

Exploring ChatGPT for Next-generation Information Retrieval: Opportunities and Challenges

Yizheng Huang and Jimmy X. Huang*

Information Retrieval and Knowledge Management Research Lab, York University, Toronto, Canada

E-mails: hyz@yorku.ca, jhuang@yorku.ca

Abstract. The rapid advancement of artificial intelligence (AI) has highlighted ChatGPT as a pivotal technology in the field of information retrieval (IR). Distinguished from its predecessors, ChatGPT offers significant benefits that have attracted the attention of both the industry and academic communities. While some view ChatGPT as a groundbreaking innovation, others attribute its success to the effective integration of product development and market strategies. The emergence of ChatGPT, alongside GPT-4, marks a new phase in Generative AI, generating content that is distinct from training examples and exceeding the capabilities of the prior GPT-3 model by OpenAI. Unlike the traditional supervised learning approach in IR tasks, ChatGPT challenges existing paradigms, bringing forth new challenges and opportunities regarding text quality assurance, model bias, and efficiency. This paper seeks to examine the impact of ChatGPT on IR tasks and offer insights into its potential future developments.

Keywords: Information Retrieval, ChatGPT, Large Language Models

1. Introduction

On November 30, 2022, OpenAI unveiled ChatGPT¹, an AI chatbot application powered by the advanced GPT-3.5 and later GPT-4 generative language models. This application quickly attracted over a hundred million users worldwide, setting a new record for rapid product dissemination [1]. ChatGPT, as an embodiment of these models, demonstrated significant advancements over its predecessors, quickly becoming a central topic in both industrial and academic circles. While some view ChatGPT as a disruptive technological innovation, predicting revolutionary changes in various sectors, others believe its success stems more from effective product and market strategies than from purely technological breakthroughs.

Indeed, ChatGPT heralded a new phase in Generative AI, distinct from previous models like GPT-3. This new generation of AI models, including ChatGPT, is capable of generating unique content, not just refining or predicting information based on training examples [2]. GPT-3.5 established a strong foundation with its robust capabilities [3–5]. GPT-4 further expanded these capabilities, offering enhanced understanding, accuracy, and contextual relevance. The evolution from GPT-3.5 to GPT-4 has shown great promise in numerous information retrieval tasks (e.g. [6,

*Corresponding author. E-mail: jhuang@yorku.ca.

¹<https://chat.openai.com>

Table 1
Comparison of pre-trained large language models in recent years.

Pre-trained Language Models	Release Data	Size of Pre-training Corpus	Parameters Size
BERT-Large [12]	2018-10	16 GB	340M
GPT-2 [13]	2019-02	40GB	1.5B
RoBERTa [14]	2019-07	161 GB	340M
XLNet-Large [15]	2019-07	142 GB	340M
T5-11B [16]	2019-10	750 GB	11B
OPT [17]	2020-05	180B tokens	175B
GPT-3 [4]	2020-06	45TB	175B
mT5-XXL [18]	2020-10	750 GB	13B
ERNIE 3.0 [19]	2021-07	375B tokens	10B
Yuan 1.0 [20]	2021-10	180B tokens	245B
PaLM [21]	2022-04	780B tokens	540B
BLOOM [22]	2022-11	366B tokens	176B
GPT-4 [1]	2023-04	About 13T tokens	About 1.76T
PaLM2 [21]	2023-05	100B tokens	16B
LlaMA2 [23]	2023-07	2T tokens	70B
Qwen-14B [24]	2023-09	2.4T tokens	14B
Skywork [25]	2023-10	3.2T tokens	13B

7)), particularly in text classification [8], document ranking [9], question-answering systems [10], and multimodal retrieval [11]. The introduction of ChatGPT, leveraging these advancements, has spurred progress in this field, highlighting the impressive abilities of large language models (LLMs) in understanding and generating semantic information.

Amid these rapid technological developments, ChatGPT has been applied in various practical settings. Notably, it powers Microsoft’s AI-driven search engine, New Bing, based on GPT-4², and integrates with other multimodal pre-trained models, enhancing the scope of IR tasks. Traditionally, supervised learning has been the main approach in IR, involving the design of statistical or probabilistic models trained on specific task-related data, parameter optimization through loss function minimization, and model inference on new data. The advent of deep neural networks shifted the focus from traditional machine learning models to deep learning models. However, the reliance on the supervised learning framework persisted. This method, training models on labeled datasets to predict or categorize unseen data, has driven significant progress in various IR applications. Nonetheless, the emergence of ChatGPT and the GPT-X models it is based on (where X represents different versions) has posed new challenges to existing IR paradigms, introducing research and application issues such as ensuring text quality, addressing model bias and ethical concerns, and improving model efficiency and practicality.

This paper delves into the opportunities and challenges brought forth by ChatGPT in IR tasks. We also offer a forward-looking view on the future development of ChatGPT and its underlying GPT-X models, aiming to provide valuable insights for research and applications in related fields.

2. Pretrained Large Language Models

The field of information retrieval has experienced a remarkable transformation with the emergence of pretrained large language models (PLLMs). This evolution, progressing from initial simplistic models to the current advanced dense retrieval models, has significantly broadened the scope and capabilities of IR and related fields. A comparison of recent pre-trained language models, including their training datasets and parameter size, can be seen in Table 1.

²<https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web>

Early Language Models The era of language models began with statistical approaches, notably n-gram models. These models predicted subsequent words based on the probability distribution of word sequences in sentences. The field advanced with the introduction of neural network-based models, such as the Neural Probabilistic Language Model [26], marking a new phase in language modeling. Following this, architectures like Convolutional Neural Networks (CNNs) [27], Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs) [28] emerged. These architectures addressed issues like data sparsity and capturing long-term dependencies but faced challenges in processing long sequences and parallelization.

The Transformer Paradigm A significant breakthrough occurred with the introduction of the Transformer architecture by Google in 2017 [29]. This architecture, featuring self-attention mechanisms, enabled efficient parallel processing of sequences and effective management of long-term dependencies, overcoming many limitations of previous models.

The evolution of OpenAI's Generative Pre-trained Transformer (GPT) series is a testament to the success of the Transformer architecture. GPT-1 laid the foundation, and subsequent versions, GPT-2 and GPT-3, dramatically expanded the scale and capabilities of these models. Notably, GPT-3, with its 175 billion parameters, demonstrated an impressive leap in generating human-like text and facilitating meaningful interactions.

ChatGPT: The New Frontiers Building on GPT-3, OpenAI developed ChatGPT based on the GPT-3.5 architecture. This model was specifically designed to overcome certain limitations of GPT-3, particularly in producing coherent and contextually relevant responses over extended dialogues. The training of ChatGPT involved a novel approach, Reinforcement Learning from Human Feedback (RLHF) [30], involving multiple iterations of model refinement using a reward model created from human-ranked responses.

