Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.052256

REVIEW

Unlocking the Potential: A Comprehensive Systematic Review of ChatGPT in Natural Language Processing Tasks

Ebtesam Ahmad Alomari*

Faculty of Computing and Information, Al-Baha University, Al-Baha, 65779, Saudi Arabia

*Corresponding Author: Ebtesam Ahmad Alomari. Email: Ealomari@bu.edu.sa

Received: 27 March 2024 Accepted: 11 June 2024 Published: 20 August 2024

ABSTRACT

As Natural Language Processing (NLP) continues to advance, driven by the emergence of sophisticated large language models such as ChatGPT, there has been a notable growth in research activity. This rapid uptake reflects increasing interest in the field and induces critical inquiries into ChatGPT's applicability in the NLP domain. This review paper systematically investigates the role of ChatGPT in diverse NLP tasks, including information extraction, Name Entity Recognition (NER), event extraction, relation extraction, Part of Speech (PoS) tagging, text classification, sentiment analysis, emotion recognition and text annotation. The novelty of this work lies in its comprehensive analysis of the existing literature, addressing a critical gap in understanding ChatGPT's adaptability, limitations, and optimal application. In this paper, we employed a systematic stepwise approach following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to direct our search process and seek relevant studies. Our review reveals ChatGPT's significant potential in enhancing various NLP tasks. Its adaptability in information extraction tasks, sentiment analysis, and text classification showcases its ability to comprehend diverse contexts and extract meaningful details. Additionally, ChatGPT's flexibility in annotation tasks reduces manual efforts and accelerates the annotation process, making it a valuable asset in NLP development and research. Furthermore, GPT-4 and prompt engineering emerge as a complementary mechanism, empowering users to guide the model and enhance overall accuracy. Despite its promising potential, challenges persist. The performance of ChatGPT needs to be tested using more extensive datasets and diverse data structures. Subsequently, its limitations in handling domain-specific language and the need for fine-tuning in specific applications highlight the importance of further investigations to address these issues.

KEYWORDS

Generative AI; large language model (LLM); natural language processing (NLP); ChatGPT; GPT (generative pre-training transformer); GPT-4; sentiment analysis; NER; information extraction; annotation; text classification

Nomenclature

AI	Artificial Intelligence
LLM	Large Language Model
NLP	Natural Language Processing
PLM	Pre-trained Language Model

GPT	Generative Pre-trained Transformers
NER	Name Entity Recognition
PoS	Part of Speech
BERT	Bidirectional Encoder Representations from Transformers
RLHF	Reinforcement Learning from Human Feedback
T5	Text-To-Text Transfer Transformer
ICL	In context Learning
CoT	Chain of Thought
SOTA	State-of-The-Art
MLMs	Masked Language Models
MNER	Multimodal Name Entity Recognition
IE	Information Extraction
DIE	Document Information Extraction
SC	Sentiment Classification
ECA	Emotion cause analysis
ABSA	Aspect-Based Sentiment Analysis
ECA	Emotion Cause Analysis
ET	Entity Typing
RC	Relation Classification
RE	Relation Extraction
ED	Event Detection
EAE	Event Argument Extraction
EE	Event Extraction
MMSE	Mini Mental Status Exam
CDR	Cognitive Dementia Rating
MEN	Malaysian English News
MSEA	Multimodal Similar Example Awareness
P	Precision
R	Recall
P@5	Precision at 5
Rprec	R-Precision
MRR	Mean Reciprocal Rank
MAE	Mean Absolute Error
CRSP	Center for Research in Security Prices
NCBI	National Center for Biotechnology Information
BC5CDR	BioCreative V CDR

1 Introduction

Large Language Models (LLMs) are powerful tools that leverage deep learning techniques, particularly transformer architectures, to process and understand natural language. The capability of LLMs to comprehend complex linguistic patterns, semantics, and context from vast amounts of textual data empowers them to excel across a wide range of tasks with remarkable performance [1–4]. Besides, the ability of LLMs to be fine-tuned on specific datasets to optimize their performance to the requirements of particular applications makes them highly versatile and applicable in various domains [5]. Furthermore, LLMs have transformed the field of Natural Language Processing (NLP) by introducing the boundaries of what is achievable in language understanding and generation tasks.

They are considered a significant milestone in the NLP field, characterized by their enormous number of parameters and their generative ability to generate human-like text based on provided input.

ChatGPT is an extension of the GPT (Generative Pre-trained Transformer) architecture developed by OpenAI [6]. It is a transformer-based language model, leveraging large-scale pre-training on various and massive text corpora, including books, articles, and web content, enhancing the understanding of language patterns, semantics, and syntax [7]. Besides, ChatGPT's functionality is mainly based on its ability to produce coherent and contextual answers by analyzing and understanding the input text's meaning and context, facilitating flexible conversations. The model's performance is improved over time by the continuous fine-tuning and refinement process [8]. Consequently, its capabilities have placed it as an essential and powerful tool for various domains, such as education [9], healthcare [10,11], business, and others [12–14]. Moreover, ChatGPT introduces a special solution for various NLP problems, including Name Entity Recognition [15,16], annotation [17], and others [18,19]. Its contextual understanding and conversational capabilities to generate responses make it a desirable option for applications that require sophisticated language processing. However, the effectiveness of ChatGPT across various NLP tasks remains an open question, requiring a systematic exploration to investigate its potential and weaknesses.

This review paper centers its investigation on revealing the role of ChatGPT in specific NLP tasks, namely i) Information Extraction and group of sub-tasks uber it includes Name Entity Recognition (NER), event extraction, relation extraction, and Part of Speech (PoS) tagging, ii) Text Classification, iii) Sentiment Analysis and Emotion Recognition, iv) and Annotation. The aim is to comprehensively analyze ChatGPT's impact, highlighting its strengths and limitations.

1.1 Research Gap, Novelty, and Contributions

ChatGPT has gained extensive attention and adoption in recent years across various domains, including solving NLP problems. Its capability to provide human-like interactions on a large scale has made it a valuable asset, which encourages researchers to conduct studies to investigate its potential. However, a comprehensive review to explore the effectiveness of ChatGPT across diverse NLP tasks is notably absent. This paper seeks to bridge this gap. To the best of our knowledge, this is the first systematic review paper offering a holistic perspective on using ChatGPT in various NLP problems, providing valuable insights for researchers and practitioners in the field.

Moreover, by examining each task separately, the paper aims to provide an in-depth analysis of ChatGPT's impact, highlighting its advantages and shortcomings. This method facilitates a nuanced comprehension of ChatGPT's performance across different tasks in the NLP field.

Furthermore, the main contributions of this review are as follows:

- Explore the primary studies based on the PRISMA framework and comprehensively investigate ChatGPT utilization in the four selected NLP problems.
- Examine the natural languages explored in the applications of ChatGPT to understand the linguistic diversity of the studies.
- Provide a comprehensive overview of the selected datasets, evaluation methods, and the evaluation metrics employed to measure the performance of ChatGPT in diverse NLP tasks.
- Evaluate the impact of ChatGPT by comparing its performance against baseline models and state-of-the-art (SOTA) methods.
- Discuss the capabilities and limitations of ChatGPT in the NLP domain.

1.2 Paper Organization

The organization of the paper is illustrated in Fig. 1. Section 2 explains the evolution of LLM and ChatGPT. Section 3 discusses the materials and methods. Section 4 shows the findings. Section 5 discusses the capabilities and challenges. Finally, we draw our conclusions in Section 6.

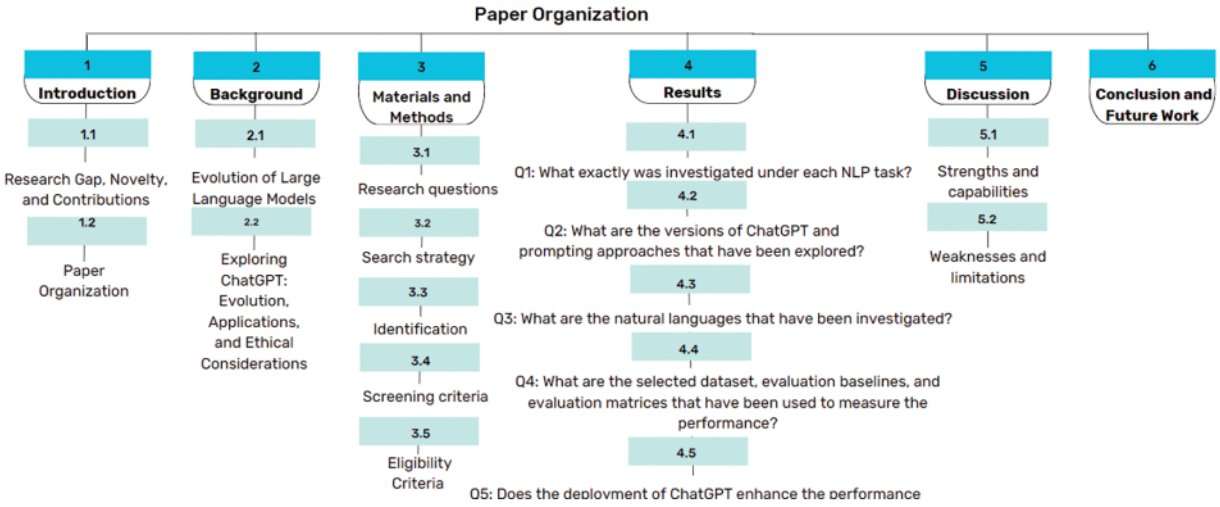

Figure 1: Paper structure

2 Background

2.1 Evolution of Large Language Models

The rapid evolution in artificial intelligence (AI) and natural language processing fields has led to the creation of increasingly sophisticated and flexible language models such as GPT-4, Claude, Bard [20], which have been involved in achieving remarkable results across various NLP tasks [21], highlighting their enormous potential in various domains [22–25]. These models rely on deep learning, which uses neural networks for in-depth analysis and processing [26,27]. Then, generative AI comes as a subclass of deep learning and enables the generation of new data in various domains based on patterns and structures learned from existing data. Additionally, generative learning refers to the process of training models to create new data samples, which facilitates LLMs to produce human-like text for tasks such as text generation and dialogue systems [28]. Several LLM families have recently been introduced, including GPT, LLaMA [29], and PaLM [30]. Fig. 2 shows the evolution of large language models and their relationship with machine learning, deep learning, and other techniques.

Furthermore, the transformer model has revolutionized the field of natural language processing [31,32], which depends on a self-attention mechanism. It is considered the backbone of state-of-the-art LLMs, including ChatGPT. The transformer uses an encoder-decoder structure. The encoder (on the left side of the architecture) maps an input sequence to a series of continuous representations. The decoder (on the right half of the architecture) receives the encoder's output and the decoder's output at a previous time step and generates an output sequence [32]. Moreover, the pre-trained language model (PLM) [33] is a sophisticated deep learning architecture primarily based on transformers and trained on extensive datasets. These models can be fine-tuned for specific tasks by supplying task-specific labeled data. Fine-tuning involves adjusting the model's parameters to optimize performance on tasks like sentiment analysis, named entity recognition, and others. PLM can be categorized into

three main classes based on their architecture: encoder-based, decoder-based, and encoder-decoder-based, as illustrated in Fig. 3. The most widespread encoder-based PLMs are BERT [34], XLNet [35], ELECTRA [35], etc. BERT (Bidirectional Encoder Representations from Transformers) was introduced by Google in 2018. It contains several layers of transformer encoders trained to identify the bidirectional context from input text. They are various models that are built upon or inspired by the BERT architecture, such as RoBERTa [36], ALBERT [37], DeBERTa [38], FineBERT [39], MentalBERT [40], BERTimbau [41], ClimateBERT [42], BioBERT [43] and KeyBERT [44].

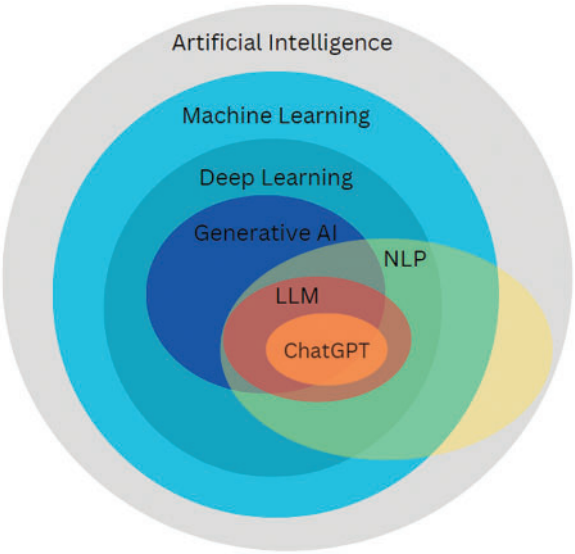

Figure 2: Evolution of LLM: Illustrating the interplay and overlap

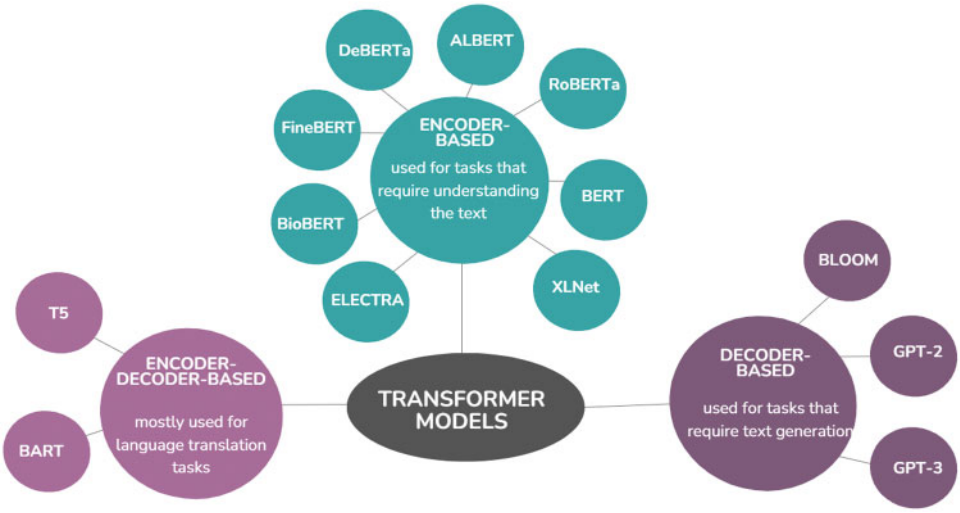

Figure 3: Transformer based models

Additionally, GPT and its subsequent versions, such as GPT-2 and GPT-3, developed by OpenAI, are the popular decoder-based PLM. They include multiple layers of transformer decoders trained to produce text autoregressively. They are widely used for NLP tasks such as text generation, dialogue

systems, etc. Finally, the well-known examples of encoder-decoder–based models are T5 (Text-To-Text Transfer Transformer) [45] introduced by Google and BART (Bidirectional and Auto-Regressive Transformers) [46] developed by Facebook.

Another vital concept in natural language processing is autoregressive language modeling, where models are trained to predict the next word in a sequence based on previously generated words [47]. Fig. 4 shows an example of a sentence illustrating the sequential generation of text using an autoregressive language model. Despite various large language models in the literature, we selected ChatGPT in this work due to its popularity and current status as a state-of-the-art conversational agent.

