# ON THE LINGUISTIC CAPABILITIES OF STATE-OF-THE-ART LARGE LANGUAGE MODELS AND HUMAN: A COMPARATIVE STUDY

Weigong Pan, Hongming Zhang

University International College, Macau University of Science and Technology

#### Introduction

Since the seminal advent of GPT-3.5 in 2022, there has been considerable scholarly debate regarding the linguistic capabilities of Large Language Models (LLMs) and their comparability to human proficiency in language use. Proponents of LLMs contend that these models have achieved a level of language acquisition and application akin to human abilities, while detractors dismiss them as merely advanced forms of autocomplete software. To date, no conclusive evidence has firmly supported either perspective. Consequently, it is imperative to undertake rigorous analytic and experimental investigations to ascertain the veracity of these claims and to understand their implications for future research in both linguistic science and machine learning.

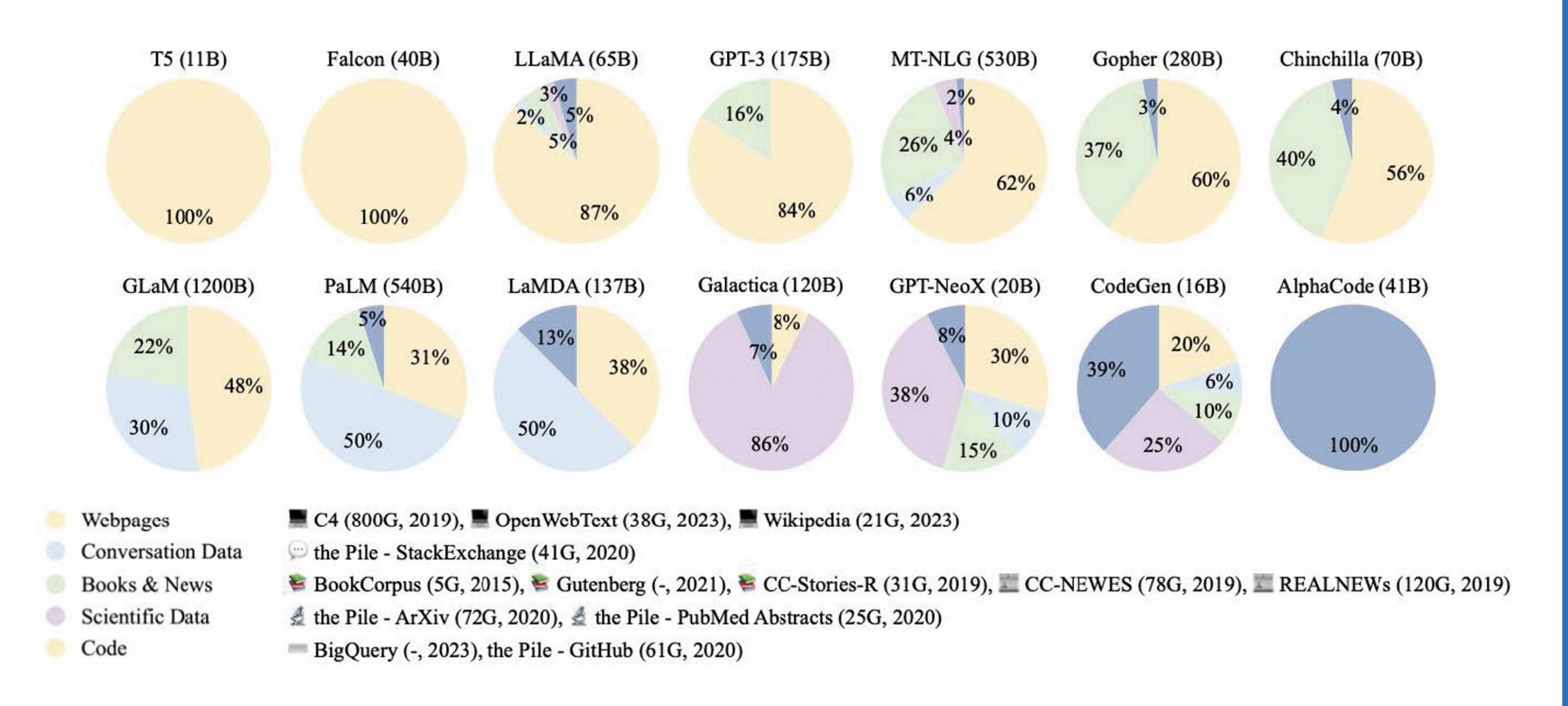


Figure 1. Data source used for the training of LLMs.(Zhao et al., 2023). Notice the obvious deficient in discourse data.