ChatGPT represents a significant advancement in creating models capable of more meaningful and context-aware user interactions. Its deployment has demonstrated potential for a wide range of real-world applications, as highlighted by recent studies and deployments [31–35].

Training Methodologies of ChatGPT ChatGPT's architecture, based on the transformer model, includes specific modifications to enhance conversational abilities. The RLHF training method is notable, involving human trainers who guide the model by ranking responses, thereby refining the model's capability to generate contextually appropriate responses. This training also incorporates safety and bias reduction measures, ensuring adherence to ethical guidelines.

Interaction Mechanisms with Prompts ChatGPT's interaction with prompts involves understanding user input's intent and context. It generates responses that are relevant, coherent, and contextually suitable by combining learned patterns from its training data with real-time input processing. This process also includes managing ambiguous or incomplete information and maintaining context over a conversation.

GPT4: Advancing ChatGPT's Capabilities GPT-4, released after ChatGPT and GPT-3.5, further pushes the boundaries of PLLMs. With an extended context window and hypothesized multimodal capabilities, GPT-4 is posited to surpass GPT-3.5 in many respects, potentially matching or exceeding human performance in various tasks. Its extensibility is evident in integrations and new services like Microsoft's Copilot³, enhancing productivity tools.

ChatGPT in IR ChatGPT significantly contributes to IR by understanding and responding to queries using its extensive internal knowledge base. Unlike traditional search engines, ChatGPT simplifies the user experience by generating useful answers without requiring users to have specific knowledge, making it an invaluable tool for various tasks. A typical scenario is ChatGPT's robustness in understanding queries that contain grammatical or spelling errors. Even when a user submits a query with such inaccuracies, ChatGPT effectively interprets the intended meaning and provides responses incorporating correct grammar and spelling. This feature enhances user experience, ensuring that communication barriers due to language proficiency or typing errors do not hinder the retrieval of accurate and relevant information.

³<https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work>

Table 2
Comparison of ChatGPT, Llama-2, Bard, and Claude.

Models	Company	Architecture	Notable Strengths	Notable Weaknesses
ChatGPT	OpenAI	Generative Pre-trained Transformer (GPT)	Creative text generation Scalability (e.g. integration with DALL-E)	Generate incoherent or incorrect text
Llama-2	Meta	Auto-regressive Language Optimized Transformer	Range of parameter sizes (7B, 13B, and 70B) Fine-Tuned Versions	No convenient bot-like interface
Bard	Google	Pathways Language Models (PaLM2) [21]	Faster Coherent responses	Significant creative limitations
Claude	Anthropic	Not fully disclosed	Large token capacity Reduced hallucinations	Strict censorship

Other Noteworthy Models The PLLM landscape features several key players besides the GPT series, as shown in Table 2. ChatGPT is renowned for its creative text generation and remarkable scalability with plugins. Nevertheless, it has the problem of producing incoherent or incorrect text. Meta’s LLaMA-2⁴ has a range of parameter sizes and offers versions that are fine-tuned for specific tasks. Despite the various parameters available, the model lacks a user-friendly bot interface, limiting access to the normal user. Google’s Bard⁵ stands out for its ability to respond consistently to varied queries but with limited creativity. Lastly, Anthropic’s Claude⁶, while not fully disclosing its architecture, has drawn attention for its extensive token capacity, facilitating the processing and generation of lengthy and complex texts. In addition, Claude is committed to reducing the generation of false or misleading information. However, it operates under a strict content review strategy, which may restrict access to legitimate information, particularly in fields like scientific research. These models reveal unique strengths and challenges, contributing to the dynamic PLLM field.

3. Potential Opportunities in Information Retrieval with ChatGPT

In the era of large models, generative models represented by ChatGPT are introducing new perspectives and methodologies for the core task of information retrieval. IR systems aim to extract relevant information from enormous amounts of textual data. Traditional IR systems often rely on keyword matching. However, with the advent of neural networks and deep learning, IR is progressively evolving towards semantic-based retrieval [36].

The deep neural networks of GPT-X enable a profound understanding of text semantics, enhancing the precision in semantic-level retrieval beyond traditional keyword-level text matching. Their generative framework allows for the formulation of precise query expressions and the generation of descriptive retrieval results, enhancing the flexibility and expressiveness of IR. With zero or few-shot learning capabilities where models require little to no training data, these models reduce the necessity for extensive annotated data, making complex retrieval tasks more manageable. The end-to-end training methodology minimizes error propagation and directly optimizes performance from input to output, improving retrieval accuracy and efficiency. Furthermore, the potential for multimodal information retrieval extends the scope beyond text to encompass images and videos, offering richer and more accurate retrieval results. Lastly, integrating knowledge graphs leverages structured knowledge in the retrieval process, simultaneously aiding in the construction and updating of knowledge graphs, thus providing a richer knowledge base for IR.

3.1. Information Extraction

Information Extraction (IE) is a fundamental task in information retrieval, encompassing sub-tasks such as named entity recognition (NER) and event extraction (EE). IE has evolved significantly over the years. Initially, the focus was on structured and semi-structured data extraction, employing various techniques, tools, and systems to extract useful information automatically [37]. Early IE systems were primarily rule-based, relied on a large amount of human involvement, and were tailored for specific domains like chemical or medical search [38–43].

⁴<https://ai.meta.com/llama>

⁵<https://bard.google.com>

⁶<https://www.anthropic.com/index/introducing-claude>

Transitioning into the contemporary period, the field has seen a shift towards employing deep learning technologies, which excel at extracting structured information from unstructured text without being confined to a particular domain [44, 45]. The core idea of deep learning is to extract features from the original data, moving from low-level to high-level and from the concrete to the abstract through a series of non-linear transformations in a data-driven manner. These methods have significantly improved the advanced levels of various fields, including speech recognition, visual object recognition, and object detection, showcasing the efficacy of deep learning in handling complex IE tasks [46].

Moreover, researchers hope these large-scale language models can process text efficiently and extract valuable information without the necessity for retraining, potentially replacing manual annotation. However, multiple extensive IE experiments on ChatGPT show a significant performance gap between ChatGPT and state-of-the-art (SOTA) results on datasets with zero/few-shot IE sub-tasks [47–50]. Although the results are unsatisfactory, they spark new research perspectives in IE, such as the possibility that IE tasks can be decomposed into multiple simpler subtasks [47], a rethinking of the evaluation strategy might reflect a more accurate performance of ChatGPT [48], and ChatGPT’s performance can be significantly improved by prompt engineering [49].

3.2. Text Classification

In exploring text classification tasks in the era of large language models, it’s pertinent first to introduce the traditional and prevalent methodologies in text classification. Traditional text classification approaches generally rely on statistical learning paradigms such as Naive Bayes and K-Nearest Neighbors [51–53]. These methods entail substantial effort in feature engineering to construct meaningful representations of text. Subsequently, with the advent of deep neural networks, models like RNNs, CNNs, and Graph Neural Networks (GNNs) [54] have emerged as mainstream paradigms, significantly automating the construction of rich semantic representations of text.

