Towards Generative Search and Recommendation: A keynote at RecSys 2023

Tat-Seng Chua
National University of Singapore
Singapore
dcscts@nus.edu.sg

Abstract
The emergence of large language models (LLMs), especially ChatGPT, has for the first time make AI known to almost everyone and affected every facet of our society. The LLMs have the potential to revolutionize the ways we seek and consume information. This has stemmed the recent trends in both academia and industry to develop LLM-based generative AI systems for various applications with enhanced capabilities. One such systems is the generative search and recommender system, which is capable of performing content retrieval, content repurposing, content creation and their integration to meet users’ information needs. However, before such systems can be widely used and accepted, we need to address several challenges. The primary challenge is the trust and safety in the generated content as we expect the LLM’s to make mistakes with hallucination. This is because of the quality of data being used for their training is often erroneous and biased. The other challenges in the search and recommendation domain include: how to teach the system to be pro-active in anticipating the needs of users and in directing the conversation towards a fruitful direction; as well as the integration of retrieved and generated content. This keynote presented a generative information seeking paradigm, and discuss key research towards a trustable generative system for search and recommendation.

Date: 21 September 2023.

1 Introduction
The emergence of large language models (LLMs), such as the ChatGPT, has ushered in a new era of AI. It has, for the first time, made AI known to everyone and affected almost every facet of our society. In particular, LLMs have exhibited strong language and common-sense capability with rich knowledge. They also possess the ability for instruction following, in-context learning, as well as planning in the form of chain of thoughts. As a result, we can now freely converse with such systems to express our intent in a fine-grained and multimodal manner, and we expect the system to recommend existing items or generate new items as necessary, and present them in a concise summarized form. This has stemmed the recent trend in both academia and industry to develop LLM-based generative AI systems for various applications with enhanced capabilities. One such example is the Search Co-Pilot introduced by many Internet Search Engines, such as the Microsoft
Bing, to leverage LLM as supplement to search. The other example is the generative search and recommender system, which can help to enrich content retrieval with content repurposing, content creation, and the generation of search and recommendation rank list, in a seamless and coherent manner. In addition, LLMs have been evolved into LMFMs (Large Multimodal Foundation Models) with the incorporation of multimodal contents, along with expanded applications in many emerging multimodal domains.

However, before such systems can be widely used and accepted, we need to address several limitations and challenges. One of the key limitations of LLMs is that although they are good at knowledge, they are not the expert in most vertical domains. Moreover, they often make mistakes along with the problem of hallucination, i.e. the tendency to make up answers that are fraudulence or non-existence. This is because the vast amount of data used to train LLMs often contain errors with biases. Also, although LLMs have superior 2D layout and reasoning capability, they are weak in 3D. They are also not good at logic and logical reasoning. As a result, LLMs are vulnerable to attacks in retuning unsafe and privacy contents. In addition, LLMs are expensive to operate especially online. Hence a lot of research is still needed before LLM-based models can be employed in many mission-critical applications that require reliable and trustable results.

Although there are many problems as listed above, it is often not an option for large corporations and government organizations not to embrace LLMs. The key is in deeply understanding the limitations and challenges to come up with solutions to alleviate them in an effective and efficient manner. Given the resource constraints of many academia environments, we identify several research directions that can be pursued academically; they are: (a) generative recommendation systems; (b) LLM-powered AI Agents for emerging applications; (c) pro-active dialogue systems; and (d) trust and safety in LLM-based systems. Of course, there are also many more fruitful research directions that could be explored, including, the extraction of basic common capability of LLMs, the injection of personality and diversity into LLMs, the evaluation and audit of LLMs, and the fine-grained visual understanding, etc. In the following, we will elaborate our research on the four research focuses.

2 Generative Recommender Systems

The boom of generative models has paved the way for significant advances in recommender systems. Recently, there have been emerging innovative ideas on integrating generative models, especially LLMs, into recommender systems to reshape the traditional recommender paradigm and shed light on generative recommender systems. As shown in Figure 1, we examine the development prospects of generative recommender systems from three aspects: 1) enhancing recommender algorithms [Bao et al., 2023b], 2) changing user-system interaction model [Liao et al., 2023a], and 3) generating personalized content [Wang et al., 2023c].

