DISSERTATION APPROVAL FORM AND SIGNATURE PAGE

Instructions: This form must be completed by all doctoral students with a dissertation requirement. This form MUST be included as page 1 of your thesis via electronic submission to ProQuest.

Dissertation Title: Large-Scale Materials Knowledge Extraction Using LLMs and Human-in-the-loop

Author's Name: Xintong Zhao

Submission Date: 06/07/2025

The signatures below certify that this dissertation is complete and approved by the Examining Committee.

Role: Chair Name: Xiaohua Hu
Title: Professor
Department: Informatics
Approved: Yes Date: 06/08/2025

Role: Co-Chair Name: Jane Greenberg
Title: Kroeger Professor
Department: Informatics
Approved: Yes Date: 06/08/2025

Role: Member Name: Yuan An
Title: Associate Professor
Department: Informatics
Approved: Yes Date: 06/09/2025

Role: Member Name: Fernando Uribe Romo
Title: Associate Professor
Institution: University of Central Florida
Approved: Yes Date: 06/08/2025

Role: Member Name: Ron Daniel
Title: Founder
Institution: AI4Science.com
Approved: Yes Date: 06/08/2025

Large-Scale Materials Knowledge Extraction Using LLMs and Human-in-the-loop

A Dissertation Defense
submitted to the College of
Computing and Informatics
of Drexel University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Information Science

By

Xintong Zhao, M.S.

Philadelphia, PA
June 7, 2025

LARGE-SCALE MATERIALS KNOWLEDGE EXTRACTION USING LLMS AND HUMAN-IN-THE-LOOP

Xintong Zhao, M.S.

Dissertation Defense Committee: Xiaohua Hu, Ph.D.(Committee Chair), Jane Greenberg, Ph.D.(Co-advisor), Yuan An, Ph.D., Fernando Uribe-Romo, Ph.D.and Ron Daniel, Ph.D.

ABSTRACT

Unstructured scientific text plays a critical role in preserving, transferring, and developing research knowledge. Valuable outputs are often recorded in forms such as patents, research articles, and project reports. Unlike generic text, scientific literature usually follows specialized formats and terminology. This significant difference leads to greater challenges and opportunities for NLP (Natural Language Processing) researchers. To automate the process of extracting and structuring domain-specific knowledge from unstructured text, this dissertation addresses these challenges by leveraging NLP methods for automated materials science knowledge extraction.

Through three case studies, this dissertation explores the use of deep learning, LLM (Large Language Model) and prompt-based techniques to extract critical materials synthesis knowledge from scientific texts. Building on these efforts, the dissertation introduces an end-to-end, cost-effective framework designed for large-scale knowledge extraction with domain experts in the loop. The framework demonstrates how combining automated methods with light human guidance enables scalable, accurate, and efficient processing of materials science literature. Together, these contributions aim to mitigate key bottlenecks in scientific knowledge extraction and support the development of AI-ready materials data.

INDEX WORDS: Large Language Models (LLMs), Knowledge Extraction, Materials Science, Information Extraction, Named Entity Recognition (NER), Relation Extraction (RE)

ACKNOWLEDGMENTS

My PhD journey would not have been possible without the support of many individuals. I am especially grateful to Xiaohua and Jane for their invaluable guidance, mentorship, and patience. I truly believe choosing them as my advisors was one of the best decisions I've ever made.

I would like to thank my committee members: Ron Daniel, for sharing valuable industry insights into domain-specific NLP and offering thoughtful feedback; Fernando Uribe-Romo, for introducing me to the world of materials science and MOF synthesis; and Yuan An, for enriching my understanding of knowledge graphs. I'm deeply grateful to have had each of you on my committee.

I'm also grateful to my friends and collaborators on research projects—Eric Toberer, Emily Freed, Diego A. Gomez-Gualdron, Semion Saikin, Steven Lopez, Elif Ertekin, David Breen, Jacob Furst, Kyle Langlois, Fernando Fajardo-Rojas, Katherine Ardila, Vanessa Meschke, Claire Porter, Rachel Orenstein, Alexander Kalinowski, Joel Pepper, Scott McClellan, Kio Polson, and many more—for their support and contributions along the way.

