A Survey of Information Extraction and Large Language Models in Natural Language Processing

www.surveyx.cn

Abstract

This survey paper explores the intricate interconnections among information extraction, large language models (LLMs), natural language processing (NLP), generative AI, and text mining, highlighting their collective impact on automated language processing. LLMs, exemplified by models like GPT-4, have significantly advanced domain-specific information extraction, demonstrating high accuracy in specialized tasks such as extracting band gap information. The integration of advanced prompt engineering and parameter-efficient adaptations in generating synthetic data underscores the utility of LLMs in low-resource settings, enhancing model performance across diverse applications. Moreover, modular architectures like GATE facilitate the seamless incorporation of NLP tools, boosting research productivity. In healthcare, LLMs show potential in real-time diagnostics, though further research is needed to ensure reliable outputs. Despite these advancements, challenges persist, particularly concerning bias, reasoning, and factuality in LLM outputs, necessitating ongoing research into user-centered design and bias mitigation strategies. The survey highlights the dual perception of LLMs as both beneficial tools and sources of bias, emphasizing the need for robust evaluation methods and universal mitigation techniques. The interconnectedness of these technologies plays a pivotal role in advancing automated language processing, enhancing the ability to process and understand human language, and driving innovation across various domains. The survey concludes with the potential of LLMs to enhance machine translation systems for low-resource languages in crisis situations, reflecting their broader impact in fostering innovation and addressing complex challenges in automated language processing.

1 Introduction

1.1 Significance of Automated Language Processing

Automated language processing plays a pivotal role in modern technology, significantly improving performance across various information extraction subtasks using large language models (LLMs) [1]. In the electric energy sector, LLMs bridge knowledge gaps, enhancing operational efficiency and decision-making processes [2]. The automation of pragma-discursive corpus annotation, traditionally a manual and error-prone task, highlights the transformative potential of these technologies [3].

Neural-guided rule generation exemplifies the adaptability of these technologies, utilizing minimal examples to enhance interpretability and flexibility [4]. Additionally, comprehensive frameworks for standardizing AI concepts and methodologies provide the necessary terminological and methodological clarity to navigate the rapidly evolving AI landscape [5].

In healthcare, LLMs are set to revolutionize Electronic Health Records (EHR) research, improve clinical documentation, and enhance patient care through more accurate data analysis [6]. The versatility of LLMs, illustrated by models like ChatGPT, underscores their significant impact on AI research and development [7].

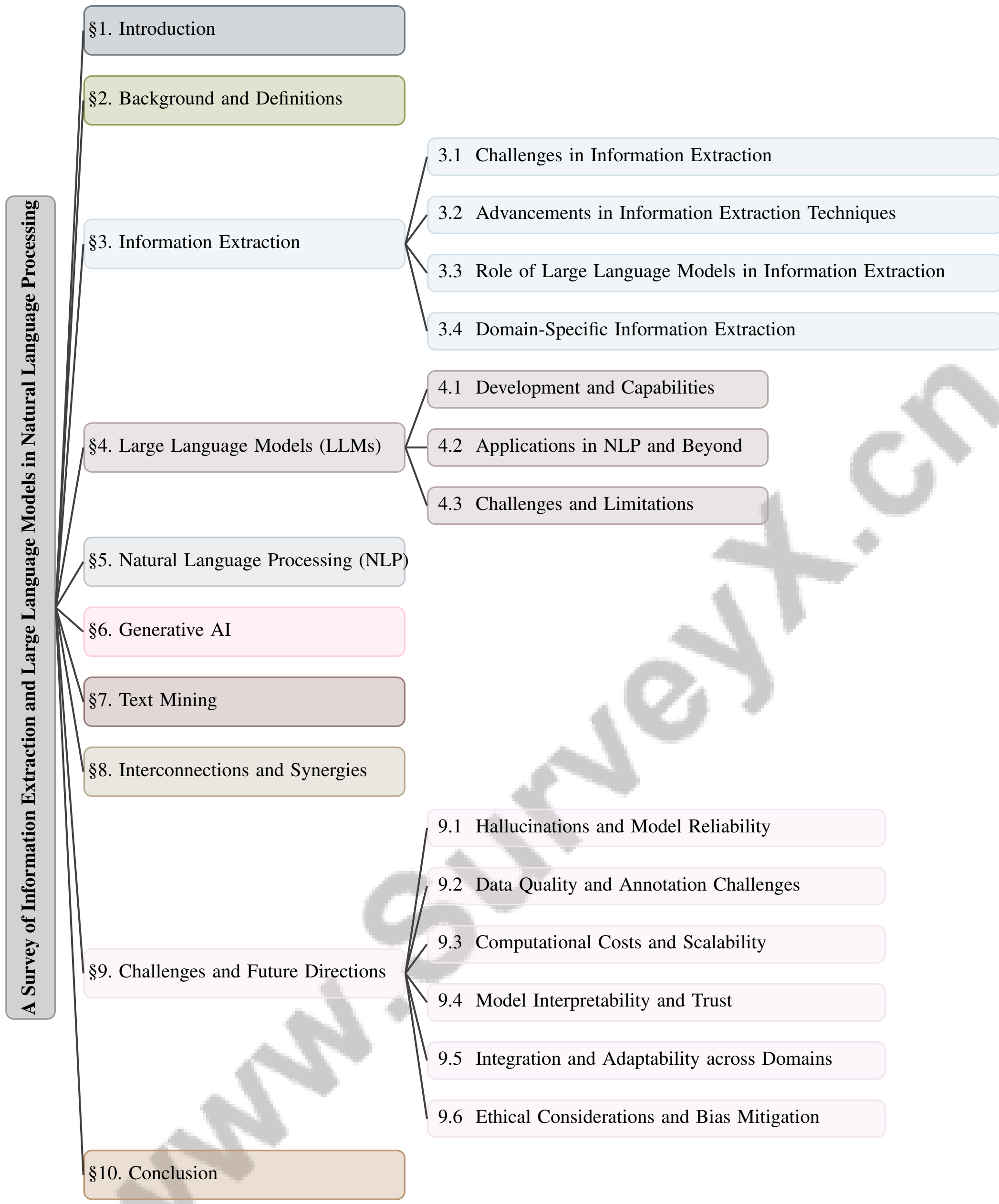


Figure 1: chapter structure

The systematic conversion of qualitative insights into quantifiable features is essential for predictive analytics, demonstrating how automated language processing translates expert intuition into actionable data [8]. Their application in music knowledge discovery further showcases the capability of these technologies to surpass traditional text search limitations [9].

The evaluation of chatbot applications, currently lacking consensus on appropriate criteria and metrics, emphasizes the need for developers to employ suitable evaluation methods to ensure the reliability and efficacy of automated language processing systems [10]. Comparative analyses of transparency approaches in LLMs reveal the necessity for context-specific solutions, as no single method is universally applicable [11].

The transformative impact of automated language processing, particularly through LLMs, is evident across various domains, including biodiversity research and academic integrity, where it enhances methodologies by automating information retrieval, improving data extraction precision, and facilitating advanced analyses. This technology streamlines extensive textual data processing and elevates AI capabilities, enabling complex tasks such as text annotation, classification, and plagiarism detection with greater accuracy and efficiency, ultimately leading to more informed decision-making and improved outcomes in critical applications [12, 13, 14].

1.2 Structure of the Survey

This survey is systematically organized to explore the interconnected fields of information extraction, large language models (LLMs), natural language processing (NLP), generative AI, and text mining. It begins with an introduction that establishes the significance of automated language processing technologies and their transformative impact across various industries, as detailed in the preceding section. Section 2 delves into background and definitions, providing precise explanations of core concepts and elucidating their interrelations.

Section 3 offers an in-depth exploration of information extraction (IE), detailing methodologies and techniques for identifying and retrieving specific data from unstructured text. It emphasizes traditional rule-based methods alongside contemporary machine learning (ML) and deep learning (DL) techniques, highlighting key sub-tasks such as Named Entity Recognition (NER) and Relation Extraction (RE), and their applications in diverse domains, including legal and medical texts. This section addresses challenges posed by varying document genres and lengths, revealing insights into the comparative effectiveness of heuristic-based versus data-driven methods in enhancing extraction accuracy. It underscores the critical role of IE in transforming unstructured data into structured formats for further NLP tasks [15, 16, 17], while also exploring recent advancements and the role of LLMs in enhancing IE processes.

Section 4 examines LLMs, highlighting their development, capabilities, and applications in NLP, along with the limitations and challenges they encounter.

In Section 5, the survey provides a comprehensive overview of various NLP techniques and their practical applications in analyzing and interpreting natural language. It discusses the integration of LLMs, such as GPT-3.5-turbo and GPT-4, which significantly enhance information extraction capabilities, particularly in low-resource scenarios and complex tasks like automated literature reviews and legal text analysis. This section underscores the effectiveness of NLP strategies, including retrieval-augmented generation and prompt-based methods, showcasing their impact on performance and robustness in extracting structured information from unstructured text [18, 19, 20, 21, 22].

Section 6 explores generative AI systems, focusing on their content creation capabilities and implications in NLP, complementing information extraction and LLMs.

Section 7 discusses text mining, emphasizing its role in extracting meaningful information and patterns from large volumes of textual data, and highlights the synergy between text mining, information extraction, and LLMs. Section 8 analyzes the interconnections among information extraction, LLMs, NLP, generative AI, and text mining, discussing how they collectively advance the automated understanding of human language.

The survey concludes with Section 9, identifying current challenges such as data quality, model interpretability, and computational costs, and discusses potential future directions and research opportunities to address these challenges. Finally, Section 10 summarizes the key points discussed in the paper, reflecting on the importance of the interconnectedness of these fields and their impact on the future of automated language processing.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

The domains of information extraction (IE), large language models (LLMs), natural language processing (NLP), generative AI, and text mining form the backbone of automated language processing technologies. IE is pivotal in converting unstructured text into structured data, facilitating machine processing and knowledge discovery [1]. This transformation is vital for NLP, which utilizes various methodologies to extract insights from specialized scientific texts, such as those in music and energetics.

LLMs, designed to comprehend and generate human-like text, enhance performance across NLP tasks by producing contextually relevant content and overcoming traditional monolingual evaluation limitations, especially in multilingual contexts. Their integration in decision-making processes, notably in complex sectors like the electric energy industry, underscores their role in augmenting data processing and decision-making [2].

NLP employs a vast array of techniques for natural language analysis and interpretation, evolving through advancements like prompt engineering, which optimizes LLM performance without extensive retraining [18, 23]. The challenge lies in enabling computers to understand and generate human language, requiring intricate layers of linguistic processing [18].

Generative AI systems automate tasks such as news generation and fact-checking, addressing digital misinformation challenges [24]. These systems also simplify complex texts into lay summaries, enhancing accessibility in specialized domains.