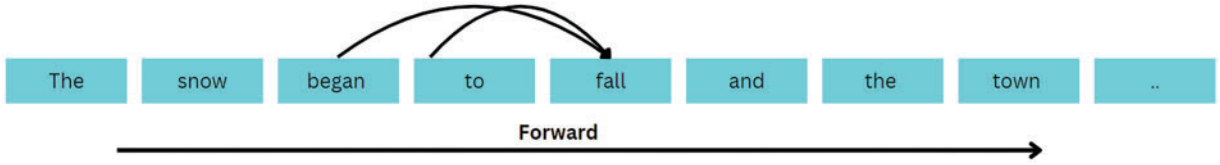

Figure 4: Sequential generation of text using an Autoregressive Language Model

2.2 Exploring ChatGPT: Evolution, Applications, and Ethical Considerations

ChatGPT stands for "Chat Generative Pre-trained Transformer". It is a powerful conversational AI system developed by OpenAI. ChatGPT has evolved rapidly from a research playground with GPT-3.5 to real-world conversational AI powered by GPT-4 [48]. Subsequently, ChatGPT leverages the capabilities of massive language models to engage in natural, open-ended dialogues, answer your questions, and even generate creative text formats [49]. ChatGPT is a subset of PLM and follows the autoregressive learning paradigm. Additionally, it was created by improving the GPT-3 language model using reinforcement learning from human feedback (RLHF), which is a technique used to enhance the performance of language models by integrating direct feedback from humans [50]. Fig. 5 explains the three stages of RLHF. The first phase is choosing a pre-trained model as the main model. After training the main model, the performance is evaluated based on human feedback to create rewards for reinforcement learning. The reward model is fine-tuned with the main model's output and obtains a quality score from testers [51].

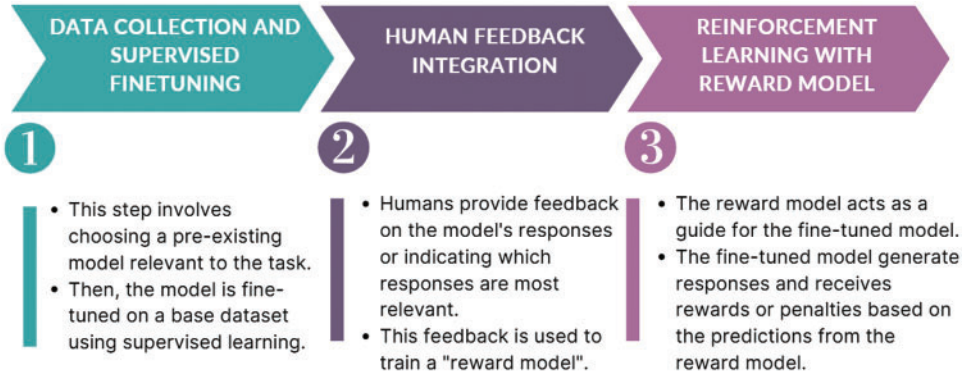

Figure 5: Overview of Reinforcement Learning from Human Feedback (RLHF)

Besides, Fig. 6 depicts a timeline of ChatGPT releases and feature enhancements over time. The naming convention for ChatGPT and its releases, such as GPT-1 [52], GPT-2, GPT-3 [53], GPT-3.5, and GPT-4 [54], is based on the series of GPT developed by OpenAI.

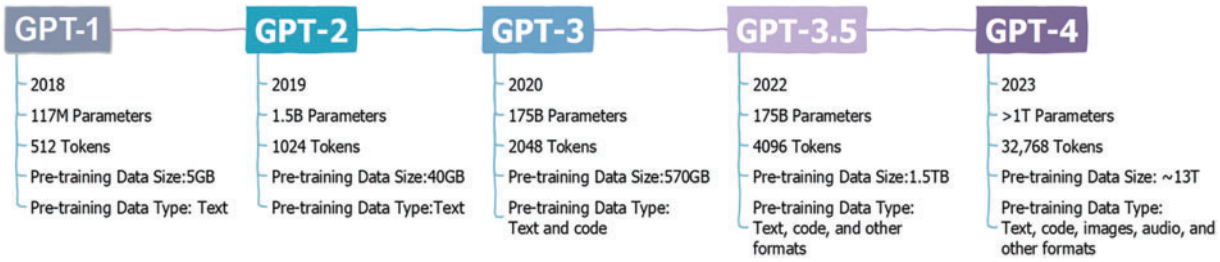

Figure 6: Evolution of ChatGPT

Moreover, Fig. 7 illustrates examples of ChatGPT applications, showcasing its versatility across various domains. Notably, in the healthcare sector, it has several applications, including developing patient-specific treatment programs [55–58], medication reminders, offering digital assistance for doctors, improving doctors' replies to insurance claims, and providing individual health advice for patients [59,60]. Subsequently, in education, it can enhance student-centric learning [61], assist in academic writing [62], provide personalized tutoring and feedback, grade student essays [63,64], act as a substitute teacher [65] and assist in learning different topics, such as programming learning [66]. Besides, businesses can utilize ChatGPT to automate customer support [67], generate marketing content [68], handle routine queries, support business operations and decision-making [69,70], address strategic business problems [71] and provide assistance to improve efficiency. Consequently, in the finance sector, ChatGPT can be used to generate financial summaries, reports, forecast outcomes, androvide personalized investment recommendations [72]. Subsequently, it can be utilized in the hospitality and tourism domain [73], in addition to playing a prospective role in the military domain [74,75].

Figure 7: Possible applications of ChatGPT

Although its applications and capabilities benefit various domains, there are several ethical concerns surrounding ChatGPT [76]. The primary concerns include but are not limited to, the potential for generating harmful or misleading content, propagating biases present in the training data, and violating privacy [77,78]. As an AI language model produces human-like content, ChatGPT can participate in spreading misinformation, hate speech, or harmful principles. Subsequently, without sufficient training or supervision, ChatGPT may unintentionally reinforce biases inherent in the

training data, leading to unjust output. Additionally, the use of ChatGPT in critical fields that require dealing with sensitive data, such as health and medical documents, raises concerns regarding privacy and security. Thus, there is a need to ensure transparency in model development and usage to address these ethical concerns.

3 Materials and Methods

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [79], which is a widely recognized and recommended approach for conducting systematic reviews. Utilizing the PRIMSA approach helps in different aspects. Primarily, PRISMA provides a standardized framework for conducting and reporting systematic reviews. Thus, it ensures transparency in the research process, making it easier for readers and reviewers to understand and evaluate the study methods and results. Subsequently, by following the PRISMA approach, researchers can minimize bias in their systematic reviews. Additionally, researchers can use PRISMA as a roadmap to replicate the study process and verify the results independently. The full stages of the paper's selection process are shown in Fig. 8.

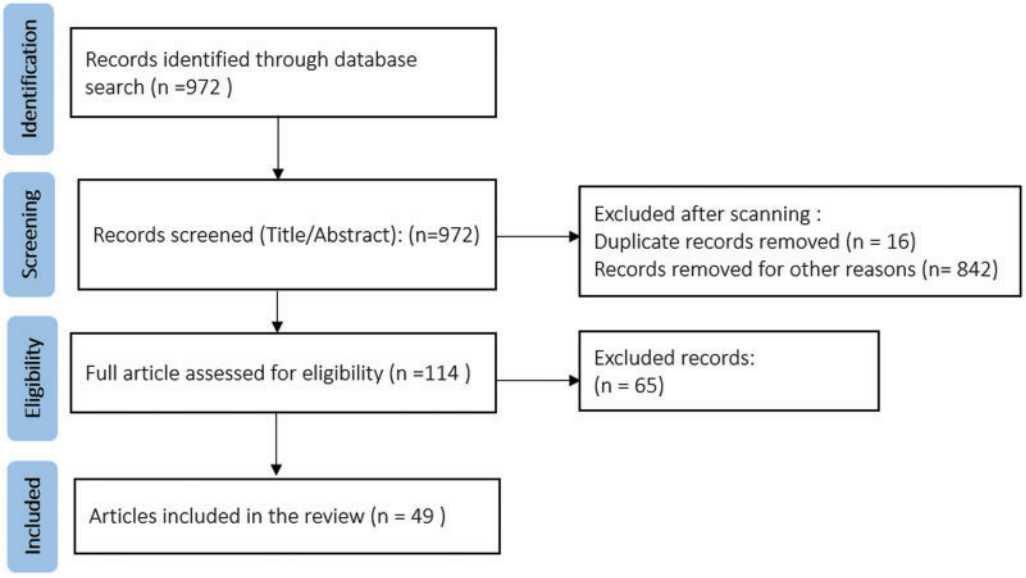

Figure 8: The article selection process: A PRISMA flow diagram

3.1 Research Questions

The research questions (RQ) of this study are as follows:

- RQ1: What exactly was investigated under each NLP task?
- RQ2: What versions of ChatGPT and prompting approaches have been explored?
- RQ3: What are the natural languages that have been investigated?
- RQ4: What are the selected dataset, evaluation baselines, and evaluation matrices that have been used to measure the performance?
- RQ5: Does the deployment of ChatGPT enhance the performance compared to the baseline models/SOTA methods?

3.2 Search Strategy

Firstly, we searched several databases, including ACM Digital Library, IEEE Xplore, ScienceDirect, Scopus, and SpringerLink. However, only a limited number of relevant works were obtained. Therefore, we used Google Scholar as an information source using Publish or Perish [80]. The final search was conducted in December 2023 using the "ChatGPT" keyword in conjunction with the following search keywords:

"Information extraction", "entity recognition", "event OR incident detection OR extraction", "annotators OR annotation", "text classification", "sentiment analysis", "emotion recognition"

3.3 Identification

The search criteria include any published scientific research or preprints that discussed the use of ChatGPT to address the following NLP tasks: (1) Information Extraction and a group of sub-tasks uber it includes Named-Entity Recognition (NER), event extraction, relation extraction, and Part of Speech (PoS) tagging, (2) Text Classification, (3) Sentiment Analysis and emotion recognition, (4) Text Annotation. The search in the first stage yielded 972 papers.

3.4 Screening Criteria

The second stage in the PRISMA search strategy process is screening (selection). The primary purpose of the screening stage is to initially assess the relevance of studies by examining their titles and abstracts. It is the first step in the process of narrowing down the group of potentially relevant papers. The following criteria were followed to select the papers in this stage (1) not duplicate, (2) written in English, (3) in the scope of the review, and (4) published in an academic source. After we applied the above-mentioned criteria, out of 972 records, 114 were accepted for further exploration in the next stage.

3.5 Eligibility Criteria

In the third stage, exclusion criteria were applied to minimize the number of studies considered for further investigation. The exclusion criteria involved studies that discussed the use of ChatGPT in areas unrelated to the paper's focus. For instance, we excluded papers that explored sentiment analysis to study users' sentiment or opinion regarding using ChatGPT, which is irrelevant to this review. We investigated leveraging ChatGPT for NLP tasks, including performing sentiment analysis. In addition, we excluded papers that mainly focused on knowledge enhancement [81], data augmentation [82,83], or content generation by ChatGPT to address NLP problems such as classification [84–86]. After applying these criteria, the final number of studies considered for future investigation in this review is 49 papers.

4 Results

4.1 Q1: What Exactly Was Investigated under Each NLP Task?

This subsection explores the works for each of the four NLP problems.

4.1.1 Information Extraction

Several works investigated the abilities of ChatGPT for name entity recognition (NER). González et al. [87] examined the potential of the zero-shot approach to detect entities in historical documents using ChatGPT. Tan et al. [88] developed a model named a unified retrieval-augmented system (U-RaNER) for fine-grained multilingual NER and evaluated ChatGPT using 3 prompting

settings, which are Single-turn, multi-turn as zero-shot learning, and Multi-ICL as few-shot learning. Other researchers [89] evaluated ChatGPT's ability to extract entities and relationships from the Malaysian English News (MEN) dataset. Xie et al. [90] decomposed the NER task from Chinese text into simpler subproblems based on labels and employed a decomposed-question-answering (Decomposed-QA) paradigm. In this approach, the model focuses on extracting entities of individual labels separately. They integrated syntactic prompting, prompting the model to initially analyze the syntactic structure of the input text and subsequently identify named entities based on this structure. Zhou et al. [91] demonstrated the utilization of ChatGPT to create instruction-tuning data for NER from extensive unlabeled web text.

Moreover, Li et al. [92] leveraged ChatGPT as a knowledge engine to generate auxiliary refined knowledge to boost the model performance in Multimodal Named Entity Recognition (MNER). They developed a framework named Prompt ChatGPT in MNER (PGIM). The first stage is generating auxiliary refined knowledge. The second stage is predicting the entities based on the generated knowledge. For the first stage, they designed a module for selecting appropriate in-context examples called Multimodel Similar Example Awareness (MSEA). These examples are then utilized to generate prompt templates of MNER and fed to ChatGPT. The prompt template consists of the following components: a fixed prompt head, some in-context examples, and a test input. The prompt head describes the MNER task in natural language according to their requirements. Besides, ChatGPT is asked to judge by itself since the input image and text are not always relevant, which guides ChatGPT in generating auxiliary refined knowledge. Further, they extended their work and proposed a framework for grounded multimodal named entity recognition [93]. Hu et al. [94] assessed the zero-shot capability of ChatGPT and GPT-3 to perform clinical NER tasks, particularly in recognizing three types of clinical entities: Medical Problems, Treatments, and Tests. Likewise, Tang et al. [95] leveraged ChatGPT in biomedical NER tasks to generate a synthetic dataset to fine-tune three pre-trained language models: BERT, RoBERTa, and BioBERT.

Moreover, other researchers explored how ChatGPT performs in extracting information regarding specific domains. Brinkmann et al. [96] studied the ability of ChatGPT to extract attribute/value pairs from product titles and descriptions. They evaluated In-Context Learning (ICL) with both zero-shot and few-shot settings under diverse prompt designs by adding a demonstration involving an example of input and output. Yuan et al. [97] explored the potential of zero-shot and chain of thought (CoT) learning for temporal relation extraction using ChatGPT. Jethani et al. [98] asked ChatGPT to analyze 765 unstructured medical notes to extract information including Mini Mental Status Exam (MMSE) and Cognitive Dementia Rating (CDR) scores and exam dates. Similarly, Peikosa et al. [99] assessed the ability of ChatGPT in patient information extraction from clinical notes as well as search query generation to retrieve clinical trials. For information extraction, they tested two approaches. The first approach is generic where ChatGPT is asked to extract a list of keywords that define the content of a given medical text. The second approach is more complex where ChatGPT acts as a medical assistant and extracts medical conditions, treatments, and related terminology. Additionally, they asked ChatGPT to act as a medical assistant and create a keyword-based query that can be used to retrieve clinical trials based on the information in the given clinical note. Additionally, He et al. [100] proposed an in-context learning framework called ICL-D3IE to address document information extraction (DIE) task, by enabling GPT-3 to predict entity labels in a test document. Unlike standard ICL, which depends on task-specific demonstrations, ICL-D3IE constructs three demonstrations to enhance in-context learning: hard demonstrations, layout-aware demonstrations, and formatting demonstrations.

Furthermore, Kartchner et al. [101] studied how ChatGPT can improve IE for clinical meta-analyses of randomized clinical trials using a zero-shot setting. Sousa et al. [102] investigated the ability of GPT-3 and GPT-3.5 to extract narrative entities, including events, participants, and time, from Portuguese news articles. Nishio et al. [103] leveraged ChatGPT to enable automatic extraction from Japanese radiology reports for CT examination of lung cancer. They extracted TNM staging, which is T (tumor size and invasiveness), N (lymph node involvement), and M (metastasis to other parts of the body). They employed LangChain API to control ChatGPT and set the temperature to zero. Their proposed system consists of three steps. In the first step, they provide ChatGPT via LangChain custom prompts with the definition of the TNM staging to improve ChatGPT's understanding of TNM staging. In the next step, they provided ChatGPT with the report and asked it to extract the required information. In the last step, they passed the output of step two to ChatGPT to extract only TNM staging.

For event detection, Gao et al. [104] conducted an experiment to test the ability of ChatGPT in zero-shot scenarios to extract events. Furthermore, other works explored the ability of ChatGPT in multi-tasking related to information extraction, Han et al. [105] evaluated the ability of ChatGPT to extract information. They mainly studied 4 Information Extraction (IE) tasks, which are Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), and Aspect-based Sentiment Analysis (ABSA). Additionally, Li et al. [106] evaluated the performance of ChatGPT on 7 information extraction (IE) tasks, which are Entity Typing (ET), Named Entity Recognition (NER), Relation Classification (RC), Relation Extraction (RE), Event Detection (ED), Event Argument Extraction (EAE) and Event Extraction (EE). Wei et al. [107] built a two-stage framework for zero-shot information extraction for English and Chinese languages using ChatGPT. In the first stage, they extracted the existing types of entities, relations, or events, while in the second stage, they found the relevant information based on what was discovered in the first stage. They focused on three tasks: entity-relation triple extract, named entity recognition, and event extraction. Lai et al. [108] tested ChatGPT's ability in different NLP tasks, including Part of Speech (POS) tagging NER. In POS tagging, the prompt involves a task description, an input, and a note for output format.