I'm deeply thankful to my husband and all my family members for their unwavering support throughout my PhD journey.

I would like to sincerely thank the NSF Office of Advanced Cyberinfrastructure (OAC #1940239, #1940307, #2118201), the Institute of Museum and Library Services (RE-246450-OLS-20), the Metadata Research Center at Drexel University, and the NSF ID4

Institute for Data Driven Dynamical Design for their generous support that made this work possible.

As a final note of gratitude, I'm sincerely thankful for the faith and inner strength that carried me through this journey.

TABLE OF CONTENTS

CHAPTER

1 Introduction 1
1.1 Motivation 2
1.2 General Challenge 3
1.3 Detailed Research Question 4

2 Related Literature 5
2.1 Development of Text Mining Techniques 5
2.2 Fine-Tuning LLMs on Labeled Dataset 10
2.3 LLM Distillation 11
2.4 Leveraging Human Efforts in the Machine Learning Framework . . . 11
2.5 NLP used in Materials Science Literature 14

3 An Exploratory Analysis: Extracting Materials Science Knowledge from Unstructured Scholarly Data 15
3.1 Chapter Summary 15
3.2 Introduction 16
3.3 Materials Science and the Status of Computational Knowledge Extraction 18
3.4 Knowledge Extraction from Textual Data 19
3.5 Research Goals and Objectives 22
3.6 Research Design and Procedures 23
3.7 Result 25
3.8 Discussion 35
3.9 Conclusion 37

4 Text to Insight - Accelerating Organic Materials Knowledge Extraction via Deep Learning 39
4.1 Chapter Summary 39
4.2 Introduction 40
4.3 Related Work 41
4.4 Method 42
4.5 Experiment Results 46
4.6 Discussion and Conclusion 47

5 Fine-Tuning BERT Model for Materials Named Entity Recognition 49

5.1 Chapter Summary . . . 49
5.2 Introduction . . . 49
5.3 Background and Related Work . . . 50
5.4 Methodology . . . 53
5.5 Experiment Details . . . 55
5.6 Discussion and Future Work . . . 56
5.7 Conclusion . . . 57

6 Large-Scale Text Mining of MOF Literature for Synthesis Codification Using LLMs and Expert-Guided Schema . . . 58
6.1 Chapter Summary . . . 58
6.2 Introduction . . . 59
6.3 Methods . . . 62
6.4 Results and Discussion . . . 69
6.5 Discussion . . . 72
6.6 Conclusion . . . 74

7 Toward a Sustainable, Scalable, and Cost-Effective Framework for Scientific Knowledge Extraction . . . 75
7.1 Chapter Summary . . . 75
7.2 Recap of Key Findings and Proposed Framework . . . 75
7.3 Implementation of "Teacher-Student" Framework . . . 77

8 Conclusions . . . 80

9 Future Directions . . . 82
9.1 Data-driven Materials Discovery Accelerated by LLMs . . . 82
9.2 Domain-Specific Expert System based on Retrieval-Augmented Generation (RAG) . . . 82

APPENDIX

A Prompts Used for Synthesis Extraction . . . 84

B Python Code to Use GPT4 . . . 86

C Example Synthesis Codification Outputs . . . 89

BIBLIOGRAPHY . . . 91

LIST OF FIGURES

2.1 Architecture Design of CBOW and SG Model 6

2.2 Probability Ratio of Word Co-Occurrence 7

3.1 Distribution of reported knowledge extraction methods 26

3.2 Topics distribution of knowledge extraction studies in materials science . . 27

3.3 Relevance evaluation of terms extracted by HIVE-4-MAT 29

3.4 Relevant terms are in the upper portion and partially relevant in the lower . 30

3.5 Sample abstract 32

3.6 Sample extraction result from HIVE-4- MAT 32

3.7 Sample extraction result from MatScholar 33

4.1 Research Framework for Automatic Knowledge Extraction using Deep Learning 42

4.2 Illustration of the BiLSTM-CNN-CRF model architecture 45

5.1 A Visualized Demonstration of Named Entity Recognition task in materials science. Different entity types are highlighted to different colors, where MAT stands for Materials, PRO stands for Material Property, DSC is Descriptor and CMT is Characterization method. The goal of NER is to automatically detect entities that fall into these pre-defined semantic types. This example is from [116] and visualized by SpaCy Python library[37]. . 51