Text mining is crucial for extracting meaningful information and patterns from large textual datasets, aiding in data benchmarking and summarization. Its synergy with IE and LLMs is vital for processing and understanding human language. User perceptions of biases in LLMs highlight the need for careful application of these technologies [25].

The integration of these fields propels automated language processing, enabling efficient management of large textual data and innovative solutions to linguistic challenges. Developing standardized ontologies to represent AI concepts and methodologies accurately is essential for clarity in this rapidly evolving landscape [5]. This integration ensures that valuable knowledge is not lost amid the vast scientific literature [26].

3 Information Extraction

Category	Feature	Method
Challenges in Information Extraction	Prompt-Based Enhancement	EMRE2llm[27]
Advancements in Information Extraction Techniques	Integrated Extraction	SLM[28], T5-IE[29], LLM-IE[30]
	Causal Understanding	CRL-GAI[31]
	Efficiency and Scalability	EYEGLAXS[32]
Role of Large Language Models in Information Extraction	Quality and Insight Enhancement	EIE-LLM[8], LUNA[33]
	Data Augmentation Techniques	LLM2LLM[34]
	Prompt and Process Strategies	GAMedX[35], FINDER[36]
	Graph and Semantic Integration	MLOE[37], KG-LLM[38]
Domain-Specific Information Extraction	Neural Network Architectures	A2I[39], Seq2Seq[40]
	Graph-Based Techniques	DWIE[41]

Table 1: This table provides a comprehensive overview of the current landscape in information extraction, categorizing advancements and challenges across various domains. It highlights the role of large language models (LLMs) in enhancing information extraction processes, detailing specific methods and features employed to address domain-specific complexities and improve data handling efficiency. The table serves as an essential reference for understanding the methodologies and innovations driving progress in this field.

In the realm of information extraction, the challenges presented by diverse and complex datasets necessitate a nuanced understanding of the specific obstacles faced across various domains. The following subsection delves into the distinct challenges encountered in information extraction, particularly focusing on the intricacies of domain-specific data and the implications for the extraction process. Table 1 presents a detailed classification of challenges, advancements, and the role of large language models in information extraction, offering insights into the methodologies employed to address domain-specific complexities. Table 2 offers a detailed classification of the challenges, advancements, and the role of large language models in information extraction, providing valuable insights into the methodologies addressing domain-specific complexities. As illustrated in **??**, this figure depicts the hierarchical structure of challenges, advancements, and domain-specific applications in information extraction, highlighting the role of large language models and methodological innovations in addressing complex data landscapes. By examining these challenges, we can better appreciate the evolving landscape of information extraction and the need for tailored methodologies that address the unique requirements of different fields.

3.1 Challenges in Information Extraction

Information extraction (IE) faces numerous challenges, particularly in handling the complexity and variability of domain-specific data, which often lacks sufficient annotated datasets for effective model training [16]. This scarcity of data is further compounded by the uneven distribution of entity pair classes, which complicates the extraction processes [27]. The high dimensionality and class imbalance present in documents, such as those related to public affairs, pose significant hurdles in accurately classifying texts with overlapping topics [42]. Additionally, the presence of unknown confounding features that correlate with treatment features leads to biased causal estimates, presenting another layer of complexity in information retrieval [31].

In specialized fields like medicine and real estate, the complexity of language and domain-specific nuances present formidable challenges. For instance, the legal ambiguity in real estate contracts and the complex medical terminologies in healthcare complicate the extraction processes. Moreover, the phenomenon of hallucination in large language models (LLMs) can lead to inaccuracies in keyword extraction, highlighting the limitations of existing models [43]. The challenges are further exacerbated by the lack of empirical research on LLM performance across diverse domains, which limits their applicability in complex fields [44].

Another significant challenge is the transformation of digital text collections into rich knowledge environments that capture the semantics of the content, which is particularly difficult in fields such as music [9]. Existing benchmarks often fail to support multi-turn dialogs over meeting content effectively, and current automated dialog generation methods lack the necessary quality in attribution [45].

These challenges are visually summarized in Figure 2, which illustrates the primary challenges in Information Extraction (IE), categorizing them into data scarcity and imbalance, domain-specific complexities, and model limitations. Each category encompasses specific issues such as insufficient datasets, legal ambiguities, and hallucination in large language models, highlighting the multifaceted nature of challenges in IE. These challenges underscore the need for innovative approaches that enhance the accuracy, efficiency, and interpretability of information extraction processes across diverse and complex data landscapes.

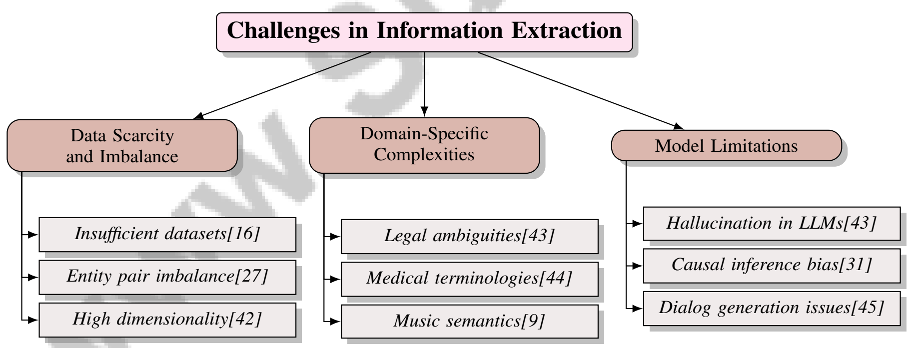


Figure 2: This figure illustrates the primary challenges in Information Extraction (IE), categorizing them into data scarcity and imbalance, domain-specific complexities, and model limitations. Each category encompasses specific issues such as insufficient datasets, legal ambiguities, and hallucination in large language models, highlighting the multifaceted nature of challenges in IE.

3.2 Advancements in Information Extraction Techniques

Recent advancements in information extraction (IE) techniques have significantly enhanced the processing of unstructured data by leveraging the synergy between large language models (LLMs) and innovative methodologies. The current landscape of IE is organized into two primary categories: Knowledge Engineering (KE) methods and Machine Learning (ML) methods, with ML further divided into supervised, semi-supervised, and unsupervised techniques [16]. This structured approach facilitates the systematic improvement of IE processes across diverse applications.

A notable innovation in the field is the end-to-end approach of relation extraction using LLMs, which contrasts with traditional multi-step methods that often struggle to maintain accuracy [38]. This

AI-generated, for reference only.

approach streamlines the extraction process by integrating multiple stages into a cohesive model, thereby improving both efficiency and accuracy.

The development of benchmarks such as those introduced by Swarup et al. provides a comprehensive analysis of multiple relation extraction algorithms across diverse datasets, emphasizing the importance of data characteristics in enhancing algorithmic performance [46]. This focus on data-driven insights is crucial for tailoring IE techniques to specific domain requirements.

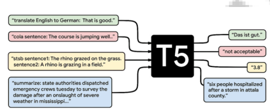
Innovative methodologies such as EYEGLAXS utilize Parameter-Efficient Fine-Tuning (PEFT) techniques, including LoRA, combined with advanced attention mechanisms like Flash Attention. These techniques enable the efficient processing of long sequences while maintaining high performance, thereby addressing the challenges associated with large-scale data [32].

The integration of document retrieval with the generative capabilities of LLMs marks a substantial improvement in the relevance and quality of extracted information. This is particularly evident in domains like real estate, where structured information extraction from contracts is automated to enhance speed and precision [30]. Such integration exemplifies the potential of LLMs to streamline complex information extraction processes.

Moreover, the introduction of LLM2LLM, a targeted iterative data augmentation strategy, enhances small seed datasets by using a teacher LLM to generate synthetic data based on the model's incorrect predictions. This method significantly improves the robustness and adaptability of IE systems [34].

The primary innovation in causal representation learning methods, as highlighted by Imai, lies in eliminating the need to learn causal representation from the data itself, thereby improving the accuracy and efficiency of causal estimates compared to existing methods [31].

These advancements in information extraction techniques, including the utilization of machine learning and deep learning methods, significantly enhance the capability to process and analyze unstructured textual data across various domains, such as legal, medical, and biodiversity literature. By integrating heuristic-based and data-driven approaches, as well as leveraging large language models, these techniques enable more precise identification of structured information, improving accuracy in tasks like named entity recognition and semantic role labeling. Furthermore, the exploration of human-machine collaboration frameworks has demonstrated that combining human insight with automated processes can yield high precision in information extraction, making these advancements crucial for effective knowledge discovery in diverse applications. [12, 14, 16, 17]. The continuous development and integration of these methodologies are essential for addressing the complex challenges inherent in extracting valuable insights from unstructured data.



(a) Translation and summarization tasks with T5 model[29]

Training Stage	It	Perplexity Training Set	Test set
2 (matched)	0	9.27	34.81
2 (matched)	1	5.81	31.25
2 (matched)	2	5.51	31.41
4 (L-matched)	0	4.71	24.39
4 (L-matched)	1	4.61	24.73
4 (L-matched)	2	4.56	24.88

(b) The table shows the perplexity of a model trained on different stages and training iterations.[28]

Documents	Total_Sentences	Total_Tokens
Long_generic (HP)	6480	77290
Short_generic (NYtimes)	54	1121
Long_domain (Neurology)	13719	235093
Short_domain (Brain Inflammation)	712	8464

nyt: new york restaurant review[27]
hp: (Book I) Harry Potter[20]
short_brain: (publication) Brain Inflammation, Degeneration, and Plasticity in Multiple Sclerosis[7]
long_brain: (textbook) Clinical Neurology 8th Edition[23]

TABLE I
DOCUMENTS STATISTICS

(c) Documents Statistics[17]

Figure 3: Examples of Advancements in Information Extraction Techniques

As shown in Figure 3, The example on "Information Extraction; Advancements in Information Extraction Techniques" showcases three distinct facets of modern information extraction methodologies through illustrative figures. The first figure highlights the application of the T5 model in translation and summarization tasks, demonstrating its ability to process and respond to text samples in multiple languages and formats, such as English, German, and Spanish. This reflects the T5 model's versatility in handling diverse linguistic inputs and generating coherent outputs. The second figure presents a table that details the perplexity of a model across various training stages and iterations, providing insight into the model's learning efficiency and accuracy during the training process. The perplexity values, associated with different training stages labeled as "2 (matched)" and "4 (L-matched)," underscore the model's performance improvements. Lastly, the third figure offers a statistical overview of documents, categorizing them by type and detailing metrics such as total sentences and tokens. This statistical analysis aids in understanding the complexity and structure of different document types, ranging from generic to domain-specific texts. Together, these examples illustrate significant advancements in information extraction techniques, emphasizing improvements in model adaptability, training efficiency, and document analysis. [**?**]royesh2024informationextractionapplicationdomain,chelba2001informationextractionusingstructured,yuan2023informationextra

3.3 Role of Large Language Models in Information Extraction

Large Language Models (LLMs) have significantly advanced the field of information extraction (IE) by leveraging their sophisticated architectures to process and organize unstructured data into structured formats. These models, including those like GPT-3, have demonstrated proficiency in tasks such as named entity recognition (NER) and relation extraction, effectively transforming complex textual data into organized structures [35]. Despite their capabilities, challenges such as inaccuracies and hallucinations persist, particularly in domains with high stakes like finance [44].