#### Questions:

To what extent do Large Language Models emulate human linguistic systems, and which significant components are they currently missing?

What methods can be employed to validate or refute the capability of LLMs as accurate models of human language systems?

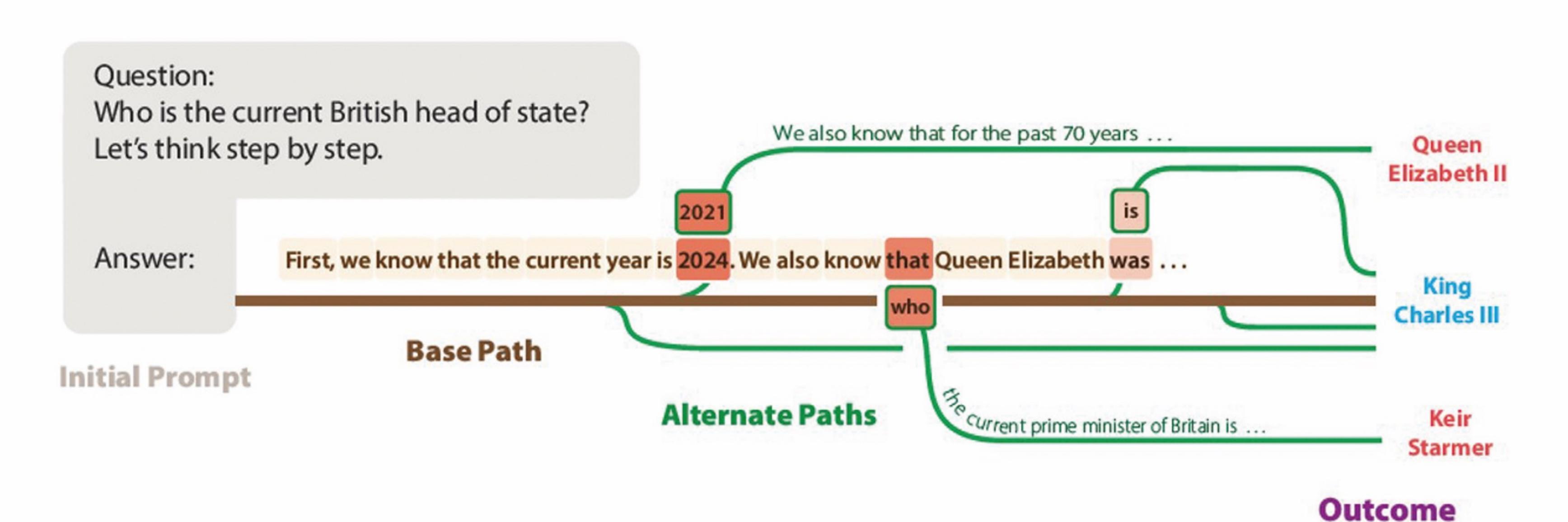


Figure 2. Example of semantical inconsistencies (Bigelow et al., 2024)

### Methods

This research systematically reviews studies that investigated the performance of large language models (LLMs) from a linguistic perspective. It includes works that employ quantitative methodologies or analyses of the models' internal structures to identify potential grammatical inconsistencies or lack of linguistic subsystems. Additionally, this review synthesizes findings from studies that utilized perceptual evaluations and qualitative analyses to assess how the absence of said linguistic subsystems in LLMs influences their linguistic output during human interactions.