4.1.2 Text Classification

Few works focus on text classification, Reiss [109] investigated the consistency of ChatGPT's zero-shot capabilities for website annotation and classification into News or not News, focusing on different model parameters, prompt variations, and repetitions of identical inputs. Zhao et al. [110] proposed a framework called ChatAgri to perform cross-linguistic text classification using ChatGPT. They ran several experiments and tested different prompts using zero-shot and few-shot approaches on GPT-4, vanilla ChatGPT, and GPT-3.5. Subsequently, Loukas et al. [111] investigated few-shot GPT-3.5 and GPT-4 abilities in financial text classification. Lamichhane [112] measured the performance of ChatGPT in mental health classification. They tested its ability in three topics: two of them are binary classification, which are stress detection, and depression detection, while the third topic (suicidality detection) is 5-class classification. Trajano et al. [113] utilized ChatGPT for multi climate-related text classification tasks, including climate detection classification, climate-related commitments and actions classification, climate change specificity classification and climate change disclosure category classification. All of the classification tasks are binary classification except the last one. Besides, they performed climate sentiment analysis. Consequently, Oliveira et al. [114] leveraged ChatGPT to classify tweets in the Portuguese language into hateful or not-hateful.

4.1.3 Sentiment Analysis and Emotion Recognition

Yang et al. [115] conducted a study to understand ChatGPT's ability to perform emotion recognition in conversation and emotional reasoning in a zero-shot setting. They asked ChatGPT to explain its reasoning from the perspectives of emotions and physical symptoms after binary/multi-class detection. Subsequently, Zhong et al. [116] utilized standard few-shot, zero-shot chain-of-thought (CoT), and manual few-shot CoT prompting for sentiment analysis. Golubev et al. [117] applied sentiment analysis and predicted sentiment towards a named entity within a single sentence in Russian news text. Wang et al. [118] evaluated the ability of ChatGPT in both zero-shot and few-shot scenarios to perform sentiment, emotion, and opinion analysis. They focused on several tasks, including sentiment classification (SC), aspect-based sentiment analysis (ABSA), and emotion cause analysis (ECA). Fatouros et al. [119] discussed whether the use of ChatGPT led to improvements in sentiment analysis within the financial domain. Similarly, Lopez-Lira et al. [120] investigated the effectiveness of ChatGPT in forecasting stock price movements, specifically focusing on return predictability using sentiment analysis of news headlines. Zhang et al. [121,122] explored the ability to detect stance events in a zero-shot setting by ChatGPT. They tested a simple prompt and asked ChatGPT to select from one of the three categories: Favor, Against, or Neither. Kocoń et al. [123] evaluated the ability of GPT-4 using zero-shot and few-shot learning in 25 tasks, including some of the tasks that we focused on in this review, which are sentiment analysis, emotion recognition, and stance detection. Khondaker et al. [124] focused on the Arabic language and studied the capability of ChatGPT on diverse NLP tasks, including sentiment analysis and emotion detection. They investigated the abilities of both ChatGPT and GPT-4 on dialectal Arabic and modern standard Arabic.

4.1.4 Annotation

Kuzman et al. [125] focused on categorizing texts into genre categories, such as New Legal, Promotion, Opinion/Argumentation, Instruction, Information/Explanation, Prose/Lyrical, Forum, and Other. The model was tested on two datasets English and Slovenian. Chen et al. [126] leveraged ChatGPT to create annotated data for event extraction tasks in the Chinese language and then trained the model using supervised learning algorithms. They followed two phases: recall and generation. In the recall phase, ChatGPT posts several questions regarding event extraction and data generation. The goal is to guide and correct the answers provided by ChatGPT. In the second phase, ChatGPT was asked to generate sentences based on specified trigger words and event arguments to provide numerous amounts of labeled corpora. However, their work was limited to data augmentation and input construction. Korini et al. [127] tested different prompt designs in zero-and few-shot settings for column-type annotation using ChatGPT. They proposed a two-step annotation pipeline by asking ChatGPT in the first step to predicting the topical domain of the table to be annotated and then involved only the labels that are related to the predicted domain to annotate the columns. Then, they tested ChatGPT's ability to determine the semantic type of single columns in the table and all columns at once. Furthermore, Belal et al. [128] assessed the ability of ChatGPT as a text annotator in zero-shot for sentiment analysis, while Koptyra et al. [129] focused on annotation for emotion recognition. Subsequently, Huang et al. [130] focused on understanding the abilities of ChatGPT to detect implicit hateful tweets by binary classification and give explanations for the reasoning. Gilardi et al. [131] studied the ability of ChatGPT on various annotation tasks, including relevance, stance detection, topics detection, general frame detection, and policy frames detection. The testing dataset includes 2382 manually annotated tweets. Alizadeh et al. [132] compared the performance of Large Language Models (HuggingChat and FLAN [133]) with ChatGPT and human-based services (MTurk) in text annotation tasks.

Moreover, Li et al. [134] conducted a study to evaluate ChatGPT and GPT-4 in zero-shot or few-shot settings on different NLP tasks, which are text classification, numerical reasoning, named entity recognition, question answering, and sentiment analysis. Qin et al. [135] assessed the ability of ChatGPT and ChatGPT 3.5 in zero-shot to perform different NLP tasks, including sentiment analysis and NER to extract the following entities: "Loc", "Per", "Org", and "Misc" stand for "Location", "Person", "Organization", and "Miscellaneous Entity". Sun et al. [136] tested the ability of ChatGPT to perform multitas, including subtasks on information extraction, which include event extraction, PoS tagging, entity-relation extraction, and sentiment analysis. Fig. 9 unravels the diverse landscape of Natural Language Processing research by illustrating the classification of NLP tasks investigated by each paper.

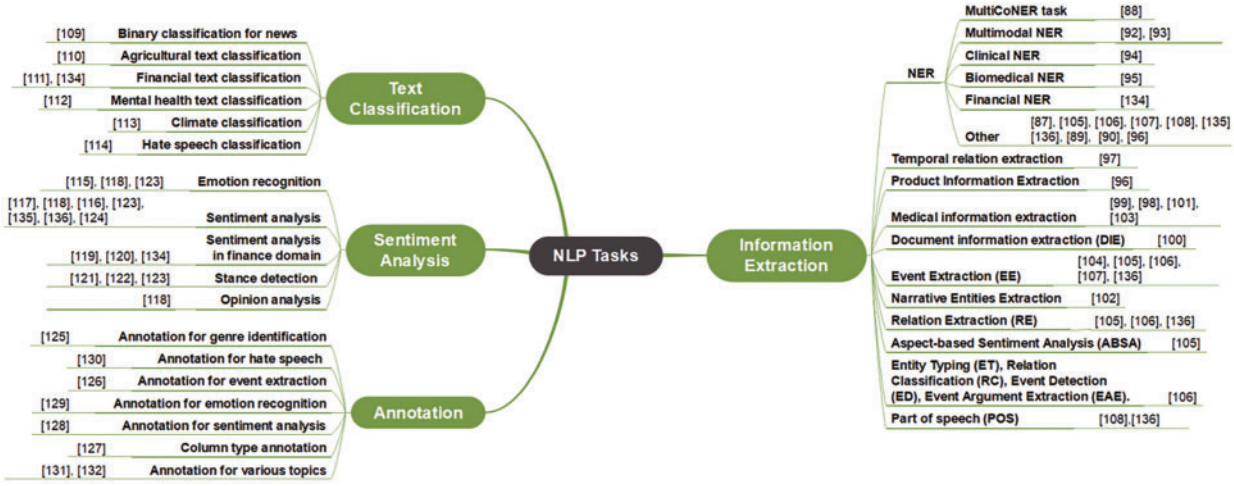

Figure 9: Classification of explored NLP tasks and sub-tasks by each paper

4.2 Q2: What Versions of ChatGPT and Prompting Approaches Have Been Explored?

Fig. 10 shows the variety of ChatGPT versions explored. Notably, only Hu et al. [94] and He et al. [100] conducted assessments using GPT-3. The majority of studies have predominantly employed ChatGPT 3.5. Besides, GPT-4 has been utilized in only six studies, with two focusing on text classification (Agricultural text classification [110], Financial text classification [111]), two under sentiment analysis task [123,124], and one study for medical information extraction [103]. In another study [134], GPT-4 was tested for multitasking in the financial domain, encompassing Named Entity Recognition (NER), text classification, and sentiment analysis. It is important to note that the figure excludes papers that did not specify the version of ChatGPT used.

Furthermore, there are distinct prompting approaches, which include the following:

- Zero-shot learning:

By using this approach, we drop ChatGPT into a new task without giving any prior training or examples on that task. Thus, its response depends on its understanding of language patterns, which is where the model leverages its internal knowledge gained from the pre-training phase on diverse datasets.

- Few-shot learning:

Provide ChatGPT with a few examples related to the task, which will guide the model in giving a response based on the provided examples.

- Chain of Thought (CoT):

This approach is like a back-and-forth brainstorming conversation with ChatGPT, where we provide it with an initial prompt and then give a follow-up prompt to guide the model.

- In-Context Learning (ICL):

This approach involves priming ChatGPT with relevant information and background details before the main prompt.

- Combining CoT with zero-shot or few-shot learning:

This approach helps to keep the conversation going with ChatGPT by asking follow-up questions to shape the outcome using a small number of examples for CoT few-shot learning or without any example for CoT zero-shot learning.

- Combining ICL with zero-shot or few-shot learning:

ICL zero-shot relies on the internal knowledge of ChatGPT and its ability to understand context without giving examples, while in ICL, a few shots provide few examples of the desired outcomes.

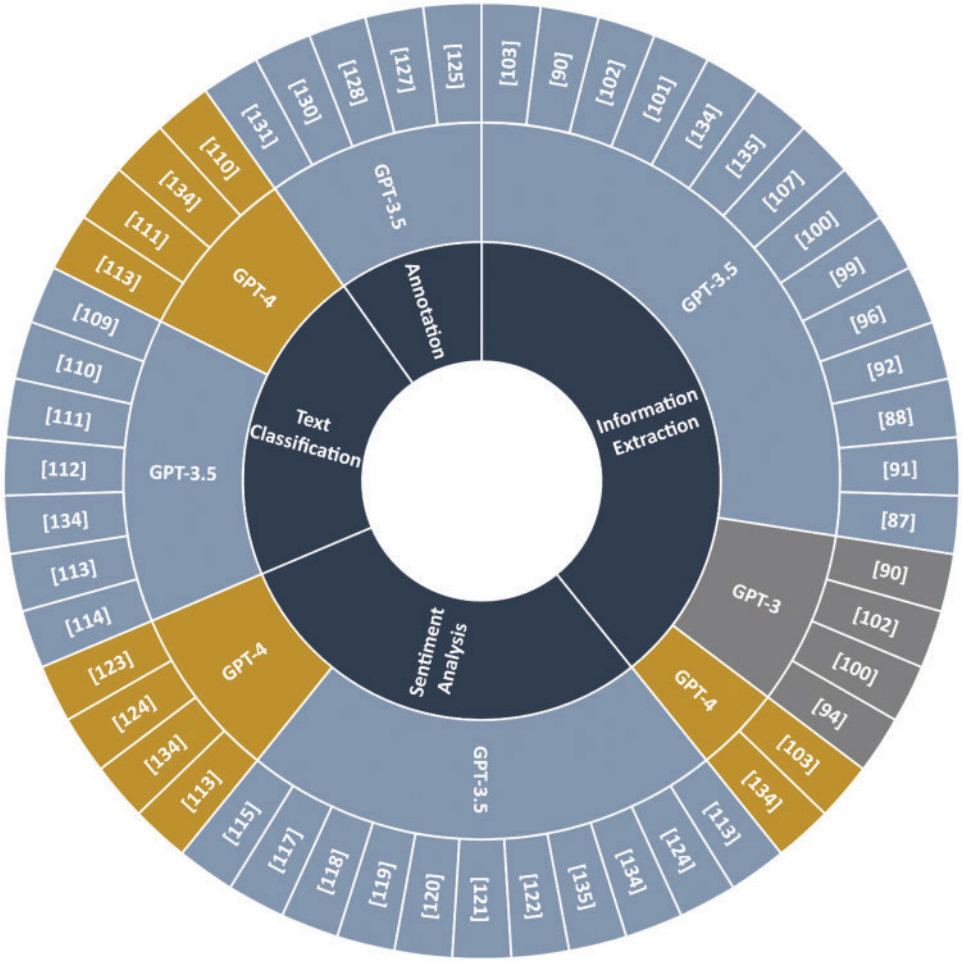

Figure 10: The utilization of different ChatGPT versions across various NLP tasks in each paper

Fig. 11 illustrates the explored prompting approaches for each NLP task. As we can see, various approaches have been used for information extraction, including zero-shot, few-shot, zero-shot ICL, few-shot ICL, few-shot CoT and CoT. Most of the works have used zero-shot while only two works investigated using Zero-shot ICL [89,96], and only [105] explored using few-shot CoT while [97] and [136] explored using CoT. Moreover, the works for sentiment analysis were assessed using zero-shot, few-shot, zero-shot CoT, few-shot CoT and CoT. Combining CoT with zero-shot or few-shot was evaluated only by [116]. Subsequently, references [122,136] investigated the utilization of CoT learning for sentiment analysis, while [113] explored both sentiment analysis and text classification using the same approach. Conversely, only zero-shot and few-shot learning approaches have been examined for annotation tasks.

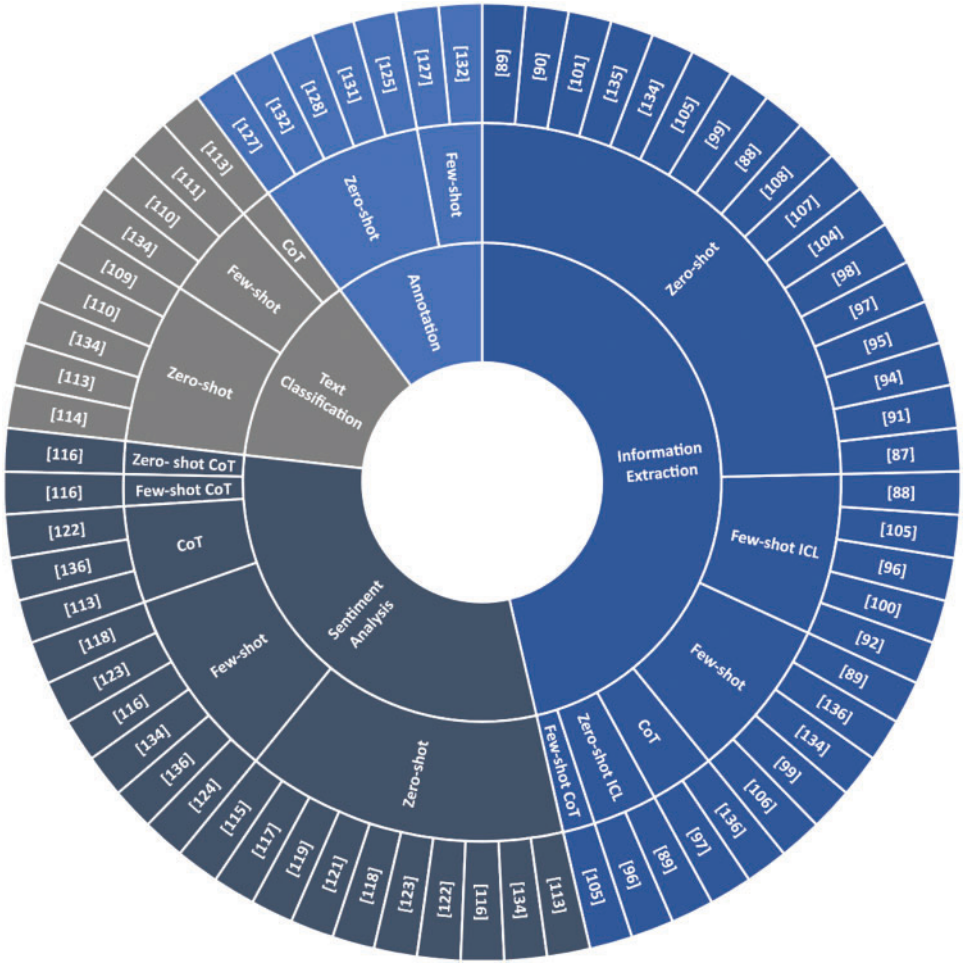

Figure 11: Exploring prompts in natural language processing tasks—a comprehensive breakdown by each paper and task

4.3 Q3: What Are the Natural Languages that Have Been Investigated?

A Pie Chart in Fig. 12 illustrates the frequency of studies in various linguistic contexts. It should be noted that certain papers may be counted multiple times depending on the number of languages investigated. It is evident that most of the studies predominantly focused on the English

language. Fig. 13 visualizes the investigated natural Languages for each NLP task. Notably, five papers specifically examined the text in the Chinese language, covering information extraction [107,108], NER [90], text classification [28], and annotation [126]. Furthermore, three papers concentrated on the French language, exploring its usage in information extraction [87,108] and text classification [110]. Additionally, the Russian language was employed in [117] for sentiment analysis, while the German language was studied in [109] for text classification; both of these languages were undertaken by researchers in [108] for information extraction. Additionally, other languages, including Spanish, Italian, Dutch, Polish, Vietnamese, Turkish, Arabic, Greek, Thai, Bulgarian, Hindi, and Urdu have attained comparatively less attention, with only [108] testing multiple languages for information extraction. Subsequently, the Malaysian language has been investigated in [89] for NER task, the Portuguese language in [102] for narrative IE, the Japanese Language in [103] for medical IE, and the Arabic language in [124] for sentiment analysis.

