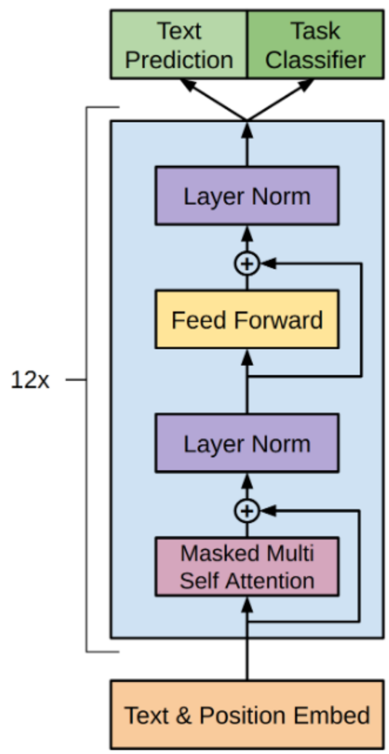
Cognitive modeling of development using AI tools

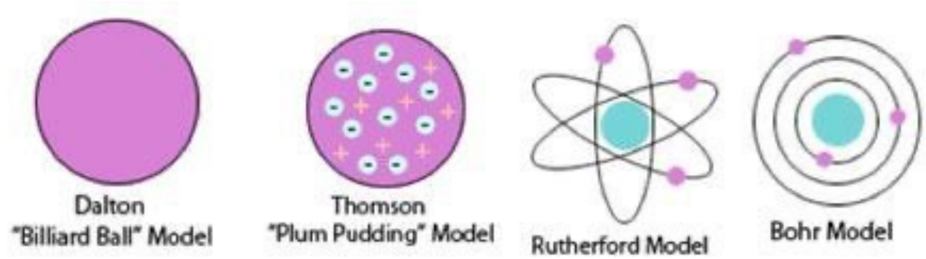
Michael C. Frank
Stanford University



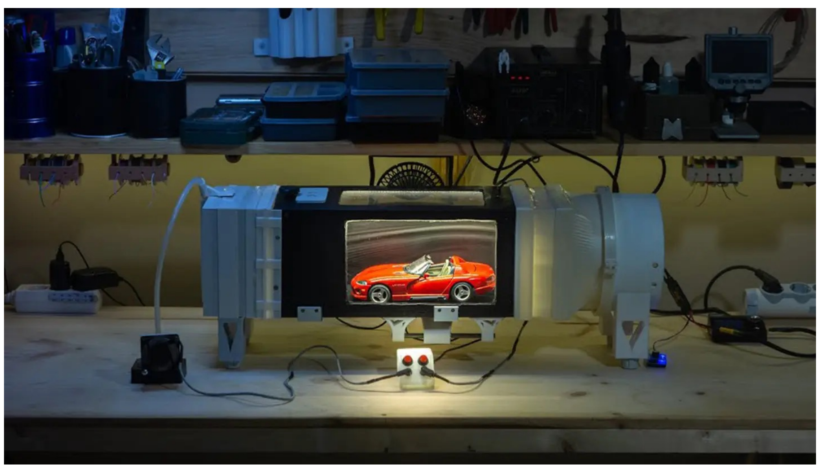
What can recent progress in AI tell us about the human mind?

AI models can act as scientific models

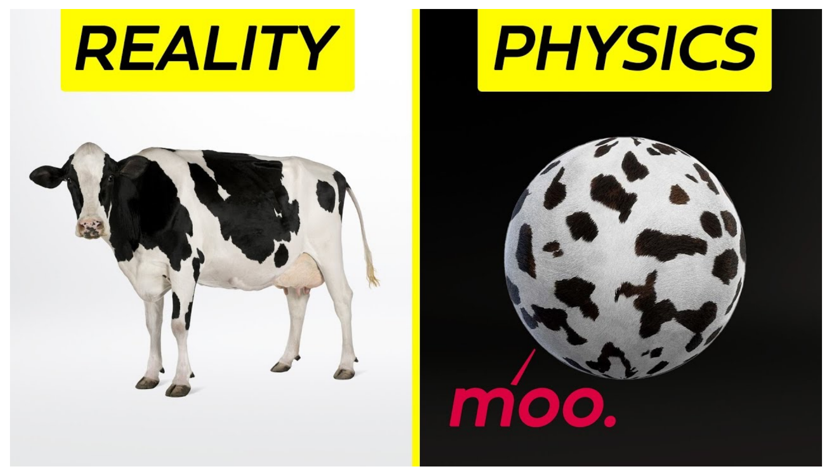
Simplification to enable explanation/reasoning