Innovative methodologies have been developed to enhance the performance of LLMs in information extraction. For instance, embedding-based retrieval (EBR) systems have been integrated with LLMs to automate the extraction of entities and attributes from agricultural documents, converting them into structured data [36]. This integration highlights the capacity of LLMs to streamline information extraction across various domains.

The development of benchmarks designed to evaluate the few-shot information extraction capabilities of LLMs, especially in comparison to smaller language models (SLMs) across multiple datasets and tasks, has provided valuable insights into their proficiency and adaptability [47]. These benchmarks are crucial for understanding the strengths and limitations of LLMs in diverse linguistic environments.

Frameworks like LUNA, a model-based universal analysis framework, have facilitated quality analysis of LLMs by incorporating abstract model construction and semantics binding [33]. This approach enhances the understanding of LLMs' capabilities in semantic parsing and structured logical forms [48].

The use of multi-LLM orchestration engines, which utilize temporal graph databases and vector databases, has further enhanced context retention and data retrieval, demonstrating the potential of LLMs in complex data environments [37]. Moreover, expert insight encoding via LLMs (EIE-LLM) has been employed to convert qualitative insights from investigators into quantifiable features for predictive analytics, showcasing the versatility of LLMs in various applications [8].

Despite these advances, challenges remain in ensuring the accuracy and reliability of LLM outputs, particularly in low-resource settings where language mismatch and repetition errors are prevalent. The ongoing advancement and integration of innovative methodologies are crucial for effectively tackling the complexities of information extraction (IE), particularly in diverse domains characterized by varying document genres and lengths. This is underscored by research indicating that no single approach, whether heuristic-based or data-driven, consistently outperforms others across different IE tasks such as named entity recognition (NER) and semantic role labeling (SRL). Moreover, the implementation of human-in-the-loop systems can enhance precision in high-stakes applications, while recent techniques leveraging large language models (LLMs) and multimodal data sources have shown promise in improving extraction quality and comprehensiveness. Collectively, these developments highlight the need for a multifaceted approach to enhance the robustness and adaptability of IE systems in processing unstructured data efficiently. [39, 17, 12, 49, 22]

Furthermore, the integration of LLMs with semantic technologies for reasoning and inference has been explored to improve methods for extracting information from unstructured text data to create Knowledge Graphs, as highlighted by Trajanoska et al. [38]. Additionally, the introduction of LLM2LLM, an iterative data augmentation framework, enhances small seed datasets by using a teacher LLM to generate synthetic data based on the model's incorrect predictions [34].

These advancements in information extraction techniques, including the utilization of machine learning and deep learning methods, significantly enhance the capability to process and analyze unstructured textual data across various domains, such as legal, medical, and biodiversity literature. By integrating heuristic-based and data-driven approaches, as well as leveraging large language models, these techniques enable more precise identification of structured information, improving accuracy in tasks like named entity recognition and semantic role labeling. Furthermore, the exploration of human-machine collaboration frameworks has demonstrated that combining human insight with automated processes can yield high precision in information extraction, making these advancements crucial for effective knowledge discovery in diverse applications. [12, 14, 16, 17]. The continuous development and integration of these methodologies are essential for addressing the complex challenges inherent in extracting valuable insights from unstructured data.

3.4 Domain-Specific Information Extraction

Domain-specific information extraction (IE) presents unique challenges and opportunities across various fields, necessitating tailored approaches to effectively process unstructured data. As illustrated in Figure 4, the hierarchical structure of domain-specific information extraction categorizes key domains such as Legal, Financial, and Conversational Systems. Each domain highlights specific methodologies and applications, including entity recognition and sequence-to-sequence (Seq2Seq) models in the Legal domain, fine-tuning of large language models (LLMs) in the Financial sector, and intent classification in Conversational Systems.

In legal contexts, domain-specific entity recognition is vital for applications such as question-answering systems and information retrieval within case law documents [50]. This involves extracting relevant legal entities and relationships, which is crucial for facilitating efficient legal research and decision-making processes. The complexity of legal language and the need for accuracy in entity recognition underscore the importance of specialized models and datasets, as demonstrated by experiments on legal documents in Portuguese, which highlight the real-world applicability of sequence-to-sequence models in extracting structured information from legal texts [40].

In the financial domain, the fine-tuning of large language models (LLMs) is critical for applications such as financial forecasting and sentiment analysis. The process involves careful dataset selection, preprocessing, and model choice, tailored to the specific requirements of financial data [51]. The variability and complexity of financial texts demand robust models capable of capturing nuanced information, which is essential for accurate financial analysis and decision-making.

The use of deep learning techniques, such as the A2I method, enhances the flexibility and scalability of information extraction processes, accommodating the variability inherent in unstructured data across domains [39]. By focusing on entities as clusters rather than isolated mentions, neural approaches can leverage contextual relationships between entities, improving performance in tasks such as entity-centric information extraction [41].

In the realm of conversational systems, intent classification for banking chatbots represents a domain-specific challenge requiring the evaluation and comparison of different models. Benchmarks in this area facilitate the assessment of model effectiveness, ensuring that chatbots can accurately interpret and respond to diverse customer inquiries [52]. Similarly, the REGEN benchmark addresses the challenge of generating rich natural language narratives from user-item interaction data in conversational recommendation tasks, focusing on capturing user preferences and producing informative outputs [53].

The extraction of skills from job descriptions is another domain-specific application that benefits from benchmarks designed to evaluate model effectiveness. This task is crucial for advancing natural language processing capabilities in human resources and recruitment, enabling the automated identification of relevant skills and competencies [54].

Overall, domain-specific information extraction requires the integration of specialized methodologies and datasets to address the unique challenges presented by different fields. The ongoing development of customized methodologies and evaluation benchmarks is crucial for improving the accuracy and efficiency of information extraction (IE) processes across various domains, as evidenced by studies highlighting the impact of document genre and length on IE tasks, the advantages of human-in-the-loop systems for high-precision extraction, and the implementation of machine learning pipelines for structured information retrieval from unstructured text. These advancements not only enhance the performance of tasks such as named entity recognition and semantic role labeling but also address the challenges posed by different document characteristics and the need for reliable, efficient extraction methods in high-stakes applications [12, 55, 17].

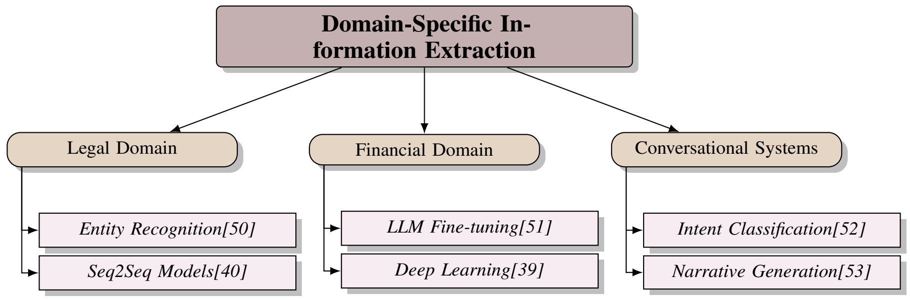


Figure 4: This figure illustrates the hierarchical structure of domain-specific information extraction, categorizing key domains such as Legal, Financial, and Conversational Systems. Each domain highlights specific methodologies and applications, such as entity recognition and Seq2Seq models in the Legal domain, LLM fine-tuning in the Financial sector, and intent classification in Conversational Systems.

Feature	Challenges in Information Extraction	Advancements in Information Extraction Techniques	Role of Large Language Models in Information Extraction
Methodology Type	Problem Identification	Innovative Techniques	Llm Integration
Application Domain	Various Domains	Unstructured Data	Multiple Domains
Key Innovation	Domain-specific Complexities	End-to-end Relation Extraction	Embedding-based Retrieval

Table 2: This table provides a comprehensive comparison of methodologies in information extraction, highlighting the challenges faced, advancements in techniques, and the role of large language models (LLMs). It categorizes the methodologies based on their type, application domain, and key innovations, offering insights into the integration of LLMs and the handling of unstructured data across multiple domains.

4 Large Language Models (LLMs)

The exploration of Large Language Models (LLMs) demands a comprehensive understanding of their development and capabilities, which have significantly evolved to address complex natural language processing tasks. This section outlines the advancements in LLMs, emphasizing their transformative impact across various applications.

4.1 Development and Capabilities

LLMs have undergone substantial evolution, marked by advancements in architecture and functionality that enable them to proficiently perform complex natural language processing tasks. The shift from millions to hundreds of billions of parameters has greatly enhanced their ability to generate coherent and contextually relevant text across diverse applications [56]. This scalability facilitates zero-shot and few-shot learning capabilities, allowing task execution without extensive task-specific training [1].

Innovative methodologies have further optimized LLM performance. The Multi-LLM Orchestration Engine (MLOE) exemplifies the ability to maintain long-term context while grounding responses in historical interactions and relevant data [37]. Additionally, advanced natural language understanding allows LLMs to extract nuanced insights from qualitative data, enhancing entity extraction accuracy from complex texts.

LLMs' integration into domain-specific applications, such as medical information extraction, demonstrates their versatility. Models like GPT-4.0 surpass predecessors like GPT-3.5 in tasks requiring higher understanding and summarization capabilities, particularly in the medical domain [44]. Moreover, LLMs facilitate the automatic generation of Knowledge Graphs from unstructured text, showcasing utility in complex data environments [38].

The development of benchmarks for analyzing complex public affairs documents highlights advancements in LLMs, improving classification accuracy and providing effective tools for various applications [42]. The employment of deep generative models in estimating causal effects while controlling for confounding features illustrates the innovative application of LLMs in causal representation learning [31].

Despite these advancements, challenges in evaluating LLM performance persist. Studies on keyword extraction using models like Llama2-7B, GPT-3.5, and Falcon-7B emphasize the need for robust assessment frameworks [43]. Continuous development of novel prompting strategies, interpretability frameworks, and architectures that enhance memory and recall capabilities is essential for maximizing LLM potential across domains. The LLM2LLM method addresses the observation that LLMs excel at simpler examples but struggle with more complex ones, emphasizing the need for targeted data augmentation [34]. Furthermore, the MISeD benchmark has been developed to leverage LLMs in automating dialog dataset creation, particularly in generating queries and responses [45].

4.2 Applications in NLP and Beyond

LLMs significantly advance natural language processing (NLP) and extend their capabilities to various domains. In NLP, they have enhanced tasks such as zero and few-shot named entity recognition (NER), exemplified by the llmNER library's effective validation using benchmark datasets [57]. Their application in document understanding tasks, including Visual Question Answering and Key Information Extraction, underscores their versatility in managing complex textual data [58].