5.2 The model architecture of fine-tuned BERT for named entity recognition. . 54

6.1 End-to-End Framework of LLM-Enhanced Synthesis Extraction and Codification. Starting from a collection of materials literature, this set of scholarly research articles were first matched with DOI numbers from MOF CSD database; then the resulting articles related to MOF synthesis in XML format were parsed and converted to text format. Among all parsed articles, paragraphs that reported detailed synthesis procedures were identified by LLM. Finally, a collection of fully codified, computer-queryable synthesis procedures were extracted based on the knowledge schema designed by domain scientists. 63
6.2 Codification Standard Design Process 65
6.3 An Example of Synthesis Procedure Codification - from Text to Action Graph 66
6.4 True Positive, False Positive, True Negative and False Negative 67
6.5 Distribution of Explored Information Types 69
7.1 An Overview of Proposed Teacher-Student Distillation Framework 77
7.2 Comparison of classification performance between GPT-4o and Distillation of BERT Model 78

List of Tables

3.1 Evaluation results of HIVE-4-MAT 29

3.2 Analysis for not relevant (NR) terms 31

3.3 Comparison of different knowledge extraction approaches 34

4.1 Description of Target Materials Entity Types 44

4.2 Performance comparison (F1 scores) of BiLSTM-CRF models with different word embeddings and character-level inputs across four entity types. 46

5.1 A Glance at MatScholar Dataset 55

5.2 Performance Evaluation on Two Fine-Tuned Bert NER Models 56

6.1 Performance metrics for keyword recognition and relation extraction tasks. 71

CHAPTER 1

INTRODUCTION

As we move through the year 2025, we find ourselves fully immersed in an era defined by the explosion of big data. With over a billion websites online and millions of videos and written documents generated every day, we are facing to an unprecedented scale of available information. Research indicates that the total volume of global data has grown exponentially in recent years and is projected to reach 163 zettabytes by 2025 [92]. The key challenge lies not only in access, but also in how we can efficiently process, analyze, and extract meaningful insights from an overwhelming abundance of information.

For example, text is one of the most prevalent and widely generated forms of data in today's digital world because of its widespread use. From everyday communications like phone messages and social media posts to highly specialized content such as scientific literature, including research papers and patent documents, different forms of text possess a large amount of information that spans across a range of platforms and audiences.

However, extracting meaningful information from domain-specific texts poses several unique challenges compared to more general content. First, domain texts are typically written to communicate specialized knowledge within a specific field, often using technical vocabulary, acronyms, and conventions that require prior expertise to interpret accurately. Second, these texts frequently follow rigid formats dictated by industry or academic standards, which can vary significantly across disciplines. Third, access to high-quality domain-specific text data—such as peer-reviewed journal articles—is often restricted due

to copyright limitations, making large-scale data collection and analysis less accessible and thus more difficult.

In addition to the challenges mentioned above, the volume of scientific literature has been accelerating at an exponential pace, largely driven by advances in digital technology. This trend is especially evident in the rapid growth of digital scholarly publications across nearly every academic discipline [46, 89, 105]. Materials science research is one of them. As of April 13th 2021, a count of 9,861,616 materials scientific outputs have been indexed by the Scopus database [126]. Facing millions of documents, manually reading related literature to extract knowledge is simply not a feasible approach for any researcher. While researchers start to use computational approaches to extract knowledge from materials science literature, the body of work is fairly limited [48, 116]. There is an urgent need for robust methods to extract highly domain-specific knowledge.

Overall, this research proposal addresses challenges in extracting scientific knowledge from highly specific documents, using materials science, especially studies in Metal-Organic Frameworks (MOFs) as a case study.