Entering the era of LLMs, models like ChatGPT have markedly impacted text classification tasks. These models achieve high-quality text semantic modeling from massive text corpora through supervised pre-training techniques, substantially enhancing the performance in text classification tasks. Particularly in addressing open-domain tasks, domain adaptation, few-shot (where models learn from a small set of labeled examples), and zero-shot (where models generalize to unseen classes) problems, these large models exhibit impressive performance and exceptional generalization capabilities [55–58].

ChatGPT can be utilized to undertake a knowledge graph extraction task to obtain refined and structured knowledge from raw data. The collected knowledge is then transformed into a graph, which is subsequently utilized to train an interpretable linear classifier to render predictions, exhibiting impressive performance [59].

In scenarios with few or zero examples, LLMs leverage pre-trained knowledge to achieve satisfactory classification outcomes, mitigating the dependency on large labeled datasets inherent in traditional methods. This capability is invaluable in domains encumbered by limited training data due to costly and labor-intensive annotation processes [60]. In addition, high-quality categorization lays a solid foundation for accurate and efficient annotation, thus potentially speeding up the annotation process, reducing costs, and improving the overall quality of the annotated data, which greatly benefits the text annotation task [61].

From an information retrieval perspective, text classification serves as a crucial mechanism for ranking and categorizing textual data, aiding in the efficient retrieval and management of information. Combining the knowledge graph and few-shot learning capabilities based on LLMs, text classification tasks can extract and utilize relevant information from extensive data, achieving more accurate and efficient categorization.

3.3. Document Ranking

Document ranking is a crucial process in information retrieval systems, determining the order in which retrieved documents are presented based on their estimated relevance to a query. Historically, the methodologies employed for document ranking have predominantly centered on term-based matching, leveraging standard techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and BM25 [62]. These traditional approaches assess the significance of terms within documents and their corresponding relevance to the query at hand [63, 64]. However,

they often fall short in capturing the semantic relationships between terms and may overlook contextual relevance, which is increasingly important in refining the precision of document retrieval.

Transitioning into the modern era, machine learning has found a foothold in document ranking through methods like Learning to Rank [65], which predicts a relevance score for each document-query pair, ranking documents accordingly. Thereafter, deep learning models started gaining traction. CNNs, RNNs, and attention-based mechanisms such as BERT [12] have been employed to enhance the representation of text data and improve the understanding of natural language queries [66]. Recently, the focus has also shifted towards dense retrieval and re-ranking models [67]. Dense retrieval models propose a more accurate approach to document ranking tasks by embedding both documents and queries in a continuous vector space. Re-rankers take an initial set of retrieved candidates and re-sort them based on relevance scores, ensuring a more reliable list of results in response to a query.

The advent of large language models has opened new possibilities for document ranking in IR. Investigations have revealed that ChatGPT can deliver competitive or even superior ranking performance compared to supervised methods on popular IR benchmarks when properly instructed [68, 69]. The emergence of GPT-4 has further pushed the boundaries, showcasing AI-driven document ranking, significantly impacting the search engine domain [69]. In addition, a human-involved experiment comparing the search performance and user experience of ChatGPT and Google Search points to practical insights. Although ChatGPT cannot always outperform Google Search, it considerably enhances work efficiency and increases user satisfaction [70].

Furthermore, domain-specific document ranking emerges as a promising area for the application of GPT-4. Presently, ranking methods heavily rely on training data and fine-tuning. However, the scarcity of high-quality annotated datasets in specialized domains such as medicine and law poses a significant challenge, impeding the efficacy of deploying pre-trained models for ranking documents [71]. LLMs like GPT-4, endowed with expansive knowledge and pronounced generalization capability due to their vast training data spectrum, present a viable solution. These models hold the potential to serve as data augmentation tools in such contexts, synthesizing pseudo-label data that could improve the performance of retrieval models in data-scarce situations [72, 73]. By generating synthetic yet relevant data, GPT-4 could significantly enhance the model's ability to accurately rank documents in domain-specific scenarios, thereby bridging the data gap and facilitating improved performance in document retrieval tasks.

3.4. Conversational Search

Conversational Search (CS) has significantly evolved over the years, transitioning from rule-based models to the more advanced machine learning and deep learning models prevalent today [74, 75]. Traditionally, it is divided into two main subtasks: task-oriented and open-dialog/interactive tasks. Task-oriented conversational IR (Information Retrieval) systems employed a pipeline approach, integrating several modules like intent recognition, dialogue management, and response generation to handle user interactions [76]. Conversely, open-domain conversational IR systems aim to engage users in more social and less goal-directed conversations. Initially, these systems relied on retrieval-based approaches, but the advent of generative models allowed for more fluid and natural responses [77, 78]. They function like an IR system, extracting related information from a pre-designed database.

The introduction of transformer-based models, such as OpenAI's ChatGPT, Anthropic AI's Claude, and Google's LaMDA [79], marked a paradigm shift in the domain of CS. These models' capability to generate human-like text based on a given context has expanded the horizons of what's possible in task-oriented and open-domain CS systems.

Several opportunities arise as the field advances with contributions from models like ChatGPT and GPT-4. Thanks to these models' impressive intent understanding, semantic parsing, and API integration capabilities, the union of task-oriented and open-domain dialogues under a single technical framework is now attainable. This union could lead to the development of CS systems that are not only functional but also emotionally intelligent, catering to the practical needs of users. Moreover, the pursuit of creating more personalized CS systems remains a significant area of research and development. Advancements in these areas are expected to push CS systems closer to delivering a truly human-like and enriching conversational experience.

3.5. Multimodal Retrieval

In the realm of multimodal retrieval, the transition from traditional methods to cutting-edge techniques showcases a remarkable development. Initially, traditional multimodal retrieval predominantly fell under the Nearest Neighbor (NN) problem [80]. However, these methods struggled to bridge the semantic gap between low-level features (such as color, texture, and shape) and users' high-level informational needs. As technology advanced, the focus shifted towards crafting unified representations for data across different modalities, such as text, images, audio, and video, aiming to foster seamless and enriched interactions between these modalities [81].

The field then embraced cross-modal retrieval, emphasizing the importance of modeling relationships between different modalities. This approach allowed users to retrieve desired information by submitting data in one modality to fetch related data in another, marking a significant stride towards enhancing accuracy and scalability in retrieval [82, 83]. Additionally, the emergence of retrieval-augmented multimodal models began integrating external knowledge more scalably and modularly. For a given input text, such models use a retriever to fetch relevant documents from external sources and a generator (often a language model) to produce predictions based on the acquired information. Typically, these external sources include text corpora and structured knowledge bases. However, retrieval-augmented methods were initially researched for text, and extending them to the multimodal domain remains challenging. The main difficulty lies in the design of the retriever and generator that can handle multimodal documents containing both images and text.

