Large Language Models and Children Have Different Learning Trajectories in Determiner Acquisition

Olivia La Fiandra[1], Nathalie Fernandez[2], Patrick Shafto[3], Naomi Feldman[1]
[1]University of Maryland, College Park. [2]University of California, Berkeley. [3]Rutgers University, Newark.

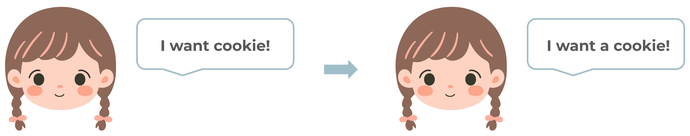
Research question

Prior research primarily compares LLMs and humans based on:

- Sample efficiency: data required to reach competency.
- End-state benchmarks: characterizing the final capabilities of the model.

These approaches provide little insight into the learning process itself, such as the order of language acquisition. Examining the learning process itself could reveal differences in the mechanisms underlying the language learning process.

Determiner Acquisition



Children's determiner acquisition is guided by both linguistic input and emerging pragmatic competence. They will often omit determiners before acquiring full competency, but otherwise use them correctly in adult-like ways.[1]

Methodology

We define determiner use as a multinomial distribution of three events:

- Definite article (DEF)
- Indefinite article (IND)
- Omission of the required determiner (OMS)

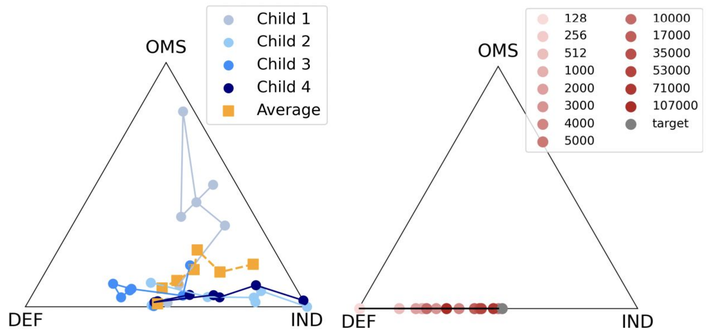
The learning trajectory captures the shift in these distributions over time.

Data

Speech samples of four children ages 18 -24 months from the Braunwald and Providence corpora. Determiner use annotated to evaluate distributions at two month intervals.

Pythia-70m[2] was evaluated at 15 checkpoints. Prompted with 100 lines of adult speech from the corpora. Model output of up to 20 tokens were annotated to evaluate distributions at each checkpoint.

Learning Process



Children: Initial distributions favor the indefinite determiner (IND) with some omissions.

Model: At the first measured point, the model exhibits a 100% definite (DEF) with no omissions recorded.

Model Prompt	Output
we're recording our voices we're recording our voices	we're recording our voices and the the the the the the the the the the the the the the the the the the the the
I did wanna hear why don't you sing it together	I did wanna hear why don't you sing it together, and the, and the, and the, and the, and the, and the, and the, and

Examples of model output at checkpoint 128.

Discussion

Learning trajectories show differences in the order of acquisition of determiners between the LLM and children. Whereas children begin omitting determiners and primarily using the indefinite determiner, the LLM does not make omissions and, instead, overuse the definite determiner.

The definite first trajectory may demonstrate a frequency bias, given that the definite determiner occurs roughly twice as often as the indefinite determiner in training data (e.g., COCA data).[3] At the earliest checkpoint (128), the model's output loops phrases like "and the," suggesting the might be one of the first words learned.

Yet children's behavior does not align with frequency-based expectations. The definite determiner imply abstract concepts like uniqueness and specificity (e.g., "I have the cookie" vs. "I have a cookie"), requiring a pragmatic competence that children may initially struggle with.[1, 4] Consequently, children may omit definite determiners rather than incorrectly using them, as the messages with omissions are often still understood (e.g., "I want cookie" vs. "I want the cookie").

References

1. Yuanfan Ying, Valentine Hacquard, Alexander Williams, and Jeffrey Lidz. 2024. Children do not overuse "the" in natural production. Proceedings of the 48th annual Boston University Conference on Language Development.
2. Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.
3. Mark Davies. 2008. The corpus of contemporary american english.
4. MP Maratsos. 1976. The use of definite and indefinite reference in young children. Cambridge University Press.

Acknowledgements

This research was supported by a University of Maryland Baggett Fellowship and by the University of Maryland Strategic Partnership: MPowering the State, a formal collaboration between the University of Maryland College Park and the University of Maryland Baltimore.