1.1 Motivation

A lot of research has been done on exploring various methods to automatically extract valuable information from text [61]. Early approaches relied heavily on hand-crafted rules, which were often labor-intensive and limited in scope. These were later replaced by kernel-based machine learning methods and deep learning models, which offered improved performance and greater flexibility. Most recently, advances in large language models (LLMs) have pushed the boundaries even further, outperforming previous techniques and establishing new state-of-the-art results across a wide range of natural language processing (NLP) tasks in general domains.

However, if we focus on domain-specific knowledge, even though state-of-the-art generative AI models such as GPT-4 and Gemini have demonstrated superior performance across a variety of general tasks, their effectiveness in scientific domains—particularly for domain-specific knowledge extraction remains under-explored and presents an important area for further investigation. In addition, the cost associated with using these models—typically priced per token or API call—poses a practical barrier for sustained or large-scale use in research settings. Another approach to collect annotated corpora from domain experts to train deep learning models or fine-tuning pre-trained language models. While some studies have shown promising results on narrowly defined domain-specific tasks, these approaches often require significant time investment and domain expertise [53, 72].

1.2 General Challenge

As previously mentioned, domain-specific knowledge discovery (e.g., in science research, financial and legal industry) remains a complex problem for several reasons. First, although LLMs have demonstrated impressive performance on various NLP tasks, a common limitation is their lack of knowledge in scientific domains. To better understand what an LLM "knows", a straightforward method is to examine its training corpus. For example, GPT-3 was trained on 45 terabytes of raw text - more than half from web crawling, along with Wikipedia, public domain books and social media posts [10]. As with many LLMs, the proportion of scientific literature in this corpus is relatively small.

Compared to generic domain text, scientific documents have distinct characteristics: (1) their terminology differs from everyday language, and even between scientific domains; (2) they contain specialized knowledge rarely encountered in non-academic contexts. These factors above hindered the performance of LLMs.

Second, large-scale human annotation is prohibitively expensive. While many studies focus on entity extraction, they often rely on large amount of human annotated data [123]. This approach has the following key drawbacks: (1) manual annotation is time-consuming; (2) the inter-annotator agreement rate is often inconsistent; (3) scientific data annotation requires domain expertise. In addition, access to full-length scientific articles is a common challenge due to copyright restrictions.

In summary, these challenges are significant bottlenecks in extracting structured knowledge from scientific documents. Thus, exploring how LLMs can enable scalable, high-quality knowledge discovery with reasonable human effort is a critical research direction.

1.3 Detailed Research Question

Therefore, the following detailed research questions are derived from the overarching inquiry:

1. How can large language models (LLMs) be used to accelerate knowledge extraction across large volumes of scientific literature?

2. How can human expertise be integrated into LLM-based frameworks to maximize extraction performance?

3. How can LLMs be adapted to better understand and incorporate domain-specific knowledge?

4. How can we design LLM-based frameworks that are cost-efficient, scalable, and easy to deploy?

CHAPTER 2

RELATED LITERATURE

2.1 DEVELOPMENT OF TEXT MINING TECHNIQUES

Text, as the writing form of natural language, is unstructured data: it cannot be used directly in computation process because its semantic meaning cannot be readily captured by machines. Researchers have been exploring methods for mining text data for decades. Broadly speaking, these methods can be categorized into five stages: Rules-based methods - manually constructed rules based on domain specific dictionaries [25, 99] or syntactic and lexical patterns[52, 122]; unsupervised algorithms (e.g., clustering, syntactic parsing, etc.)[61, 77, 122]; feature-based machine learning, which uses hand-crafted or n-gram features [11] with ML algorithms such as Support Vector Machine (SVM) [2, 103]; task-specific deep learning models (e.g., LSTM+CRF) [14, 68, 121] and Large language models (LLMs)[19, 57, 65, 112, 113] with prompt learning.

Compared to early techniques, the advent of LLMs has significantly transformed the landscape of natural language processing. In the related work section, we will primarily focus on research developed after the emergence of deep learning.

2.1.1 WORD REPRESENTATION AND LINGUISTIC MODELS

As mentioned in Section 2.1, machines do not inherently understand the semantic meaning of natural language. To address this, researchers have proposed converting texts into numerical and linguistically meaningful representations. These numerical representations (i.e.,

vectors consisting of real numbers) are machine-readable, and words with similar meanings should be positioned close to each other in vector space.

