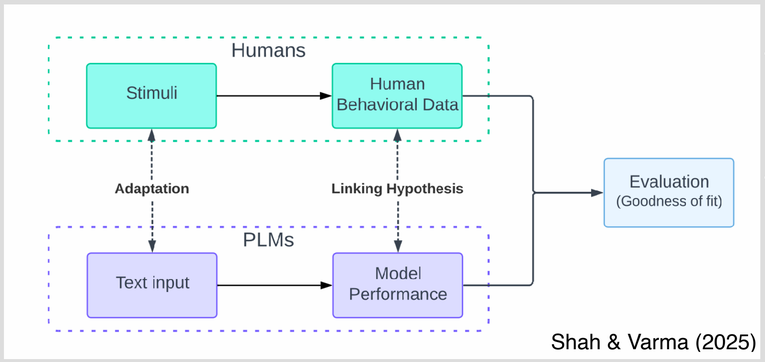
# Developmentally informed large language models for interdisciplinary collaboration

Yaxin Liu[1], Adam Green[1], & Stella Lourenco[2]
Department of Psychology, [1]Georgetown University & [2]Emory University

## How do we currently evaluate LLM-human alignment?

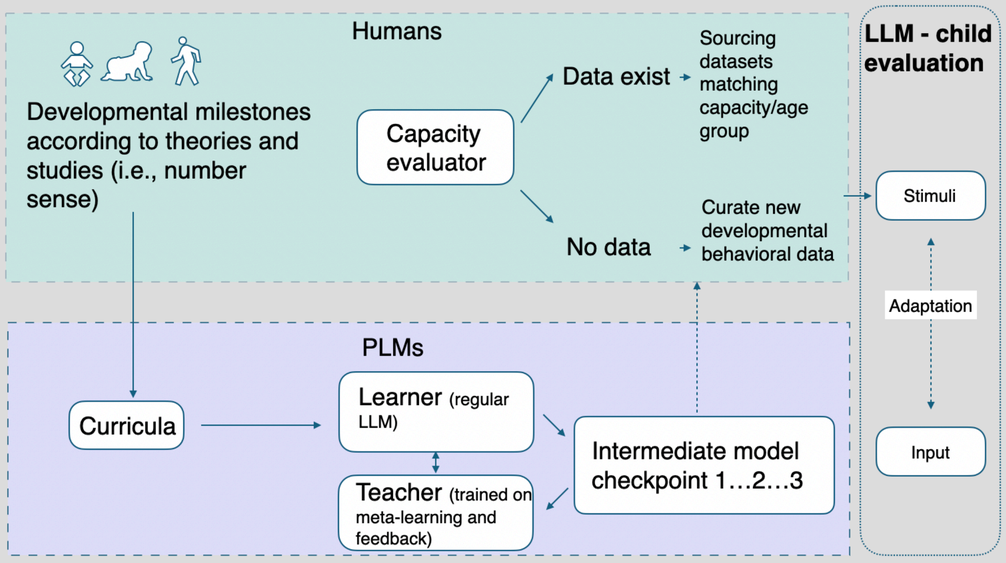



Shah & Varma (2025)

- Developmental progression during intermediate checkpoints is overlooked
- To understand how LLMs learn, examine how just final model weights, but also how abilities develop along the way

## How do we evaluate *developmentally* aligned LLM?

- Developmental alignment: as a language model is trained, its performance on core cognitive tasks may mirror the milestones of cognitive development in humans
- Why are "child-like" LLM learning stages needed? To make models more transparent, interpretable, and human-like (Shah et al., 2024)