Models afford causal intervention

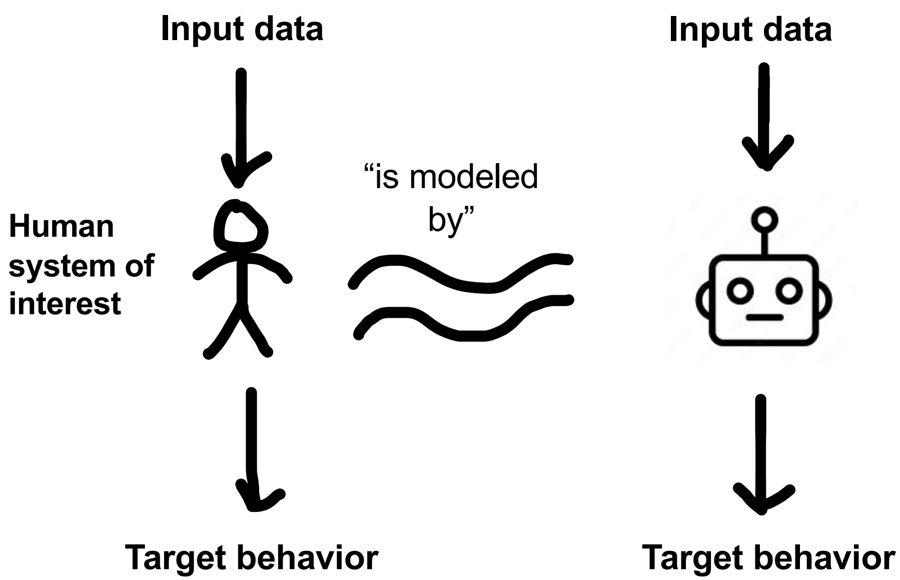


Models can risk oversimplification



Frigg & Hartmann (2020), SEP

Cognitive modeling with LLMs



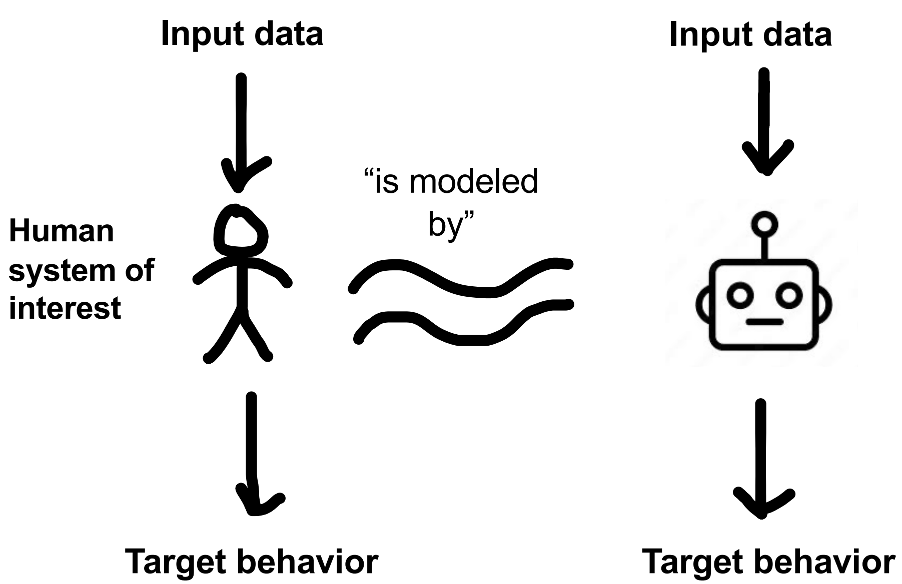
Controlled rearing: Intervene on input data and measure effects on behavior

Probing: investigate causal dependency between internal state and observed behavior

Behavioral evaluation: Measure behavior under different experimental conditions

Frank (2024), *Nature Human Behavior*

Cognitive modeling with (open) LLMs



OpenAI

Can't retrain the model
Don't know what the data are

Can't examine the representations
Can't intervene on them

Can evaluate new scenarios
No guarantee of reproducibility

RESEARCH ARTICLE | COMPUTER SCIENCES

ONLY

Using cognitive psychology to understand GPT-3

Marcel Binz and Eric Schulz Authors Info & Affiliations

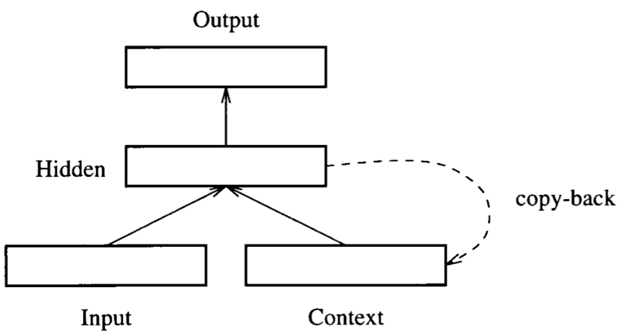
Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received October 29, 2022; accepted November 27, 2022

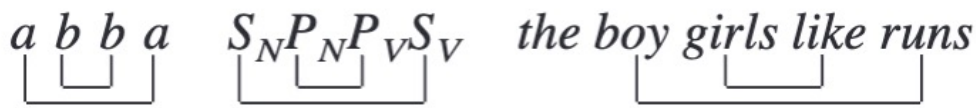
February 2, 2023 | 120 (6) e2218523120 | https://doi.org/10.1073/pnas.2218523120

Frank (2024), *Nature Human Behavior*

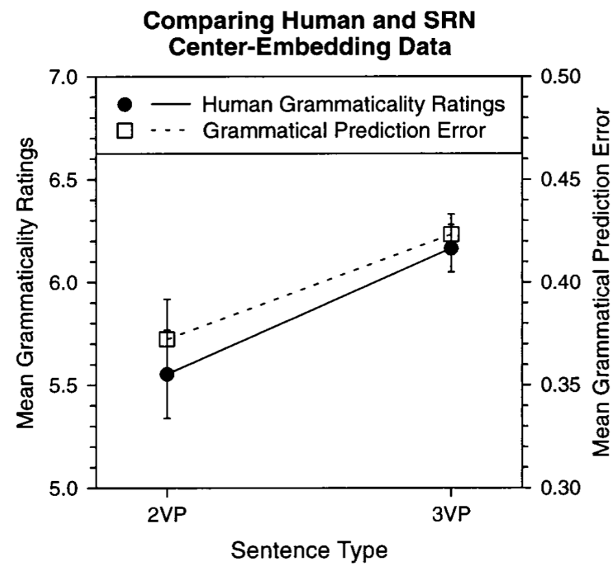
Cognitive modeling: example

Simple recurrent network (Elman, 1990)