Over the past decade, various methods for generating numerical word representations have been developed. These methods differ significantly. In terms of granularity, some algorithms treat individual words as atomic unit of texts (word-level embeddings) [70, 71, 86]; others use partial word or subword tokens as the smallest unit [7, 44]; and some rely on individual characters to build representations (character-level embeddings) [97].

Word2Vec, a well-known word-level embedding method proposed by Mikolov et al. [70, 71], include two model architectures for computing word vectors from large corpora: Continuous-Bag-of-Words (CBOW) and Skip-Grams (SG) (see figure 2.1). The central idea behind both architectures is to perform a prediction task: CBOW predicts the current word based on its surrounding context, whereas the skip-gram model predicts the surrounding words given the current word within a defined context window.

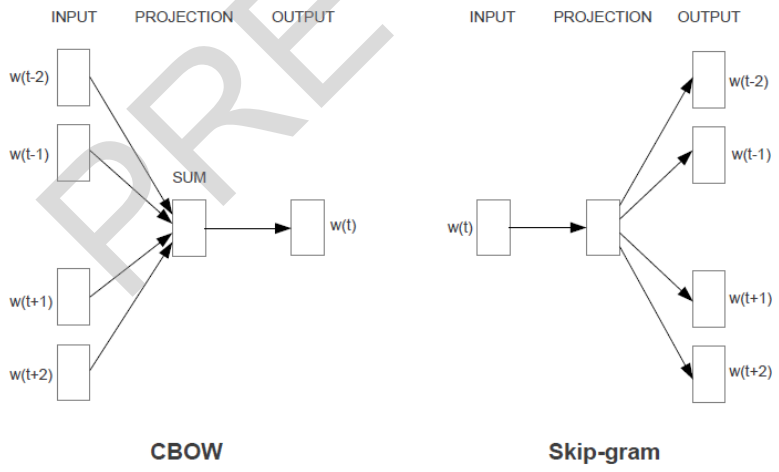


Figure 2.1: Architecture Design of CBOW and SG Model (Source [70, 71])

Instead of predicting words based on context, Pennington et al. [86] proposed the *GloVe* algorithm, which generates word-level representations by computing the probability ratio

of word co-occurrence and applying regression (see figure 2.2). Although there is no conclusive evidence that one word-level embedding method significantly outperforms another, counting-based algorithm may offer better efficiency [86].

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figure 2.2: Probability Ratio of Word Co-Occurrence (Source [86])

Up to this point, all of the approaches for generating word representations introduced above produce *static embeddings* - meaning that each word is assigned a single, unchanging vector under a given algorithm. These methods have proven effective in embedding semantic information into numerical vectors, but they share a limitation - a one-to-one mapping between a word and its vector cannot capture multiple meanings of the same word (Polysemy). For example, whether the word *mouse* refers to a small animal or a computer accessory, traditional algorithms generate only one representation for this polysemous word. This inability to distinguish between different semantic contexts limits the performance of NLP models. Soon after this, large language models (LLMs) dramatically transformed the landscape of natural language processing research.

2.1.2 Exploration of Large Language Models (LLMs)

The emergence of large language models (LLMs) have quickly drawn attention, establishing them as the new state-of-the-art in natural language processing (NLP) studies. Among the earliest and most influential LLMs [19, 65, 91], Devlin et al. [19] introduced a new language representation model called BERT (Bi-directional Encoder Representations from Transformers). One of the key differences between BERT and earlier task-specific models that used static embeddings as input is its adaptability: BERT can be fine-tuned for a variety of language tasks by adding a simple output layer. Compared to previously

