G., and Sharkey, A. J. 2019. ***Concepts and Categories: Foundations for Sociological and Cultural Analysis***. Columbia University Press.

Helpman, E. and Trajtenberg, M. 1994. A Time to Sow and a Time to Reap: Growth Based on General Purpose Technologies. ***NBER Working Papers***. Number: 4854 Publisher: National Bureau of Economic Research, Inc.

Kengpol, A. and O'Brien, C. 2001. The development of a decision support tool for the selection of advanced technology to achieve rapid product development. ***International Journal of Production Economics***, 69: 177–191.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., Crawford, M., Downey, D., Dunkelberger, J., Etzioni, O., Evans, R., Feldman, S., Gorney, J., Graham, D., Hu, F., Huff, R., King, D., Kohlmeier, S., Kuehl, B., Langan, M., Lin, D., Liu, H., Lo, K., Lochner, J., MacMillan, K., Murray, T., Newell, C., Rao, S., Rohatgi, S., Sayre, P., Shen, Z., Singh, A., Soldaini, L., Subramanian, S., Tanaka, A., Wade, A. D., Wagner, L., Wang, L. L., Wilhelm, C., Wu, C., Yang, J., Zamarron, A., Van Zuylen, M., and Weld, D. S. 2023. The Semantic Scholar Open Data Platform. arXiv:2301.10140 [cs].

Klinger, J., Mateos-Garcia, J., and Stathoulopoulos, K. 2021. Deep learning, deep change? Mapping the evolution and geography of a general purpose technology. ***Scientometrics***, 126(7): 5589–5621.

Krenn, M. and Zeilinger, A. 2020. Predicting research trends with semantic and neural networks with an application in quantum physics. ***Proceedings of the National Academy of Sciences***, 117(4): 1910–1916.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. ***Nature***, 521(7553): 436–444.

Lee, J., Suh, T., Roy, D., and Baucus, M. 2019. Emerging Technology and Business Model Innovation: The Case of Artificial Intelligence. ***Journal of Open Innovation: Technology, Market, and Complexity***.

Lenz, D. and Winker, P. 2020. Measuring the diffusion of innovations with paragraph vector topic models. ***PLOS ONE***, 15(1): e0226685. Publisher: Public Library of Science.

Liu, J., Chang, H., Forrest, J. Y.-L., and Yang, B. 2020. Influence of artificial intelligence on technological innovation: Evidence from the panel data of china's manufacturing sectors. ***Technological Forecasting and Social Change***, 158: 120142.

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., and Lee, I. 2018. Artificial Intelligence in the 21st Century. ***IEEE Access***, 6: 34403–34421. Conference Name: IEEE Access.

Liu, Y., Wu, H., Huang, Z., Wang, H., Ning, Y., Ma, J., Liu, Q., and Chen, E. 2023. TechPat: Technical Phrase Extraction for Patent Mining. ***ACM Transactions on Knowledge Discovery from Data***, 17(9): 1–31.

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Ikamura, K., Gerber, G., Liang, I., Le, L. P., Ding, T., Parwani, A. V., and Mahmood, F. 2023. A Foundational Multimodal Vision Language AI Assistant for Human Pathology. arXiv:2312.07814 [cs].

Maaten, L. v. d. and Hinton, G. 2008. Visualizing Data using t-SNE. ***Journal of Machine Learning Research***, 9(86): 2579–2605.

Makridakis, S. 2017. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. ***Futures***, 90: 46–60.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. ***AI Magazine***, 27(4): 12–12. Number: 4.

McInnes, L. and Healy, J. 2017. Accelerated Hierarchical Density Clustering. In ***2017 IEEE International Conference on Data Mining Workshops (ICDMW)***, 33–42. arXiv:1705.07321 [stat].

McInnes, L., Healy, J., and Astels, S. 2017. hdbscan: Hierarchical density based clustering. ***Journal of Open Source Software***, 2(11): 205.

Mihalcea, R., Corley, C., and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In ***Proceedings of the 21st national conference on Artificial intelligence - Volume 1***, AAAI'06, 775–780, Boston, Massachusetts. AAAI Press.