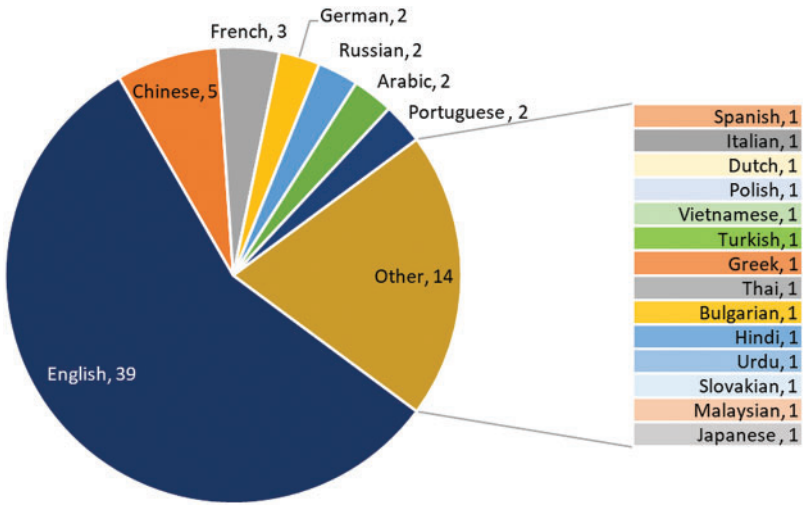

Figure 12: Distribution of papers across natural languages

4.4 Q4: What Are the Selected Dataset, Evaluation Baselines, and Evaluation Matrices that Have Been Used to Measure the Performance?

4.4.1 Information Extraction

González et al. [87] trained ChatGPT on historical data using three datasets: NewsEye, hipe-2020 and ajmc. The datasets included data from different languages, but this study focused on the French language. Besides, they only consider the coarse-grained entities. Subsequently, they compared ChatGPT performance against two state-of-the-art systems, which are the Stacked NERC model-based pretrained on BERT, and the Temporal NERC model that relies on Stacked NERC. Zhou et al. [91] used a benchmark involving 43 NER datasets that cover different domains, which are general, biomedical, clinical, STEM, programming, social media, law, finance, and transportation domains. They compare their proposed model (UniNER) with ChatGPT, Vicuna model, and InstructUIE model.

Furthermore, Tan et al. [88] compared the ability of ChatGPT against BERT-CRF and their developed model named a unified retrieval-augmented system (U-RaNER) for fine-grained multilingual NER using Wikipedia. Chanthran et al. [89] employed two Malaysian English news datasets, MEN-Dataset and DocRED: they tested various learning approaches, including zero-shot, zero-shot ICL

and few-shot. Xie et al. [90] used a group of benchmarks for NER from English and Chinese text, which are PowerPlantFlat (PPF) and PowerPlantNested (PPN), MSRA, Weibo NER, OntoNotes, ACE05, and ACE04. For evaluation, they utilized GPT-3.5-turbo for GPT-3.5, text-davinci-003 for GPT-3, and 13B chat model for Llama. Li et al. [92] used two datasets for the experiment, which are Twitter-2015 and Twitter-2017. They compared their model against several models, including BiLSTM-CRF, CNN-BiLSTM, BERT-CRF, BERT-span, RoBERTa-span, UMT, UMGF, MNER-QG, R-GCN, ITA, PromptM-NER, CAT-MNER and MoRe. Hu et al. [94] compared the performance of ChatGPT and GPT-3 against a supervised model trained using BioClinicalBERT. Likewise, Tang et al. [95] compared the performance of zero-shot ChatGPT with fine-tuning the three models which are BERT, RoBERTa, and BioBERT on the synthetic data and the original training data, which are National Center for Biotechnology Information disease *corpus* (NCBI) and the BioCreative V CDR *corpus* (BC5CDR).

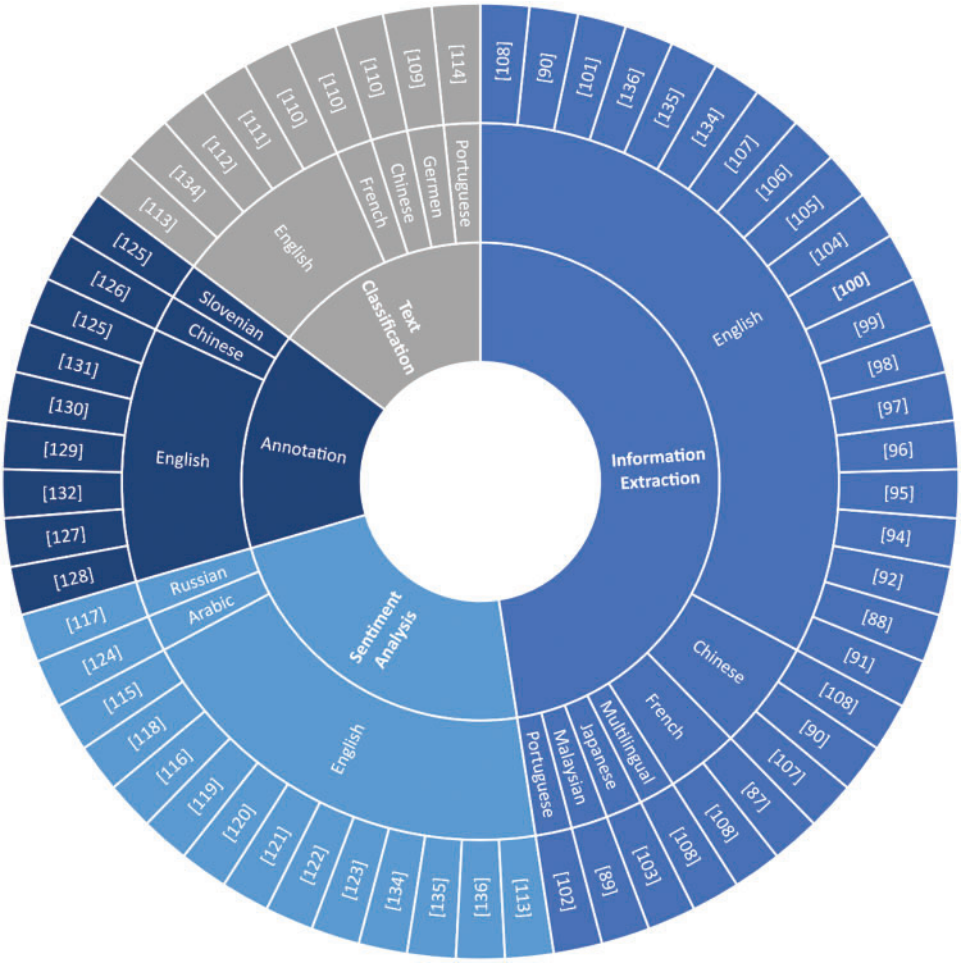

Figure 13: The multilingual landscape explored by each paper

Besides Brinkmann et al. [96] used Mave, which is a product dataset. They trained two pre-trained language models based on a PLM named AVEQA and NER. Yuan et al. [97] designed three different prompts and used three datasets MATRES, TB-Dense, and TDDMan. Due to the absence of zero-shot learning approaches for temporal relation extraction, they compared ChatGPT with

the following advanced supervised methods, which are CAEVO, SP+ILP, Bi-LSTM, Joint, Deep, UCGraph, TIMERS, SCS-EERE, RSGT, FaithTRE, DTRE and MulCo. Besides, Jethani et al. [98] used Fleiss' Kappa to measure the inter-rater agreements between reviewers. For evaluation, 22 medically expert reviewers were asked to judge the correctness of the answers produced by ChatGPT.

Similarly, Peikosa et al. [99] utilized datasets that included clinical information and compared the results with multiple baselines, including KeyBERT, BM25, RM3, QGGT, and QGMT. He et al. [100] compared the performance of their proposed framework with several baselines including BERT, LiLT, BROS, LayoutLM, XYLayoutLM, LayoutLMv2, and LayoutLMv3. The experiment includes three datasets: FUNSD, CORD, and SROIE. Kartchner et al. [101] tested two datasets CML and Remedy. Then, they compared the results obtained by utilizing GPT-3.5 Turbo with GPT-JT. Sousa et al. [102] compared the results against different baselines, which were SRL and TEFE for event extraction, HeidelTime and TEI2GO for time extraction, and SRL baseline for the participant extraction task. Further, they selected the Text2Story dataset since it included annotations for all the entities and had not been used in training the models. Nishio et al. [103] compared the results of the zero-shot approach using GPT-3.5-turbo against GPT-4 by deploying the RR-TNM Japanese dataset for extracted TNM staging.

Gao et al. [104] used the ACE 2005 *corpus* for event extraction and compared the results with two domain-specific models, EEQA and Text2Event. Moreover, the evaluation process in [105] was done using various datasets under zero-shot prompts, few-shot in-context learning (ICL) prompts, and few-shot chain-of-thought (COT) prompts. The datasets based on the NLP tasks are as follows: for the NER task, the datasets include CoNLL03, FewNERD, ACE04, ACE05-Ent, and GE-NIA. For the RE task, the datasets include CoNLL04, NYT-multi, TACRED, and SemEval 2010. Additionally, the datasets include ACE05-Evt, ACE05+, CASIE, and Commodity News EE for the EE task. Finally, for the ABSA task, the datasets include D17, D19, D20a, and D20b. Subsequently, Li et al. [106] compared the performance of ChatGPT among different datasets. The datasets for NER include CoNLL2003 and ACE, while ET, BBN, and OntoNotes datasets were used for ET. For RC, the datasets include TACRED, RIFRE, and SemEval 2010. For RE, the datasets include ACE05-R, SciERC, OntoNotes 5.0, and PL-Marker. Additionally, ACE05-E and ONEIE are used for ED. Subsequently, they tested both Standard-IE and the OpenIE settings, where ChatGPT prompts comprised the words to give the required keys while OpenIE did not give any candidate labels to ChatGPT. After that, they compared ChatGPT with BERT, RoBERTa, and SOTA methods. Wei et al. [107] used six datasets, which are NYT11-HRL, DuIE2.0, conllpp, MSRA DuEE1.0, and ACE05. Additionally, the XGLUE-POS dataset was used in [108] to compare ChatGPT capabilities on 17 multilingual, which are English, Russian, German, Chinese, French, Spanish, Italian, Dutch, Polish, Vietnamese, Turkish, Arabic, Greek, Thai, Bulgarian, Hindi and Urdu in POS tagging against the supervised model (XLM-R).

4.4.2 Text Classification

For text classification, Reiss [109] used 234 website texts that were manually annotated into news and not news. Zhao et al. [110] tested different datasets, which are Amazon-Food-Comments, PestObserver-France, and Agri-News-Chinese which represent English, Chinese, and French languages. They considered the flowing baselines in their experiment: Support Vector Machine (SVM), random forest, TextCNN, TextRNN, T5-based prompt-tuning, BART-based prompt-tuning and PLM-based fine-tuning, in which the PLMs include BERT [34], BART [46] and T5 [45]. In the experiment, the version "bert-base-uncased" has been adopted for the PLM BERT, the version "t5-base"11 for the PLM T5, and the version "facebook/bart-base"12 for the PLM. Subsequently, the experimental hardware environment consists of a CPU Intel Core i9-9900k, and a single Nvidia GPU

of GTX 1080Ti. Additionally, Loukas et al. [111] compared the results with two fine-tuned pre-trained masked language models (MLMs), S-MPNet-v2 and P-MPNet-v2. Lamichhane [112] evaluated the performance of ChatGPT by comparing the results with various models, including fine-tuned BERT, LDA, and CNN. Trajano et al. [113] used a dataset provided by ClimateBERT covering both climate-related and non-climate-related text from several sources, including news articles, Wikipedia articles, and climate reports. Oliveira et al. [114] compared the classification results of hate tweets in Portuguese against the BERTimbau model. They employed the GPT-3.5-turbo model and set the temperature to zero. The ToLD-Br dataset has been used as a primary dataset and HLPHSD as a test dataset. To ensure reliability, they conducted the experiments five times.

4.4.3 Sentiment Analysis and Emotion Recognition

Furthermore, for sentiment analysis, Yang et al. [115] compared several traditional neural networks such as CNN and GR and BERT, RoBERTa, MentalBER, and MentalRoBERTa. Golubev [117] have used the RuSentNE *corpus* and compared ChatGPT in zero-shot with different models such as RuBERT, RuRoBERTa, XLM-RoBERTa, and RemBERT. Wang et al. [118] used 18 datasets and compared the results with BERT, corresponding state-of-the-art (SOTA) models. Besides, they performed human evaluation. Moreover, Zhong et al. [116] assessed and compared the language understanding capabilities of ChatGPT and Fine-tuned BERT on sentiment analysis and question-answering tasks on the GLUE benchmark. For evaluation, they used the following evaluation metrics: Accuracy, Pearson, Spearman correlation, and F1 score. Fatouros et al. [119] collected news headlines from ForexLive and FXstreet websites to generate the dataset. They employed a zero-shot setting by testing various promptings. Further, they compared the performance of the different ChatGPT prompts against the FinBERT model across the following metrics: accuracy, precision, recall, F1, and Mean Absolute Error (MAE).

Correspondingly, the utilized datasets in [120] are the Center for Research in Security Prices (CRSP) daily returns, news headlines, and RavenPack. They compared the results conducted by ChatGPT against BERT, GPT-1, and GPT-2. Models using accuracy, precision, recall, and F1 score metrics. Zhang et al. [121] conducted experiments on two English tweets datasets, which are SemEval-2016 and P-Stance. They suggested investigating performance improvement by prompt engineering and multi-round conversation. Kocoń et al. [123] used different datasets based on the task they tested. For sentiment analysis and emotion recognition, they used ClarinEmo, PolEm, TweetEmoji, TweetSent, GoEmo, and GoEmoPer. They compared the results of ChatGPT and GPT-4 with the SOTA results using the selected datasets. Then, they computed the loss, which indicates how much ChatGPT is worse than the SOTA methods. Khondaker et al. [124] conducted an experiment using various datasets. Additionally, they compared the capabilities of ChatGPT on modern standard Arabic and dialectical Arabic against BLOOMZ and MARBERT.