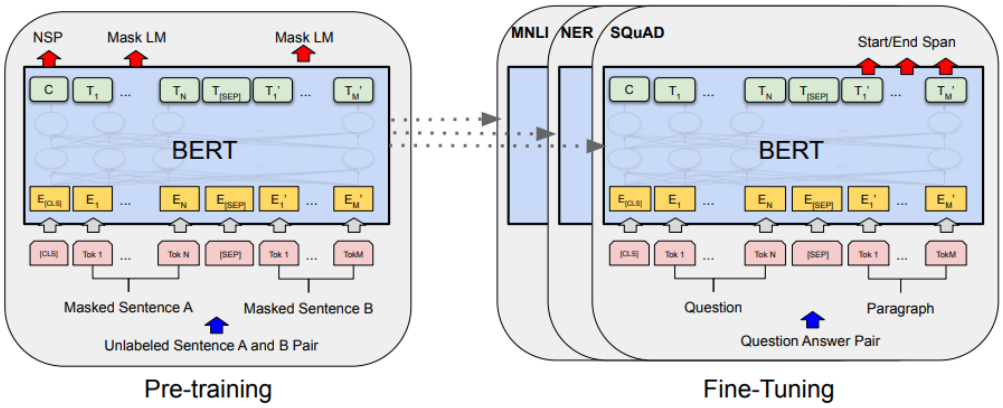


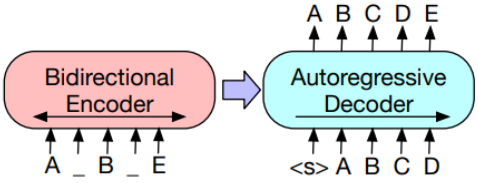
Figure 2.3: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers). (source [19])

released models, BERT enables deep **bi-directional** pre-training for language representations. In contrast, earlier models such as ELMo[87] are **unidirectional**, as they concatenate independently trained left-to-right and right-to-left representations.

To date, researchers have proposed many different large language models (LLMs). A fundamental distinction between them lies in their pre-training objectives. Both BERT [19] and RoBERTa [65] use a masked language modeling (MLM) objective, which involves randomly masking a proportion of words in the input text and training the model to predict the masked tokens. These models use only the encoder component of the Transformer architecture. Another popular objective is autoregressive (casual) language modeling (CLM), which requires models to predict the next word(s) based on preceding context. A typical CLM-based LLM is GPT [91], which uses only the decoder portion of the Transformer. Recent LLMs from the LLama series designed by Meta AI [108, 109] also follow this

(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



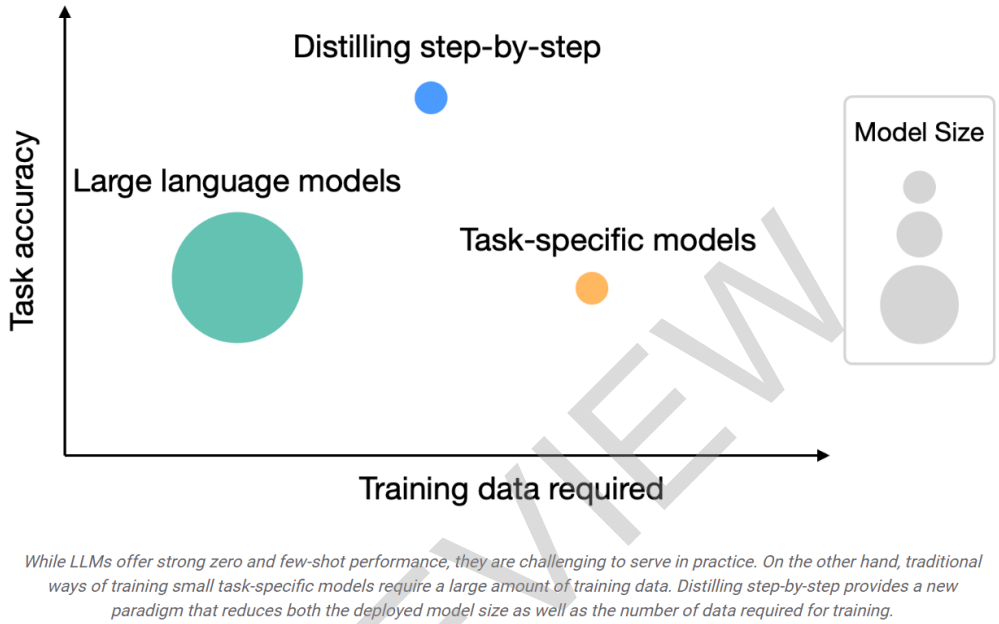
(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 2.4: A schematic comparison of BART[57] with BERT[19] and GPT[91] (source [57])

decoer-only architecture and training objective. In addition to these two objectives, BART [57] employs both encoder and decoder components of the Transformer and is trained using a "Span Corruption" objective. Similar to MLM, this task masks parts of the input text; however, rather than masking individual words, it masks entire phrases - sequences of words, which requires the model to reconstruct the missing span. A visual comparison of these three pre-training objectives is presented in figure 2.4.