Recent research on using LLMs to enhance recommender algorithms falls into two groups: LLMs-enhanced and LLMs-based recommender systems [Wu et al., 2023]. LLMs-enhanced recommender systems primarily leverage the rich world knowledge, language understanding and generation capabilities of LLMs to enhance existing recommendation systems [Bao et al., 2023a; Wu et al., 2023]. For example, prior studies utilize LLMs to enrich the semantic representations of users and items in existing recommendation systems [Xi et al., 2023a]. In contrast, LLM-based recommender systems directly adopt LLMs to rank items for recommendations [Wu et al., 2023].
Since the LLM pretraining phase does not involve recommendation tasks, directly applying LLMs for recommendations often yields limited results [Bao et al., 2023b]. It is thus crucial to utilize recommendation data for instruction tuning on LLMs. For instance, TALLRec tunes LLMs for recommendation using few-shot recommendation samples, finding that a small number of samples can significantly improve Llama’s recommendation performance [Bao et al., 2023b]. BIGRec achieves comparable or even better recommendation results than the traditional recommendation models through instruction tuning and generation grounding [Bao et al., 2023a]. TransRec discovers that using multi-facet item identifiers for instruction tuning can better bridge the recommendation data and the language space of LLMs [Lin et al., 2023b]. In addition to LLMs, previous work has also explored other generative models like diffusion models for recommendation [Wang et al., 2023d; Li et al., 2023c], validating the promising direction of generative item rankings.

Generative models also have immense potential to revolutionize the existing user-system interaction paradigm [Liao et al., 2023a]. Traditional recommender systems often rely on various forms of implicit user feedback such as the clicks and likes to model user preference [Covington et al., 2016] which may not quickly capture changes in user preference or understand users’ explicit information needs [Wang et al., 2023c]. LLMs, due to their strong language understanding, reasoning and generation capabilities, can significantly enhance existing conversational recommender systems, revolutionize the previous user-system interaction paradigm, and better capture users’ diverse information needs.

Generative models possess powerful content generation capabilities to complement existing items and meet users’ personalized needs [Wang et al., 2023c]. Existing recommender systems often rank and recommend pre-generated items from the item corpus. However, the items within the corpus may not always fulfill users’ specific information needs. For instance, a user might desire clothing with a particular pattern. In such cases, generative models can be incorporated to produce AI-generated content (AIGC), i.e., repurposing existing items or creating new items [Wang et al., 2023c]. As such, in the recommendation ecosystem, it is likely to envision AIGC gradually supplementing human-generated content.

Looking ahead to the future of generative recommender systems, we expect to see generative models gradually playing an important role in multiple aspects of the recommendation ecosystem. This includes, but not limited to, improving item retrieval and ranking, diversifying user-system interaction, enhancing user-system engagement, and addressing specific information needs.
interactions, and enriching content generation. Additionally, there will be increasing efforts to consider the trustworthiness of generative recommender systems from various aspects such as bias, fairness, and safety [Wang et al., 2023c; Zhang et al., 2023b].

3 LLM-powered AI Agents for Recommendations

Autonomous AI agents have long been regarded as stepping stones towards artificial general intelligence (AGI), with capabilities for self-guided task execution. Recent advances in Large Language Models (LLMs) have demonstrated a closer approximation to human-level intelligence in self-directed planning and instructions [OpenAI, 2023]. This advancement has spurred a growing trend in integrating LLMs as central components in developing autonomous AI agents [Qin et al., 2023; Xi et al., 2023b]. In the domain of recommender systems - vital for navigating today’s vast information landscape - LLM-powered AI agents exhibit superior performance in autonomous interaction and in discerning user preference [Lin et al., 2023a]. Such impressive capability can be harnessed to simulate authentic human behavior within recommender systems, reflecting both individual and population levels patterns [Yang et al., 2023].

Simulating user behavior in recommender systems is a complex endeavor that requires a deep understanding of human preference and behavior patterns [Chen et al., 2023a; Zhang et al., 2023c]. To bridge this gap, the design of agent modules that are tailored for recommendation context is crucial [Wang et al., 2023b; Zhang et al., 2023a]. Figure 2 depicts a framework for LLM-powered agents in recommendations; in which the profile module encapsulates personalized social traits and the user’s historical preferences, the memory module records past interaction behaviors and emotional memories, and the action module empowers agents to interact directly with the recommendation environment.

The use of LLM-empowered agents to emulate human behavior has yielded encouraging results but also exposed inherent limitations. These explorations have revealed that: (1) agents...
can preserve personalized attributes and distinct social traits embedded within the user profiles, acting in accordance with the delineated user personas; (2) leveraging the data generated from simulations to retrain recommenders can boost their performance; and (3) certain challenges in the recommendation domain, previously unresolved due to data scarcity, are now feasible to be explored. However, several limitations remain, notably, (1) current attempt at AI Agents does not adequately mirror the diverse scenarios and multimodal nature of recommender systems, compelling demand for agents with advanced multi-modal comprehension [Huang et al., 2023b]; and (2) purely prompting-based Agents fall short in effectively and accurately simulating human behavior, thereby making it essential to fine-tune these agents to better reflect the user preferences [Zhang et al., 2023a; Li et al., 2023b].