The boy girls dogs bite like runs

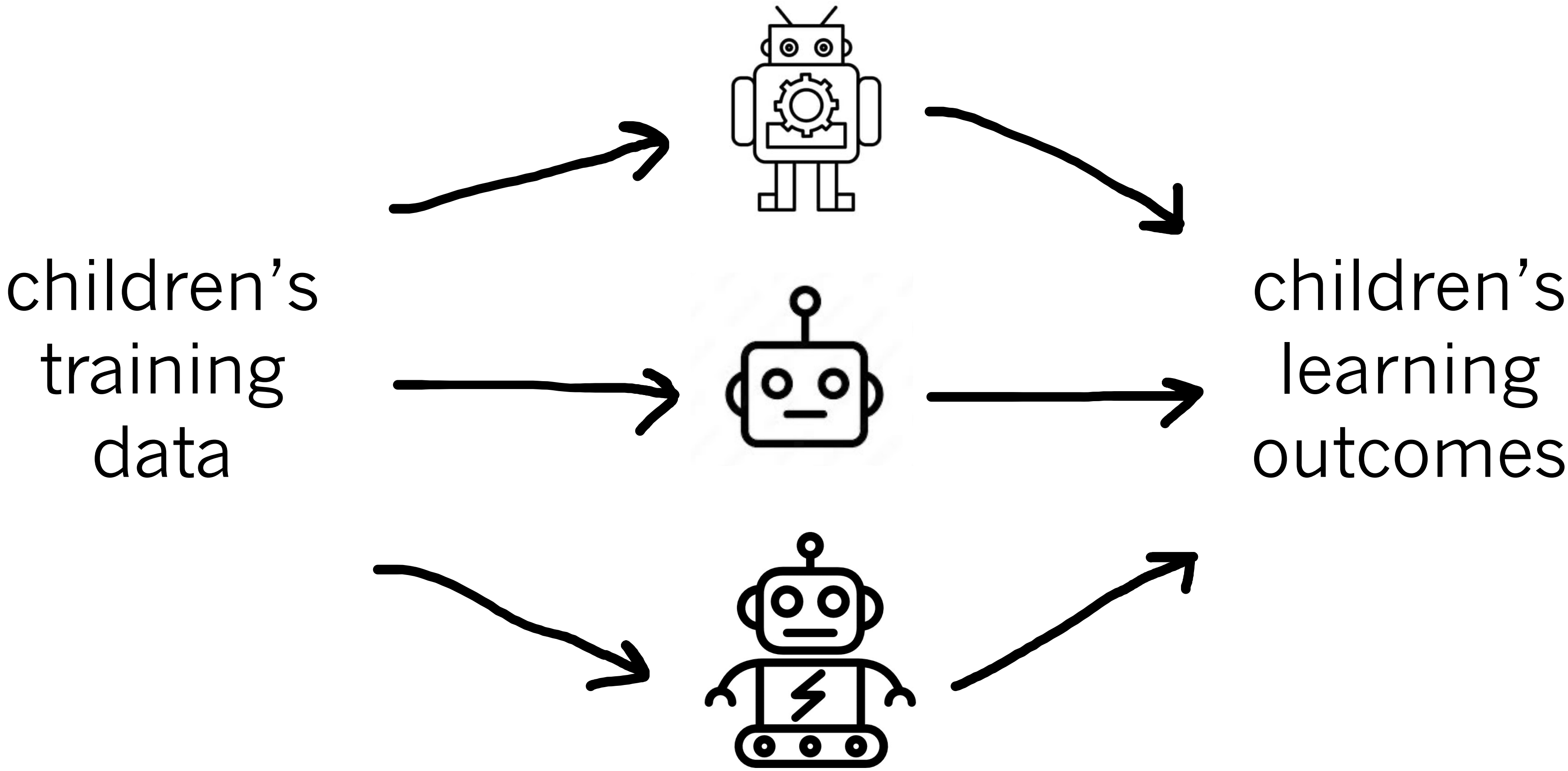


Proof of concept strategy: gradient increases in processing difficulty without prespecification

Contrasting training data produced different results

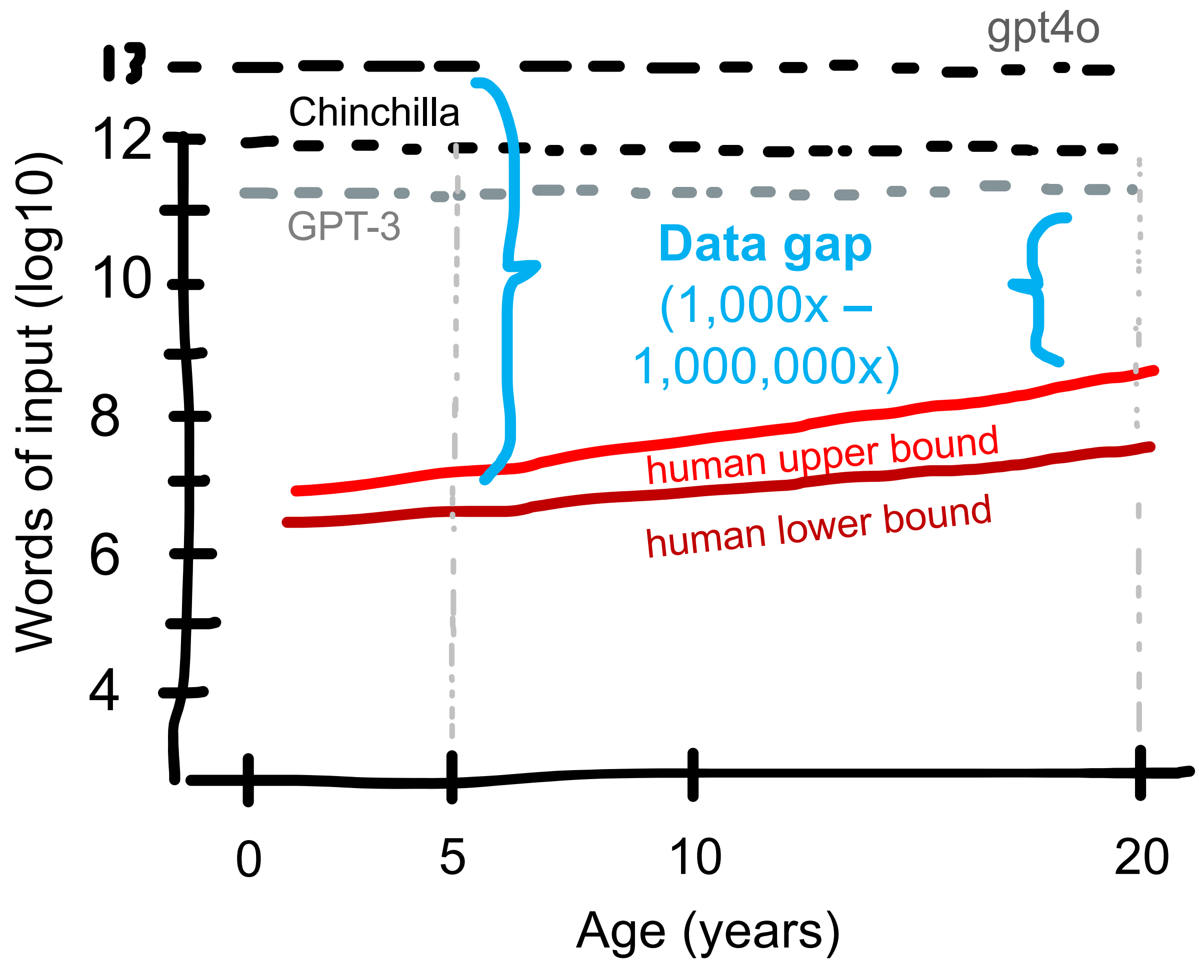
Christiansen & Chater (1999)