Addressing this challenge, the Retrieval-Augmented Text-to-Image Generator (Re-Imagen) [84] represents a significant advancement. Utilizing a diffusion-based method, this model generates high-fidelity images that are remarkably accurate, even when depicting entities not previously encountered. The process hinges on the effective use of information retrieved from external sources, enabling the creation of visually precise representations. Similarly, the Multimodal Retrieval-Augmented Transformer (MuRAG) [85] focuses on answering natural language questions using image retrieval methods. Although these works concentrate on generating single modalities (text or image), RA-CM3 [86] proposed a comprehensive and unified model capable of retrieving and generating both images and text. Notably, the generator model develops capabilities such as controlled image generation in a contextual learning framework through retrieval-enhanced training.

The debut of GPT-4 notably impacted the field of multimodal retrieval, ushering in an era closer to human-like AI. GPT-4 is a large multimodal model capable of processing both text and image inputs while delivering text outputs, pushing closer to human-level performance on various benchmarks, albeit with certain limitations in real-world scenarios [87]. Conversely, ChatGPT has been empowered by GPT-4V(ision) [88], boosting its multimodal capabilities. For instance, the integration of DALL-E 3⁷ with ChatGPT facilitates smoother interaction, where ChatGPT aids in crafting precise prompts for DALL-E 3, turning user ideas into vibrant AI-generated art.

The arrival of large language models marked a significant milestone in bridging the semantic gap between multiple types of information, paving the way for more intuitive and rich interactions across diverse data modalities. In recommendation systems, LLMs have shown immense promise [89]. They foster a more comprehensive understanding of user preferences and behaviors by integrating information from various sources and modalities. For example, a recommendation system powered by a multimodal LLM can analyze textual reviews, image-based preferences, and purchase histories to generate more accurate and personalized product recommendations. Moreover, by understanding the semantic relationships between different items and user interactions, these models can provide a more enriched and personalized user experience, thereby enhancing user satisfaction and engagement.

Similarly, the medical field has seen substantial advancements by incorporating LLMs [90, 91]. In clinical settings, multimodal LLMs can assist in synthesizing information from diverse sources such as electronic health records, medical imaging, and genomic data to provide more comprehensive and personalized insights. This holds vast potential to support diagnostic processes, treatment planning, and personalized medicine. For instance, integrating textual clinical notes with medical imaging data can empower clinicians with a more holistic understanding of a patient's condition, enabling better-informed decision-making.

⁷<https://openai.com/dall-e-3>

4. Unresolved Challenges in Information Retrieval with ChatGPT

ChatGPT proves the duality of technological advances in AI. On the one side, it can greatly enhance the productivity of users from all walks of life thanks to its excellent language comprehension and generation capabilities. Whether in education, business, or personal assistance, ChatGPT is a powerful tool that facilitates task completion, inspires creativity, and helps make correct decisions.

On the flip side, it reveals the ethical dilemmas associated with misinformation, disinformation, and the potential misuse possibilities of fabricating deceptive or harmful content. Its remarkable ability to produce realistic text blurs the boundaries of information authenticity, making it challenging for individuals to discern between real and fake content. These potential risks highlight the limitless possibilities of ChatGPT while also emphasizing the need to navigate the ethical regulation that accompanies such groundbreaking innovations.

4.1. Hallucination

The challenge of hallucination in large language models, underscored by Google AI researchers in 2018 [92], presents a formidable hurdle to their deployment. Hallucination, a phenomenon where models generate convincing yet factually incorrect or misleading content, harbors serious risks. This is particularly concerning in critical applications such as decision-making, where the propagation of false information can lead to adverse outcomes [93, 94]. OpenAI, the developer of ChatGPT, has recognized the concerns regarding the model's propensity for factual inaccuracies and is actively pursuing measures to mitigate this issue⁸.

Information retrieval strategies are poised to be instrumental in addressing the hallucination challenge. A viable approach could be establishing a continuous feedback loop wherein the model's outputs are rigorously evaluated, and refinements are made based on identified inaccuracies. This iterative process aims to bolster the model's accuracy and reliability over time. Specifically, integrating IR models to work in tandem with LLMs could present a robust solution [95]. By augmenting LLMs with updated and accurate information extracted from external sources, IR models can potentially curtail the generation of factually inaccurate responses, thus mitigating the occurrence of hallucinations.

4.2. Ethical Issues and Safety

The ethical and safety concerns surrounding ChatGPT are multi-faceted, arising from their profound language understanding and generation capabilities. As advanced iterations of language models developed by OpenAI, these models harbor significant expectations alongside concerns due to their potential transformative impact on society [96].

The expansive training data and the complex nature of these models introduce risks associated with bias and fairness. The training material, sourced from human-generated content, may inadvertently perpetuate existing societal biases. Instances where models exhibited gender or racial biases are emblematic of this problem. These biases can manifest across various applications, potentially leading to unfair or discriminatory outcomes [97].

Moreover, the emergence of generative AI poses challenges related to misinformation and abuse. Their ability to generate text can be leveraged to fabricate misleading information, contribute to online misinformation campaigns, or even generate harmful or abusive content. The lack of source attribution in responses generated by ChatGPT exacerbates this issue, as users may struggle to discern the veracity of the generated content [98]. The potential misuse of generative AI for criminal activities such as fraud or harassment is another significant concern. LLMs can be employed to create realistic fake content for nefarious purposes, thereby reducing costs and increasing the efficiency of executing fraudulent activities.

In an IR system, while relevance is often prioritized, this can lead to insufficient diversity in the results [42, 43]. Frequently, the most prominent subtopic groups dominate the search results, marginalizing minority topics. This imbalance can cause users to exert extra effort to find items related to less common topics, leading to a partial and skewed information result. Moreover, the tendency of users to click primarily on top search results can facilitate

⁸<https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai>

a cycle of unfairness. Ranking or recommendation algorithms incorporating user feedback tend to maintain these items' top positions, creating a positive feedback loop for unfairness issues. This is particularly problematic in systems like ChatGPT, where an initial response with an ethical issue can be challenging to correct internally. If users trust such a response, repeated interactions can exacerbate the issue, contributing to the development of bias. Quantifying bias in terms of gender and age can be beneficial to address these challenges [99]. A fairness-aware ranking algorithm that accounts for these factors can have a positive impact. Additionally, considering fairness as an optimization problem opens up new approaches [100]. Implementing a fairness-constrained reinforcement learning algorithm can help balance relevance with the need for diversity and fairness in IR systems [101].

Furthermore, the scalability of these models amplifies privacy concerns. As models become larger and necessitate more computational resources, the need to offload processing to cloud servers escalates. This centralization can heighten the risk of data breaches and misuse of personal information, especially if adequate measures are not in place to secure user data.

Overall, the array of ethical and safety concerns emanating from the deployment of ChatGPT underscores the imperative of diligent oversight, robust regulatory frameworks, and continuous dialogue among stakeholders to ensure the responsible development and use of these transformative technologies.

