Appraising Success of LLM-based Dialogue Agents

Ritika Wason
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)
New Delhi, India
ritika.wason@bvicam.in

Parul Arora
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)
New Delhi, India
parul.arora@bvicam.in

Devansh Arora
Dept. of Computer Sc. and Engineering
Indraprastha Institute of Information Technology (IIIT), Delhi
New Delhi, India
devansh20053@iiitd.ac.in

Jasleen Kaur
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)
New Delhi, India
jasleenkaur.s1mca21@bvicam.in

Sunil Pratap Singh
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)
New Delhi, India
sunil_pratap@rediffmail.com

M. N. Hoda
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)
New Delhi, India
mca@bvicam.ac.in

***Abstract*— Large language models (LLM) can be prompted to exhibit human like dialogue. However, LLM-based dialogue agentsmay in many senses mimic a human being as well as differentiate from the same. In this manuscript we try to unveil the intricacies of an LLM-based dialogue agent. The capabilities of these models have shaken the business as well as research world. However, the success of such models in any given domain depends on a number of considerations. In this manuscript we outline the same.**

Keywords*— *Large Language Models (LLM), Dialogue Agents, GPT, Generative Model, Artificial General Intelligence (AGI).

I. INTRODUCTION

Large language models popularly known as LLMs are an attempt towards attaining artificial general intelligence (AGI) system that have transformer network as its core computational engine [1]. These systems are a disembodied neural network trained on a large data corpus of human-generated text to predict the next word technically termed as a token [2].They exhibit varied capabilities beyond text generation like conducting conversations, guiding tasks as well as reasoning [3]. A suitably trained and tested LLM can be easily substituted to mimic human language convincingly given a sequence of words as input token or context [4]. This capability allows numerous applications of LLMs as dialogue agents in user facing applications like customer support, conversational agents, digital personal assistants etc. [5]. However, such tasks demand an adaptable dialogue agent that can quickly adapt to new tasks exhibiting some degree of autonomous behavior [6].In present day such LLMs exhibit several capabilities including language understanding, generation, interaction as well as reasoning [7], [8]. Popular examples of such models in the current times are GPT4, Synthetic Interactive Persona Agent (SIPA), OpenAI's ChatGPT, etc [9]. The rise of ChatGPT has put the competences and in competencies of these models under the scanner [10]. An LLM-based dialogue agent is an extension of an AI-based agent with three main conceptual parts namely: brain, perception and action [11], [12]. These are depicted in Fig. 1.

As depicted in Fig. 1, any LLM based agent has 3 conceptual parts, namely the brain, perception and action [13]. The brain is the fundamental of an AI-based agent as it not only stores the vital data, information and knowledge but conducts the essential tasks of information processing, decision making, reasoning and planning [5]. Thus, the LLM is the brain of an AI based agent. The next part perception serves a role comparable to that of human sensory organs. It primarily expands the agent's perceptual space from text-based to a multimodal space that may include varied sensory modalities like text, sound, visuals, touch, smell etc. Thus, the agent is able to better perceive information from the external environment. The action part is the outcome of the agent. The agent is expected to process textual output, take embedded actions and apply tools to better respond to the environmental changes and provide feedback [5].

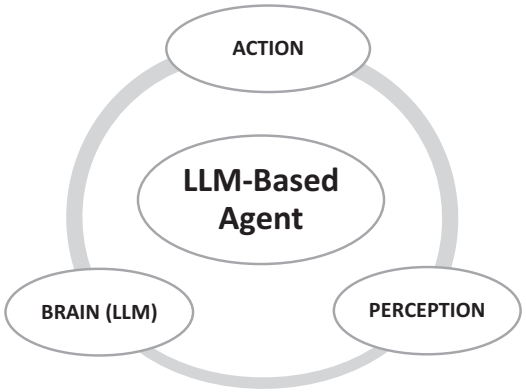

Fig. 1. Conceptual Parts of an LLM based Agent

This manuscript is an effort to trace the LLM-based dialogue agents to appreciate their internal capabilities, applications and unveil the reasons for their success. The rest of the manuscript is organized as follows: Section ii investigates the general mechanism of appropriately training an LLM-based dialogue agent. Section iii briefs the design of these agents that empowers them to make the right decision. Section iv highlight the unintentional risks associated with the use of these agents. Section v concludes by evaluating the current achievements in this rapidly evolving field.