Opportunity: using AI as models of human learning



Frank (2024), *Nature Human Behavior*

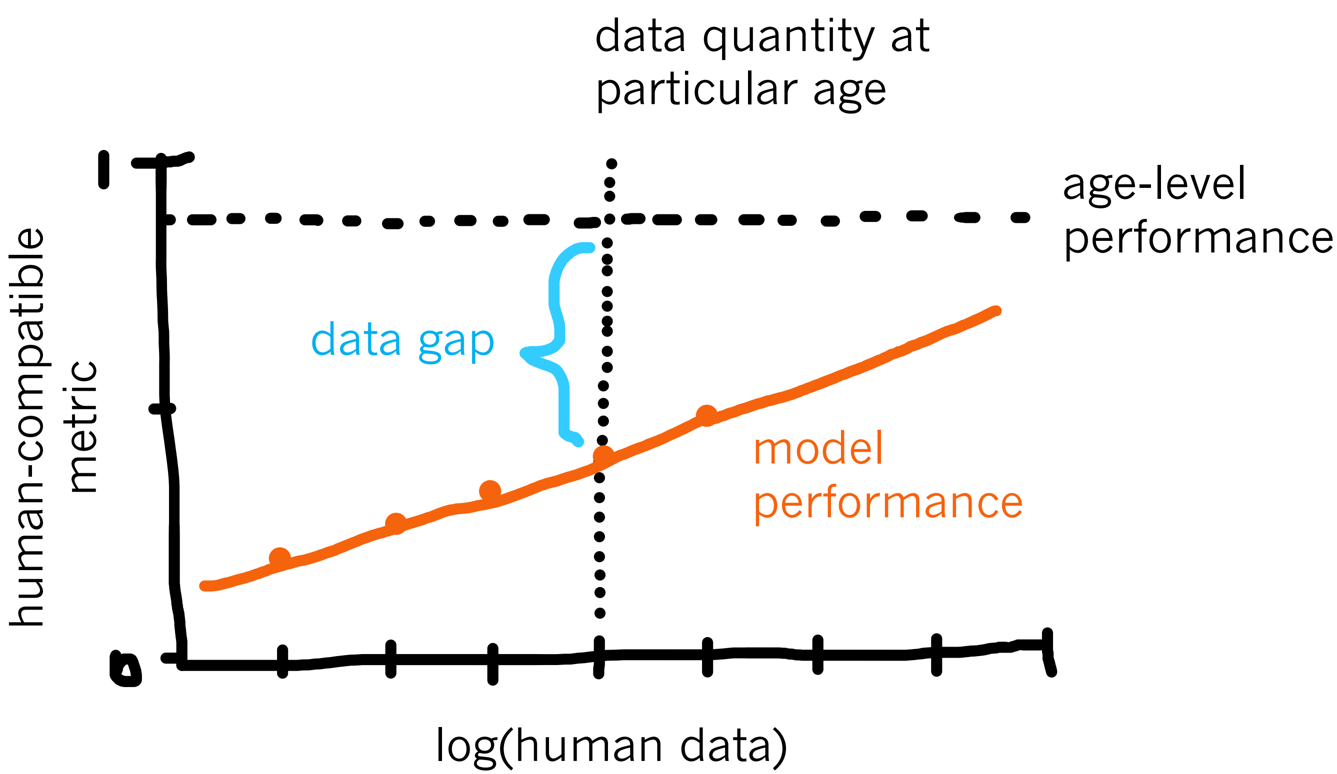
Data gaps in language



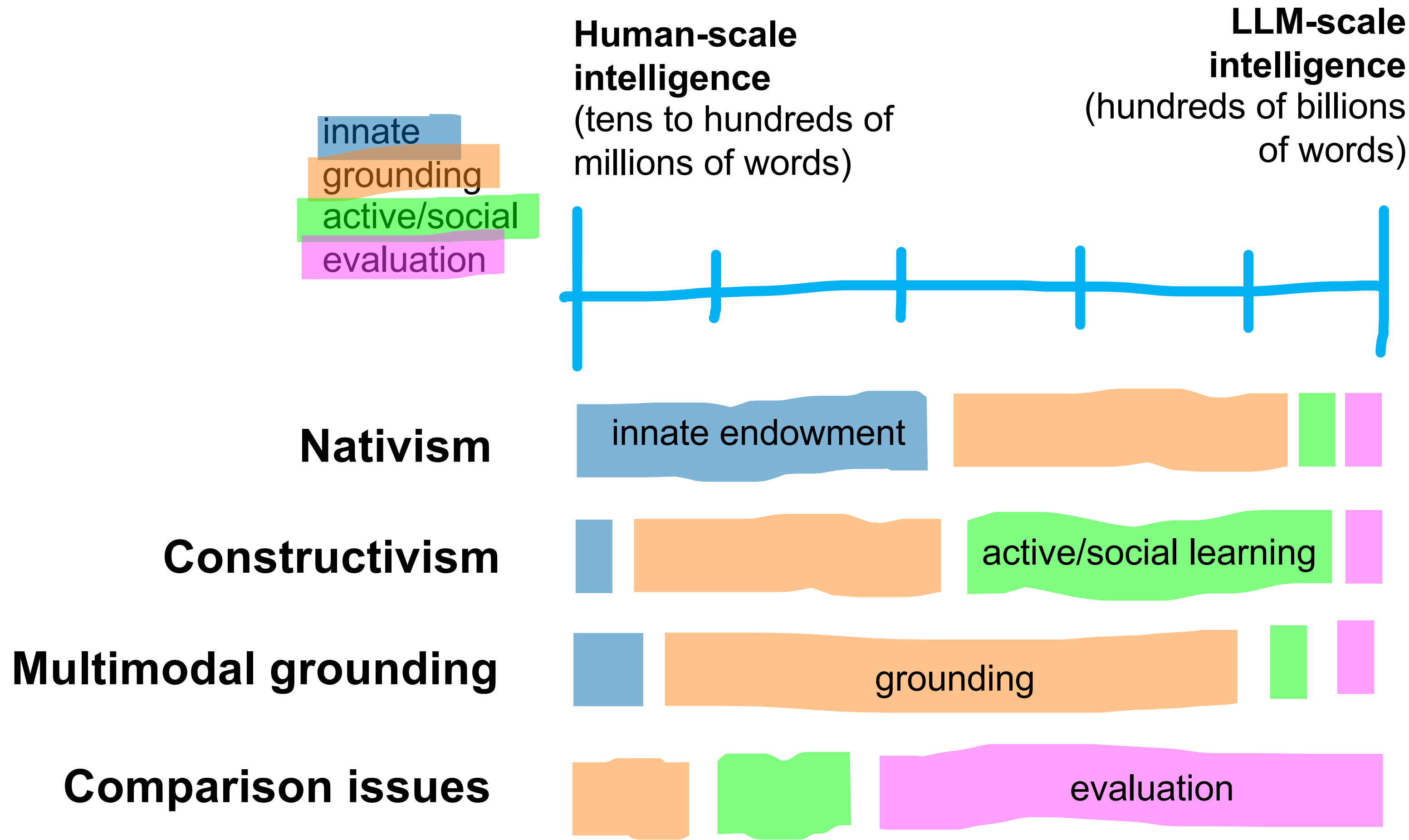
Frank (2024), *Trends in Cog Sci*

Defining task-specific data gaps

For a specific task, e.g., grammar learning, visual categorization...