4.3. Interpretability

As language models become more complex with increased parameters and depth, their decision-making processes become less interpretable. This complexity also challenges understanding the vector and parameter representations within deep neural networks [102].

Characterizing large language models as “black-box” models summarizes this fundamental challenge in deploying and trusting these systems [103]. While the user can observe the inputs and outputs, the intricacies of the processes in between remain hidden, preventing a clear understanding of how the model derives a particular output from a given input. This opacity extends to an inability to discern what aspects of the input data the model considers important, obscuring interpretability.

The main reason for this challenge is that while LLMs are good at recognizing patterns and correlations in data, they lack a grasp of causality [104]. This inadequacy is particularly evident in decision-making. Moreover, LLMs are prone to inherit biases present in the training data, which highlights another dimension of the interpretability challenge. Any bias may permeate the model's behavior, leading to anomalous or unfair results. Diagnosing and mitigating these biases becomes difficult without a clear window into the model's inner workings.

The challenge of interpretability is further exacerbated by the unpredictability of LLMs in the face of new or adversarial inputs. These models may exhibit erratic behavior in the face of unexpected input scenarios, which is difficult to address without an interpretability framework. Improving the interpretability of LLMs is, therefore, not just an academic exploration but a pragmatic need to ensure responsible and credible deployment of these models, especially as they enter increasingly sensitive and critical domains. Uncovering the “black box” nature of LLM and building robust interpretability frameworks is necessary for developing machine learning and AI.

Retrieval-Enhanced Machine Learning [11] presents a promising approach to addressing the issue of interpretability. In pre-trained language models, the training knowledge is embedded within the learned model parameters, making it difficult to understand model predictions. In contrast, when the reasoning process relies on retrieved information, predictions can be directly linked to specific data, typically stored in an accessible text format. This feature improves the interpretability of the model's outputs. Additionally, Aspect Learning [105] can further enhance interpretability. By incorporating aspects, the model not only grasps general language semantics, like other pre-trained models, but also acquires domain-specific knowledge, enabling it to identify aspects relevant to a particular domain. These “explicit aspects” significantly improve interpretability, as the retrieved documents are expected to share similar aspects (or categories) with the input query.

OpenAI has initiated efforts to automate the interpretability of large language models by using GPT-4 itself to generate and score explanations of neuron behavior in other language models [106]. This initiative aims to uncover how different parts of the neural network operate, although the technique still struggles with larger models, indicating room for improvement. The initiative by OpenAI represents a significant stride towards demystifying the operations of LLMs, hoping to foster more responsible and effective use of these powerful tools in various domains.

5. Conclusions and Future Directions

ChatGPT signifies a remarkable stride in Generative AI, enriching multiple information retrieval tasks. They excel in understanding and generating textual content, with applications extending to various practical and academic domains such as healthcare, education, and programming, thus reshaping traditional paradigms. However, this advancement isn't without challenges.

Ethical dilemmas such as misinformation, disinformation, and potential misuse for harmful content generation pose serious concerns. The issue of hallucination, generating incorrect or misleading content, highlights the need for robust mechanisms to ensure accuracy and reliability. Furthermore, the challenge of interpretability remains a substantial hurdle. The “black box” nature of these models hinders transparency in their decision-making processes, which is essential for responsible AI deployment, especially in critical domains.

In addressing these challenges, recent works in IR have made strides in these areas. We note that fairness retrieval methods have shown the potential to mitigate biases in PLLMs, promoting more equitable and unbiased content generation. Additionally, the application of retrieval-enhanced learning methods has been identified as beneficial in tackling interpretability issues. By integrating context-rich information into the learning process, these methods can provide insights into the decision-making mechanisms of these complex models.

The advent of ChatGPT embodies the broader narrative of AI development, filled with promises of technological innovation and the imperative of addressing ethical, safety, and privacy challenges. Continued research and proactive steps to mitigate these challenges while exploring new ways to harness the power of these models responsibly will help navigate the complexities of AI. Collaborative efforts among researchers, practitioners, and policymakers are pivotal in realizing a future where AI significantly enhances human capabilities while preserving ethical and social values.

Acknowledgements

We express our sincere gratitude to the reviewers for their insightful comments and to the editor for their valuable assistance, both of which have significantly contributed to the enhancement of this paper. This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chairs (YRC) program.

References

- [1] OpenAI, GPT-4 Technical Report, *PREPRINT* (2023).
- [2] J. Deng and Y. Lin, The Benefits and Challenges of ChatGPT: An Overview, *Frontiers in Computing and Intelligent Systems* **2**(2) (2022), 81–83.
- [3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, Proximal Policy Optimization Algorithms, *arXiv preprint arXiv: 1707.06347* (2017).
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language Models Are Few-shot Learners, *Advances in neural information processing systems* **33** (2020), 1877–1901.
- [5] A. Neelakantan, T. Xu, R. Puri, A. Radford, J.M. Han, J. Tworek, Q. Yuan, N. Tezak, J.W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T.E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F.P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder and L. Weng, Text and Code Embeddings by Contrastive Pre-Training, *arXiv preprint arXiv: 2201.10005* (2022).
- [6] J.X. Huang, J. Miao and B. He, High Performance Query Expansion Using Adaptive Co-training, *Inf. Process. Manag.* **49**(2) (2013), 441–453. doi:10.1016/J.IPM.2012.08.002. <https://doi.org/10.1016/j.ipm.2012.08.002>.
- [7] Z. Ye, J.X. Huang and H. Lin, Finding a good query-related topic for boosting pseudo-relevance feedback, *J. Assoc. Inf. Sci. Technol.* **62**(4) (2011), 748–760. doi:10.1002/ASI.21501. <https://doi.org/10.1002/asi.21501>.
- [8] X. Huang, Y.R. Huang, M. Wen, A. An, Y. Liu and J. Poon, Applying Data Mining to Pseudo-Relevance Feedback for High Performance Text Retrieval, in: *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*, IEEE Computer Society, 2006, pp. 295–306. doi:10.1109/ICDM.2006.22.
- [9] X. Huang, M. Zhong and L. Si, York University at TREC 2005: Genomics Track, in: *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, E.M. Voorhees and L.P. Buckland, eds, NIST Special Publication, Vol. 500-266, National Institute of Standards and Technology (NIST), 2005. http://trec.nist.gov/pubs/trec14/papers/yorku-huang2_geo.pdf.