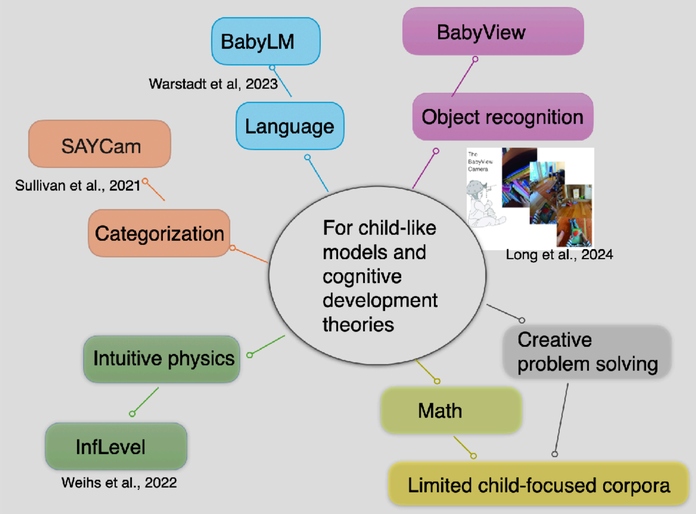

## Example datasets of LLMs that resembles adults and children

| Domain | Dataset/tasks | Cognitive alignment | Possibility of Developmental alignment |
|---|---|---|---|
| Linguistic competence | BLiMp | ✓ | ✓ |
| Conceptual understanding | Typicality effect | ✓ | ✓ |
| Numerical abilities | Mental number line | ✗ | ✗ |
| Fluid reasoning | Ravens Progressive Matrices | ? | ✓ |

Shah et al., 2024

| Pitfalls | Solutions |
|---|---|
| Using human and model similarities as alignment (i.e., neglecting developmental alignment and only final checkpoints are tested) | - Developmentally realistic training corpora or progressive "curricula"<br>- Evaluate models at multiple checkpoints during training |
| Mapping between model outputs and human performance lacks control | - Use small-scale empirical human data to establish direct evidence of human performance (Ivanova, 2025)<br>- Control for context and prompt wordings (instructions) |
| Mismatch in data scale and modality (e.g., Pre-trained LLMs trained on vast data vs. human only require few-shots learning) | - Curate developmentally plausible training data that matches human learning experiences<br>- Use smaller, structured datasets reflecting child input |
| Limited interpretability and individual variation | - Parse trees, attention analysis<br>- Simulate individual differences |

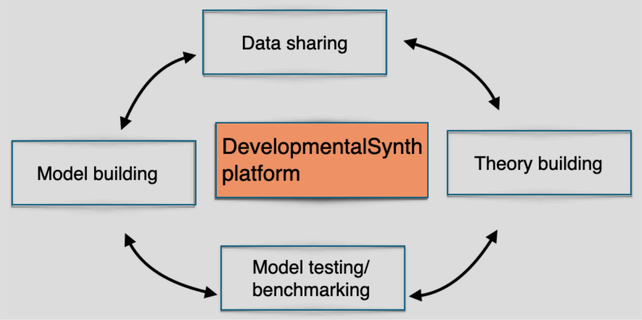
## Current Developmental Datasets and Initiatives




- Open source
- Developmental
- For research in cognitive science

## Resources for community building and collaboration

**A hypothetical platform for developmental researchers and ML scientists**




- MetaLab
- ManyBabies

References

Ivanova, A. (2025). *How to evaluate the cognitive abilities of LLMs*.
Long, B., et al. (2024). *BabyView: Object recognition and early visual learning from an infant's perspective*.
ManyBabies Consortium. (2020). *Quantifying sources of variability in infancy research using the infant-directed speech preference*.
MetaLab. https://langcog.github.io/metalab/
Shah, P., et al. (2024). *Development of Cognitive Intelligence in Pre-trained Language Models*.
Shah, P., & Varma, S. (2025). *The potential – and the pitfalls – of using pre-trained language models as cognitive science theories*.
Sullivan, J., et al. (2021). *SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective*.
Warstadt, A., et al. (2023). *BabyLM: Training language models on child-directed speech for cognitive development research*.
Weihs, L., et al. (2022). *Benchmarking Progress to Infant-Level Physical Reasoning in AI*.