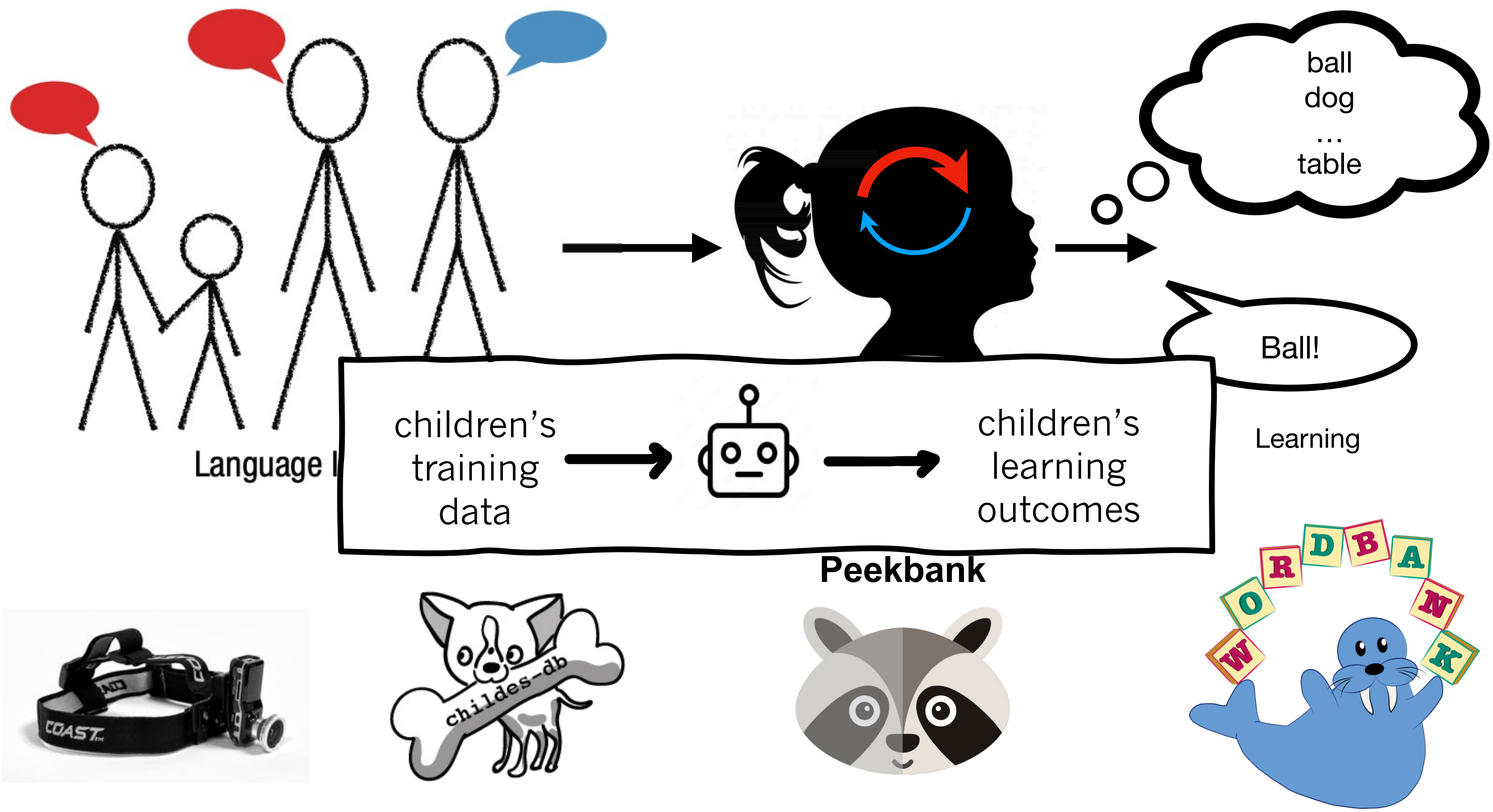


How could we explain data gaps?



Frank (2024), *Trends in Cog Sci*

My lab builds [**open**] data resources

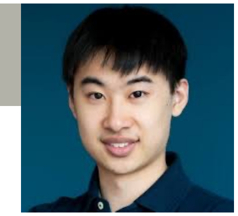


Outline

1. **Developmental training data** (x axis)

2. Developmental evaluation (y axis)

Is kids' language input especially good?

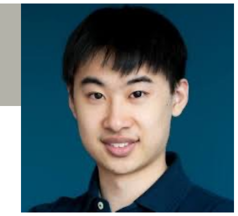


- Trained GPT-2-small on 29M words of:
  - CHILDES
  - TinyDialogues (new developmental dialogue corpus)
  - BabyLM (Wikipedia, transcripts, books, etc.)
- Evaluated on:
  - Zorro – 2AFC minimal pairs with child-directed vocab
  - Word similarity (pairwise distance)
- Child-directed speech resulted in LOWER performance...

Model	Zorro	WS
CHILDES	77.77%	0.24
TD	79.42%	0.41
BabyLM	81.75%	0.42

Feng, Goodman, & Frank (2024), *EMNLP*

Is the developmental ordering of language useful?

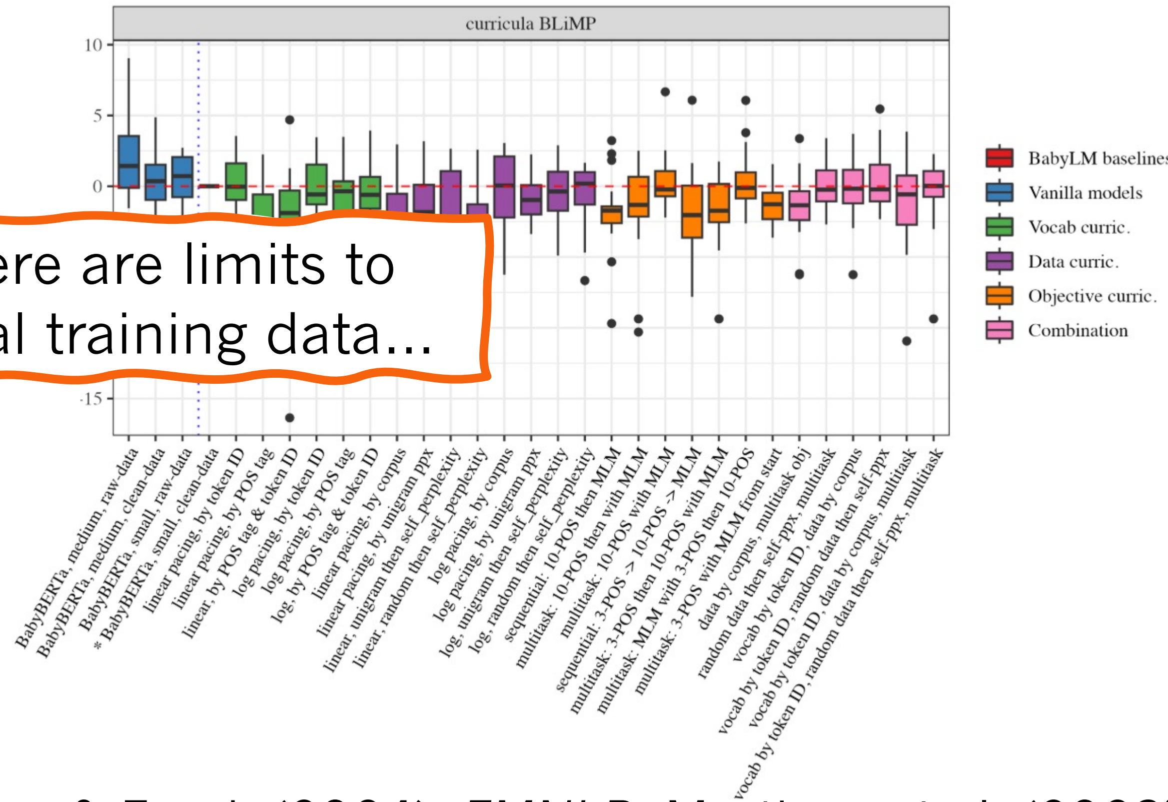


Ours

Model	Order	Zorro	WS
CHILDES	Age	73.54%	0.19
CHILDES	Reverse	74.18%	0.18
CHILDES	Random	74.76%	0.18
TD	Age	77.	
TD	Reverse	75.	
TD	Random	77.54%	0.32

No effects of age-ordering in children's data...