- [10] Q. Chen, Q. Hu, J.X. Huang, L. He and W. An, Enhancing Recurrent Neural Networks with Positional Attention for Question Answering, in: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, N. Kando, T. Sakai, H. Joho, H. Li, A.P. de Vries and R.W. White, eds, ACM, 2017, pp. 993–996. doi:10.1145/3077136.3080699.
- [11] H. Zamani, F. Diaz, M. Dehghani, D. Metzler and M. Bendersky, Retrieval-Enhanced Machine Learning, in: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2875–2886. ISBN 9781450387323. doi:10.1145/3477495.3531722.
- [12] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.
- [13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., Language Models Are Unsupervised Multitask Learners, *OpenAI blog* 1(8) (2019), 9.
- [14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, Roberta: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv:1907.11692* (2019).
- [15] Z. Yang, Z. Dai, Y. Yang, J.G. Carbonell, R. Salakhutdinov and Q.V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding., in: *Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 32, 2019, pp. 5754–5764.
- [16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P.J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2023.
- [17] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X.V. Lin et al., Opt: Open Ore-trained Transformer Language Models, *arXiv abs/2205.01068* (2022).
- [18] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua and C. Raffel, mT5: A Massively Multilingual Pre-trained Text-to-text Transformer, 2021.
- [19] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu and H. Wang, ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation, 2021.
- [20] S. Wu, X. Zhao, T. Yu, R. Zhang, C. Shen, H. Liu, F. Li, H. Zhu, J. Luo, L. Xu and X. Zhang, Yuan 1.0: Large-Scale Pre-trained Language Model in Zero-Shot and Few-Shot Learning, 2021.
- [21] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T.S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov and N. Fiedel, PaLM: Scaling Language Modeling with Pathways (2022).
- [22] B. Workshop, T.L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilıcak, D. Hesslow, R. Castagné, A.S. Luccioni, F. Yvon et al., Bloom: A 176b-parameter Open-access Multilingual Language Model, *arXiv abs/2211.05100* (2022).
- [23] H. Touvron, L. Martin, K.R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D.M. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A.S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I.M. Kloumann, A.V. Korenev, P.S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E.M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J.X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov and T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, *ArXiv abs/2307.09288* (2023). <https://api.semanticscholar.org/CorpusID:259950998>.
- [24] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou and T. Zhu, Qwen Technical Report, 2023.
- [25] T. Wei, L. Zhao, L. Zhang, B. Zhu, L. Wang, H. Yang, B. Li, C. Cheng, W. Lü, R. Hu, C. Li, L. Yang, X. Luo, X. Wu, L. Liu, W. Cheng, P. Cheng, J. Zhang, X. Zhang, L. Lin, X. Wang, Y. Ma, C. Dong, Y. Sun, Y. Chen, Y. Peng, X. Liang, S. Yan, H. Fang and Y. Zhou, Skywork: A More Open Bilingual Foundation Model, 2023.
- [26] Y. Bengio, R. Ducharme and P. Vincent, A Neural Probabilistic Language Model, *Advances in neural information processing systems* 13 (2000).
- [27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai et al., Recent Advances in Convolutional Neural Networks, *Pattern recognition* 77 (2018), 354–377.
- [28] A. Sherstinsky, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network, *Physica A: Statistical Mechanics And Its Applications* (2018). doi:10.1016/j.physa.2019.132306.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin, Attention Is All You Need, *Advances in neural information processing systems* 30 (2017).
- [30] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., Training Language Models to Follow Instructions with Human Feedback, *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

- [31] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang and Y. Zhang, ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge, *CUREUS* (2023). doi:10.7759/cureus.40895.
- [32] J.S. Park, J.C. O'Brien, C.J. Cai, M.R. Morris, P. Liang and M.S. Bernstein, Generative Agents: Interactive Simulacra of Human Behavior, *arXiv preprint arXiv: Arxiv-2304.03442* (2023).
- [33] M.T.R. Laskar, M.S. Bari, M. Rahman, M.A.H. Bhuiyan, S. Joty and J.X. Huang, A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets, in: *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J.L. Boyd-Graber and N. Okazaki, eds, Association for Computational Linguistics, 2023, pp. 431–469. doi:10.18653/v1/2023.findings-acl.29.
- [34] I. Jahan, M.T.R. Laskar, C. Peng and J.X. Huang, Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers, in: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, BioNLP@ACL 2023, Toronto, Canada, 13 July 2023*, D. Demner-Fushman, S. Ananiadou and K. Cohen, eds, Association for Computational Linguistics, 2023, pp. 326–336. doi:10.18653/v1/2023.bionlp-1.30.
- [35] M.T.R. Laskar, M. Rahman, I. Jahan, E. Hoque and J. Huang, Can Large Language Models Fix Data Annotation Errors? An Empirical Study Using Debatapepia for Query-Focused Text Summarization, in: *Findings of The Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino and K. Bali, eds, Association for Computational Linguistics, Singapore, 2023, pp. 10245–10255. doi:10.18653/v1/2023.findings-emnlp.686. <https://aclanthology.org/2023.findings-emnlp.686>.
- [36] M. Pan, J. Wang, J.X. Huang, A.J. Huang, Q. Chen and J. Chen, A Probabilistic Framework for Integrating Sentence-level Semantics via BERT into Pseudo-relevance Feedback, *Inf. Process. Manag.* **59**(1) (2022), 102734. doi:10.1016/J.IPM.2021.102734. <https://doi.org/10.1016/j.ipm.2021.102734>.
- [37] J. Cowie and W. Lehnert, Information Extraction, *Communications of the ACM* **39**(1) (1996), 80–91.
- [38] S. Sarawagi et al., Information Extraction, *Foundations and Trends® in Databases* **1**(3) (2008), 261–377.
- [39] W.-T. Balke, Introduction to Information Extraction: Basic Notions and Current Trends, *Datenbank-Spektrum* **12** (2012), 81–88.
- [40] M. Lupu, J.X. Huang, J. Zhu and J. Tait, TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC, *SIGIR Forum* **43**(2) (2009), 63–70. doi:10.1145/1670564.1670576.
- [41] M. Lupu, F. Piroi, X. Huang, J. Zhu and J. Tait, Overview of the TREC 2009 Chemical IR Track, in: *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, E.M. Voorhees and L.P. Buckland, eds, NIST Special Publication, Vol. 500-278, National Institute of Standards and Technology (NIST), 2009. <http://trec.nist.gov/pubs/trec18/papers/CHEM09.OVERVIEW.pdf>.
- [42] X. Huang and Q. Hu, A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval, in: *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, J. Allan, J.A. Aslam, M. Sanderson, C. Zhai and J. Zobel, eds, ACM, 2009, pp. 307–314. doi:10.1145/1571941.1571995.
- [43] X. Yin, J.X. Huang, Z. Li and X. Zhou, A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia, *IEEE Trans. Knowl. Data Eng.* **25**(6) (2013), 1201–1212. doi:10.1109/TKDE.2012.24.
- [44] K. Adnan and R. Akbar, An analytical Study of Information Extraction from Unstructured and Multidimensional Big Data, *Journal of Big Data* **6**(1) (2019), 1–38.
- [45] A.-u. Rahman, D. Musleh, M. Nabil, H. Alubaidan, M. Gollapalli, G. Krishnasamy, D. Almoqbil, M.A.A. Khan, M. Farooqui, M.I.B. Ahmed et al., Assessment of Information Extraction Techniques, Models and Systems., *Mathematical Modelling of Engineering Problems* **9**(3) (2022).
- [46] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo and Z. Wang, A Survey of Information Extraction Based on Deep Learning, *Applied Sciences* **12**(19) (2022), 9691.
- [47] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang et al., Zero-shot Information Extraction via Chatting with ChatGPT, *arXiv preprint arXiv:2302.10205* (2023).
- [48] R. Han, T. Peng, C. Yang, B. Wang, L. Liu and X. Wan, Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors, *arXiv preprint arXiv:2305.14450* (2023).
- [49] C. Yuan, Q. Xie and S. Ananiadou, Zero-shot Temporal Relation Extraction with ChatGPT, *arXiv preprint arXiv:2304.05454* (2023).
- [50] B. Li, G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao and S. Zhang, Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness, *arXiv preprint arXiv:2304.11633* (2023).
- [51] F. Peng, X. Huang, D. Schuurmans and S. Wang, Text Classification in Asian Languages without Word Segmentation, in: *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, 2003, Sapporo, Japan, July 7, 2003*, J. Adachi, ed., ACL, 2003, pp. 41–48. <https://dl.acm.org/citation.cfm?id=1118941>.
- [52] A. An, Y. Huang, X. Huang and N. Cercone, Feature Selection with Rough Sets for Web Page Classification, *Trans. Rough Sets* **2** (2004), 1–13. doi:10.1007/978-3-540-27778-1_1.
- [53] J. Zhou, Q. Chen, J.X. Huang, Q.V. Hu and L. He, Position-aware Hierarchical Transfer Model for Aspect-level Sentiment Classification, *Inf. Sci.* **513** (2020), 1–16. doi:10.1016/J.INS.2019.11.048. <https://doi.org/10.1016/j.ins.2019.11.048>.
- [54] L. Yao, C. Mao and Y. Luo, Graph convolutional networks for text classification, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 2019, pp. 7370–7377.
- [55] M. Soni and V. Wade, Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms, *arXiv preprint arXiv: 2303.17650* (2023).
- [56] X. Chen, J. Ye, C. Zu, N. Xu, R. Zheng, M. Peng, J. Zhou, T. Gui, Q. Zhang and X. Huang, How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks, *arXiv preprint arXiv: 2303.00293* (2023).

