Challenges for Language Models Beyond Stochastic Parrots

Du Jiang
School of Intelligence Science and Technology
Peking University
dujiang@pku.edu.cn

Abstract

This essay discusses the challenges faced by statistical language models beyond stochastic parrots. Key challenges include contextual understanding, commonsense reasoning, handling ambiguity, conversational depth, and emotional intelligence. Designing evaluation tasks that test these properties is essential for advancing language models and enhancing their proficiency beyond imitation.

1 Introduction

Asking how to design tasks or evaluations to differentiate between simple imitation and true communication may not fit our true requirements, since the performance of ideal imitation might approach that of true communication with sufficient data and unlimited resources. Here we only focus on a relevant and crucial perspective: given limited resources and data, what are the most important challenges faced by statistical language models?

Language models have made remarkable advancements in recent years[7][4], but the journey to creating AI systems capable of genuine communication is far from over. When we talk about efficient language models beyond stochastic parrots[1], a series of challenges must be addressed to bring language models closer to human-level proficiency. This essay explores five key challenges: contextual understanding, commonsense Reasoning, handling ambiguity, conversational depth, and emotional intelligence.

2 Challenges and Evaluation Tasks

2.1 Contextual Understanding

Language models need to do more than string together words; they must grasp the context of a conversation. Designing tasks that evaluate a model's ability to maintain, adapt, and respond coherently within a given context is vital. Context-aware conversations and contextual questions can be used to test this property[6][2].

Scenario-Based Conversations Provide a scenario or context at the beginning of a conversation and require the model to generate responses that are consistent with the given context. The model's ability to maintain and adapt to context should be tested.

Contextual Questions The model is asked to answer questions that depend on prior parts of the conversation. The model's understanding of the conversation's context and ability to recall relevant information is evaluated.

Causal Reasoning Present tasks that involve causal relationships or events unfolding over time. The model should be able to infer causality based on contextual information.

2.2 Commonsense Reasoning

True communication requires models to possess commonsense reasoning abilities. This challenge involves creating tasks that assess the model's capacity to make logical inferences, answer common-sense questions, and engage in analogical reasoning. As studied in previous works[5][3], a language model should go beyond memorization and demonstrate genuine reasoning skills.

Multiple-Choice Questions The task requires the model to choose answers based on common-sense reasoning. These questions should be related to everyday situations or general knowledge.

Fill in the Blanks Create tasks where the model needs to complete sentences with missing information. Common-sense knowledge is essential for filling in the gaps accurately.

Analogies Develop analogy-based tasks that require the model to make logical and common-sense inferences. For example, presenting analogies like "A is to B as C is to ?" to evaluate its reasoning abilities.

2.3 Handling Ambiguity

Effective communication necessitates the ability to handle and resolve ambiguity[8]. Task design can incorporate sentences or phrases with multiple interpretations, requiring the model to disambiguate based on context. Evaluations may also encompass humor and puns, as detecting linguistic ambiguity is a key element of language understanding.

Ambiguity Resolution Construct tasks with ambiguous language or multiple interpretations. The model should be able to recognize the ambiguity and seek clarifications or provide responses that address different possible interpretations.

Humor and Puns Include tasks that involve humor, wordplay, or puns. Evaluating the model's ability to detect and generate humor or puns can reveal its capacity to handle linguistic ambiguity.

Sentence Disambiguation Provide sentences with ambiguous phrases and ask the model to disambiguate them based on the surrounding context.

2.4 Conversational Depth

Communication goes beyond isolated sentences. It involves extended, multi-turn conversations that maintain context and coherence. Tasks should challenge models to engage in deeper, contextually relevant interactions over multiple turns. Sequential reasoning and role-playing simulations can be employed to evaluate conversational depth.

Extended Conversations The goal is to handle extended, multi-turn conversations on a given topic or scenario. This requires the model's ability to maintain coherent and contextually relevant discussions over several interactions.

Sequential Reasoning Given a sequence of events or instructions over multiple conversation turns, the model should reason sequentially and provide responses accordingly.

Role-Playing Games Create interactive role-playing games or simulations where the model engages in multi-turn conversations with a human evaluator. The depth and quality of the interaction can be evaluated.

2.5 Emotional Intelligence

Emotional intelligence is essential for authentic communication. Although not a definitive indicator, it can help distinguish true understanding from stochastic parroting. We can evaluate a model's ability to recognize emotions in text, generate empathetic and contextually relevant responses, and maintain emotional consistency throughout a conversation[9].

Emotion Recognition Create tasks where the model identifies and acknowledges emotions expressed in user input.

Empathetic Responses Test the model's ability to generate empathetic and contextually appropriate responses to emotionally charged content.

Consistency Ensure that the model maintains emotional consistency throughout a conversation, matching the user's emotional tone.

3 Discussion

In the pursuit of building language models that move beyond stochastic parroting, it is evident that the challenges identified in this essay are interconnected and collectively crucial for achieving genuine communication. Contextual understanding is foundational, as models must comprehend and adapt to the ever-evolving conversation context. Commonsense reasoning allows them to engage in logical inferences, fostering richer and more informative interactions. Handling ambiguity and maintaining conversational depth ensure the coherency and relevance of responses over extended interactions. Emotional intelligence adds authenticity to these communications. Successfully navigating these challenges is a substantial step towards developing language models that truly understand and engage in meaningful conversations, enabling their use across diverse domains, from natural language interfaces to human-robot interactions.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922. 1

[2] Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. Contextual representation learning beyond masked language modeling, 2022. 1

[3] Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense knowledge in large language models, 2022. 2

[4] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017, 2023. ISSN 2950-1628. doi: https://doi.org/10.1016/j.metrad.2023.100017. URL https://www.sciencedirect.com/science/article/pii/S2950162823000176. 1

[5] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners, 2022. 2

[6] Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. Large language models know your contextual search intent: A prompting framework for conversational search, 2023. 1

[7] OpenAI. Gpt-4 technical report, 2023. 1

[8] Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. Task ambiguity in humans and language models, 2022. 2

[9] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. Emotional intelligence of large language models, 2023. 2