The integration of LLMs with retrieval-augmented generation techniques has improved performance across applications. Models like GPT-3.5 Turbo and LLaMA2, when fine-tuned with these methodologies, demonstrate enhanced synthesis of information from multiple sources, enriching context for tasks such as entity description and summarization. This capability is crucial in domains requiring high precision and recall, as evidenced by human evaluators assessing LLM outputs in agricultural documents from the AHDB Agriculture and Board database [36].

Beyond NLP, LLMs contribute significantly to fields like recommender systems. Their integration into pre-training, fine-tuning, and prompting paradigms has improved recommendation accuracy and user satisfaction [59]. The REGEN benchmark illustrates the potential of LLMs to incorporate user-generated narratives into datasets, such as Amazon Product Reviews, enhancing conversational contexts in recommendation tasks [53].

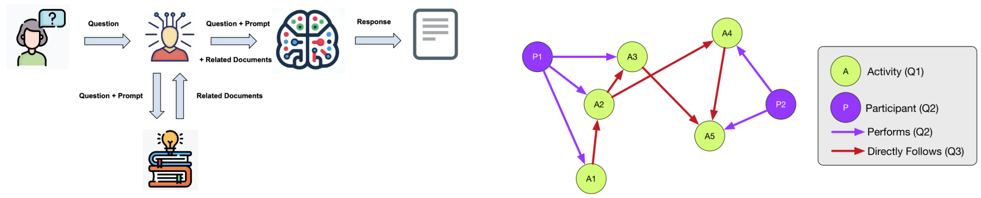
In specialized domains, LLMs have shown adaptability in tasks like Open Information Extraction (OpenIE) for Portuguese texts, improving extraction performance through the fine-tuning of both open-source and commercial models [60]. Their utility in interactive applications is further demonstrated by models like RoLlama, capable of engaging in multi-turn conversations [61].

The medical field has benefited from LLM applications, particularly in enhancing diagnostic processes. Studies involving surgical pathology reports have employed models like LLaMA and BERT to improve information extraction and analysis, showcasing the efficiency of LLMs in medical literature reviews and their ability to accurately extract critical information. Additionally, the competitive performance of ChatGPT-3.5-turbo in Chinese open-source models highlights its advantages in specific linguistic contexts [62].

Furthermore, LLMs have been instrumental in creating Knowledge Graphs focused on sustainability, utilizing advanced models like ChatGPT and REBEL for joint entity and relation extraction [38]. Their application in data visualization tasks, as demonstrated by ChartGPT, offers users more intuitive and flexible interactions for generating visualizations [56].

The diverse applications of LLMs across NLP and beyond illustrate their transformative potential in various industries. Their integration with other technologies is expected to expand their impact, driving advancements in data processing, optimization, and information retrieval. The ability of LLMs to extract true internal representations of treatment texts in generative AI applications further exemplifies their role in addressing confounding issues [31].

As illustrated in Figure 5, LLMs have revolutionized the field of Natural Language Processing (NLP) and have applications that extend beyond traditional language tasks. A prime example is the development of sophisticated question-answering systems, as depicted in the provided flowchart. This system processes user-inputted questions through a complex neural network, integrating prompts and relevant documents to generate comprehensive responses, showcasing LLMs' intricate processing capabilities. Additionally, the network diagram highlights their versatility in managing complex relationships and activities within a system, with nodes representing various activities and participants interconnected by arrows denoting different relationship types. These examples underscore LLMs' transformative impact in enhancing the efficiency and scope of NLP applications and beyond [63, 64].



(a) A Flowchart of a Question-Answering System[63]

(b) The image depicts a network of activities and participants, with arrows indicating the relationships between them.[64]

Figure 5: Examples of Applications in NLP and Beyond

4.3 Challenges and Limitations

LLMs face a variety of challenges and limitations that influence their deployment and effectiveness across applications. A significant hurdle is the substantial computational resources required for training and operation, which can impede accessibility for research labs with limited funding [7]. This challenge is compounded by scalability issues, as the complexity and size of LLMs necessitate advanced infrastructure for efficient management [37].

Another critical limitation is the insufficient safety measures in many LLMs, which raises concerns about their potential misuse in generating misleading or harmful content without effective detection mechanisms [65]. This highlights the urgent need for enhanced safety protocols and monitoring systems to mitigate risks associated with LLM outputs.

The ability of LLMs to generalize across unseen tasks remains a challenge, particularly in complex reasoning and multi-step decision-making scenarios, where traditional deep neural network (DNN) methods often falter [59]. Additionally, reliance on guiding information from smaller models can adversely affect performance if the guiding model is inadequately trained, emphasizing the importance of high-quality data and training processes [27].

Data augmentation methods for LLMs also encounter limitations, as existing techniques do not effectively expand datasets for specialized tasks, leading to suboptimal performance during fine-tuning [34]. This issue is exacerbated by syntactic and lexical overfitting, which undermines the diversity and quality of datasets generated by LLMs [66].

The integration of LLMs with other technologies, such as graph and vector databases, presents additional challenges, including managing the complexity of these systems and addressing inefficiencies in the reflection phase [37]. These challenges underscore the necessity for continuous innovation in methodologies that enhance the adaptability and efficiency of LLMs in processing complex data environments.

Addressing the challenges associated with the effective application of LLMs is crucial for enhancing their capabilities and ensuring reliable and ethical use across diverse fields. This includes improving the extraction of relevant tasks, machine learning models, and datasets from academic literature to streamline the identification of suitable methods for specific applications. Developing frameworks like Missing Information Guided Retrieval-Augmented Generation (MIGRES) can enhance LLMs' performance in complex reasoning and retrieval tasks, thereby improving the quality of information obtained. Leveraging LLMs to assess the value of scientific ideas can facilitate the discernment of impactful research, ultimately advancing scientific inquiry and innovation [67, 68, 69]. Continued research and development efforts are essential to overcome these limitations, focusing on improving computational efficiency, safety measures, and data augmentation techniques to enhance the performance and applicability of LLMs.

5 Natural Language Processing (NLP)

5.1 Overview of NLP Techniques

Natural Language Processing (NLP) encompasses techniques that enable machines to process human language, categorized into syntactic, semantic, and pragmatic processing. Syntactic processing involves parsing and part-of-speech tagging to understand grammatical structures. Semantic processing focuses on meaning through methods like word sense disambiguation and semantic role labeling, while pragmatic processing considers context for tasks such as discourse analysis and sentiment detection. Machine learning models, including layer-wise relevance propagation (LRP), classify extensive text collections by identifying semantic categories and sentiments, revealing word influence on classification decisions. The advent of large language models (LLMs) has automated the annotation of complex pragmatic features, enhancing corpus-based analyses [70, 3, 71, 72].

As illustrated in Figure 6, which depicts the hierarchical structure of NLP techniques, these techniques are categorized into processing types, machine learning models, and applications. This figure highlights the integration of traditional NLP with advanced models like LLMs, showcasing their impact on various applications such as sentiment analysis and automated literature reviews. LLMs have revolutionized NLP by improving task accuracy and efficiency. Using deep learning architectures like transformers, they excel in text summarization, producing concise summaries from extensive datasets [73]. Techniques such as named entity recognition (NER), topic modeling, and machine translation have benefited from LLMs, providing robust frameworks for linguistic challenges. Neural networks, especially sequence-to-sequence models, have enhanced machine translation by maintaining contextual relevance [16, 72]. LLMs refine topic modeling techniques like Latent Dirichlet Allocation (LDA) for better topic categorization.

Sentiment analysis, a vital NLP application, uses lexicon-based methods and classifiers like CNNs and SVMs to evaluate text emotion. LRP enhances model explainability by showing word contributions to sentiment classification [18, 69, 72]. LLMs' ability to process vast data and detect emotional cues boosts sentiment analysis accuracy, aiding market research and social media monitoring.

Integrating traditional NLP with LLMs expands language processing scope and effectiveness. This hybrid approach supports sophisticated textual data analyses, as shown in studies automating literature reviews and optimizing legal decision support systems. For instance, employing various NLP strategies, including retrieval-augmented generation with models like GPT-3.5-turbo, improves literature review generation and legal text analysis, offering valuable insights [19, 20]. Continued development promises further advancements in human language understanding and generation, driving innovation across applications and industries.

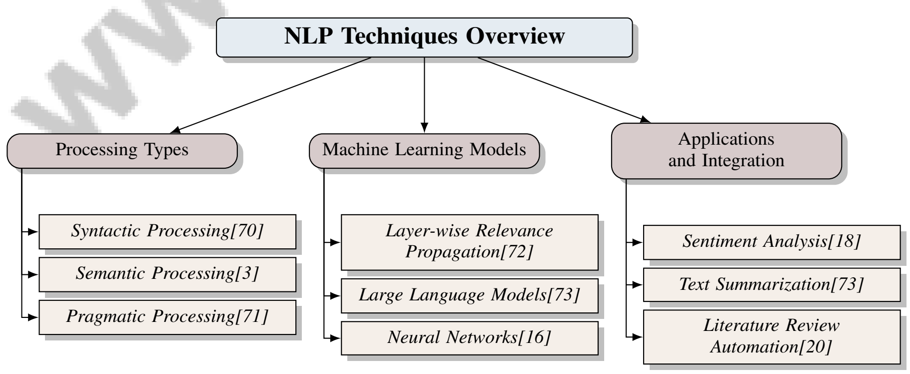


Figure 6: This figure illustrates the hierarchical structure of NLP techniques, categorizing them into processing types, machine learning models, and applications. It highlights the integration of traditional NLP with advanced models like LLMs, showcasing their impact on various applications such as sentiment analysis and automated literature reviews.

5.2 Applications in Analyzing Natural Language

NLP facilitates insights from textual data through diverse applications. One significant use is skill extraction from job postings, identifying explicit and implicit skills required across industries, exemplified by the SkillSpan dataset [54].

In sentiment analysis, NLP evaluates and categorizes text emotional tone, crucial for public opinion understanding, customer feedback, and user experience enhancement in marketing and social media monitoring. Advanced techniques like NER and LLMs discern sentiment nuances, offering deeper emotional context insights [54, 18, 20, 12, 74]. LLMs enhance sentiment analysis accuracy by capturing subtle emotional cues and efficiently processing large data volumes.

In machine translation, NLP ensures accurate text conversion between languages, preserving meaning and context. Sophisticated algorithms analyze semantic and syntactic structures, ensuring linguistically correct and culturally relevant translations. Machine learning techniques enhance translation quality by identifying key concepts and sentiments [75, 13, 18, 72]. Neural networks and sequence-to-sequence models improve translation accuracy, bridging language barriers.

NER identifies and classifies key entities within unstructured text, transforming documents into structured formats for analysis. Recent advancements in NER use machine learning and deep learning to enhance accuracy across document genres, improving downstream NLP tasks like information retrieval and question answering [16, 17, 41]. Topic modeling categorizes large text corpora topics, providing insights into document thematic structures.