- [57] R. Nogueira, W. Yang, K. Cho and J. Lin, Multi-stage document ranking with BERT, *CoRR* **abs/1910.14424** (2019).
- [58] H. Zamani, S. Dumais, N. Craswell, P. Bennett and G. Lueck, Generating Clarifying Questions for Information Retrieval, in: *Proceedings of The Web Conference 2020*, ACM, 2020.
- [59] Y. Shi, H. Ma, W. Zhong, G. Mai, X. Li, T. Liu and J. Huang, Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs, *arXiv preprint arXiv:2305.03513* (2023).
- [60] B. Zhao, W. Jin, J. Ser and G. Yang, ChatAgri: Exploring Potentials of ChatGPT on Cross-linguistic Agricultural Text Classification, *NEUROCOMPUTING* (2023). doi:10.48550/arXiv.2305.15024.
- [61] F. Gilardi, M. Alizadeh and M. Kubli, ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks, *Proceedings of the National Academy of Sciences of the United States of America* (2023). doi:10.1073/pnas.2305016120.
- [62] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford et al., Okapi at TREC-3, *Nist Special Publication Sp* **109** (1995), 109.
- [63] J. Zhao, J.X. Huang and Z. Ye, Modeling Term Associations for Probabilistic Information Retrieval, *ACM Trans. Inf. Syst.* **32**(2) (2014), 7:1–7:47. doi:10.1145/2590988.
- [64] J. Zhao, J.X. Huang and B. He, CRTER: Using Cross Terms to Enhance Probabilistic Information Retrieval, in: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, W. Ma, J. Nie, R. Baeza-Yates, T. Chua and W.B. Croft, eds, ACM, 2011, pp. 155–164. doi:10.1145/2009916.2009941.
- [65] T.-Y. Liu et al., Learning to Rank for Information Retrieval, *Foundations and Trends® in Information Retrieval* **3**(3) (2009), 225–331.
- [66] M.T.R. Laskar, J.X. Huang and E. Hoque, Contextualized Embeddings Based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task, in: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association, 2020, pp. 5505–5514. <https://aclanthology.org/2020.lrec-1.676/>.
- [67] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L.Y. Wu, S. Edunov, D. Chen and W.-t. Yih, Dense Passage Retrieval for Open-Domain Question Answering, *Conference on Empirical Methods in Natural Language Processing* (2020). doi:10.18653/v1/2020.emnlp-main.550.
- [68] S. Wang, H. Scells, B. Koopman and G. Zucco, Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?, *arXiv preprint arXiv:2302.03495* (2023).
- [69] W. Sun, L. Yan, X. Ma, P. Ren, D. Yin and Z. Ren, Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent, *arXiv preprint arXiv:2304.09542* (2023).
- [70] R. Xu, Y. Feng and H. Chen, Chatgpt vs. Google: A Comparative Study of Search Performance and User Experience, *arXiv preprint arXiv:2307.01135* (2023).
- [71] Y. Huang and J. Huang, Diversified Prior Knowledge Enhanced General Language Model for Biomedical Information Retrieval, in: *ECAI 2023 - 26th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications, Vol. 372, IOS Press, 2023, pp. 1109–1115. doi:10.3233/FAIA230385.
- [72] L. Wang, N. Yang and F. Wei, Query2doc: Query Expansion with Large Language Models, *arXiv preprint arXiv: 2303.07678* (2023).
- [73] L. Gao, X. Ma, J. Lin and J. Callan, Precise Zero-Shot Dense Retrieval without Relevance Labels, *Annual Meeting of the Association for Computational Linguistics* (2022). doi:10.48550/arXiv.2212.10496.
- [74] K. Keyvan and J.X. Huang, How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges, *ACM Comput. Surv.* **55**(6) (2023), 129:1–129:40. doi:10.1145/3534965.
- [75] J. Zou, J.X. Huang, Z. Ren and E. Kanoulas, Learning to Ask: Conversational Product Search via Representation Learning, *ACM Trans. Inf. Syst.* **41**(2) (2023), 45:1–45:27. doi:10.1145/3555371.
- [76] J. Zhao, J.X. Huang, H. Deng, Y. Chang and L. Xia, Are Topics Interesting or Not? An LDA-based Topic-graph Probabilistic Model for Web Search Personalization, *ACM Trans. Inf. Syst.* **40**(3) (2022), 51:1–51:24. doi:10.1145/3476106.
- [77] L. Zou, L. Xia, Y. Gu, X. Zhao, W. Liu, J.X. Huang and D. Yin, Neural Interactive Collaborative Filtering, in: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, J.X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen and Y. Liu, eds, ACM, 2020, pp. 749–758. doi:10.1145/3397271.3401181.
- [78] L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin and J.X. Huang, Hypergraph Contrastive Collaborative Filtering, in: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J.S. Culpepper and G. Kazai, eds, ACM, 2022, pp. 70–79. doi:10.1145/3477495.3532058.
- [79] R. Thoppilan, D.D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H.S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M.R. Morris, T. Doshi, R.D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi and Q. Le, LaMDA: Language Models for Dialog Applications, *arXiv preprint arXiv: 2201.08239* (2022).
- [80] W. Cao, W. Feng, Q. Lin, G. Cao and Z. He, A Review of Hashing Methods for Multimodal Retrieval, *IEEE Access* **8** (2020), 15377–15391. doi:10.1109/ACCESS.2020.2968154.
- [81] R. Zhao, H. Chen, W. Wang, F. Jiao, X.L. Do, C. Qin, B. Ding, X. Guo, M. Li, X. Li and S. Joty, Retrieving Multimodal Information for Augmented Generation: A Survey, *arXiv preprint arXiv: 2303.10868* (2023).
- [82] K. Wang, Q. Yin, W. Wang, S. Wu and L. Wang, A Comprehensive Survey on Cross-modal Retrieval, *arXiv preprint arXiv: 1607.06215* (2016).