4.4.4 Annotation

Moreover, for annotation, Kuzman et al. [125] tested their model on two datasets in English and Slovenian. Then, they compared the ChatGPT model with a massively multilingual base-sized XLM-RoBERTa Transformer-based model (X-GENRE classifier). Additionally, Gilardi et al. [131] compared ChatGPT with MTurk crowd-workers platform. Huang et al. [130] annotated tweets using Amazon Mechanical Turk (Mturk). After that, they used Informativeness and Clarity tools to capture the relevance and measure the clarity of the generated explanation. Chen et al. [126] ran an experiment for event extraction and compared their model with other models in the two datasets, DuEE1.0 and Title2Event.

Koptyra et al. [129] compared emotion recognition using three datasets: CLARIN-Emo, ChatGPT-Emo, and Stockbrief-GPT. CLARIN-Emo consists of human-written Polish reviews that have been manually annotated with emotions, while the Stockbrief-GPT dataset is written by humans and annotated with emotions using Chat-GPT. Additionally, in the ChatGPT-Emo dataset, both text and labels were produced completely by ChatGPT. Subsequently, the analysis in [132] relies on zero-shot and few-shot approaches and uses four datasets which include tweets and news articles. Korini et al. [127] compared the results with a fine-tuned RoBERTa model and the DODUO model. Belal et al. [128] used manually annotated tweets and Amazon reviews and then compared the results with two lexical tools, which are VADER and TextBlob.

Furthermore, Li et al. [134] evaluated GPT-4 using different finance-related datasets on different NLP tasks: text classification, named entity recognition, and sentiment analysis. For evaluation, they compared the results with several pre-trained models (BloombergGPT, GPT-Neo, OPT66B, BLOOM176B). Qin et al. [135] compared the F1 score of ChatGPT in NER with three fine-tuned baselines, which are Flair, LUKE, and ACE, using the CoNLL03 dataset. For sentiment analysis, they compared ChatGPT performance with FLAN (zero-shot), and T5-11B (fine-tuned) and used the SST2 dataset. Sun et al. [136] employed different datasets based on the tasks. WSJ Treebank and Tweets dataset have been used for PoS tagging, ACE2005 dataset for event extraction, ACE2004 and ACE2005 datasets for entity extraction and both CoNLL2003 and OntoNotes5.0 have been used for NER. For sentiment analysis, they utilized SST-2, IMDb, and Yelp datasets. They compared the results with RoBERTa-Large. Further, they tested different strategies, including using CoT, demonstrations strategy with KNN, and one-input-multiple-prompts strategy, to address the issue of having limited tokens in the input.

Table 1 and Table 2 show the dataset, evaluation baseline, and evaluation metrics across different papers, providing insights into comparative analyses and evaluation approaches. As we can see, the datasets are varied, and social data is rarely used. ACE05 [137] and ACE04 [138] dataset, which is a Multilingual Training *Corpus* in three languages (English, Chinese, and Arabic) has been used by [104–107,136] for event extraction, NER and entity extraction tasks. All of them used it for English text except [107], which focused on the Chinese language. CoNLL2003 [139] dataset for NER has been used by [105–107,135,136]. Additionally, OntoNotes5.0 [140], which is a large *corpus* in English, Arabic, and Chinese languages has been employed in [106], and [136] for NER and relation extraction. Moreover, Tweets datasets played a role in [92,136,121,123] for sentiment analysis and in [128,132] for annotation. Additionally, News served as a dataset for NER [87], text classifications [110,123,134], sentiment analysis [117,120] and annotation [132]. Additionally, Reddit comments have been utilized for classification [112] while Wikipedia is used for sentiment analysis and emotion recognition [123]. Likewise, the used baselines for evaluation are varied including deep learning models, machine learning algorithms, topic modeling, and LLM and they depend on the type of the task. Subsequently, the commonly used metrics in all tasks are Accuracy, Precision (P), Recall (R), and F1 score (Micro-F1 and Macro-F1, and weighted F1). Other metrics have been used such as P@5 (Precision at 5), P@10, R-Precision (Rprec), and mean reciprocal rank (MRR).

Table 1: The explored datasets in various NLP tasks across multiple papers

NLP tasks	Ref.	Datasets
Information extraction	[87]	NewsEye, hipe-2020 and AJMC

(Continued)

Table 1 (continued)

NLP tasks	Ref.	Datasets
	[91]	43 NER datasets that cover different domains, which are general, biomedical, clinical, STEM, programming, social media, law, finance, and transportation domains
	[88]	MultiCoNER
	[92]	Twitter-2015 and Twitter-2017
	[94]	Synthetic clinical notes from MTSamples, and the 2010 i2b2 challenge
	[95]	NCBI and BC5CDR *corpus*
	[96]	Mave
	[97]	MATRES, TB-Dense and TDDMan
	[89]	MEN-Dataset and DocRED
	[90]	PPF, PPN, MSRA, Weibo NER, OntoNotes, ACE05, BC5CDR, and ACE04
	[103]	RR-TNM
	[98]	Clinical notes
	[99]	TREC 2021 and TREC 2022
	[100]	FUNSD, CORD, SROIE
	[104]	ACE05 *corpus*
	[105]	CoNLL2003, FewNERD, ACE04, ACE05-Ent, GE-NIA, CoNLL04, NYT-multi, TACRED, SemEval 2010, ACE05-Evt, ACE05+, CASIE, Commodity New, D17, D19, D20a, and D20b
	[106]	CoNLL2003, BBN, TACRED, RIFRE, SemEval 2010, SciERC, OntoNotes 5.0, PL-Marker, ACE, ACE05-R, ACE05-E and ONEIE
	[107]	NYT11-HRL, DuIE2.0, conllpp, MSRA DuEE1.0, ACE05
	[108]	XML-R
	[134]	Financial NER,
	[135]	CoNLL2003
	[136]	CoNLL2003, OntoNotes5.0, ACE04, ACE05, WSJ Treebank, and Tweets dataset
	[101]	CML and remedy
	[102]	Text2Story
Text classification	[109]	Manually annotated websites
	[110]	Amazon-Food-Comments, PestObserver-France, Natural-Hazards-Twitter, and Agri-News-Chinese
	[111]	Banking77
	[112]	Manually labeled user posts in different Reddit groups
	[134]	Headline
	[113]	ClimateBERT

(Continued)

Table 1 (continued)

NLP tasks	Ref.	Datasets
	[114]	ToLD-Br, HLPHSD
Sentiment analysis	[115]	IEMOCAP, MELD, and EmoryNLP datasets
	[117]	RuSentNE *corpus* (Russian news)
	[118]	18 benchmark datasets
	[116]	GLUE
	[119]	Collected from ForexLive and FXstreet websites
	[120]	CRSP, daily returns, news headlines, and RavenPack
	[121]	Tweets datasets (SemEval-2016 and P-Stance)
	[122]	Tweets datasets (SemEval-2016 and P-Stance)
	[123]	ClarinEmo, PolEm, TweetEmoji, TweetSent, GoEmo and GoEmoPer
	[134]	Financial PhraseBank
	[135]	SST2
	[136]	SST-2, IMDb, and Yelp
	[124]	ORCA
Annotation	[125]	EN-GINCO and GINCO
	[131]	Manually annotated tweets
	[130]	LatentHatred dataset
	[126]	DuEE1.0, Title2Event
	[129]	CLARIN-Emo, ChatGPT-Emo, Stockbrief-GPT
	[132]	Tweets and news articles
	[127]	SOTAB benchmark
	[128]	Tweets, amazon reviews

Table 2: Evaluation baselines and metrics explored in NLP tasks

NLP task	Ref.	Evaluation baselines/SOTA	Metrics
Information extraction	[87]	Stacked NERC and temporal NERC models	P, R, and F1 score
	[91]	ChatGPT, vicuna model and InstructUIE	Averaged F1 scores
	[88]	BERT-CRF, U-RaNER, RaNER, RaNER, MSF	P, R, and micro-averaged F1 scores

(Continued)

Table 2 (continued)

NLP task	Ref.	Evaluation baselines/SOTA	Metrics
	[92]	BiLSTM-CRF, CNN-BiLSTM, BERT-CRF, BERT-span, RoBERTa-span, UMT, UMGF, MNER-QG, R-GCN, ITA, PromptM-NER, CAT-MNER and MoRe.	P, R, and micro-averaged F1 scores
	[94]	BioClinicalBERT	P, R, and micro-averaged F1 scores
	[95]	BERT, RoBERTa, and BioBERT	P, R and F1 score
	[96]	Pre-trained models named, AVEQA and NER	P, R, and F1 score
	[89]	–	F1 score
	[90]	GPT-3, Llama	F1 score
	[103]	GPT-4	Accuracy
	[97]	CAEVO, SP+ILP, Bi-LSTM, Joint, Deep, UCGraph, TIMERS, SCS-EERE, RSGT, FaithTRE, DTRE and MulCo	P, R and F1 score
	[98]	Humen-reviewers	Fleiss' Kappa
	[99]	KeyBERT, BM25, RM3, QGGT, QGMT	nDCG@10, P@5, P@10, Rprec, and MRRs
	[100]	BERT, LiLT, BROS, LayoutLM, XYLayoutLM, LayoutLMv2, and LayoutLMv3	F1 score
	[104]	EEQA and Text2Event	P, R and F1 score
	[105]	SOTA methods	Micro-F1
	[106]	BERT and RoBERTa	Accuracy
	[107]	Vanilla ChatGPT prompt	F1 score
	[108]	XGLUE-POS	Accuracy
	[134]	BloombergGPT, GPT-NeoX, OPT66B, BLOOM176B	Accuracy, F1
	[135]	Flair, LUKE and ACE	F1 score
	[136]	RoBERTa-Large	F1 score, accuracy
	[101]	GPT-JT	Accuracy, P, R and F1 score
	[102]	SRL,TEFE, HeidelTime and TEI2GO	P, R and F1 score

(Continued)

Table 2 (continued)

NLP task	Ref.	Evaluation baselines/SOTA	Metrics
Text classification	[109]	Two temperature settings and prompt instruction of ChatGPT	Krippendorff's Alpha
	[110]	SVM, random forest, TextCNN, TextRNN, BERT-based fine-tuning, T5-based prompt-tuning, BART-based prompt-tuning	Micro-F1, macro-F1, and weighted-F1
	[111]	S-MPNet-v2 and P-MPNet-v2.	Micro-F1 and Macro-F1
	[112]	fine-tuned BERT, LDA, and CNN.	F1 score, accuracy
	[134]	BloombergGPT, GPT-NeoX, OPT66B, BLOOM176B	Accuracy, F1
	[113]	ClimateBERT	Accuracy, F1
	[114]	BERTimbau	P, R and F1 score
Sentiment analysis	[115]	CNN, GRU, BiLSTM_Att, fastText, BERT, RoBERTa, MentalBERT, MentalRoBERTa	weighted-F1 scores
	[117]	RuBERT, RuRoBERTa, XLM-RoBERTa, RemBERT	F1-PN-macro
	[118]	BERT, corresponding SOTA models and human evaluation	Accuracy and macro F1 score
	[116]	Fine-tuned BERT	Accuracy, P, Spearman correlation, and F1
	[119]	FinBERT,	Accuracy, P, R, F1, and MAE
	[120]	BERT, GPT-1, and GPT-2. models,	Accuracy, P, R, F1
	[121]	Conducted experiments on SemEval-2016 and P-Stance datasets	F1-avg and macro F1
	[122]	Bicond, CrossNet, SEKT, MemNet, AOA, TAN, AS-GCN, Bert_spc, Bert-GCN, and PT-HCL	Favg and macro-F1 score
	[123]	Compared the results with SOTA results using the selected datasets	Macro-F1, accuracy

(Continued)

Table 2 (continued)

NLP task	Ref.	Evaluation baselines/SOTA	Metrics
	[134]	FinBERT BloombergGPT, GPT-NeoX, OPT66B, BLOOM176B	Accuracy, F1
	[135]	FLAN (zero-shot), and T5-11B (fine-tuned)	Accuracy
	[136]	RoBERTa-Large	F1
	[124]	BLOOMZ and MARBERT	Macro-F1
Annotation	[125]	Compared with XLM-RoBERTa	Micro F1, macro F1, accuracy
	[131]	MTurk	Accuracy
	[130]	MTurk	Informativeness and clarity scores
	[126]	CPEE, ERNIE + CRF, MRC, Bert4keras, BERT-CRF, Seq2seqMRC	P, R, and F1-measure
	[129]	Classification of different datasets	F1 score
	[132]	HugginChat, FLAN, and MTurk	Accuracy
	[127]	RoBERTa and DODUO	P, R, and micro-F1 score.
	[128]	lexical-based tools (VADER and TextBlob)	Accuracy, P, R, and F1 score

4.5 Q5: Does the Deployment of ChatGPT Enhance the Performance Compared to the Baseline Models/SOTA Methods?

4.5.1 Information Extraction

The works that showed that harnessing ChatGPT in the NER task can provide positive results are as follows: Tang et al. [95] found that ChatGPT can produce high-quality synthetic data in the biomedical NER task. Subsequently, the findings demonstrated significant improvements when fine-tuning the models on the synthetic data generated by ChatGPT compared to using ChatGPT in the zero-shot scenario. Subsequently, the findings of Li et al. [92] showed that their proposed framework that leveraged ChatGPT as a knowledge engine to boost the model performance in MNER outperforms the SOTA methods. Li et al. [134] found that GPT-4 outperforms the other models in NER for financial text using zero-shot and few-shot learning. Xie et al. [90] findings indicated that the proposed decomposed-QA and syntactic augmentation approaches improve the performance of NER in both English and Chinese text using zero-shot. It outperformed Llama2 13B, and achieved improvements, with 19.72% and 17.51% F1 on ACE05 and BC5CDR datasets, respectively.

Furthermore, Jethani et al.'s [98] results indicated that ChatGPT successfully extracted the information from the clinical note. Similarly, Peikosa et al. [99] found that queries generated by ChatGPT enhance retrieval performance from clinical notes compared to SOTA methods as well as

human-generated queries. He et al. [100] enabled GPT-3 to predict entity labels in a test document. The results indicated that the proposed in-context learning framework (ICL-D3IE) that enables ChatGPT surpassed the performance of Standard ICL and all the other baselines on the three datasets, which are FUNSD, CORD, and SROIE. Kartchner et al. [101] illustrated that ChatGPT has the capacity to improve the effectiveness of clinical meta-analyses of randomized clinical trials, highlighting its superior performance compared with the GPT-JT-6B model using zero-shot learning. Nishio et al. [103] evaluated the results of their proposed system for TNM staging extraction from Japanese reports and showed that GPT-4 achieved higher results than version 3.5. Besides, they found that the English prompt achieved higher results than the Japanese prompt. However, they did not compare their results with the baselines to show whether their system performed better or worse. Sousa et al. [102] found that GPT-3 and GPT-3.5 models excel in extracting participants and time compared to the baseline, but they do not achieve comparable results in event extraction. GPT-3.5 demonstrates superior performance over GPT-3 in extracting participants and events, while GPT-3 exhibits better effectiveness in extracting time.

Moreover, Li et al. [106] evaluated the performance of ChatGPT on ET, NER, RC, RE, ED, EAE, and EE. Their results showed that ChatGPT achieved better results using the OpenIE setting compared to the Standard-IE setting. Wei et al. [107] findings revealed that using the proposed two-stage prompt for information extraction in English and Chinese languages using ChatGPT helped in extracting information compared with using the vanilla prompt. Furthermore, Brinkmann's et al. [96] evaluated In-Context Learning (ICL) with both zero-shot and few-shot settings by providing the task description and the task demonstrations where each task demonstration includes an example input and an example output. Further, they tested closed extraction (extract value of a specific attribute) and open extraction (extract all attribute-value pairs) from a product title. The results indicate that ChatGPT in zero-shot prompt designs achieved performance similar to that of the pre-trained model, NER, for both open and closed extraction. Subsequently, the results are further improved by adding demonstrations to the prompts; in addition to that, one-shot (single example) is sufficient to enhance the performance. Gao et al. [104] noticed that the performance of ChatGPT in zero-shot scenarios to extract events varied based on the prompt styles. Sun et al. [136] found that combining the proposed strategies with ChatGPT enables achieving comparable results with the RoBERTa-Large model in PoS tagging, event extraction, and entity-relation extraction tasks. However, it failed to achieve higher results in the NER task.