4 LLM-based Proactive Conversational Systems

LLM-based conversational systems have demonstrated exceptional proficiency in context understanding and response generation in various dialogue problems. However, as LLMs are trained to passively follow users’ instructions, conversational agents built upon them typically prioritize accommodating users’ intention. Therefore, LLM-powered conversational agents often face challenges in handling proactive dialogue problems [Deng et al., 2023c; Liao et al., 2023b] that require the conversational agent to strategically take the initiative to steer the conversation towards an anticipated goal [Deng et al., 2023b]. In recent years, we have witnessed many advanced designs to make the conversational agents proactive in meeting a wide range of challenging dialogue problems, such as the clarification in information-seeking dialogues [Deng et al., 2022], target-guided dialogues [Lei et al., 2022], and non-collaborative dialogues [Zhan et al., 2022], etc.

Recent works investigate prompt-based methods to trigger the proactivity of LLMs [Chen et al., 2023b; Deng et al., 2023d]. For example, Chen et al. [2023b] suggest using LLMs to generate controllable responses based on pre-defined dialogue strategies, though these strategies are not always available in dynamic real-world scenarios. To this end, Deng et al. [2023d] evaluate the capability of LLMs in making strategic decisions for different types of proactive dialogues from the perspective of prompting. The Proactive Chain-of-Thought (ProCoT) prompting scheme (Figure 3) is proposed to plan the next dialogue action with chain-of-thoughts, i.e., intermediate steps of reasoning and planning the next action to take in order to reach the conversational goal. The empirical analysis reveals that while ProCoT is successful in certain dialogue types, there are still many limitations: (1) it falls short of handling domain-specific dialogue applications, such as Finance; (2) it tends to make aggressive topic transitions due to the strong capability of controllable text generation; and (3) it fails to make strategic decisions to lead the conversation in non-collaborative dialogues.

The future holds great promise for LLM-based proactive conversational systems, as research continues to bridge the gap between passive response generation and proactive dialogue planning [Deng et al., 2023e]. This evolution will spark in a new era of human-computer interactions, marked by more intuitive, strategic and dynamic conversations tailored to individual user needs and broader applications.
Figure 3. Examples of three kinds of prompting schemes for proactive dialogues [Deng et al., 2023d].
5 Trust Evaluation of LLM

Despite the impressive capabilities of LLM across a wide spectrum of language tasks, LLM is not infallible and is susceptible to issues related to trust. LLM has been observed to produce hallucinated or erroneous content, such as the inclusion of fabricated information or flawed reasoning [Zhang et al., 2023d; Wang et al., 2022]. Additionally, LLM is vulnerable to adversarial and backdoor attacks [Wang et al., 2023a]. Hence, it is important to identify and address trust-related problems in LLM-generated content to ensure its safe and dependable use. Approaches for detecting trust issues in LLM-generated content can be categorized into two main types by the source of the verification process: 1) fact-checking or external verification, and 2) consistency checking or self evaluation. An overview of these approaches is summarized in Figure 4.

The external verification entails the validation of generated content with reliable knowledge sources, systems, and trusted websites. For instance, the utilization of the reputable knowledge engine Wolfram Alpha\(^1\) in conjunction with ChatGPT serves as a mean to accessing reliable information. Additionally, Gao et al. [2023a] and Zhao et al. [2023] leverage trusted websites for addressing verification questions and performing trustworthy edits on the LLM-generated content. Furthermore, when dealing with numerical questions, Gao et al. [2023b] advocate shifting the numerical calculation to a dependable Python interpreter instead of relying on LLM. Nevertheless, the external verification approach has its limitations, as certain trusted sources may only be effective for specific categories of generated content and they may not contain up-to-date information.

The self evaluation aims to explore the LLM’s inherent capability to express confidence in its generation, providing a more general solution for all LLM outputs. Researchers employ self-consistency-based methods to obtain calibrated confidence scores [Kuhn et al., 2022; Kadavath et al., 2022; Tian et al., 2023], which is adopted to enhancing accuracy in domains like molecular reactions [Shi et al., 2023]. Moreover, self evaluation is also claimed to be accomplished by multi-aspects verification, e.g., multi-agent debate [Du et al., 2023] or LLM critique [Cohen et al., 2023]. However, the effectiveness of such methods is questionable, since the correctness and confidence of these LLM outputs remain unclear [Huang et al., 2023a]. Additionally, self evaluation can be achieved through counterfactual verification, involving asking counterfactual questions and examining answer consistency or answer confidence.

\(^1\)https://www.wolframalpha.com/.

**Figure 4.** A summary for different trust evaluation approaches on LLM.
In summary, the trust evaluation in LLM generation will continue to be a thriving area of research, as the pursuit of dependable generated content from LLMs remains a long-term objective for researchers and practitioners across diverse fields.