Mitchell, R. K. and Dino, R. N. 2011. ***In Search of Research Excellence: Exemplars in Entrepreneurship***. Edward Elgar Publishing.

Murphy, K. P. 2022. ***Probabilistic machine learning: An introduction***. MIT Press.

Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., and Rehm, G. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. arXiv:2202.06671 [cs].

Park, M., Leahey, E., and Funk, R. J. 2023. Papers and patents are becoming less disruptive over time. ***Nature***, 613(7942): 138–144.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. ***Journal of Machine Learning Research***, 12(85): 2825–2830.

Pelletier, P. and Wirtz, K. 2022. Novelpy: A Python package to measure novelty and disruptiveness of bibliometric and patent data. arXiv:2211.10346 [cs].

Priem, J., Piwowar, H., and Orr, R. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs].

Prince, S. J. 2023. ***Understanding deep learning***. MIT Press.

Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In ***2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 10674–10685, New Orleans, LA, USA. IEEE.

Russell, S. J. and Norvig, P. 2016. ***Artificial intelligence: a modern approach***. Prentice Hall series in artificial intelligence. Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Pearson, third edition, global edition edition.

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., and Motta, E. 2020. The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas. ***Data Intelligence***, 2(3): 379–416.

Sarker, I. H. 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. ***SN Computer Science***, 2(6): 420.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., and Han, J. 2017. Automated Phrase Mining from Massive Text Corpora. arXiv:1702.04457 [cs].

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. 2023. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. arXiv:2211.13308 [cs].

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. 2007. Text mining techniques for patent analysis. ***Information Processing & Management***, 43(5): 1216–1247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. Attention is All you Need. In ***Advances in Neural Information Processing Systems***, volume 30. Curran Associates, Inc.

Wani, S. U. D., Khan, N. A., Thakur, G., Gautam, S. P., Ali, M., Alam, P., Alshehri, S., Ghoneim, M. M., and Shakeel, F. 2022. Utilization of Artificial Intelligence in Disease Prevention: Diagnosis, Treatment, and Implications for the Healthcare Workforce. ***Healthcare***, 10(4): 608. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652 [cs].

Yang, Y., Akers, L., Klose, T., and Barcelon Yang, C. 2008. Text mining and visualization tools – Impressions of emerging capabilities. ***World Patent Information***, 30(4): 280–293.

APPENDIX A: SAMPLE OF THE AI CONCEPT LIST

In this appendix section, we present a sample excerpt of 30 entries from our AI Concept List in table A1. We provide each concept phrase with its regression coefficient and importance weight. The whole semantic network can be found in the corresponding GitHub Repository and Huggingface Dataset.

TABLE A1
Sample excerpt of 30 lines from the AI Concept List

Concept	Regression Coefficient	Importance Weight
graph neural network	8.589677	1.000000
neural network	8.512252	0.990986
multi-task learning manner	6.824995	0.794558
normalizing flow	5.564290	0.647788
binary classification decision	5.254628	0.611738
point cloud analysis	4.991500	0.581104
smart city	4.749547	0.552937
open information extraction	4.673741	0.544111
fuzzy logic	4.551451	0.529875
graph convolutional	4.449122	0.517962
global black-box optimization	4.447448	0.517767
adversarial training process	4.266508	0.496702
crelu	3.890238	0.452897
preceding dialogue context	3.882794	0.452030
humanoid robot control	3.472160	0.404225
adversarial training mechanism	3.447482	0.401352
simultaneous mutual information	3.447327	0.401334
grammar induction	3.447202	0.401319
semantic relatedness	3.447172	0.401316
standard machine translation	3.446959	0.401291
k nearest neighbor method	3.156036	0.367422
slightly lower accuracy	3.112892	0.362399
agent based	3.111854	0.362278
typical weakly supervised	3.111457	0.362232
original feature map	2.840155	0.330647
channel attention	2.828382	0.329277
zero shot dst setting	2.677841	0.311751
important observation	2.537400	0.295401
invariant local feature	2.534656	0.295082
deep nonparametric clustering	2.534377	0.295049