Conversely, the remaining studies indicate that ChatGPT falls short of outperforming the other models. González et al. [87] found that ChatGPT in zero-shot learning faced several issues in identifying entities, including entity complexity, inconsistent entity annotation guidelines, and multilingualism. Subsequently, Tan et al. [88] findings indicate that ChatGPT performance was poor compared to their proposed tool U-RaNER for NER even though they used different prompting settings, which are Single-turn, Multi-turn as zero-shot learning and Multi-ICL as few-shot learning. Qin et al. [135] found that ChatGPT 3.5 could not achieve satisfactory results on NER compared to previous fine-tuning baselines. Similarly, Zhou et al. [91] found that their model (UniNER) surpasses ChatGPT and other state-of-the-art systems and attained remarkable accuracy. Additionally, the evaluation result by Hu et al. [94] indicates that ChatGPT performance surpasses GPT-3. However, none of them can beat the supervised model. Thus, they suggested fine-tuning ChatGPT with domain-specific corpora to improve the performance. Yuan et al. [97] found that leveraging ChatGPT for temporal relation extraction underperforms the existing models, such as Bi-LSTM. Further, they found that their designed Chain-of-thought Prompt achieved better results than the other prompt, which demonstrates the importance of proper prompt engineering. Han et al. [105] findings demonstrated that ChatGPT

cannot outperform SOTA methods for the following tasks: NER, RE, EE, and ABSA using zero-shot, few-shot ICL, and few-shot COT prompts. Lai et al. [108] tested the ability of ChatGPT in different NLP tasks, including part-of-speech (POS) tagging and NER. In POS tagging, the prompt involves a task description, an input, and a note for output format. The XGLUE-POS dataset was used to compare ChatGPT capabilities on 17 multilingual POS against the supervised model (XLM-R). Their findings indicated that ChatGPT performed significantly worse than SOTA methods in NER. However, ChatGPT can achieve competitive results in some languages, such as English, Thai, Vietnamese, Bulgarian, Hindi, and Urdu. Chanthran et al. [89] found that ChatGPT failed to achieve good results for NER from Malaysian English news articles.

4.5.2 Text Classification

For text classification, three studies demonstrate that ChatGPT surpasses the other models. Zhao et al. [110] tested several experiments and evaluated different prompts using zero-shot and few-shot approaches on GPT-4, vanilla ChatGPT, and GPT-3.5. They compared the results with several baseline models, including traditional ML methods such as SVM and Random Forest. The finding indicated that the proposed framework, ChatAgri which used ChatGPT for cross-linguistic text classification, performs significantly better than other methods in zero-shot setting regardless of different classification category topics and counts. Subsequently, they found that GPT-4 achieved better performance than vanilla ChatGPT, GPT-3.5. Li et al.'s [134] results indicated that GPT-4 outperforms the other models in classifying news headlines that include price information and NER. Loukas et al. [111] showed that GPT-3.5 and GPT-4 for financial text classification gave competitive results with fine-tuned pre-trained masked language models (MLMs), and achieved good results even when presented with fewer examples. Trajano et al. [113] executed multiple experiments to test the ability of ChatGPT 3.5 and GPT-4 in multiple classification tasks for limited text. They tested two approaches using manually designed prompts and ChatGPT, a Zero-Shot Classifier from the Scikit-LLM library. The findings indicated that ChatGPT 3.5 outperformed GPT-4 and ClimateBERT model in the three binary classification tasks (climate detection classification, climate-related commitments and actions classification, climate change specificity classification) while GPT-4 surpassed the others in the four-class classification problem (climate change disclosure category). The best results for climate detection classification were achieved using ChatGPT 3.5 with Scikit LLM Library. Deploying simple prompts helped get higher results for climate-related commitments and actions classification. Additionally, CoT improved the results of climate change specificity classification. Oliveira et al. [114] findings demonstrated the ability of ChatGPT as a heat speech classifier in the Portuguese language. It achieved good results using zero-shot learning.

Conversely, Reiss's [109] finding indicated that ChatGPT does not achieve acceptable results for annotation and binary classification into news and not news. Further, it gives different results using the same configurations, which questions its reliability. Likewise, even though Lamichhane [112] found that ChatGPT attained good results in binary classification for mental health, the results indicated that the fine-tuned BERT outperformed ChatGPT in stress detection. Similarly, combining bigrams and LDA helped in achieving higher results compared to ChatGPT for depression detection. Subsequently, the result obtained by ChatGPT for the suicidality detection task, which is 5-a class classification problem, is deficient, while using another approach, such as the CNN model achieved better results.

4.5.3 Sentiment Analysis and Emotion Recognition

For sentiment analysis, Wang et al. [118] results indicated that ChatGPT in a zero-shot scenario showed remarkable performance on the sentiment classification task. ChatGPT demonstrates superior

performance compared to BERT on SST datasets when addressing negation and speculation linguistic phenomena. Specifically, ChatGPT achieves accuracy scores of 91.00 and 92.00 for negation and speculation, respectively, whereas BERT achieves scores of 90.68 and 92.05 for the same phenomena. Subsequently, they found that a few-shot setting significantly improved the performance. Additionally, ChatGPT achieved better results than the SOTA model on the emotion cause extraction task. Even though ChatGPT shows low accuracy on sentiment information extraction tasks, human evaluation indicates that it can give good results in these tasks. Fatouros et al. [119] ChatGPT achieved higher results than the FinBERT model for sentiment analysis within the financial domain. Additionally, they highlighted the importance of prompt design in instructing ChatGPT since the results vary between different tested prompts. Similarly, Lopez-Lira et al. [120] emphasized the effectiveness of ChatGPT in forecasting stock price movements using sentiment analysis of news headlines where ChatGPT outperforms the other models. Zhang et al. [121] investigated the ability of ChatGPT to detect stance events in a zero-shot setting by ChatGPT. They found that ChatGPT using the CoT [122] approach achieved better performance than all the baselines. Sun et al. [136] found that combining the proposed demonstration strategy with ChatGPT enables higher results than RoBERTa-Large in binary classification into positive and negative. Trajano et al. [113] found that ChatGPT using a simple prompt outperformed GPT-4 and ClimateBERT in classifying paragraphs into neutral sentiment, opportunity sentiment, or risk sentiment. Zhong et al. [116] results showed that ChatGPT achieves comparable performance of 78.7% compared to BERT-base, which scored 79.2% on sentiment analysis. Additionally, ChatGPT significantly benefits from manual CoT compared with other prompting, where Few-shot CoT leads to enhancements in ChatGPT's performance, resulting in an improvement of 86.2%.

On the contrary, the remaining studies illustrate that ChatGPT cannot beat the other models. Yang et al.'s [115] findings indicated that although ChatGPT achieved good results compared to the traditional neural networks (such as CNN and GRU) for emotion recognition using zero-shot learning, its performance is still not comparable with the advanced models such as RoBERTa and MentalRoBERTa. Additionally, they highlighted the weaknesses of ChatGPT in emotion-related subjective tasks such as unstable perdition and inaccurate reasoning. Similarly, Golubev [117] compared ChatGPT in zero-shot for sentiment analysis in Russian news text with different models such as RuBERT, RuRoBERTa, XLM-RoBERTa, and RemBERT. They found that ChatGPT comes in fourth place in evaluation with 60% of the F-measure. Khondaker et al. [124] results showed that ChatGPT cannot surpass MARBERT on sentiment analysis and emotion detection from Arabic text using zero-shot and few-shot. ChatGPT achieved comparable results with BLOOMZ using a 5-shot while BLOOMZ needed 10-shot. Further, they found that both ChatGPT and GPT-4 failed to achieve good results in dialectal Arabic. Besides, employing GPT-4 as an evaluator emphasizes that it can be a good alternative for human evaluation.

Li et al. [134] found that ChatGPT and GPT-4 performance in sentiment analysis tasks in the financial domain, which required an understanding of financial knowledge and terminology, is lower than that of domain-specific models such as FinBert. Kocoń et al. [123] findings indicated that ChatGPT and GPT-4 do not outperform the SOTA method in sentiment analysis, emotion recognition, and stance detection tasks. They computed the loss, which shows how much ChatGPT is worse than the SOTA methods. Notably, datasets focused on emotions with subjective interpretations showed the highest Loss. Specifically, GoEmotions, PolEmo, and TweetEmoji exhibited Loss values of 51.56 and 43.51, respectively. Meanwhile, the Loss values for GoEmoPer0, GoEmoPer1, GoEmoPer2, and GoEmoPer3 were 56.44, 71.26, 69.23, and 64.58, respectively. Qin et al. [135] found that chatgpt3.5

achieved higher accuracy than ChatGPT in classifying text into positive and negative. However, they cannot beat the baseline method.

4.5.4 Annotation

For text annotation, several works indicate that ChatGPT can surpass the existing models. Gilardi et al. [131] focused on various annotation tasks, including relevance, stance detection, topics detection, general frame detection, and policy frames detection. They found that ChatGPT zero-shot annotation achieved better results than MTurk. Huang et al. [130] results indicate that ChatGPT was able to identify 80% of the implicit hateful tweets. On top of that, the explanations given by ChatGPT are generally considered to be clearer than those written by humans. Chen et al. [126] leveraged ChatGPT to create annotated data for event extraction tasks in the Chinese language and then trained the model using supervised learning algorithms. The findings demonstrate that the proposed model using ChatGPT outperforms the baseline models. Belal et al. [128] assessed the ability of ChatGPT as a text annotator in zero-shot for sentiment analysis. The results show that ChatGPT significantly outperforms the lexicon-based and can recognize emojis as well as sarcasm. Korini et al. [127] results indicate that ChatGPT achieved competitive results for the column type annotation task in zero and few shots, while a RoBERTa model needs to be fine-tuned with 356 shots to achieve similar results with ChatGPT in zero-shot.

However, three works show that ChatGPT struggles to achieve good results. Kuzman et al. [125] found that the fine-tuned X-GENRE classifier performs significantly better than ChatGPT on the Slovenian test set. Additionally, ChatGPT performs poorly when the Slovenian prompt is used rather than the English prompt for text annotation into genre categories. Alizadeh et al. [132] compared the performance of Large Language Models (HugginChat and FLAN) with ChatGPT and human-based services (MTurk) in text annotation tasks. The results indicated that open-source LLMs achieved higher results than ChatGPT in some tasks and both beat MTurk. Koptyra et al.'s [129] findings for emotion recognition showed that their classifier achieved the best results on the CLARIN-Emo dataset that has been manually annotated compared with the other two datasets (ChatGPT-Emo and Stockbrief-GPT), which indicates that high-quality data can be obtained via manual annotation.

Fig. 14 illustrates a bar chart comparing the performance of ChatGPT across various publications. The chart is divided into categories representing different scenarios. The light blue color represents the number of publications where ChatGPT demonstrated superior performance compared to baseline models, while the gray color represents the number of publications where ChatGPT achieved results comparable to baseline models. Besides, the dark orange color represents the number of publications where ChatGPT faced challenges and did not perform as well as baseline models. This visual representation provides insights into the distribution of outcomes across the examined publications, shedding light on the effectiveness of ChatGPT in comparison to baselines/SOTA in the specified NLP tasks.

5 Discussion

This section illustrates the strengths and weaknesses of utilizing ChatGPT. It starts with discussing the general strengths of ChatGPT and then explains its capabilities for each of the discussed NLP tasks in this review. Then, it outlines the challenges in leveraging ChatGPT for NLP tasks.

5.1 Strengths and Capabilities

Generally, the capabilities of ChatGPT in NLP tasks can be summarized as follows:

- **General language understanding:** the ability to possess a vast knowledge of language patterns, which makes it suitable for tasks that require a deep language understanding. Besides, it is trained on massive text corpora, as shown in Fig. 7, facilitating a broad understanding of natural language and enabling the generation of text across various domains and topics such as healthcare, finance, etc.
- **Adaptability of context:** It can adapt its responses based on the context of the conversation or prompt and produce relevant outputs, making it practical for various NLP tasks, including sentiment analysis and text classification. Subsequently, it relies on transformer architecture, which enables the effective capture of contextual information and adjustment of responses based on preceding dialogue.
- **Availability of information:** a considerable volume of information regarding this model including prompt engineering is publicly available due to the widespread popularity of ChatGPT compared to other domain-specific LLMs and given its online accessibility and the massive amount of data it's trained on.
- **Effectiveness in learning from a few examples:** the ability to learn effectively from a relatively few training examples, reducing the need for extensive training data and making it ideal for tasks like named entity recognition and information extraction.
- **Potential enhancements achieved through ChatGPT-4 and prompt engineering:** utilizing ChatGPT-4 along with effective prompt engineering can enhance the model's performance and adapt its responses to specific tasks or contexts, such as extracting particular information or understanding the sentiment and emotions.

Figure 14: Comparison of ChatGPT performance in relation to baseline models: number of publications

Furthermore, Table 3 provides an overview of the studies where ChatGPT has shown superior performance compared to baseline models and state-of-the-art (SOTA) approaches. It includes details such as the study reference, the NLP task addressed, the learning approach, the ChatGPT version, and the natural language explored.

Moreover, from this review, we can notice that several studies have investigated the use of ChatGPT in information extraction, and the results have been promising. It surpasses the other baselines using zero-shot for NER [90]. Besides, it outperforms the baselines and SOTA methods for financial NER [134]. In the health domain, ChatGPT beat the other models in [95] biomedical NER and medical IE [98,101,103]. All these studies used zero-shot learning, except [99], which tested both zero-shot and few-shot learning for patient information extraction. However, it fails in clinical NER [94] and we assume the reason is the use of GPT-3. Additionally, GPT-4 in [103] achieved higher results compared with version 3.5.

Table 3: Details of studies demonstrating ChatGPT outperformance over baselines and SOTA

Task	Ref.	Subtask	GPT version	Prompt leaning approach	Language
Information extraction	[90]	NER	GPT-3.5, GPT-3	Zero-shot	English and Chinese
	[107]	NER, EE	GPT-3.5	Zero-shot	Chinese
	[92]	Multimodal NER	GPT-3.5	Few-shot ICL	English
	[95]	Biomedical NER	Na	Zero-shot	English
	[134]	Financial NER	GPT-3.5, GPT-4	Zero-shot, few-shot	English
	[100]	Document IE	GPT-3	Few-shot ICL	English
	[99]	Medical IE	GPT-3.5	Zero-shot, few-shot	English
	[101]	Medical IE	GPT-3.5	Zero-shot	English
	[98]	Medical IE	NA	Zero-shot	English
	[103]	Medical IE	GPT-3.5, GPT-4	Zero-shot	Japanese
	[102]	Narrative IE	GPT-3.5, GPT-3	NA	Portuguese
	[106]	NER, EE, RE	NA	Few-shot	English
Text classification	[110]	Agricultural text classification	GPT-3.5, GPT-4	Zero-shot, few-shot	English, Chinese, and French
	[134]	Financial text classification	GPT-3.5, GPT-4	Zero-shot, few-shot	English
	[113]	Climate	GPT-3.5, GPT-4	Zero-shot, CoT	English
	[114]	Hate speech	GPT-3.5	Zero-shot	Portuguese

(Continued)

Table 3 (continued)

Task	Ref.	Subtask	GPT version	Prompt leaning approach	Language
Sentiment analysis and emotion recognition	[118]	Sentiment analysis, emotion recognition, and opinion analysis	GPT-3.5	Zero-shot, few-shot	English
	[119]	Sentiment analysis in the finance domain	GPT-3.5	Zero-shot	English
	[120]	Sentiment analysis in the finance domain	GPT-3.5	NA	English
	[121]	Stance detection	GPT-3.5	Zero-shot	English
	[122]	Stance detection	GPT-3.5	zero-shot, CoT	English
	[136]	Sentiment analysis	GPT-3.5	Few-shot, CoT	English
	[113]	Climate classification	GPT-3.5, GPT-4	Zero-shot, CoT	English
Annotation	[131]	Various topics	GPT-3.5	Zero-shot	English
	[130]	Hate speech	GPT-3.5	NA	English
	[126]	Event extraction	NA	NA	Chinese
	[128]	Sentiment analysis	GPT-3.5	Zero-shot	English

Furthermore, utilizing ChatGPT using few-shot ICL for multimodal NER [92] and document information extraction [100] beats the other approaches. However, limited studies investigated using Few-shot COT, CoT and zero-shot ICL and they failed to outperform the baselines. Moreover, prompt engineering helped to outperform the SOTA methods in event extraction and NER using zero-shot [107] and using few-shot for multiple tasks, including ET, NER, RE, EE RC, ED, and EAE [106].

