LLMs at the Phonetics-Phonology-Orthography Interface

Zoe Tholen
University of Amsterdam

The popularization of Large Language Models (LLMs) has sparked discourse both about if an LLM could ever attain "true" knowledge, and about how this knowledge could be utilized to aid the sciences. The current study followed the behavioral testing procedures suggested by Ribeiro et al. (2020) to see if Google's Gemini Flash and OpenAi's GPT-5.1 employ human-like (productive, layered) phonetic and phonological "knowledge" while transcribing into and from IPA, as well as English and Korean orthography, or if they merely replicate IPA transcriptions from their training dataset. In the first test, models were asked to transcribe 14 English pseudowords (presented in IPA) into English orthography. The second test assessed whether the same pseudowords in English orthography could be implemented into Korean orthography, while retaining English phonotactics, rather than adapting Korean phonotactics. Lastly, a third test introduced an additional syllable (occurring in both languages) to the second test's stimuli and checked whether this needlessly affected the rest of the transcription. The tests were evaluated for the LLMs' abilities to recognize and rank language-specific phonotactic demands to implement phonetic information into the requested orthographical representations. For example, when English loanwords are adapted into Korean phonotactics, they sometimes experience syllabification as this allows for the maintenance of finer phonetic details in coda plosives (e.g.: notebook → 노트북 [no.tʰɯ.buk]), whereas a native English speaker might prefer to sacrifice phonetic detail in order to maintain the correct syllable structure (e.g.: notebook → 놑북 [not̚.buk]). The second test assessed whether such a non-canonical, English-favouring transcription could be applied by the LLMs.

Generally, the models performed well on all tasks, with failure rates between about 7% and 29% across models and tests. However, both LLMs scored worse on the third than on the second test, despite the simplicity of the additional task, perhaps indicating that the LLMs did not productively apply phonological rules in a human-like manner. The specific mistakes made and explanations offered by the LLMs further suggested that they underprioritized language-specific phonology and inappropriately mimicked input from their training datasets. These findings imply that publicly available LLMs may mostly be able to imitate productive phonological knowledge, but might not readily be used as aids in research demanding phonetic and phonological transcription or analysis without extensive prompting and human oversight.

References

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. *Annual Meeting of the Association for Computational Linguistics*, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442