NLP's applications in machine translation, email spam detection, information extraction, and summarization illustrate its transformative potential. It enables computational language analysis, enhances text understanding through machine learning, and supports structured information extraction from unstructured data [17, 18, 72]. LLM integration promises further NLP capability enhancement, driving data analysis and interpretation innovation.

As illustrated in Figure 7, this figure highlights key NLP applications, emphasizing skill extraction, sentiment analysis, and machine translation. Each category is supported by specific techniques and datasets, showcasing NLP's potential to transform language processing tasks across various domains. The first example evaluates agroecological practices' climate change impact using a decision tree model, while the second example employs a text-based NER prompt on the CoNLL03 dataset to identify and categorize entities. This further demonstrates NLP's versatility in processing natural language data and its potential to advance fields like environmental science and information extraction [76, 21].

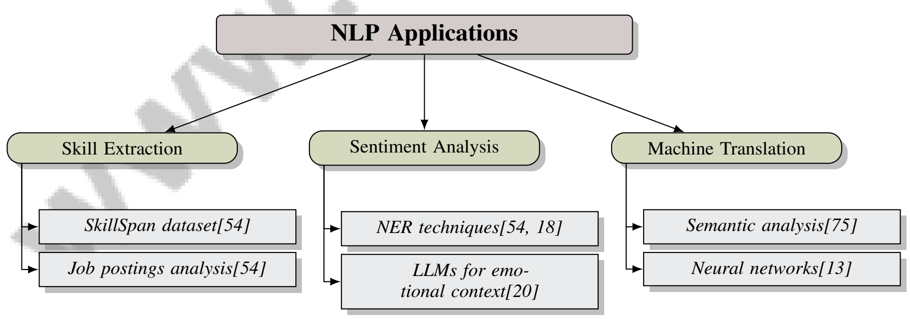


Figure 7: This figure illustrates key NLP applications, highlighting skill extraction, sentiment analysis, and machine translation. Each category is supported by specific techniques and datasets, showcasing NLP's potential to transform language processing tasks across various domains.

5.3 Enhancement of Optimization Processes

NLP integration into optimization processes advances applications by enhancing data-driven decision-making efficiency. It processes vast unstructured textual data, extracting insights to optimize algorithms and workflows. Information extraction (IE) techniques convert text into structured information, and machine learning models map documents to abstract concepts. LLMs improve information extraction accuracy, identifying key elements in complex texts, optimizing operations and decision-making from business to scientific research [77, 78, 17, 72].

NLP enhances sentiment analysis and opinion mining, extracting sentiment from datasets to inform marketing and customer relationship management. Machine learning models and IE techniques identify sentiment indicators, systematically analyzed for marketing strategies and customer engagement [17, 69, 46, 22, 72]. Advanced NLP models optimize strategies based on sentiment analysis, improving customer satisfaction.

In machine translation, NLP optimizes language models for accurate, context-aware translations. Sequence-to-sequence models and attention mechanisms adapt to linguistic intricacies, enhancing translation processes. Studies show smaller pre-trained models effectively translate low-resource languages, emphasizing context type and retrieval methods for accuracy. These models streamline information extraction by generating structured outputs, reducing reliance on rule-based systems, bridging communication gaps across languages [79, 64, 40].

NLP improves information retrieval with IE and retrieval-augmented generation (RAG), automating literature reviews from datasets. These methodologies enhance accuracy in tasks like NER and semantic role labeling (SRL), extracting relevant information from document repositories. Human validation within automated systems maintains high precision, making NLP essential for information retrieval optimization in high-stakes applications [17, 72, 12, 20, 70]. NLP-driven semantic analysis enhances search algorithms, delivering precise results, improving user experience, and information accessibility, crucial in legal and medical retrieval.

NLP optimizes topic modeling by categorizing thematic structures within text corpora, enhancing data organization and analysis. Applications like content recommendation systems personalize experiences by suggesting relevant materials, and document classification categorizes based on concepts like semantic category or sentiment. Advanced machine learning and human-in-the-loop approaches improve structured information extraction accuracy, offering insights into categorization decision-making processes [17, 72, 12, 16, 70].

The evolution of NLP, particularly through LLMs, is set to enhance optimization processes across applications. Research shows LLMs effectively extract process descriptions, achieving up to 8% F1 score improvements over traditional methods. This advancement enhances decision-making by integrating LLMs with optimization algorithms, refining modeling strategies, and elevating output quality, paving the way for innovative AI solutions across domains [78, 80, 18]. By integrating sophisticated language processing, NLP continues to drive innovation and efficiency in decision-making, optimizing workflows across industries.

6 Generative AI

6.1 Synthetic Data Generation Frameworks

Synthetic data generation frameworks have become pivotal in advancing NLP and related fields by creating artificial datasets that mimic real-world data. This facilitates the training and evaluation of machine learning models while addressing data scarcity and privacy issues. A key approach involves prompt engineering and parameter-efficient adaptations that optimize these processes [81]. Prompt engineering refines generative model prompts to enhance synthetic data quality and relevance, leveraging LLMs to produce contextually diverse datasets crucial for robust NLP system development. This approach automates information retrieval from extensive biodiversity literature, improving data collection scalability and efficiency [82, 14]. Parameter-efficient adaptations, on the other hand, focus on optimizing model parameters for efficient data generation in resource-constrained scenarios.

These frameworks significantly impact NLP by creating diverse datasets, enhancing model generalization and performance in tasks like NER and SRL. Automating methodology selection tailored to specific domains through advanced task-model-dataset relationship analysis improves NLP application accuracy, addressing bias and interpretability challenges [75, 17, 69, 72]. Generating substantial synthetic data volumes not only combats data scarcity but also enhances NLP solution scalability and mitigates biases, promoting equitable AI systems.

Beyond NLP, synthetic data generation frameworks enhance large-scale data analysis and machine learning model training, especially where real-world data is scarce. Recent generative AI advancements, particularly through LLMs, have shown effectiveness in producing high-quality, task-specific training data, addressing data availability and model performance challenges [75, 81]. The ability to generate high-quality synthetic datasets supports innovative solutions, driving AI research and

AI-generated, for reference only.

deployment advancements. Integrating these frameworks with existing NLP techniques promises to enhance language processing systems' capabilities, fostering continued innovation in the field.

6.2 Linguistic Creativity and CREATIVITY INDEX

Linguistic creativity in generative AI has gained significant attention, particularly through frameworks quantifying and enhancing creative outputs. A notable advancement is the CREATIVITY INDEX, measuring linguistic creativity by reconstructing text from web snippets [83]. This index evaluates the novelty and diversity of generated content, providing insights into AI systems' creative capabilities.

Linguistic creativity involves producing contextually relevant and coherent narratives that resonate with human creativity. The CREATIVITY INDEX captures these dimensions by analyzing AI-generated text's divergence from conventional patterns, introducing innovative linguistic constructs. This contributes to advancing AI systems to emulate human-like creativity, enabling original and meaningful outputs comparable to professional human authors' creativity [83, 68, 13, 12].

Integrating linguistic creativity measures into generative AI systems impacts content creation, storytelling, and artistic expression. Quantifying creativity through metrics like the CREATIVITY INDEX provides insights into elements driving AI systems' creative outputs, allowing for LLM refinement and optimization for innovative content generation. Studies show the CREATIVITY INDEX differentiates human authors' creative capabilities from LLMs, revealing professional human writers typically have higher creativity scores. Advanced algorithms like DJ SEARCH enhance text originality evaluation against existing content databases, streamlining scientific idea evaluation [83, 75, 68]. This creativity focus emphasizes developing AI systems that replicate human language and contribute to language and creativity evolution in digital contexts.

Future research should refine methodologies for quantifying linguistic creativity, exploring creativity's interplay with other AI performance dimensions. Ethical considerations surrounding creative AI systems, such as generating misleading or culturally insensitive content, warrant careful examination [5]. Addressing these challenges can foster responsible generative AI technology advancement, promoting systems enhancing human creativity while adhering to ethical standards.

6.3 Revolutionizing Recommender Systems

Generative AI has transformed recommender systems, enhancing their ability to provide personalized and contextually relevant suggestions across various domains. The integration of LLMs and generative techniques has led to sophisticated recommender systems utilizing user interactions and content features for tailored recommendations. This transformation is evident through pre-training, fine-tuning, and prompting paradigms, enhancing recommendation accuracy and user satisfaction [59].

A key advancement is retrieval-augmented generation techniques, enriching information synthesis from multiple sources, enhancing context for recommendation tasks. This approach allows recommender systems to incorporate diverse user-generated narratives into datasets, like Amazon Product Reviews, improving recommendation relevance and informativeness [53]. The REGEN benchmark exemplifies this potential by generating rich natural language narratives from user-item interaction data, capturing user preferences and producing engaging outputs.

Generative AI's application in recommender systems transcends traditional content-based and collaborative filtering methods. By harnessing generative models' creative capabilities, these systems produce innovative content and tailored suggestions dynamically adapting to users' changing preferences and interests, enhancing user engagement and satisfaction. This adaptability is significant in conversational recommendation systems, where integrating user narratives and product endorsements leads to more personalized and contextually relevant recommendations [75, 53, 84, 1, 68]. This capability is valuable in dynamic environments where user preferences evolve rapidly, making fresh and relevant recommendations crucial for maintaining engagement.

Generative AI integration into recommender systems enhances recommendation personalization and fosters interactive and conversational interfaces. Leveraging advanced LLM capabilities for natural language generation and understanding addresses challenges in incorporating user preferences and contextual information effectively. Datasets like REGEN enrich product recommendations with user narratives, and innovative fusion architectures combining collaborative filtering and content-based embeddings improve language metrics and user engagement [59, 53, 1]. These advancements allow users to interact with systems more naturally and intuitively, enhancing the overall user experience. By incorporating generative models capable of understanding and generating human-like text, recommender systems provide nuanced, context-aware recommendations, further personalizing user experiences.

Generative AI integration into recommender systems marks a transformative leap in design and functionality, leveraging LLM capabilities to create personalized and contextually relevant user experiences. This advancement enhances narrative generation explaining product recommendations, as seen with the REGEN dataset, and facilitates incorporating user-specific data and preferences into the recommendation process. Emerging techniques like rubric-enabled generative AI (REGAI) further optimize LLM performance through structured evaluation methods, showcasing generative AI's potential to improve recommender systems' effectiveness across applications [85, 53, 86, 20]. By enabling personalized, contextually relevant, and creative recommendations, generative AI revolutionizes user interactions with digital content, driving innovation and enhancing satisfaction across various applications. Ongoing development and refinement of these technologies promise to further elevate recommender systems' capabilities, ensuring their relevance and effectiveness in an increasingly digital landscape.