All studies mentioned above focus on the English language except [107] and [90], which also explored the Chinese language, while [103] examined the Japanese language. Following this, ChatGPT 3.5 was utilized by all studies mentioned, except for [134] and [103], which employed version 4. Only one study has tested multiple languages and found that ChatGPT in zero-shot can achieve competitive performance in PoS tagging [108] for English, Thai, Vietnamese, Bulgarian, Hindi, and Urdu languages.

Furthermore, in text classification tasks, ChatGPT performs significantly better than other methods in Financial text classifications [134], Agricultural text classification [110], climate text classification [113] and hate speech classification from tweets [114]. All studies utilized GPT-4 except for [114], which used version 3.5. Also, only [113] tested the CoT learning approach. Additionally, reference [114] concentrated on the Portuguese language. While the remaining studies explored the English language, reference [110] also investigated Chinese and French languages.

Moreover, for sentiment analysis, ChatGPT outperforms the baselines and SOTA methods and shows remarkable performance using few-shot and zero-shot for both sentiment classification and emotion recognition tasks [118] and binary sentiment classification [136] using few-shot and CoT. Besides, ChatGPT outperforms other models in sentiment classification within the climate domain using CoT and zero-shot [113], and the finance domain [119,120]. For stance detection, ChatGPT achieved good results using zero-shot [121] and CoT learning [122]. All these studies were centered on English and utilized ChatGPT 3.5, except for [113], indicating that using GPT-4 improves performance. Consequently, it can be seen from Table 3 that the combination of CoT with zero-shot, as seen in [113,122], and with few-shot, as observed in [136], yielded improved outcomes.

Subsequently, for annotation tasks, ChatGPT outperforms the existing approaches, particularly in annotation for implicit hateful text [130], sentiment classification [43], and stance detection [131]. Table 4 summarizes the capabilities of ChatGPT for each of the discussed NLP tasks in this review.

Table 4: ChatGPT's capabilities in NLP tasks—insights from reviewed papers

Information extraction
• Identifies and classifies named entities with good performance. • Adapts to domain-specific terminology and entity types such as biomedical terminology, increasing its versatility. • Extracts key information regarding different domains such as business or medical information extraction from text with good accuracy. • Identifies and extracts key events within a text. • Handles diverse text structures and formats, including news articles, medical documents, and social data.
Text classification
• Categorizes text accurately into predefined classes, enabling efficient organization and retrieval.
Sentiment analysis
• Identifies sentiment accurately across varied text, including social media posts, reviews, websites, and news articles. • Ensure accuracy in specific domains including finance as well as the ability to differentiate positive, negative, and neutral sentiment.

(Continued)

Table 4 (continued)

Annotation
• Generates diverse and accurate annotations for text data, reducing manual effort in labeling tasks. • Speeds up the annotation process, accelerating research and development in NLP.

5.2 Weaknesses and Limitations

Although leveraging ChatGPT achieved promising results for information extraction, particularly for the medical domain, a primary concern arises regarding privacy. This stems from the necessity to directly upload patient data to ChatGPT. Besides, limited studies investigated using ICL and CoT learning methods, indicating a need for further investigation in the future to study their impact on performance. Furthermore, studies employing ChatGPT for text classification remain scarce, and few managed to surpass baseline models. For example, tasks like news website binary classification [109] and 5-class suicidality detection [112] yielded underwhelming results with ChatGPT. Notably, both used ChatGPT 3.5 and focused on English language texts. Therefore, further studies are required to handle multi-label classification tasks effectively and assign multiple categories to a single text.

Moreover, while ChatGPT holds immense promise for sentiment analysis and emotion recognition, relying on English-centric studies limits their true potential. More research is needed to explore ChatGPT's ability to perform these tasks in diverse languages. By examining deeper for different languages, we can unlock a richer understanding of human sentiment and emotion across cultures, evaluating the capabilities of ChatGPT to be a truly global tool for emotional intelligence and communication. Testing different prompt learning approaches and datasets, including short unstructured text such as social data, offers a powerful avenue for this exploration.

Furthermore, the existing works on text annotation focused on specific domains. Further studies are required to evaluate ChatGPT's ability to perform various tasks and test its reliability and efficiency as a text annotator on larger datasets. Additionally, there is a need to assess its ability to work on short, unstructured text in different languages.

Subsequently, we can summarize the challenges of utilizing ChatGPT for the NLP tasks as follows:

- **Limited Contextual Understanding:** ChatGPT might face difficulties in comprehending context-specific details, potentially resulting in misinterpretations.
- **Knowledge Gap:** the model might lack the specific knowledge needed for complex domains, leading to inaccurate or misleading understanding since its knowledge is based on pre-existing data.
- **Dependency on Input Quality:** feeding ChatGPT poor data quality can lead to inaccurate or misleading outcomes. ChatGPT's performance depends heavily on the training data it receives; thus, biased data can lead ChatGPT to stumble into inaccurate conclusions.
- **Static Knowledge Base:** ChatGPT may not integrate real-time information, hindering its ability to handle dynamic situations and rapidly evolving contexts.
- **Incapability to Handle Unstructured Data:** this model's performance can suffer when faced with unstructured formats.

- **Lack of Explainability:** ChatGPT's internal decision-making process is a "black box.". Thus, the lack of transparency raises doubts about reliability, particularly in sensitive domains like healthcare or finance.
- **Struggle with Non-English Language:** the lack of extensive training data in languages other than English contributes to this limitation, making it less suitable for seamless adaptation to a diverse range of linguistic contexts.
- **Limited Control over Outcomes:** The limited control over the information extracted by ChatGPT raises the challenge of ensuring only relevant and accurate data is retrieved.
- **Security and Privacy Concerns:** Uploading private sensitive data, such as patient data introduces potential privacy risk from unauthorized parties.

Addressing these challenges is essential for optimizing the application of ChatGPT in various contexts, including NLP problems. Even though ChatGPT excels in general conversation, domain-specific tasks often demand additional fine-tuning. Subsequently, without understanding its reasoning, relying on ChatGPT's conclusions can be risky.

In conclusion, this systematic review synthesizes key findings and identifies common trends, challenges and opportunities in leveraging ChatGPT for NLP tasks. The outlined insights serve as a foundation for future research, guiding the exploration of novel applications and methodologies that capitalize on the capabilities of ChatGPT in the dynamic landscape of natural language processing.

6 Conclusion and Future Work

This systematic review investigates the utilization of ChatGPT across four main NLP tasks, which are 1) information extraction and multiple subtasks under it, including NER, event extraction and detection, and POS tagging, 2) text classification, 3) sentiment analysis and emotion recognition, and 4) annotation. The review investigated its applications and exposed a dynamic landscape of opportunities and challenges.

ChatGPT's potential to improve various natural language processing tasks makes it a valuable tool for researchers, developers, and practitioners in the field of NLP. In the domain of information extraction, ChatGPT showed its adaptability in various subtasks, highlighting its ability to comprehend and extract significant details from diverse contexts. Subsequently, ChatGPT can improve sentiment analysis by providing insights into the emotions, attitudes, and opinions expressed in text. In the domain of text classification, the model provides a solution for categorizing and labeling textual content. Likewise, the adaptability of ChatGPT for annotation tasks emphasizes its ability to reduce manual effort in labeling tasks as well as speed up the annotation process. Additionally, even though the studies that used ChatGPT-4 are limited, we noticed that it introduces inherent improvements in language understanding. Similarly, prompt engineering serves as a complementary mechanism, allowing users to guide the model and improve accuracy.

While the potential is promising, challenges endure. The model's limitations in handling domain-specific language, potential biases, and the need for fine-tuning in certain applications underscore the importance of continued research to address these limitations. Following this, the reliability of ChatGPT needs to be tested using more extensive datasets and diverse data structures, including short-text samples from social media. Further investigations are necessary, particularly in the realm of text classification, since the works on this subject are very limited. Our observations indicate that results vary based on the classification topics, and there is a significant correlation wherein higher numbers of classes correspond to lower achieved results.

The future trends and directions for research include further investigation of fine-tuning strategies and optimization techniques to enhance ChatGPT's performance in specific NLP tasks; exploration of ChatGPT's applicability and effectiveness in domain-specific NLP tasks, such as financial or medical text analysis; examination of ChatGPT's cross-lingual and multilingual capabilities; development of standardized evaluation metrics and benchmarks for evaluating ChatGPT's performance across different NLP tasks, providing fair comparison with other models; and investigation of ethical considerations of ChatGPT, such as bias and privacy. Finally, the insights gathered from this review offer a roadmap for the researchers who are interested in improving NLP tasks. Inspire further collaboration, innovation, and discussion about the significant role of ChatGPT in shaping the landscape of natural language understanding and generation.

Acknowledgement: The author thanks anonymous reviewers and journal editors for their valuable comments, which significantly improved this paper.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: No new data were created during this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares that they have no conflicts of interest to report regarding the present study.

References

1. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems. 2023 Jul 1;11(7):1–37.
2. Skondras P, Zervas P, Tzimas G. Generating synthetic resume data with large language models for enhanced job description classification. Future Internet. 2023 Nov 1;15(11):1–12.
3. He X, Li S, Zhao G, Han X, Zhuang Q. Multi-task biomedical overlapping and nested information extraction model based on unified framework. In: CCF International Conference on Natural Language Processing and Chinese Computing; 2023; Cham: Springer Nature Switzerland.
4. Zhao J, Liu C, Liang J, Li Z, Xiao Y. A novel cascade instruction tuning method for biomedical NER. In: ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024; Seoul, Korea; IEEE. p. 11701–5.
5. Mohapatra H, Mishra SR. Unlocking insights: exploring data analytics and AI tool performance across industries. Stud Big Data. 2024;145:265–88.
6. OpenAI. Available from: https://openai.com/. [Accessed 2024].
7. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering; 2024 Mar 28. Available from: http://arxiv.org/abs/2303.16416. [Accessed 2024].
8. Javaid M, Haleem A, Singh RP. A study on ChatGPT for industry 4.0: background, potentials, challenges, and eventualities. J Econ Technol. 2023;1:127–43.
9. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci. 2023;13(4):1–15.
10. Panwar P, Gupta S. A review: exploring the role of ChatGPT in the diagnosis and treatment of oral pathologies. Oral Oncol Rep. 2024 Jun;10:100225.

11. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. In: Computer methods and programs in biomedicine. UK: Elsevier Ireland Ltd. 2024. vol. 245.
12. Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and ChatGPT: applications, challenges, and AI-human collaboration. J Inf Technol Case Appl Res. Routledge; 2023;25:277–304.
13. Filippo C, Vito G, Irene S, Simone B, Gualtiero F. Future applications of generative large language models: a data-driven case study on ChatGPT. Technovation. 2024 May 1;133:103002.
14. Sallam M. Practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11(6):887.
15. Pakhale K. Comprehensive overview of named entity recognition: models, domain-specific applications and challenges; 2023 Sep 25. Available from: http://arxiv.org/abs/2309.14084. [Accessed 2024].
16. Moscato V, Postiglione M, Sperlí G. Few-shot named entity recognition: definition, taxonomy and research directions. ACM Trans Intell Syst Technol. 2023 Oct 9;14(5):1–46.
17. Ollion É., Shen R, Macanovic A, Chatelain A. ChatGPT for text annotation? Mind the hype! 2023;1–10.
18. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI models: a preliminary review. Future Internet. 2023 Jun 1;15(6):192.
19. Zaremba A, Demir E. ChatGPT: unlocking the future of NLP in finance. Modern Financ. 2023 Nov 1;1(1):93–8.
20. Borji A, Mohammadian M. Battle of the wordsmiths: comparing ChatGPT, GPT-4, claude, and bard. SSRN Electron J. 2023.
21. Ding B, Qin C, Liu L, Chia YK, Li B, Joty S, et al. Is GPT-3 a good data annotator? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics; 2023; Toronto, America: Association for Computational Linguistics (ACL). p. 11173–95.
22. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Jul 17;29(8):1930–40.
23. Turan SC, Yildiz K, Büyüktanir B. Comparison of LDA, NMF and BERTopic topic modeling techniques on amazon product review dataset: a case study. Stud Comput Intell. 2024;1145:23–31.
24. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol. 2023 Oct 1;19(5):607.
25. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High-Confid Comput. 2014;4:100211. doi:10.1016/j.hcc.2024.100211.
26. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey; 2024 Feb 9. Available from: http://arxiv.org/abs/2402.06196. [Accessed 2024].
27. Bandi A, Adapa PVSR, Kuchi YEVPK. The power of generative AI: a review of requirements, models, input–output formats, evaluation metrics, and challenges. In: Future Internet. Basel, Switzerland: Multidisciplinary Digital Publishing Institute (MDPI); 2023. vol. 15.
28. Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT; 2023 Mar 7. Available from: http://arxiv.org/abs/2303.04226. [Accessed 2024].
29. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023.
30. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra Adam Roberts Paul Barham Hyung Won Chung G, Sutton Sebastian Gehrmann C, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res. 2023;24:1–13.
31. Courant R, Edberg M, Dufour N, Kalogeiton V. Transformers and visual transformers. In: Neuromethods. New York City: Humana Press Inc.; 2023. p. 193–229.

32. Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:30.
33. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. In: Engineering. Amsterdam, The Netherlands: Elsevier Ltd.; 2023. vol. 25, p. 51–65.
34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies-Proceedings of the Conference; 2019. p. 4171–86.
35. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019.
36. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach; 2019 Jul 26. Available from: http://arxiv.org/abs/1907.11692. [Accessed 2024].
37. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations; 2019 Sep 26. Available from: http://arxiv.org/abs/1909.11942. [Accessed 2024].
38. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention; 2020 Jun 5. Available from: http://arxiv.org/abs/2006.03654. [Accessed 2024].
39. Yang Y, Christopher M, Uy S, Huang A. FinBERT: a pretrained language model for financial communications; 2020. Available from: https://arxiv.org/abs/2006.08097v2. [Accessed 2024].
40. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: publicly available pretrained language models for mental healthcare. In: 2022 Language Resources and Evaluation Conference; 2022;LREC, Marseille, France.
41. Souza F, Nogueira R, Lotufo R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics); New York City: Springer International Publishing; 2020.
42. Webersinke N, Kraus M, Bingler JA, Leippold M. ClimateBert: a pretrained language model for climate-related text. SSRN Electron J. 2021.
43. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
44. KeyBERT. KeyBERT. Available from: https://maartengr.github.io/KeyBERT/api/keybert.html. [Accessed 2024].
45. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21:1–67.
46. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics; 2020; p. 7871–80.
47. Teräsvirta T, Yang Y. Specification, estimation and evaluation of vector smooth transition autoregressive models with applications. 2014. doi:10.1002/9780470996430.ch21.
48. Kalla D, Kuraku S. Study and analysis of chat GPT and its impact on different fields of study. Int J Innov Sci Res Technol. 2023;8:827–33.
49. Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat Lang Process J. 2024 Mar;6:100048.
50. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Int Things Cyb-Phys Syst. 2023 Jan 1;3:121–54.
51. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44.