Curricularization also not useful in BabyLM data



But there are limits to unimodal training data...

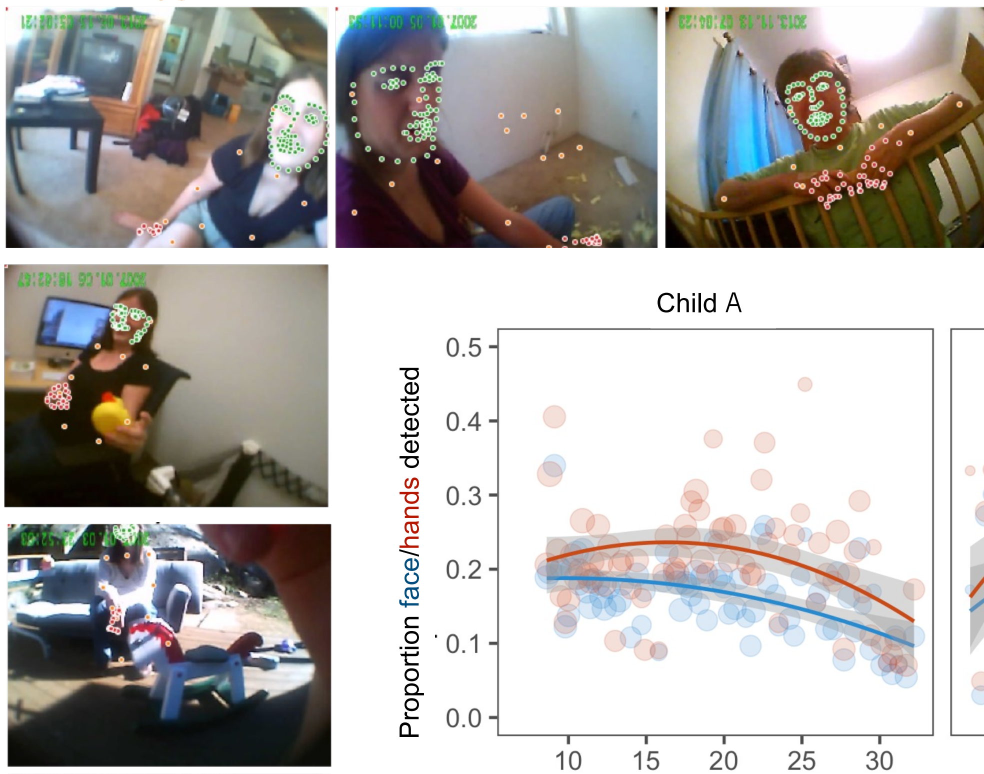
Feng, Goodman, & Frank (2024), *EMNLP*; Martinez et al. (2023)

SAYCam – a dataset of egocentric video

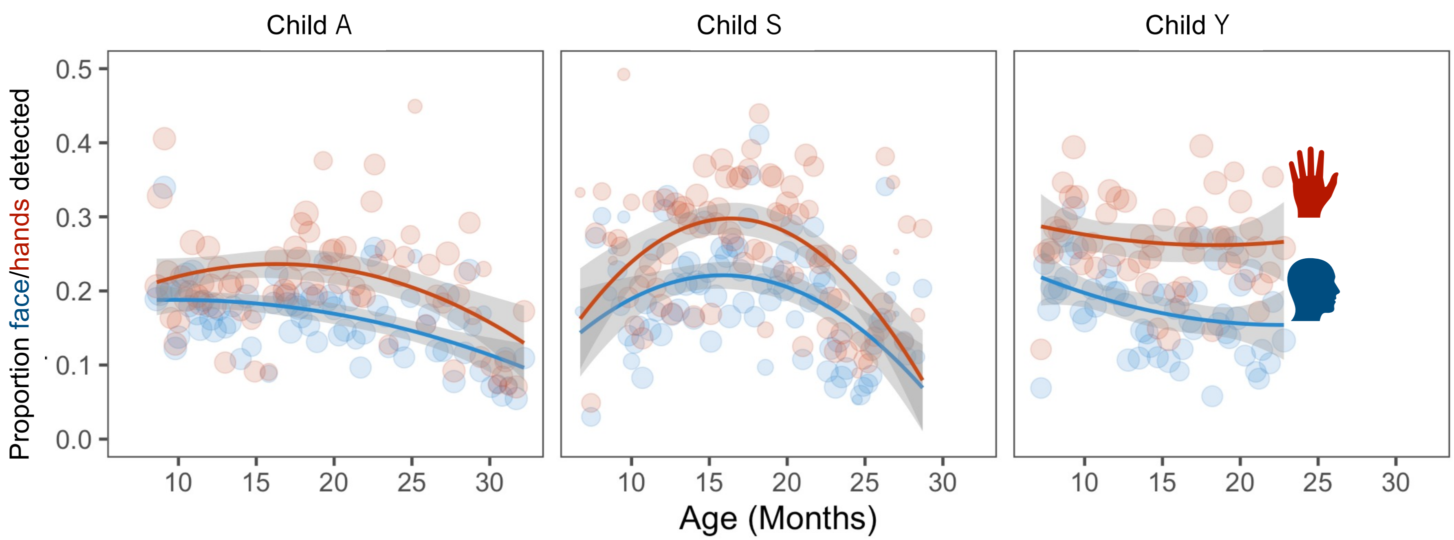
- 3 children, ~470 hours
  - 2 hrs/week from 6 – 30 months – 2013 - 2016
- 640x480, fisheye lens, relatively low-quality audio
- Children of developmental psychologists
- Shared openly through Databrary.org
  - Accessible to researchers via institutional agreement