II. Training the LLM-Based Dialogue Agents

Conversational interfaces demand flexibility, accuracy, and contextual memory along with chain-of-thought reasoning. Simply put, an LLM based dialogue agent is meant to answer questions just like humans [14]. The questions are made up of tokens that may be a sequence of words, punctuation marks, emojis and so on [15]. The agent is expected to guess the token expected to come next. This sequence is made from the vast corpus of public data available on the Internet [16].

Training the LLM based dialogue agent involves a dataset constituted of (INSTRUCTION, OUTPUT) pairs in a supervised mode, thus altering the gap between the next-word prediction objective of LLM-based dialogue agent [17]. The INSTRUCTION in this case is the input and the OUTPUT is the expected token. The basic pipeline would involve Instruction dataset construction followed by some mechanism for fine-tuning the dataset. Wide array of techniques has been applied to achieve this fine-tuning of the agents for enhancing their capabilities and controllability. Some of these are listed in Table I:

TABLE I. Techniques For Specializing Dialogue Agents

S. No.	LLM Dialogue Agents		
	Name	***Characteristic***	***Limitation***
1	Instruction Tuning (IT) [17]	Work on human generated instruction and sample response pairs.Ex: InstructGPT	Numerous data samples required for training may be unavailable or costly to generate. It may only capture surface-level patterns in the output instead of comprehending the task.
2	Self-Talking LLM[18]	To bootstrap task-specific data LLMs themselves.	Accuracy is good but dialog success is questionable
3	Goal Oriented Dialogue (GOD)[13]	AI chat agent needs to proactively pose questions and guide users towards specific goals or task completion.	Tradeoff between fluent language generation and task-specific control exists.

Table I above is not complete but it reports the most common techniques for fine-tuning the input to the LLMs of dialogue agents. However, limitations of each mechanism remain as listed in table I. Further it should also be noted that though self-talking as known as self-correcting technique holds a lot of potential, a completely self-correcting system is yet to be realized.

III. Briefing Design of LLM Dialogue Agents

LLMs generally refer to transformer-based models with billions of parameters trained on trillions of tokens [19]. Transformers empower neural networks to process large chunks of text simultaneously in order to establish stronger relationship between words and their context of appearance. The underlying language model of such dialogue agents can be mathematically represented as a conditional probability distribution expressed as [20]:

$$P(w_{n+1}|w_1 \dots w_n) \quad (1)$$

where,

$w_1 \dots w_n$ = is a sequence of tokens (the context)

w_{n+1} = predicted next token

However, for any such dialog agent a major challenge is dialog state management [21]. The system should be capable of deciding the next dialog state and the system response to the user. The current LLM agents exhibit a high level of autonomy. Varied techniques have been applied to respond to user input by choosing the next dialog state. Table II below depicts the same:

TABLE II. Techniques For Deciding the Next Dialog State

S. No.	LLM Decision Making			
	Name	***Characteristic***	***Limitation***	***Use Case***
1	Chain-of-Thought [7]	Decompose user question into subtasks. User input thus directs chain creation on the fly basis.	Has shown performance gains with models of approx.. 100 billion parameters. However, smaller models tend to result in illogical chains leading to lower accuracy	Symbolic Reasoning, Arithmetic Reasoning, Commonsense Reasoning
2	Prompt Chaining [18]	Generates a predetermined chain for an anticipated use-case.	Success depends heavily on initial prompt, context maintenance in long chains is a challenge, ethical questions in sensitive areas cannot be ignored.	Quantum computing, Self-Improving AI systems, Personalized Healthcare/ Academic Predictions
3	Prompt Pipelines [22]	Extend prompt templates by automatically injecting contextual reference data for each prompt.	These models can generate coherent text, as well as they may not provide clear explanations.	Content Generation, Dialogue Systems, Virtual Assistants etc.
4	QuickReply Intents [23]	Generate a user intent taxonomy and apply it to do log analysis, etc.	Chance of creating a feedback loop without clear evaluation.	Taxonomy generation from logs, Search and recommender systems

Irrespective of the decision-making technique employed, all LLM based decision making agents require large scale pre-training on massive text corpora along with reinforcement learning from human feedback. Their iterative prompt cycle is

key to facilitating natural conversations between users and LLM agents. Dialogue agents based on LLMs demand leveraging the capabilities of these models to generate human-like responses in a conversation. A generic design for an LLM-based dialogue agent may contain the following components as detailed below and depicted in Fig. 2.