7 Text Mining

Text mining extends beyond simple data extraction, encompassing a range of applications that leverage textual information for enhanced decision-making and analysis. A key application is in benchmarking and text summarization, where text mining is essential for evaluating and condensing vast information into manageable summaries. This aids in comprehending extensive datasets and supports further analysis across diverse fields. The following subsection delves into benchmarking and text summarization methodologies and the impact of advanced natural language processing (NLP) techniques.

7.1 Benchmarking and Text Summarization

Benchmark	Size	Domain	Task Format	Metric
LLM-SUM[73]	300,000	Text Summarization	Abstractive Summarization	BLEU Score, ROUGE Score
VINE[87]	10,000	Knowledge Graph Construction	Knowledge Extraction	F1-score
LLM-as-reference[88]	280,000	News Summarization	Abstractive Summarization	ROUGE-1, ROUGE-2
Kleptotrace[89]	11,152	Financial Crime	Named Entity Recognition	Accuracy, F1 Score
Kiwi[90]	1,588	Clinical Information Extraction	Named Entity Recognition	F1
LLMs4OL[91]	3,000,000	Biomedical	Term Typing	MAP@1, F1-score
IE-Benchmark[17]	10,708	Clinical Neurology	Named Entity Recognition	F1-score, Accuracy
SIBB[52]	300	Banking	Intent Classification	In-scope Accuracy, Out-of-scope False Positive Rate

Table 3: This table presents a comprehensive overview of various benchmarks used in the evaluation of large language models (LLMs) across multiple domains. It details the size, domain, task format, and evaluation metrics for each benchmark, highlighting the diversity and specificity of tasks such as text summarization, knowledge graph construction, and named entity recognition.

Text mining is crucial for benchmarking and summarization, extracting meaningful patterns from large textual datasets. Advanced NLP techniques, notably large language models (LLMs), have significantly enhanced text summarization, enabling concise and coherent summaries that improve information retrieval and content generation [73]. Datasets like CNN/Daily Mail and XSum are vital for benchmarking summarization models, providing standards to evaluate various methodologies [73]. These datasets ensure models are tested against diverse inputs, enhancing robustness and adaptability.

Benchmarking involves systematically evaluating algorithms to assess their effectiveness in processing and summarizing textual data, identifying strengths and weaknesses, and facilitating the development of more efficient text mining solutions. Insights from model explainability and task interrelationships contribute to this process [69, 72]. Establishing benchmarks enables comparison of model performance, driving innovation in text summarization techniques. Table 3 provides an extensive comparison of benchmarks utilized in assessing the performance of large language models in text summarization and other related tasks, emphasizing their role in advancing NLP techniques.

LLMs have advanced benchmarking and summarization by employing deep learning architectures that capture intricate language details, leading to high-quality summaries. LLM-generated summaries are often preferred by human evaluators over traditional ones, aligning with human expectations. Contrastive learning techniques surpass standard supervised fine-tuning, enhancing summary quality and relevance [13, 88, 72]. Consequently, text mining is indispensable for managing large data volumes, supporting decision-making and knowledge discovery across domains.

Text mining's role in benchmarking and summarization advances NLP techniques, enhancing textual information accessibility. Machine learning models improve usability in tasks like topic categorization and information extraction. Innovative methods, such as layer-wise relevance propagation (LRP), enhance classification accuracy and model explainability. Approaches that extract relationships among tasks, models, and datasets facilitate automatic methodology recommendations, propelling progress and making complex information more comprehensible [17, 69, 72].

7.2 Skill Extraction from Job Postings

Text mining's application in extracting skills from job postings is crucial for efficiently matching job seekers with opportunities. LLM-enhanced techniques improve parsing and interpreting unstructured job advertisement data, identifying both explicit and implicit skills across industries [54]. Specialized datasets like SkillSpan refine skill extraction methodologies by fine-tuning LLMs like Skill-LLM, enhancing extraction accuracy from job descriptions. This advancement addresses challenges posed by traditional Named Entity Recognition (NER) techniques and facilitates recommending appropriate machine learning models and datasets for specific tasks, streamlining hiring processes [55, 54, 69]. These datasets provide a structured framework for evaluating model effectiveness in capturing nuanced skill information.

Text mining for skill extraction automates recruitment processes, reducing time and effort needed to analyze job advertisements. Advanced NLP techniques and fine-tuned LLMs enhance talent acquisition strategies, efficiently identifying candidates with requisite skills. This automation improves recruitment efficiency and candidate-job matching accuracy, leading to better employment outcomes [12, 54, 17, 20].

Advancements in text mining, particularly through LLM integration, will further refine skill extraction. As technologies evolve, they will transform recruitment by precisely identifying skills in the job market. Machine learning models, such as convolutional neural networks and data-driven information extraction methods, will enhance candidate evaluation accuracy by analyzing vast unstructured data. This evolution enables recruiters to distill relevant information from various sources, incorporating human validation for high precision in candidate selection, ultimately leading to more effective hiring processes aligned with organizational needs [39, 17, 12, 49, 72].

8 Interconnections and Synergies

This section examines the complex interactions among advanced technologies, focusing on the synergy between Large Language Models (LLMs) and information extraction methods. Understanding these connections is crucial for improving data processing efficiency and accuracy, enabling innovative applications across various fields. The following subsection explores the transformative impact of LLMs on data analysis and interpretation in conjunction with information extraction.

8.1 Synergy with Information Extraction and LLMs

The integration of LLMs with information extraction techniques marks a significant advancement, enhancing the extraction of structured data from unstructured text. This synergy is exemplified by frameworks that convert qualitative insights into quantifiable metrics, highlighting the complementary relationship between qualitative analysis and LLMs [8]. By utilizing LLMs' capabilities to process complex data, these frameworks derive nuanced insights from qualitative data, thus improving information extraction processes.

Benchmarks like REGEN are pivotal in refining conversational recommenders, showcasing how LLMs enhance contextual relevance and accuracy in information extraction tasks [53]. Tailored reasoning environments further boost LLM performance in Open Information Extraction (OIE) tasks by enabling generalization across diverse scenarios. A human-centered approach to transparency underscores the importance of aligning with stakeholders' goals in transparency development, crucial for integrating LLMs into reliable and trustworthy information extraction systems [11]. Frameworks like ALLURE propose systematic auditing processes to incorporate significant deviations into LLM evaluation mechanisms, ensuring continuous refinement to meet dynamic information extraction demands.

Despite challenges such as potential disinformation generation by LLMs, analysis reveals variability in their susceptibility to harmful narratives, emphasizing the need for robust safety measures [65]. The synergistic relationship between information extraction and LLMs promises advancements in automated language understanding and processing. Continuous innovation and adaptation of these technologies are crucial for developing robust information extraction systems capable of navigating complex data environments.

8.2 LLMs and Knowledge Graphs

The integration of LLMs with Knowledge Graphs (KGs) marks a pivotal advancement in NLP, enhancing LLMs' factual reasoning and contextual understanding. This synergy enables LLMs to leverage structured knowledge from KGs, addressing limitations in fact recall and knowledge-grounded content generation. Studies demonstrate the effectiveness of this integration in improving tasks like entity and relation extraction, link prediction, and question-answering, while mitigating hallucinations [92, 93, 94, 87, 38].

Incorporating KGs into LLMs significantly improves factual reasoning, allowing access to structured knowledge repositories and generating accurate, contextually relevant responses [94]. This integration facilitates the development of intelligent language models, enhancing information extraction and question-answering capabilities. Research shows that LLMs utilizing KGs provide factual context, helping mitigate memory limitations and domain-specific hallucinations. Advancements in extracting structured entities and their relationships from diverse data sources highlight the potential of LLMs and KGs to deliver precise responses in applications like education and research [95, 69, 92].

The combination of LLMs and KGs enhances models' ability to handle ambiguity, providing a robust framework for differentiating entities and concepts, minimizing errors, and enhancing content quality. Machine learning models like GPT-3 have proven effective in extracting structured information from unstructured text, improving extraction accuracy and enabling nuanced performance evaluation through metrics such as Structured Entity Extraction [15, 95, 96]. These advancements contribute to knowledge base development, enhancing efficiency and reliability for applications like chatbots and recommendation systems.

The ongoing integration and refinement of LLMs with KGs are expected to advance NLP capabilities, driving innovation in information retrieval, knowledge management, and automated reasoning. As machine learning and LLMs evolve, they will play a vital role in the future of language processing, enhancing information extraction and summarization capabilities for more sophisticated, context-aware interactions with digital content [12, 49, 17, 72].

8.3 Interdisciplinary Applications and Innovations

The integration of technologies such as information extraction, LLMs, NLP, generative AI, and text mining has catalyzed transformative interdisciplinary applications across sectors. These technologies enhance data retrieval from diverse sources and enable coherent summaries and automated literature reviews. For instance, LLMs streamline information extraction from biodiversity literature and academic papers, facilitating the identification of relevant machine learning models and datasets, thereby empowering researchers and practitioners [49, 14, 69, 20].

In education, LLMs and NLP revolutionize personalized learning by analyzing students' interactions to offer tailored content, improving outcomes and engagement. Generative AI enhances this by using retrieval-augmented generation (RAG) to produce adaptive materials tailored to individual needs, leveraging pretrained models for contextually relevant content delivery [63, 97, 86].

In healthcare, NLP and Information Extraction (IE) advancements, particularly with Transformers-based models and LLMs, enhance the extraction of vital information from unstructured clinical data. These innovations streamline annotation processes and improve data interpretation accuracy, transforming healthcare practitioners' access to patient information [6, 98, 99, 100, 16]. LLMs in processing electronic health records (EHRs) enhance clinical decision-making, while generative AI optimizes patient care by generating treatment recommendations.

In the legal sector, integrating information extraction techniques and LLMs enhances legal research and document analysis efficiency by automatically identifying critical elements from legal texts, thus improving accuracy and reliability in building dynamic legislation networks [15, 69]. These technologies reduce the time and effort required for legal professionals, enhancing legal process efficiency.

The financial industry benefits from text mining and NLP, facilitating sentiment analysis and market predictions. Machine learning technologies process extensive datasets of financial news, providing investors with insights into market trends and investment opportunities. Information extraction techniques discern relevant entities from unstructured data, supporting informed decision-making [39, 29, 69, 72]. Generative AI aids in creating predictive models for financial forecasting and risk management.

Advancements in fairness certification in natural language applications emphasize the importance of ethical considerations in AI deployment. Future research should refine certification criteria, explore fairness implications, and establish best practices for implementation [101]. This focus on fairness ensures responsible AI development across sectors.

The interdisciplinary applications of these technologies underscore their transformative potential. As they evolve, they promise to enhance AI systems' capabilities in addressing complex challenges across industries. Information extraction methods advance to convert unstructured data into structured formats, improving decision-making in fields like behavioral sciences and trust in AI. Techniques like layer-wise relevance propagation (LRP) enable deeper understanding of machine learning models, allowing for accurate text categorization while maintaining explainability. Innovations like GPT-3 streamline entity extraction from text corpora, improving knowledge base development efficiency. Collectively, these advancements pave the way for new innovation avenues and enhance AI's ability to address complex problems effectively [85, 39, 96, 72].