6 Safety Evaluation of LLM

LLMs play a crucial role in information dissemination and communication, making safety evaluation on LLM-generated content of utmost importance. The safety evaluation aims to prevent the generation of harmful content, including but not limited to biased, unfaithful, and discriminatory responses [Sun et al., 2023, 2021; Zhang et al., 2023e]. In this research line, one crucial branch is to design diverse attacks to assess LLM safety [Deng et al., 2023a]. By designing transparent and controllable attack methods, we can examine the safety of LLM-generated content and also provide various attack instances to help align LLMs with safety objectives.

As shown in Figure 5, we introduce two attack paradigms for LLM safety evaluation. The first is using red teaming attacks to poison prompts, aiming to induce LLMs to generate harmful responses [Perez et al., 2022]. By evaluating the harmfulness of responses to diverse poisoned prompts, we can measure LLM safety from diverse aspects such as faithfulness, discrimination, and viruses. For example, previous research has explored using human-generated poisoned prompts to instruct LLMs through in-context learning to generate more poisoned prompts for attacks [Deng et al., 2023a]. To poison prompts, the popular strategies involve role play, payload splitting, and reverse prompts [Kang et al., 2023; Li et al., 2023a; Perez and Ribeiro, 2023]. The second paradigm is to poison external data to attack LLMs. It is common to retrieve external data (e.g., the latest events or documents) to assist LLMs in generating responses [Greshake et al., 2023; Pan et al., 2023b]. We can inject harmful information into external data to induce the retrieval-based LLMs to generate harmful content. For instance, prior work has considered injecting documents with misinformation to induce LLMs to generate incorrect answers [Pan et al., 2023a].

![Figure 5. Illustration of safety evaluation by poisoning prompts and external data.](image)

LLM safety evaluation requires extensive future endeavors from various aspects, including expanding attack targets, refining attack methods, and enhancing evaluation tools to inspect generated content from multiple perspectives. Besides, how to better utilize attack instances to...
enhance LLM safety is also an important concern. Moreover, moving beyond LLM safety evaluation, there is a broad landscape in LLM evaluation, including performance measures (e.g., speed, accuracy, and diversity), content measures (e.g., commonsense knowledge, domain expertise, and weaknesses), trust measures (e.g., authenticity, bias and fairness, privacy, safety, and legal compliance). It signifies promising avenues for future research.

7 Summary

The popularity of ChatGPT, or more generally LLM, has made AI known to almost everyone and affected every facet of our society. As discussed, LLM possesses strong natural language, common sense and reasoning capabilities, and is capable of comprehending a huge amount of data corpus within any organization, correlating them with relevant external data via retrieval, and providing a summary or answering almost any question with some reasoning capabilities. Because of this, it has the potential to be used in many domains, ranging from E-Commerce, Finance, Medical, Media and Education, etc. However, it has many shortcomings, including the problem of trust, safety and privacy, and is also costly to operate especially for online applications. Hence our research must not focus only on its strengths in applications like generative research and recommendation, and proactive dialogs, but must address its weaknesses on trust and safety too. In line with this, we have discussed four research directions that we can fruitfully carry out and contribute in the academic environment. In addition, there are many more challenges that we can do and contribute that are not covered in this keynote.

In every decade, there are some technological innovations, such as the Internet, PC, Deep-Blue, Alpha-Go, and now LLM, that have disrupted some aspects of our work and life. And each time there were doomsday predictions that the machines are going to take over the world and rule over humans. However, as we observed over the decades, and in this case, the LLM-based AI systems, the machines are not good at handling complex tasks that require human-level intelligence and common sense. Thus, I see a future where the machines do the mundane jobs that humans hate to do, and leaving the more interesting higher-level jobs that require natural reasoning and creativity to humans. Of course, in the future, the AI systems will become more and more intelligent, and we should have the common sense of not to allow machines to make critical decisions. In summary, the crises that the machines are taking over the world are happening over and over, but we always have the common sense to weather through them with better outcomes for our society.

Acknowledgments

I would like to thank An Zhang, Yang Deng, Moxin Li, and Wenjie Wang for their contribution to this keynote and paper. This work is supported by NExT++ Research Center.
References


Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender ai agent: Integrating large language models for interactive recommendations, 2023b.


Zihao Li, Aixin Sun, and Chenliang Li. Diffurec: A diffusion model for sequential recommendation. *TOIS*, 2023c.


Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *SIGIR*. ACM, 2023d.


Zhengyi Yang, Jiancan Wu, Yanchen Luo, Jizhi Zhang, Yancheng Yuan, An Zhang, Xiang Wang, and Xiangnan He. Large language model can interpret latent space of sequential recommender, 2023.


Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Agentcf: Collaborative learning with autonomous language agents for recommender systems, 2023c.