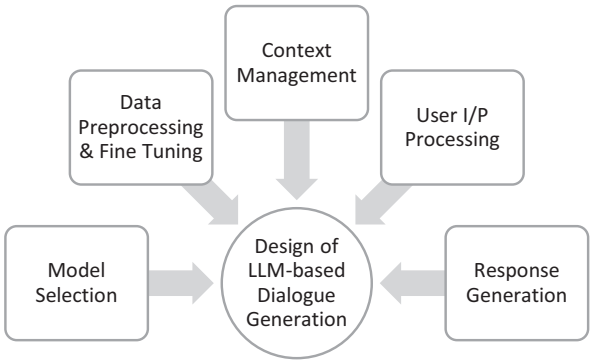

Fig. 2. Generic Design of LLM-Based Dialogue Agent

a) Model Selection: A suitable pre-trained LLM model like GPT-3 can be chosen or created as per requirements. Consider factors like model size, computational resources, and the specific use case.

b) Data Preprocessing and Fine Tuning: The training data is cleaned to remove noise and inconsistencies and brought in a format the LLM can understand. This is essential to allow the model to adapt to domain-specific language and behavior.

c) Context Management: Conversation context should be controlled in LLMs. They typically generate responses based on the preceding context during a conversation.

d) User Input Processing: Appropriate mechanism to process user inputs is required totokenize and encode the input text for the LLM. Relevant information from user queries should be extracted to get relevant context for generating responses.

e) Response Generation: The LLM shall generate responses to the user input. Different strategies for response generation may be experimented with to improve the diversity and quality of responses.

The design is only a base framework for any LLM-based dialogue agent. Carefully combining these core components can enable users to efficiently prompt anthropomorphic LLM based machines.

IV. Unintentional Risks of LLM Based Dialogue Agents

Overestimating LLMs can lead to unreliable applications [24]. Like two sides of a coin, it is worthwhile to investigate the underlying risks of having an LLM based dialogue agent as one of the predefined characters in role play with a human participant.

- *Data Scarcity*: Dialogue is a complex task. Any dialogue agent is expected to be trained in grammar, syntax, dialog reasoning as well as language generation. Mastering these skills calls for large amount of domain-specific data. Pre-trained can help reduce this challenge but their success may vary as per the domain.
- *Model Bias*: Large language models often echo the biases of the data they are trained on. Further any external tools used may also introduce bias. It is critical to mitigate this bias in the model to ensure fairness. Addressing bias effectively would lead to unbiased language models.
- *Societal Dangers*: LLMs have a negative potential to deceive human users, for instance in the form of personalized phishing attacks. Training agents which do not rely on human-generated training data anymore could therefore simplify the creation of deceptive dialogue models by malicious actors
- *Hallucination*: Hallucination is a major concern when it comes to the response of LLMs. Some individuals argue that LLMs cannot be utilized in serious scenarios such as healthcare or legal domains due to hallucinations. Besides, we acknowledge that there are some harmful responses from LLMs.
- *Mismatch between the training objective and user objective*: LLMs-based dialogue agents shall be trained to minimize the contextual word prediction error on a large corpus. However, the user would expect the model to comply by their instructions safely. This mismatch can be overcome by the techniques proposed in Section ii above. However, this is also demanding in terms of drafting high-quality instructions to directly map to the desired target behaviors.
- *Generating Flawed Code*: Bias-driven LLM is susceptible to generating flawed code reducing the overall system accuracy.

V. Conclusion and Future Work

LLM-based dialogue agents offer numerous applications. However, their vast potential is yet to be explored. Creating a successful LLM-based dialogue agent is not just getting the input data corpus and the model right. There are many other intricacies involved. The generated responses should be analyzed to ensure coherence, correctness, and appropriateness to the specific situation. The overall user experience should also be considered. The dialogue agent should be user-friendly, with clear communication and appropriate feedback loops. Proper mechanisms should be implemented to handle errors or misunderstood inputs gracefully. Thorough testing is crucial to identify and address any issues with the dialogue agent. Performance using relevant metrics, based on feedback and test results should be evaluated for improvement. One should also consider the scalability of the dialogue agent, especially if it's intended to handle a large number of users or conversations. Optimize the system architecture to accommodate varying levels of demand. Also implement

security measures to protect user data and ensure privacy. Be aware of the potential risks associated with deploying conversational agents, such as unintentional disclosure of sensitive information.

Besides the above consideration it is important to note that success of domain-specific dialog agents is governed by the availability of domain-specific data which can be concern. To overcome this, work in zero-shot domain adaptation if the future course of action of LLM-based dialogue agent. This shall ensure application of a dialogue agent to any novel domain without the availability of any domain-specific training data. The metrics and benchmarks for LLMs shall mature with time.

REFERENCES

[1] H. Strobelt et al., "Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models," IEEE Trans Vis Comput Graph, vol. 29, no. 1, pp. 1146–1156, Jan. 2023, doi: 10.1109/TVCG.2022.3209479.