Sullivan et al. (2021), *Open Mind*

Characterizing the social information in the infant view



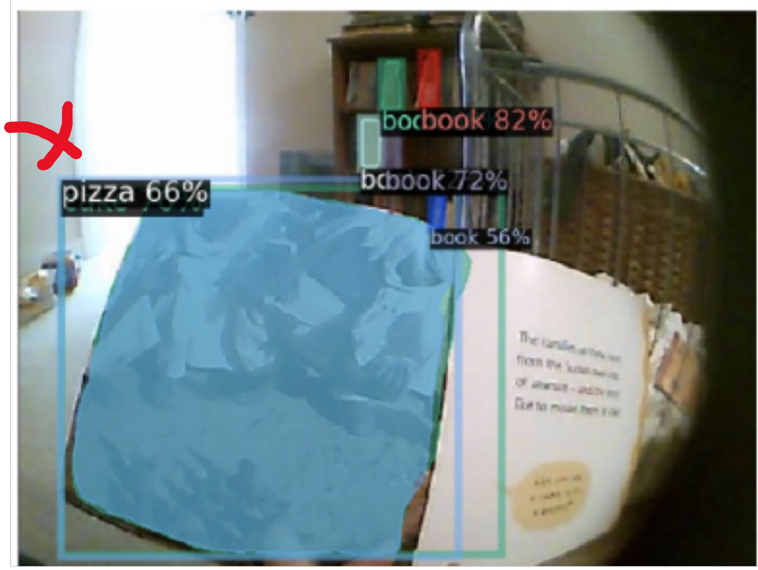
OpenPose (2017) detections
Faces: P=.70, R=.58 (Nose keypoint)
Hands: P=.73, R=.40 (Wrist keypoint)



Long, Kachergis, Agrawal, & Frank (2022), *Dev Psych*

Insufficient for understanding (supervised) category learning

1. Videos too low-resolution for segmentation models
2. View angle often obscures what children are interacting with
3. Infants see some objects way more than others (Clerkin et al., 2017)



Before fine-tuning

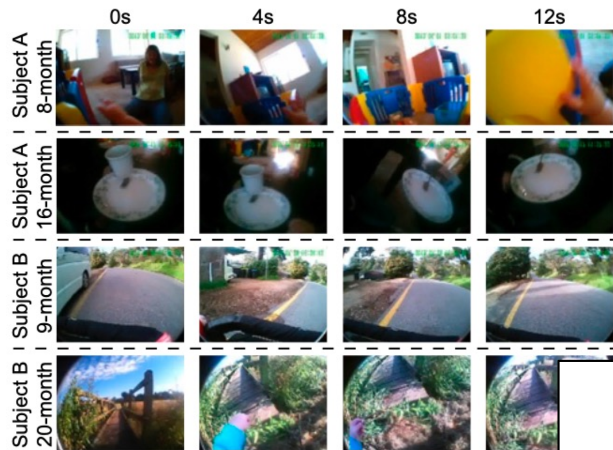
After fine-tuning

Mask R-CNN (ResNet + FPN); He et al., 2017
2215 frame segmentations for 10 frequent categories

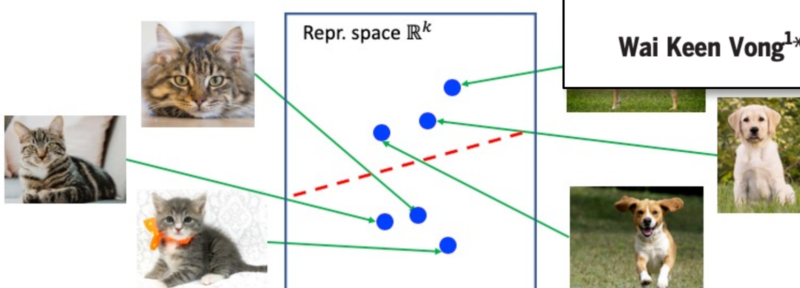
Long et al. (2021), *Proc. CogSci*

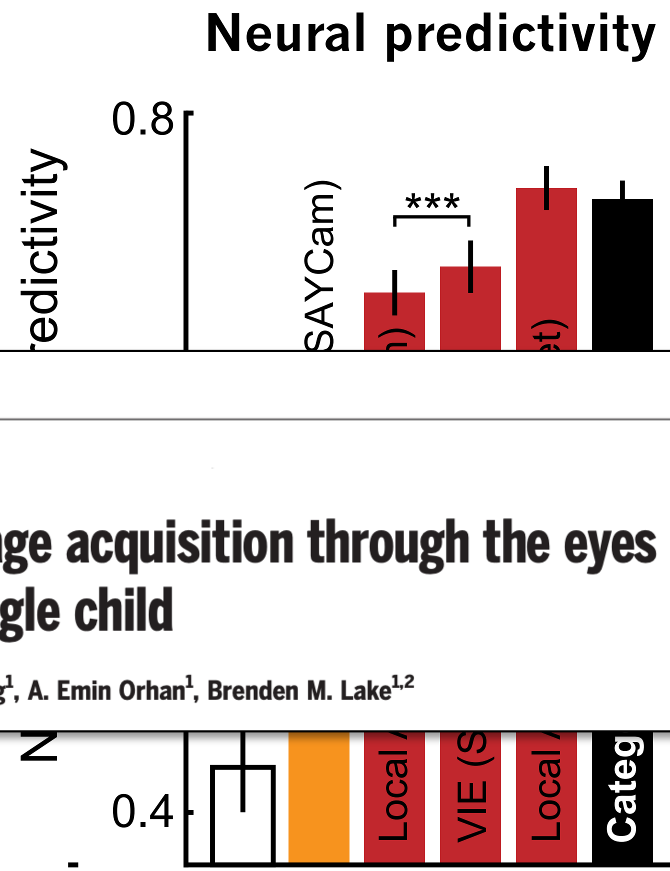
Can we train models to learn from these data?

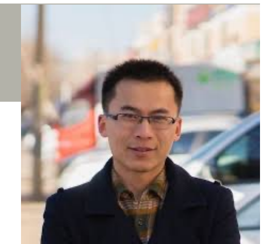
SAYCam headcam data

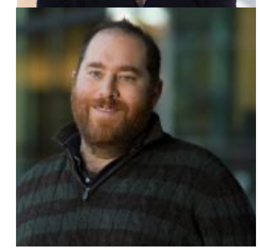


Contrastive learning alg









RESEARCH

MACHINE LEARNING

Grounded language acquisition through the eyes and ears of a single child

Wai Keen Vong[1]*, Wentao Wang[1], A. Emin Orhan[1], Brenden M. Lake[1,2]

Zhuang et al. (2020), *PNAS*

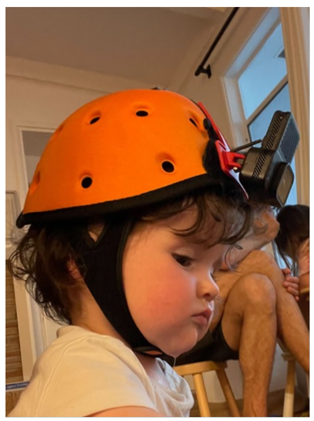
BabyView: high-resolution camera design

Based on ultra Lightweight GoPro Hero Bones

Easy to build, easy to deploy (3D printed mount + components: ~$400/unit)