52. Openai AR, Openai KN, Openai TS, Openai IS. Improving language understanding by generative pre-training; 2018.
53. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. In: Minds and machines. USA: Springer Science and Business Media B.V.; 2020. vol. 30, p. 681–94.
54. Open AI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report; 2023 Mar 15. Available from: http://arxiv.org/abs/2303.08774. [Accessed 2024].
55. Kozel G, Gurses ME, Geçici NN, Gökalp E, Bahadir S, Merenzon MA, et al. Chat-GPT on brain tumors: an examination of artificial intelligence/machine learning's ability to provide diagnoses and treatment plans for example neuro-oncology cases. Clin Neurol Neurosurg. 2024 Apr 1;239:108238.
56. Ma Y. The potential application of ChatGPT in gastrointestinal pathology. Gastroenterol Endosc. 2023 Jul 1;1(3):130–1.
57. Cloesmeijer M, Janssen A, Koopman S, Cnossen M, Mathot R. ChatGPT in pharmacometrics? Potential opportunities and limitations; 2023. Available from: https://doi.org/10.22541/au.168235933.39569649/v1. [Accessed 2024].
58. Sorin V, Glicksberg BS, Barash Y, Konen E, Nadkarni G, Klang E. Applications of large language models (LLMs) in breast cancer care. medRxiv. 2023. doi:10.1101/2023.11.04.23298081.
59. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):1–5.
60. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks, Stand Evaluat. 2023;3(1):100105.
61. Yu H. The application and challenges of ChatGPT in educational transformation: new demands for teachers' roles. In: Heliyon. Netherlands: Elsevier Ltd.; 2024. vol. 10.
62. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min. 2023;16(1):1–11.
63. Baidoo-Anu D, Leticia Owusu A. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. J AI. 2023;7(1):52–62.
64. Fuchs K. Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? In: Frontiers in education. Lausanne, Switzerland: Radiological Society of North America Inc.; 2023. vol. 8.
65. Al Ghazali S, Zaki N, Ali L, Harous S. Exploring the potential of ChatGPT as a substitute teacher: a case study. Int J Inf Educ Technol. 2024;14(2):271–8.
66. Yilmaz R, Karaoglan Yilmaz FG. Augmented intelligence in programming learning: examining student views on the use of ChatGPT for programming learning. Comput Hum Behav: Art Hum. 2023;1(2):100005.
67. Agarwal M, Raju V, Nanda I. Descriptive research on AI-based tools to aid personalized customer service: case of ChatGPT. J Reprod Res. 2023;1(1):140–6.
68. Rivas P, Zhao L. Marketing with ChatGPT: navigating the ethical terrain of GPT-based chatbot technology. AI. 2023;4(2):375–84.
69. Raj R, Singh A, Kumar V, Verma P. Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations. Bench Council Trans Benchmarks, Stand Evaluat. 2023 Sep 1;3(3):100140.
70. Chuma EL, De Oliveira GG. Generative AI for business decision-making: a case of ChatGPT. Manag Sci Business Decis. 2023 Jul 1;3(1):5–11.
71. Jarco D, Sulkowski L. Is ChatGPT better at business consulting than an experienced human analyst? An experimental comparison of solutions to a strategic business problem. Forum Scientiae Oeconomia. 2023 Jun 30;11(2):87–109.
72. Khan MS, Umer H. ChatGPT in finance: applications, challenges, and solutions. Heliyon. 2024 Jan 30;10(2):1–8.

73. Dwivedi YK, Pandey N, Area M, Currie W, Micu A. Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges, and research agenda. Int J Contemp Hosp Manag. 2024;36(1):1–12.
74. Guo D, Chen H, Wu R, Wang Y. AIGC challenges and opportunities related to public safety: a case study of ChatGPT. J Safety Sci Resilien. 2023 Dec 1;4(4):329–39.
75. Biswas S. Prospective role of chat GPT in the military: according to ChatGPT; 2023. doi:10.32388/8WYYOD.
76. Chowdhury MNUR, Haque A. ChatGPT: its applications and limitations. In: 2023 3rd International Conference on Intelligent Technologies, CONIT 2023; 2023; Hubli, India.
77. Islam I, Islam MN, Muhammad I, Islam N. Opportunities and challenges of ChatGPT in academia: a conceptual analysis; 2023. doi:10.22541/au.167712329.97543109/v1.
78. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information. 2023 Aug 16;14(8):462. doi:10.3390/info14080462.
79. PRISMA. Available from: http://www.prisma-statement.org/. [Accessed 2024].
80. Publish or perish. Available from: https://harzing.com/resources/publish-or-perish. [Accessed 2023].
81. Yang C, Zhang P, Qiao W, Gao H, Zhao J. Rumor detection on social media with crowd intelligence and chatgpt-assisted networks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec; Singapore. p. 5705–17.
82. Feng P, Wu H, Yang Z, Wang Y, Ouyang D. Leveraging prompt and Top-K predictions with ChatGPT data augmentation for improved relation extraction. Appl Sci. 2023 Nov 28;13(23):12746. doi:10.3390/app132312746.
83. Bhavya B, Isaza PT, Deng Y, Nidd M, Azad AP, Shwartz L, et al. Exploring large language models for low-resource IT information extraction. In: IEEE International Conference on Data Mining Workshops, ICDMW; 2023; Shanghai, China. p. 1203–12.
84. Shushkevich E, Alexandrov M, Cardiff J. Improving multiclass classification of fake news using BERT-based models and ChatGPT-augmented data. Inventions. 2023 Oct 1;8(5):112.
85. Skondras P, Psaroudakis G, Zervas P, Tzimas G. Efficient resume classification through rapid dataset creation using ChatGPT. In: 14th International Conference on Information, Intelligence, Systems and Applications, IISA 2023; 2023; Volos, Greece.
86. Li B, Ju J, Wang C, Pan S. How does ChatGPT affect fake news detection systems? In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). New York City: Springer, Midtown Manhattan; 2023. vol. 14177, p. 565–80.
87. González-Gallardo CE, Boros E, Girdhar N, Hamdi A, Moreno JG, Doucet A, et al. Yes but.. can ChatGPT identify entities in historical documents?. In: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL); 2023 Jun 26; Santa Fe, NM, USA; IEEE. p. 184–9.
88. Tan Z, Huang S, Jia Z, Cai J, Li Y, Lu W, et al. DAMO-NLP at semeval-2023 task 2: a unified retrieval-augmented system for multilingual named entity recognition. arXiv preprint arXiv:2305.03688. 2023 May 5.
89. Chanthran MR, Soon LK, Ong HF, Selvaretnam B. How well ChatGPT understand Malaysian English? An evaluation on named entity recognition and relation extraction; 2023 Nov 20, Available from: http://arxiv.org/abs/2311.11583. [Accessed 2024].
90. Xie T, Li Q, Zhang J, Zhang Y, Liu Z, Wang H. Empirical study of zero-shot NER with ChatGPT; 2023 Oct 15. Available from: http://arxiv.org/abs/2310.10035. [Accessed 2024].
91. Zhou W, Zhang S, Gu Y, Chen M, Poon H. UniversalNER: targeted distillation from large language models for open named entity recognition. arXiv preprint arXiv:2308.03279. 2024.

92. Li J, Li H, Pan Z, Sun D, Wang J, Zhang W, et al. Prompting ChatGPT in MNER: enhanced multimodal named entity recognition with auxiliary refined knowledge; 2023 May 20. Available from: http://arxiv.org/abs/2305.12212. [Accessed 2024].
93. Li J, Li H, Sun D, Wang J, Zhang W, Wang Z, et al. LLMs as bridges: reformulating grounded multimodal named entity recognition. arXiv preprint arXiv:2402.09989. 2024.
94. Hu Y, Ameer I, Zuo X, Peng X, Zhou Y, Li Z, et al. Zero-shot clinical entity recognition using ChatGPT. arXiv preprint arXiv:230316416. 2023 Nov.
95. Tang R, Han X, Jiang X, Hu X. Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360. 2023 Mar 8.
96. Brinkmann A, Shraga R, Der RC, Bizer C. Product information extraction using ChatGPT. arXiv preprint arXiv:2306.14921. 2023 Jun 23.
97. Yuan C, Xie Q, Ananiadou S. Zero-shot temporal relation extraction with ChatGPT. arXiv preprint arXiv:2304.05454. 2023 Apr 11.
98. Jethani N, Jones S, Genes N, Major VJ, Jaffe IS, Cardillo AB, et al. Evaluating ChatGPT in information extraction: a case study of extracting cognitive exam dates and scores. medRxiv. 2023.
99. Peikos G, Symeonidis S, Kasela P, Pasi G. Utilizing ChatGPT to enhance clinical trial enrollment. arXiv preprint arXiv:2306.02077. 2023.
100. He J, Wang L, Hu Y, Liu N, Liu H, Xu X, et al. ICL-D3IE: in-context learning with diverse demonstrations updating for document information extraction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023; Paris, France. p. 19485–94.
101. Kartchner D, Al-Hussaini I, Kronick O, Ramalingam S, Mitchell C. Zero-shot information extraction for clinical meta-analysis using large language models. In: The 22nd Workshop Biomed Natural Lang Process BioNLP Shared Tasks; 2023; Toronto, Canada. p. 396–405.
102. Sousa H, Guimarães N, Jorge A, Campos R. GPT struct me: probing GPT models on narrative entity extraction. 2023 Nov 24. Available from: http://arxiv.org/abs/2311.14583. [Accessed 2024]
103. Nishio M, Fujimoto K, Rinaldi F, Matsuo H, Rohanian M, Krauthammer M, et al. Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC. In: NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies 2023; NTCIR, Tokyo, Japan.
104. Gao J, Zhao H, Yu C, Xu R. Exploring the feasibility of ChatGPT for event extraction. arXiv preprint arXiv:230303836. 2023.
105. Han R, Peng T, Yang C, Wang B, Liu L, Wan X. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. arXiv preprint arXiv:2305.14450. 2023.
106. Li B, Fang G, Yang Y, Wang Q, Ye W, Zhao W, et al. Evaluating ChatGPT's information extraction capabilities: an assessment of performance, explainability, calibration, and faithfulness. arXiv preprint arXiv:2304.11633. 2023.
107. Wei X, Cui X, Cheng N, Wang X, Zhang X, Huang S, et al. Zero-shot information extraction via chatting with ChatGPT. arXiv preprint arXiv:2302.10205. 2023.
108. Lai VD, Ngo NT, Ben Veyseh AP, Man H, Dernoncourt F, Bui T, et al. ChatGPT beyond english: towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613. 2023.
109. Reiss MV. Testing the reliability of ChatGPT for text annotation and classification: a cautionary remark. arXiv preprint arXiv:2304.11085. 2023.
110. Zhao B, Jin W, Del Ser J, Yang G. ChatAgri: exploring potentials of ChatGPT on cross-linguistic agricultural text classification. Neurocomputing. 2023;557:126708.

111. Loukas L, Stogiannidis I, Malakasiotis P, Vassos S, Ai H. Breaking the bank with ChatGPT: few-shot text classification for finance. In: Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting; 2023; Macao, China. p. 75–80.
112. Lamichhane B. Evaluation of ChatGPT for NLP-based mental health applications. arXiv preprint arXiv:2303.15727. 2023.
113. Trajanov D, Lazarev G, Chitkushev L, Vodenska I. Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks. In: E3S Web of Conferences; 2023; Athens, Greece; EDP Sciences.
114. Oliveira AS, Cecote TC, Silva PHL, Gertrudes JC, Freitas VLS, Luz EJS. How good is ChatGPT for detecting hate speech in Portuguese? In: Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana SBC 2023; 2023.
115. Yang K, Ji S, Zhang T, Xie Q, Ananiadou S. On the evaluations of ChatGPT and emotion-enhanced prompting for mental health analysis; 2023. doi:10.48550/arXiv.2304.03347.
116. Zhong Q, Ding L, Liu J, Du B, Tao D. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT; 2023. doi:10.48550/arXiv.2302.10198.
117. Golubev A. RuSentNE-2023: evaluating entity-oriented sentiment analysis on Russian news texts anton. arXiv preprint arXiv:230517679. 2023.
118. Wang J, Liang Y, Meng F, Shi H, Li Z, Xu J, et al. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.043392023. 2023.
119. Fatouros G, Soldatos J, Kouroumali K, Makridis G, Kyriazis D. Transforming sentiment analysis in the financial domain with ChatGPT. Mach Learn Appl. 2023;14:100508. doi:10.1016/j.mlwa.2023.100508.
120. Lopez-Lira A, Tang Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. SSRN Electron J. 2023.
121. Zhang B, Ding D, Jing L. How would stance detection techniques evolve after the launch of ChatGPT? arXiv preprint arXiv:221214548. 2023.
122. Zhang B, Fu X, Ding D, Huang H, Li Y, Jing L. Investigating chain-of-thought with ChatGPT for stance detection on social media; 2023. doi:10.48550/arXiv.2304.03087.
123. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: jack of all trades, master of none. Inf Fusion. 2023;99:101861.
124. Khondaker MTI, Waheed A, Nagoudi EMB, Abdul-Mageed M. GPTAraEval: a comprehensive evaluation of ChatGPT on Arabic NLP; 2023 May 24. Available from: http://arxiv.org/abs/2305.14976. [Accessed 2024].
125. Kuzman T, Mozetič I, Ljubešić N. ChatGPT: beginning of an end of manual linguistic data annotation? Use case of automatic genre identification; 2023. doi:10.48550/arXiv.2303.03953.
126. Chen J, Chen P, Wu X. Generating Chinese event extraction method based on ChatGPT and prompt learning. Appl Sci. 2023;13(17):9500.
127. Korini K, Bizer C. Column type annotation using ChatGPT. arXiv preprint arXiv:2306.00745. 2023.
128. Belal M, She J, Wong S. Leveraging ChatGPT as text annotation tool for sentiment analysis. arXiv preprint arXiv:230617177. 2023.
129. Koptyra B, Ngo A, Radliński Ł., Kocoń J. CLARIN-Emo: training emotion recognition models using human annotation and ChatGPT. In: International Conference on Computational Science; 2023; Switzerland, Cham; Springer Nature. p. 365–79.

130. Huang F, Kwak H, An J. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In: Companion Proceedings of the ACM Web Conference 2023 (WWW'23 Companion); 2023 April 30–May 4. Austin, TX, USA; Association for Computing Machinery.
131. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd-workers for text-annotation tasks. Proc Natl Acad Sci. 2023;120(30):e2305016120.
132. Alizadeh M, Kubli M, Samei Z, Dehghani S, Bermeo JD, Korobeynikova M, et al. Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. arXiv preprint arXiv:2307.02179. 2023.
133. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned language models are zero-shot learners; 2021 Sep 3. Available from: http://arxiv.org/abs/2109.01652. [Accessed 2024].
134. Li X, Zhu X, Ma Z, Liu X, Shah S. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? An examination on several typical tasks; 2023. doi:10.48550/arXiv.2305.05862.
135. Qin C, Zhang A, Zhang Z, Chen J, Yasunaga M, Yang D. Is ChatGPT a general-purpose natural language processing task solver? arXiv preprint arXiv:230206476. 2023.
136. Sun X, Dong L, Li X, Wan Z, Wang S, Zhang T, et al. Pushing the limits of ChatGPT on NLP tasks. arXiv preprint arXiv:2306.09719. 2023.
137. ace-2005. Available from: https://paperswithcode.com/dataset/ace-2005. [Accessed 2024].
138. ace-2004. Available from: https://paperswithcode.com/dataset/ace-2004. [Accessed 2024].
139. CONLL2003. Available from: https://huggingface.co/datasets/conll200. [Accessed 2024].
140. OntoNotes-501. Available from: https://paperswithcode.com/dataset/ontonotes-5-0. [Accessed 2024 Jan 5].