Large Language Models (LLMs) have reshaped the landscape of Natural Language Processing by enabling data-efficient learning approaches such as few-shot and even zero-shot learning. However, their enormous model size makes them costly and challenging to deploy. According to Hsieh et al. [38], serving a single LLM with 175 billion parameters requires at least 350 GB of GPU memory with specialized hardware infrastructure. Several

models even exceed 175 billion parameters, which further increases the resource burden. This demand of high computational resources poses a significant barrier for most individuals and organizations.



While LLMs offer strong zero and few-shot performance, they are challenging to serve in practice. On the other hand, traditional ways of training small task-specific models require a large amount of training data. Distilling step-by-step provides a new paradigm that reduces both the deployed model size as well as the number of data required for training.

Figure 2.5: A comparison between different NLP Models (source [38])

To address this challenge, one potential solution is to deploy smaller LLMs. To ensure that these smaller models maintain strong performance, they are usually trained by two main approaches: (1) fine-tune [19] and (2) distillation [33, 62].

2.2 Fine-Tuning LLMs on Labeled Dataset

A common approach to fine-tune LLMs is to attach an additional linear layer to the model architecture and adjust its weight using a set of labeled training data. The amount of required labeled data varies across tasks and domains. By fine-tuning on such datasets, LLMs can be adapted to specific NLP task, such as named entity recognition [125], text

classification and more. This approach is often considered a promising solution - particularly for scientific text mining. However, it is important to be aware that fine-tuning results may be affected by over-fitting [30, 106]. Moreover, as discussed earlier, acquiring high-quality, manually labeled datasets present a significant challenge itself.

2.3 LLM Distillation

Fine-tuning and distillation share similarities - both methods aim to improve model performance by leveraging labeled data to refine LLM-based models. However, the key distinction is that in distillation, the labeled dataset is automatically generated by a larger, more flexible and robust LLM [38]. To produce labeled data for a smaller downstream model, few-shot learning and prompt-based techniques are commonly employed. The purpose of using these techniques is to further minimize the need for manually crafted "seed examples" when working with LLMs.

2.4 Leveraging Human Efforts in the Machine Learning Framework

To address the research questions outlined in the earlier section, one complementary approach is to more effectively leverage human effort throughout the machine learning process. Human expertise is typically required at multiple stages, including framework design, data annotation, model development and evaluation.

However, in real-world applications, machine learning (ML) models often encounter a range of issues. They may underperform due to limitations in framework design, scope or shifts in context [36]. In light of this, studying the interactions between human experts and the ML development process holds substantial potential - not only to achieve more robust AI performance, but also to optimize the use of expert time and efforts.

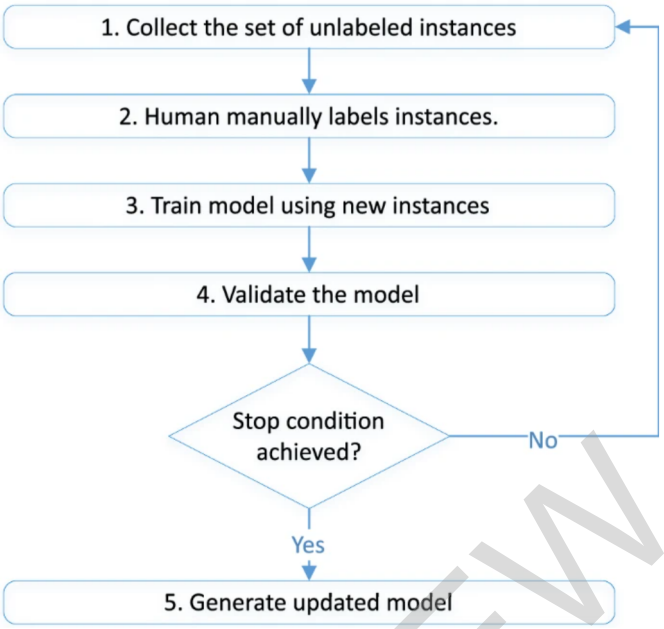


Figure 2.6: An Example of Leveraging Human Efforts in AI Model (source [75])