- [83] S. Hofstätter, J. Chen, K. Raman and H. Zamani, FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation, in: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1437–1447. ISBN 9781450394086. doi:10.1145/3539618.3591687.
- [84] W. Chen, H. Hu, C. Saharia and W.W. Cohen, Re-Imagen: Retrieval-Augmented Text-to-Image Generator, *International Conference on Learning Representations* (2022). doi:10.48550/arXiv.2209.14491.
- [85] W. Chen, H. Hu, X. Chen, P. Verga and W.W. Cohen, MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text, *Conference on Empirical Methods in Natural Language Processing* (2022). doi:10.48550/arXiv.2210.02928.
- [86] M. Yasunaga, A. Aghajanyan, W. Shi, R. James, J. Leskovec, P. Liang, M. Lewis, L. Zettlemoyer and W. Yih, Retrieval-Augmented Multimodal Language Modeling, in: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett, eds, Proceedings of Machine Learning Research, Vol. 202, PMLR, 2023, pp. 39755–39769. <https://proceedings.mlr.press/v202/yasunaga23a.html>.
- [87] J. Li, H. Li, Z. Pan and G. Pan, Prompt ChatGPT In MNER: Improved multimodal named entity recognition method based on auxiliary refining knowledge from ChatGPT, *arXiv preprint arXiv: 2305.12212* (2023).
- [88] OpenAI, GPT-4V(ision) System Card, *PREPRINT* (2023).
- [89] W. Wang, X. Lin, F. Feng, X. He and T.-S. Chua, Generative Recommendation: Towards Next-generation Recommender Paradigm, *arXiv preprint arXiv: 2304.03516* (2023).
- [90] J. Li, A. Dada, J. Kleesiek and J. Egger, ChatGPT in Healthcare: A Taxonomy and Systematic Review, *medRxiv* (2023), 2023–03.
- [91] H. Huang, O. Zheng, D. Wang, J. Yin, Z. Wang, S. Ding, H. Yin, C. Xu, R. Yang, Q. Zheng et al., ChatGPT for Shaping The Future of Dentistry: The Potential of Multi-modal Large Language Model, *International Journal of Oral Science* **15**(1) (2023), 29.
- [92] K. Lee, O. Firat, A. Agarwal, C. Fannjiang and D. Sussillo, Hallucinations in Neural Machine Translation (2018).
- [93] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung et al., A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, *arXiv preprint arXiv:2302.04023* (2023).
- [94] M. Sallam, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on The Promising Perspectives and Valid Concerns, in: *Healthcare*, Vol. 11, MDPI, 2023, p. 887.
- [95] L. Gao, Z. Dai, P. Pasupat, A. Chen, A.T. Chaganty, Y. Fan, V.Y. Zhao, N. Lao, H. Lee, D.-C. Juan and K. Guu, RARR: Researching and Revising What Language Models Say, Using Language Models, *arXiv preprint arXiv: 2210.08726* (2022).
- [96] T.Y. Zhuo, Y. Huang, C. Chen and Z. Xing, Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity, *arXiv preprint arXiv: 2301.12867* (2023).
- [97] B. Dash and P. Sharma, Are ChatGPT and Deepfake Algorithms Endangering The Cybersecurity Industry? A Review, *International Journal of Engineering and Applied Sciences* **10**(1) (2023).
- [98] P.P. Ray, ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future scope, *Internet of Things and Cyber-Physical Systems* (2023).
- [99] S.C. Geyik, S. Ambler and K. Kenthapadi, Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2221–2231. ISBN 9781450362016. doi:10.1145/3292500.3330691.
- [100] R. Gao and C. Shah, How Fair Can We Go: Detecting the Boundaries of Fairness Optimization in Information Retrieval, in: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 229–236. ISBN 9781450368810. doi:10.1145/3341981.3344215.
- [101] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou and Y. Zhang, Towards Long-Term Fairness in Recommendation, in: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 445–453. ISBN 9781450382977. doi:10.1145/3437963.3441824.
- [102] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas and P. Sen, A Survey of the State of Explainable AI for Natural Language Processing, in: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, ACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, K. Wong, K. Knight and H. Wu, eds, Association for Computational Linguistics, 2020, pp. 447–459. <https://aclanthology.org/2020.aacl-main.46/>.
- [103] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, A Survey of Methods for Explaining Black Box Models, *ACM computing surveys (CSUR)* **51**(5) (2018), 1–42.
- [104] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao and J. Zhu, Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges, in: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, Springer, 2019, pp. 563–574.
- [105] W. Kong, S. Khadanga, C. Li, S.K. Gupta, M. Zhang, W. Xu and M. Bendersky, Multi-Aspect Dense Retrieval, in: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 3178–3186. ISBN 9781450393850. doi:10.1145/3534678.3539137.
- [106] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu and W. Saunders, Language Models Can Explain Neurons in Language Models, URL <https://openaiublic.blob.core.windows.net/neuron-explainer/paper/index.html>. (Date accessed: 14.05. 2023) (2023).
- [107] J. Piskorski and R. Yangarber, Information Extraction: Past, Present and Future, *Multi-source, multilingual information extraction and summarization* (2013), 23–49.
- [108] D. Adiwardana, M.-T. Luong, D.R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu et al., Towards a Human-like Open-domain Chatbot, *arXiv preprint arXiv:2001.09977* (2020).

This figure "iospress.png" is available in "png" format from:

<http://arxiv.org/ps/2402.11203v1>