[2] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and Applications of Large Language Models," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.10169

[3] "A Very Gentle Introduction to Large Language Models without the Hype | by Mark Riedl | Medium." Accessed: Jan. 21, 2024. [Online]. Available: https://mark-riedl.medium.com/a-very-gentle-introduction-to-large-language-models-without-the-hype-5f67941fa59e

[4] D. Zhang, L. Chen, S. Zhang, H. Xu, Z. Zhao, and K. Yu, "Large Language Models Are Semi-Parametric Reinforcement Learning Agents," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.07929

[5] D. Ulmer, E. Mansimov, K. Lin, J. Sun, X. Gao, and Y. Zhang, "Bootstrapping LLM-based Task-Oriented Dialogue Agents via Self-Talk," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.05033

[6] H. Lakkaraju, D. Slack, Y. Chen, C. Tan, and S. Singh, "Rethinking Explainability as a Dialogue: A Practitioner's Perspective," Feb. 2022, Accessed: Jan. 21, 2024. [Online]. Available: http://arxiv.org/abs/2202.01875

[7] S. Avasthi, R. Chauhan, and D. P. Acharjya, "Extracting information and inferences from a large text corpus," International Journal of Information Technology (Singapore), vol. 15, no. 1, pp. 435–445, Jan. 2023, doi: 10.1007/S41870-022-01123-4/FIGURES/8.

[8] M. Maree, R. Al-Qasem, and B. Tantour, "Transforming legal text interactions: leveraging natural language processing and large language models for legal support in Palestinian cooperatives," International Journal of Information Technology (Singapore), vol. 16, no. 1, pp. 551–558, Oct. 2023, doi: 10.1007/S41870-023-01584-1/METRICS.

[9] 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, vol. 3. KeAi Communications Co., pp. 121–154, Jan. 01, 2023. doi: 10.1016/j.iotcps.2023.04.003.

[10] V. Muthusamy, Y. Rizk, A. Gulati, and P. Dube, "Towards large language model-based personal agents in the enterprise: Current trends and open problems."

[11] D. Bajaj, A. Goel, S. C. Gupta and H. Batra, "Muce: a multilingual use case model extractor using gpt-3," Int J Inform Technol, vol. 14, no. 3, pp. 1543–1554, 2022.

[12] K. Boehner, J. Vertesi, P. Sengers, and P. Dourish, "How HCI interprets the probes," Conference on Human Factors in Computing Systems - Proceedings, pp. 1077–1086, 2007, doi: 10.1145/1240624.1240789.

[13] "DiagGPT: An LLM-based Dialogue System with Automatic Topic Management for Goal-Oriented Dialogue Anonymous ACL submission." [Online]. Available: https://openai.com/blog/chatgpt.

[14] Z. Xi et al., "The Rise and Potential of Large Language Model Based Agents: A Survey," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.07864

[15] M. Shanahan, K. McDonell, and L. Reynolds, "Role play with large language models," Nature, vol. 623, no. 7987, pp. 493–498, Nov. 2023, doi: 10.1038/s41586-023-06647-8.

[16] "LLM - Deliberation: Evaluating LLMS with Interactive Multi-Agent Negotiation Games," ICLR, 2024.

[17] S. Zhang et al., "Instruction Tuning for Large Language Models: A Survey," 2023, [Online]. Available: http://arxiv.org/abs/2308.10792.

[18] J. Huang et al., "Large Language Models Cannot Self-Correct Reasoning Yet," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.01798

[19] B. D. Nye, D. Mee, and M. G. Core, "Generative Large Language Models for Dialog-Based Tutoring: An Early Consideration of Opportunities and Concerns," 2023. [Online]. Available: www.github.com/opentutor/.

[20] Y. Shao, L. Li, J. Dai, and X. Qiu, "Character-LLM: A Trainable Agent for Role-Playing." [Online]. Available: https://github.com.

[21] P. Gupta and J. P. Bigham, "Improving Reliability in Dialogue Systems," 2023. [Online]. Available: www.lti.cs.cmu.edu.

[22] J. Huang et al., "Large Language Models Can Self-Improve," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 1051–1068.

[23] C. Shah et al., "Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies," Sep. 2023, [Online]. Available: http://arxiv.org/ abs/2309.13063

[24] S. Bhowmik, S. Sultana, A. A. Sajid, S. Reno, and A. Manjrekar, "Robust multi-domain descriptive text classification leveraging conventional and hybrid deep learning models," International Journal of Information Technology, pp. 1–13, Oct. 2023.