In addition to open-source LLMs, there are also close-source models that have demonstrated significant potential to influence everyday life. After the release of ChatGPT [81] in 2022, this LLM-powered question-answering tool became one of the most widely discussed technologies within the scientific research community. Its ability to interact with users and deliver comprehensive responses based on multiple input sources (e.g., text, images) presents major opportunities across a variety of domains. Fields such as education, healthcare, and marketing, which rely on literature analysis, record keeping, or interpersonal communication, stand to benefit greatly from this innovation. However, ChatGPT also poses risks due to hallucinations and limitations in domain-specific knowledge.

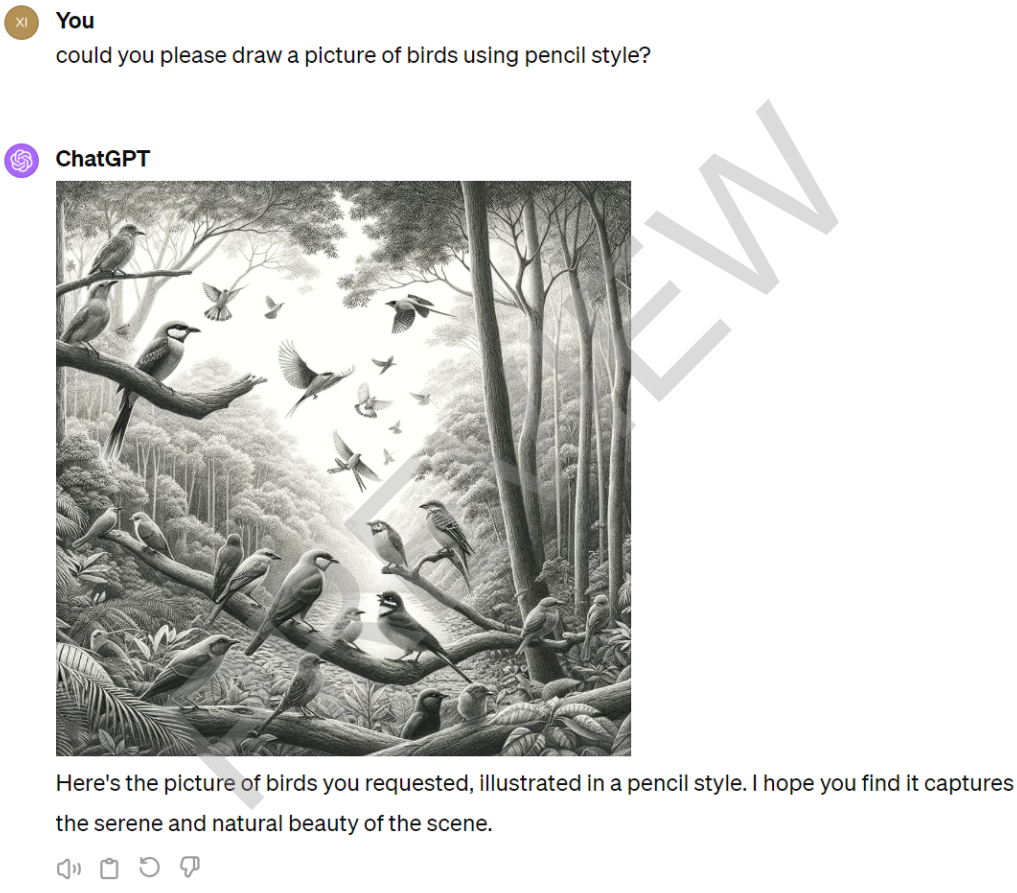


Figure 2.7: A Demonstration of Using ChatGPT to Generate Picture[81]