9 Challenges and Future Directions

The challenges and future prospects of large language models (LLMs) are multifaceted, impacting their reliability and applicability. A significant issue is the occurrence of hallucinations, where models produce outputs that deviate from factual accuracy, raising concerns about their trustworthiness. This section delves into the complexities of hallucinations, model reliability, and the strategies proposed for mitigating these issues.

9.1 Hallucinations and Model Reliability

Hallucinations in LLMs pose a critical challenge to their reliability, especially in contexts requiring high factual accuracy. These occur when models generate content not rooted in the input data, leading to potentially misleading outputs [43]. The risk increases with complex inputs [44], and natural language ambiguity complicates mapping user utterances to specific tasks, such as visualization [56]. Strategies to enhance reliability include integrating external knowledge bases to ground outputs and employing retrieval-augmented generation techniques for improved accuracy [38, 37]. Frameworks like LUNA address trustworthiness by focusing on robustness and mitigating hallucinations [33]. Despite these advancements, challenges like disinformation generation and error propagation in multi-step pipelines persist [65, 38]. Addressing these requires a multifaceted approach, including external knowledge integration and continuous model refinement, to ensure outputs are accurate and trustworthy.

9.2 Data Quality and Annotation Challenges

Data quality and annotation are pivotal challenges in NLP and information extraction, affecting model accuracy and reliability. The scarcity of quality datasets, particularly in non-English languages, limits model performance [102, 62]. In specialized fields like medicine, the reliance on quality training datasets underscores the need for comprehensive data collection [35]. Manual annotation is resource-intensive, often resulting in small datasets that fail to capture real-world diversity, limiting generalizability [99]. Class imbalance further complicates model performance on less frequent topics [42]. Future research should refine evaluation methods and increase evaluator numbers, particularly in chatbot applications, to enhance robustness and reliability [10]. By prioritizing these areas, the field can achieve greater accuracy and reliability across diverse applications.

9.3 Computational Costs and Scalability

LLM deployment faces significant computational costs and scalability challenges due to high resource demands. The substantial expense of embedding and processing hinders scalability [32]. Existing knowledge distillation methods often fail to fully leverage LLM capabilities, leading to suboptimal resource use [103]. High costs of advanced commercial LLMs and environmental impacts from energy consumption further constrain deployment [104, 105]. Innovative approaches like AdaInfer and the LLM2LLM framework aim to reduce computational costs and enhance scalability [106, 34]. Despite these advancements, resource-intensive nature remains a barrier, necessitating ongoing innovation in model design and optimization to facilitate broader deployment.

9.4 Model Interpretability and Trust

Enhancing the interpretability and trustworthiness of LLMs is crucial for their reliable application. Interpretability ensures stakeholders understand and trust model outputs, particularly in sensitive areas like scientific hypothesis generation [107]. The integration of external knowledge and reasoning mechanisms is vital for improving performance in specialized fields [108]. Techniques like document-wise memories and Layer-wise Relevance Propagation (LRP) promote transparency and reliability [109, 72]. Future research should focus on developing frameworks that dynamically adjust to new tasks while maintaining interpretability and trust [110]. Exploring meta-pretraining algorithms can further enhance LLM transparency and reliability [97]. Advancing model interpretability and trust requires continuous innovation to ensure AI systems are transparent and reliable.

9.5 Integration and Adaptability across Domains

Integrating and adapting information extraction technologies across domains is vital for advancing NLP and LLM capabilities. Enhancing cross-domain applicability is crucial for meeting diverse field requirements. Developing multilingual support systems for automated contract analysis can enhance adaptability in legal and financial domains [30]. Improving joint relation extraction methods and developing better prompting strategies can facilitate LLM integration into complex data environments [46]. In healthcare, expanding systems like GAMedX to address various NLP tasks can improve adaptability and patient care [35]. Exploring causal representation learning methodologies for unstructured data can further expand adaptability across domains [31]. The integration and adaptability of IE technologies present substantial opportunities for enhancing NLP and LLM applications, driving innovation and efficiency [17, 22].

9.6 Ethical Considerations and Bias Mitigation

The deployment of LLMs necessitates a thorough examination of ethical considerations and bias mitigation strategies. Biases in synthetic datasets can affect AI applications' reliability and fairness, particularly in critical sectors like healthcare [111, 6]. Addressing these challenges involves developing transparency approaches tailored for LLMs and enhancing methods for communicating uncertainty and model behavior. The potential for LLMs to produce toxic content emphasizes the need for ethical filtering mechanisms [13, 112]. Enhancing AI systems' explainability and selecting model-friendly formal languages can reduce biases and improve reliability [32]. By focusing on transparency, accountability, and robust bias mitigation methodologies, the AI community can foster ethical AI deployment that respects diverse perspectives and promotes fairness across applications.

10 Conclusion

The survey underscores the pivotal influence of large language models (LLMs) in revolutionizing automated language processing by seamlessly integrating with various natural language processing

(NLP) technologies. Models such as GPT-4 demonstrate exceptional proficiency in domain-specific information extraction, highlighting their applicability in specialized fields, including the precise retrieval of technical data. The strategic use of prompt engineering and parameter-efficient techniques for generating synthetic data exemplifies the adaptability of LLMs in resource-constrained settings, thereby enriching dataset construction and enhancing model efficacy across multiple domains.

Furthermore, modular frameworks like GATE exemplify the synergy between diverse NLP tools, fostering innovation and efficiency by streamlining research processes. LLMs hold significant promise in healthcare for real-time diagnostics and decision-making, although further exploration is needed to address existing limitations and ensure reliability. The development of standardized ontologies, such as AIO, enhances communication and collaboration within the AI community, promoting a unified understanding of AI concepts.

Despite these advancements, challenges such as bias, reasoning, and factual accuracy in LLM outputs persist, necessitating continued research into user-centered design and bias mitigation. The dual role of LLMs as both beneficial tools and potential sources of bias emphasizes the need for critical evaluation of their outputs. The survey highlights the importance of robust evaluation methods and universal mitigation strategies to bolster the reliability and trustworthiness of AI systems.

The integration of information extraction, LLMs, NLP, generative AI, and text mining is crucial for advancing automated language processing. These technologies collectively enhance the capacity to process and understand human language, driving innovation and efficiency across various sectors. The survey concludes that LLMs have the potential to significantly enhance machine translation systems for low-resource languages during crises, offering a collaborative framework for rapid development and deployment, thereby underscoring the broader impact of LLMs and related technologies in addressing complex challenges in automated language processing.

AI-generated, for reference only.

References

[1] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey, 2024.

[2] Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathi, Kevin Ding, Anupam A. Thatte, Na Li, and Le Xie. Exploring the capabilities and limitations of large language models in the electric energy sector, 2024.

[3] Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology, 2024.

[4] Robert Vacareanu, Marco A. Valenzuela-Escarcega, George C. G. Barbosa, Rebecca Sharp, and Mihai Surdeanu. From examples to rules: Neural guided rule synthesis for information extraction, 2022.

[5] Marcin P. Joachimiak, Mark A. Miller, J. Harry Caufield, Ryan Ly, Nomi L. Harris, Andrew Tritt, Christopher J. Mungall, and Kristofer E. Bouchard. The artificial intelligence ontology: Llm-assisted construction of ai concept hierarchies, 2024.

[6] Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L. Assimes, Libby Hemphill, and Siyuan Ma. A scoping review of using large language models (llms) to investigate electronic health records (ehrs), 2024.

[7] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. A glimpse in chatgpt capabilities and its impact for ai research, 2023.

[8] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[9] Sergio Oramas, Luis Espinosa-Anke, Francisco Gómez, and Xavier Serra. Natural language processing for music knowledge discovery, 2018.

[10] Bhashithe Abeysinghe and Ruhan Circi. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches, 2024.

[11] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[12] Bradley Butcher, Miri Zilka, Darren Cook, Jiri Hron, and Adrian Weller. Optimising human-machine collaboration for efficient high-precision information extraction from text documents, 2023.

[13] Mujahid Ali Quidwai, Chunhui Li, and Parijat Dube. Beyond black box ai-generated plagiarism detection: From sentence to document level, 2023.

[14] Vamsi Krishna Kommineni, Waqas Ahmed, Birgitta Koenig-Ries, and Sheeba Samuel. Automating information retrieval from biodiversity literature using large language models: A case study. *Biodiversity Information Science and Standards*, 8:e136735, 2024.

[15] Neda Sakhaee and Mark C Wilson. Information extraction framework to build legislation network, 2018.

[16] Parisa Naderi Golshan, HosseinAli Rahmani Dashti, Shahrzad Azizi, and Leila Safari. A study of recent contributions on information extraction, 2018.

[17] Shiyu Yuan and Carlo Lipizzi. Information extraction in domain and generic documents: Findings from heuristic-based and data-driven approaches, 2023.

[18] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges, 2017.

[19] Ram Mohan Rao Kadiyala, Siddartha Pullakhandam, Kanwal Mehreen, Subhasya Tippareddy, and Ashay Srivastava. Augmenting legal decision support systems with llm-based nli for analyzing social media evidence, 2024.

[20] Nurshat Fateh Ali, Md. Mahdi Mohtasim, Shakil Mosharrof, and T. Gopi Krishna. Automated literature review using nlp techniques and llm-based retrieval-augmented generation, 2024.

[21] Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. An empirical study on information extraction using large language models, 2024.

[22] Shumin Deng, Yubo Ma, Ningyu Zhang, Yixin Cao, and Bryan Hooi. Information extraction in low-resource scenarios: Survey and perspective, 2024.

[23] Shubham Vatsal and Harsh Dubey. A survey of prompt engineering methods in large language models for different nlp tasks, 2024.

[24] Xirui Peng, Qiming Xu, Zheng Feng, Haopeng Zhao, Lianghao Tan, Yan Zhou, Zecheng Zhang, Chenwei Gong, and Yingqiao Zheng. Automatic news generation and fact-checking system based on language processing, 2024.

[25] Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. People's perceptions toward bias and related concepts in large language models: A systematic review, 2024.

[26] Boshko Koloski, Nada Lavrač, Bojan Cestnik, Senja Pollak, Blaž Škrlj, and Andrej Kastrin. Aham: Adapt, help, ask, model – harvesting llms for literature mining, 2023.

[27] Wentao He, Hanjie Ma, Shaohua Li, Hui Dong, Haixiang Zhang, and Jie Feng. Using augmented small multimodal models to guide large language models for multimodal relation extraction. *Applied Sciences*, 13(22):12208, 2023.

[28] Ciprian Chelba and Milind Mahajan. Information extraction using the structured language model, 2001.

[29] Abuzar Royesh and Olamide Oladeji. Information extraction: An application to the domain of hyper-local financial data on developing countries, 2024.