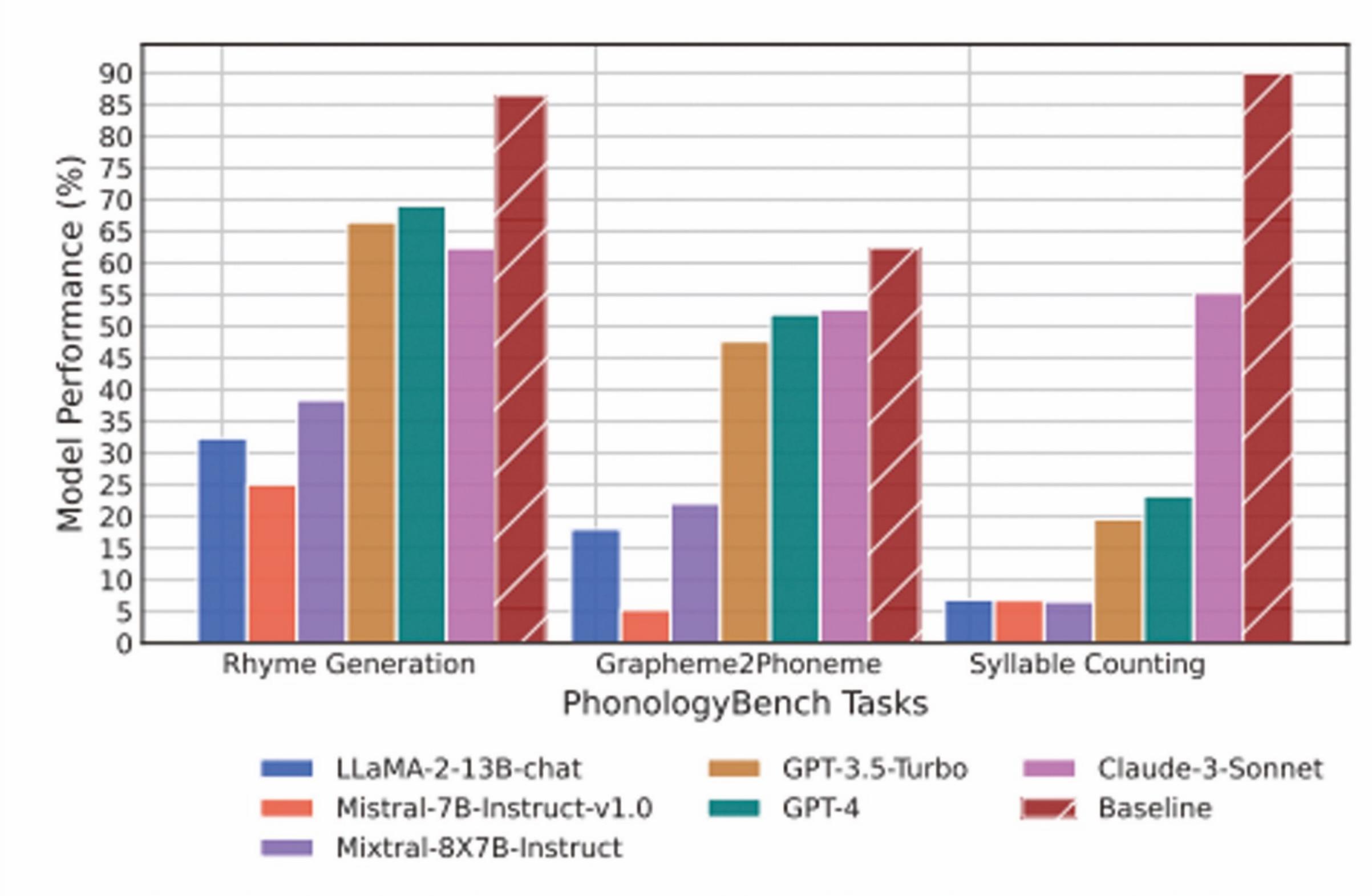


Figure 3. Phonological Performance of SotA LLMs (Suvarna et al., 2024)

## Current Findings

The study has recently discovered that Large Language Models (LLMs) exhibit deficiencies linguistic subsystems, several across phonology, encompassing semantics, and quantitative Further research pragmatics. suggests that these shortcomings may stem from prevailing tokenization. methods of Comparative analyses also indicate that, even when focusing solely on the fine-tuning process, the volume of data required for an LLM to acquire a new language is substantially greater than the amount needed by a human child.



- = U+5496
- = 11100101 10010010 1001101
- = 111001101 10010010 1001101

## 啡

- = U+5561
- = 111001101 100101 10100001
- = 11100101 10010101 10100001

Figure 4. An example on information loss in the tokeniazation process

#### References:

Bigelow, E., Holtzman, A., Tanaka, H., & Ullman, T. (2024). Forking Paths in Neural Text Generation. arXiv preprint arXiv:2412.07961. Suvarna, A., Khandelwal, H., & Peng, N. (2024). PhonologyBench: Evaluating Phonological Skills of Large Language Models. arXiv preprint arXiv:2404.02456.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.