[30] Yu Zhao and Haoxiang Gao. Utilizing large language models for information extraction from real estate transactions, 2024.

[31] Kosuke Imai and Kentaro Nakamura. Causal representation learning with generative artificial intelligence: Application to texts as treatments, 2024.

[32] Léo Hemamou and Mehdi Debiane. Scaling up summarization: Leveraging large language models for long text extractive summarization, 2024.

[33] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. Luna: A model-based universal analysis framework for large language models, 2024.

[34] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement, 2024.

[35] Mohammed-Khalil Ghali, Abdelrahman Farrag, Hajar Sakai, Hicham El Baz, Yu Jin, and Sarah Lam. Gamedx: Generative ai-based medical entity data extractor using large language models, 2024.

[36] Ruoling Peng, Kang Liu, Po Yang, Zhipeng Yuan, and Shunbao Li. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *arXiv preprint arXiv:2308.03107*, 2023.

[37] Sumedh Rasal. A multi-llm orchestration engine for personalized, context-rich assistance, 2024.

[38] Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. Enhancing knowledge graph construction using large language models. *arXiv preprint arXiv:2305.04676*, 2023.

[39] Aditya Parikh. Information extraction from unstructured data using augmented-ai and computer vision, 2023.

[40] Ramon Pires, Fábio C. de Souza, Guilherme Rosa, Roberto A. Lotufo, and Rodrigo Nogueira. Sequence-to-sequence models for extracting information from registration and legal documents, 2022.

[41] Klim Zaporojets. Neural approaches to entity-centric information extraction, 2023.

[42] Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova. Leveraging large language models for topic classification in the domain of public affairs, 2023.

[43] Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024.

[44] Yiyi Tang, Ziyan Xiao, Xue Li, Qingpeng Zhang, Esther W Chan, Ian CK Wong, and Research Data Collaboration Task Force. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: A comparative study of gpt-3.5 and gpt-4. *medRxiv*, pages 2024–03, 2024.

[45] Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, and Ido Dagan. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts, 2024.

[46] Anushka Swarup, Avanti Bhandarkar, Olivia P. Dizon-Paradis, Ronald Wilson, and Damon L. Woodard. Maximizing relation extraction potential: A data-centric study to unveil challenges and opportunities, 2024.

[47] Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!, 2023.

[48] Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lunyiu Nie, Linmei Hu, Lei Hou, and Juanzi Li. How proficient are large language models in formal languages? an in-depth insight for knowledge base question answering, 2024.

[49] Pranav Janjani, Mayank Palan, Sarvesh Shirude, Ninad Shegokar, Sunny Kumar, and Faruk Kazi. Converging dimensions: Information extraction and summarization through multisource, multimodal, and multilingual fusion, 2024.

[50] Atin Sakkeer Hussain and Anu Thomas. Large language models for judicial entity extraction: A comparative study, 2024.

[51] Cheonsu Jeong. Fine-tuning and utilization methods of domain-specific llms, 2024.

[52] Bibiána Lajčinová, Patrik Valábek, and Michal Spišiak. Intent classification for bank chatbots through llm fine-tuning, 2024.

[53] Krishna Sayana, Raghavendra Vasudeva, Yuri Vasilevski, Kun Su, Liam Hebert, James Pine, Hubert Pham, Ambarish Jash, and Sukhdeep Sodhi. Beyond retrieval: Generating narratives in conversational recommender systems, 2024.

[54] Amirhossein Herandi, Yitao Li, Zhanlin Liu, Ximin Hu, and Xiao Cai. Skill-llm: Repurposing general-purpose llms for skill extraction, 2024.

[55] Huichen Yang, Carlos A. Aguirre, Maria F. De La Torre, Derek Christensen, Luis Bobadilla, Emily Davich, Jordan Roth, Lei Luo, Yihong Theis, Alice Lam, T. Yong-Jin Han, David Buttler, and William H. Hsu. Pipelines for procedural information extraction from scientific literature: Towards recipes using machine learning and data science, 2019.

[56] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language, 2023.

[57] Fabián Villena, Luis Miranda, and Claudio Aracena. llmner: (zero|few)-shot named entity recognition, exploiting the power of large language models, 2024.

[58] Marcel Lamott and Muhammad Armaghan Shakir. Leveraging distillation techniques for document understanding: A case study with flan-t5, 2024.

[59] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models (llms), 2024.

[60] Bruno Cabral, Daniela Claro, and Marlo Souza. Exploring open information extraction for portuguese using large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 127–136, 2024.

[61] Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. Openllm-ro – technical report on open-source romanian llms, 2024.

[62] Yida Cai, Hao Sun, Hsiu-Yuan Huang, and Yunfang Wu. Assessing the performance of chinese open source large language models in information extraction tasks, 2024.

[63] Samira Ghodratnama and Mehrdad Zakershahrak. Adapting llms for efficient, personalized information retrieval: Methods and implications. In *International Conference on Service-Oriented Computing*, pages 17–26. Springer, 2023.

[64] Patrizio Bellan, Mauro Dragoni, and Chiara Ghidini. Leveraging pre-trained language models for conversational information seeking from text, 2022.

[65] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. Disinformation capabilities of large language models, 2024.

[66] Emily Reif, Minsuk Kahng, and Savvas Petridis. Visualizing linguistic diversity of text datasets synthesized by large language models, 2023.

[67] Keheng Wang, Feiyu Duan, Peiguang Li, Sirui Wang, and Xunliang Cai. Llms know what they need: Leveraging a missing information guided framework to empower retrieval-augmented generation, 2024.

[68] Yi Xu, Bo Xue, Shuqian Sheng, Cheng Deng, Jiaxin Ding, Zanwei Shen, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. Good idea or not, representation of llm could tell, 2024.

[69] S. Nishio, H. Nonaka, N. Tsuchiya, A. Migita, Y. Banno, T. Hayashi, H. Sakaji, T. Sakumoto, and K. Watabe. Extraction of research objectives, machine learning model names, and dataset names from academic papers and analysis of their interrelationships using llm and network analysis, 2024.

[70] K. R. Chowdhary. Natural language processing for word sense disambiguation and information extraction, 2020.

[71] Thierry Poibeau. Automatic extraction of paraphrastic phrases from medium size corpora, 2005.

[72] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "what is relevant in a text document?": An interpretable machine learning approach, 2016.

[73] Lochan Basyal and Mihir Sanghvi. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, 2023.

[74] Hamish Cunningham, Yorick Wilks, and Robert J. Gaizauskas. New methods, current trends and software infrastructure for nlp, 1996.

[75] Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges, 2024.

[76] Felipe Urrutia, Cristian Buc, and Valentin Barriere. Deep natural language feature learning for interpretable prediction, 2023.

[77] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models, 2022.

[78] Julian Neuberger, Lars Ackermann, Han van der Aa, and Stefan Jablonski. A universal prompting strategy for extracting process model information from natural language text using large language models, 2024.

[79] Sara Court and Micha Elsner. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem, 2024.

[80] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[81] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future, 2024.

[82] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[83] Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text, 2025.

[84] Dragomir R. Radev and Kathleen R. McKeown. Building a generation knowledge source using internet-accessible newswire, 1997.

[85] Melanie McGrath, Harrison Bailey, Necva Bölücü, Xiang Dai, Sarvnaz Karimi, and Cecile Paris. Can ai extract antecedent factors of human trust in ai? an application of information extraction for scientific literature in behavioural and computer sciences, 2024.

[86] Zach Johnson and Jeremy Straub. Development of regai: Rubric enabled generative artificial intelligence, 2024.

[87] Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, 2024.

[88] Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. On learning to summarize with large language models as references, 2024.

[89] Panagiotis Koletsis, Panagiotis-Konstantinos Gemos, Christos Chronis, Iraklis Varlamis, Vasilis Efthymiou, and Georgios Th. Papadopoulos. Entity extraction from high-level corruption schemes via large language models, 2024.

[90] Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K. Keloth, Vincent J. Zhang, Ruey-Ling Weng, Qingyu Chen, Xiaoqian Jiang, Kirk E. Roberts, and Hua Xu. Information extraction from clinical notes: Are we ready to switch to large language models?, 2025.

[91] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning, 2023.

[92] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut, 2024.

[93] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[94] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling, 2024.

[95] Haolun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. Learning to extract structured entities using language models, 2024.

[96] Ritabrata Roy Choudhury and Soumik Dey. Gpt-3 powered information extraction for building robust knowledge bases, 2024.

[97] Xin Cong. Bowen Yu, Mengcheng Fang, Tingwen Liu, Haiyang Yu, Zhongkai Hu, Fei Huang, Yongbin Li, and Bin Wang. Universal information extraction with meta-pretrained self-retrieval, 2023.

[98] Marcos Fernández-Pichel, Juan C. Pichel, and David E. Losada. Search engines, llms or both? evaluating information seeking strategies for answering health questions, 2024.

[99] Claudio Crema, Tommaso Mario Buonocore, Silvia Fostinelli, Enea Parimbelli, Federico Verde, Cira Fundarò, Marina Manera, Matteo Cotta Ramusino, Marco Capelli, Alfredo Costa, Giuliano Binetti, Riccardo Bellazzi, and Alberto Redolfi. Advancing italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application, 2024.

[100] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. Llms accelerate annotation for medical information extraction, 2023.

[101] Vincent Freiberger and Erik Buchmann. Fairness certification for natural language processing and large language models, 2024.

[102] Xuanfan Ni, Piji Li, and Huayang Li. Unified text structuralization with instruction-tuned language models, 2023.

[103] Chengyuan Liu, Yangyang Kang, Fubang Zhao, Kun Kuang, Zhuoren Jiang, Changlong Sun, and Fei Wu. Evolving knowledge distillation with large language models and active learning, 2024.

[104] Sergio Consoli, Peter Markov, Nikolaos I. Stilianakis, Lorenzo Bertolini, Antonio Puertas Gallardo, and Mario Ceresa. Epidemic information extraction for event-based surveillance using large language models, 2024.

[105] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model, 2023.

[106] Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. Not all layers of llms are necessary during inference, 2024.

[107] Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, and Ross D. King. Scientific hypothesis generation by a large language model: Laboratory validation in breast cancer treatment, 2024.

[108] Aishik Nagar, Viktor Schlegel, Thanh-Tung Nguyen, Hao Li, Yuping Wu, Kuluhan Binici, and Stefan Winkler. Llms are not zero-shot reasoners for biomedical information extraction, 2024.

[109] Bumjin Park and Jaesik Choi. Memorizing documents with guidance in large language models, 2024.

[110] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey, 2024.

[111] Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. "im not racist but...": Discovering bias in the internal knowledge of large language models, 2023.

[112] Hosein Hasanbeig, Hiteshi Sharma, Leo Betthauser, Felipe Vieira Frujeri, and Ida Momennejad. Allure: Auditing and improving llm-based evaluation of text using iterative in-context-learning, 2023.