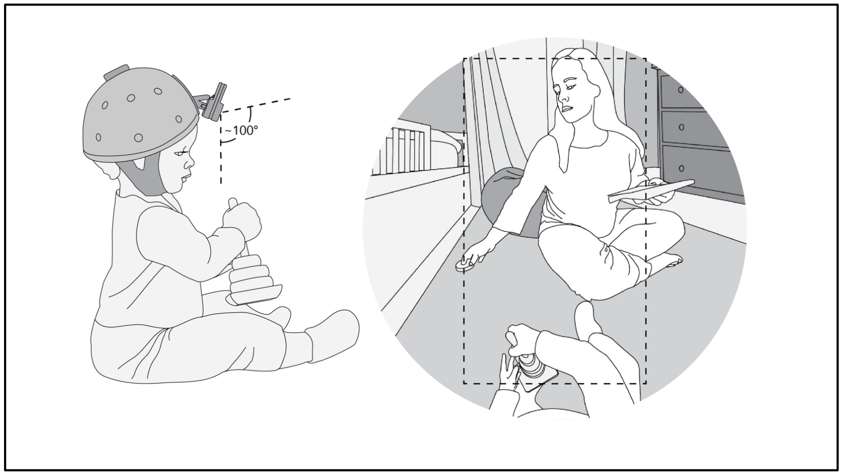
Accelerometer and gyroscope to identify self-motion

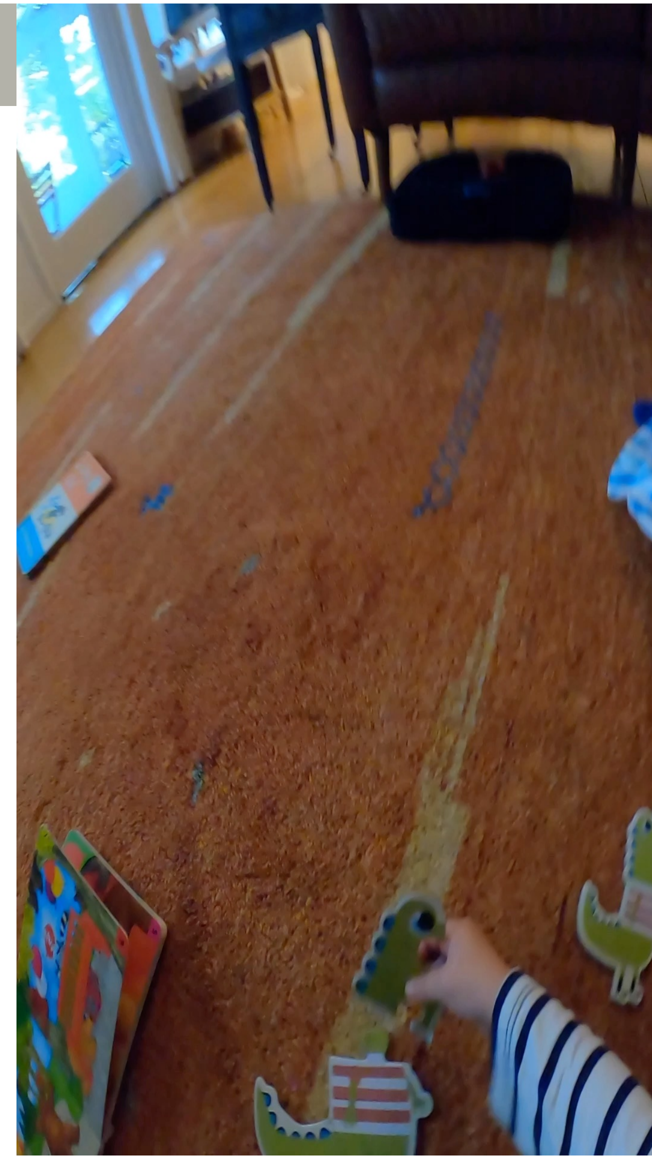
Daylight



Long et al. (2023), *BRM*

The BabyView dataset (1/25)



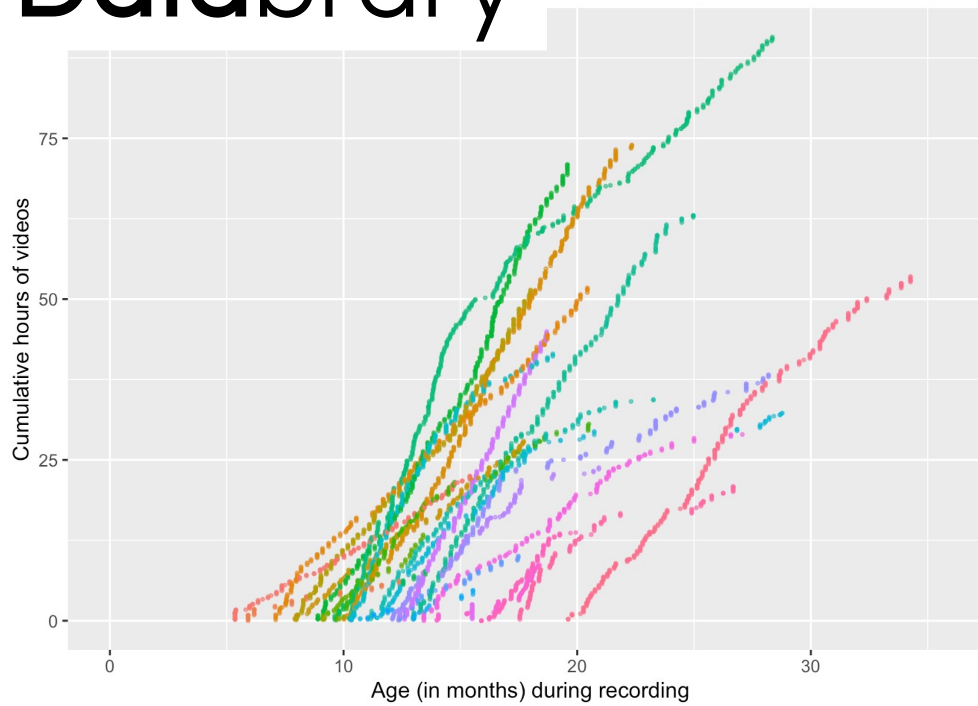


- **BV-main** (home recording): 882 hrs
- **Luna** (Lake daughter, different camera): 71 hrs
- **Bing** (preschool): 111 hrs
- **Total to date**: 1065 hrs

BabyView details

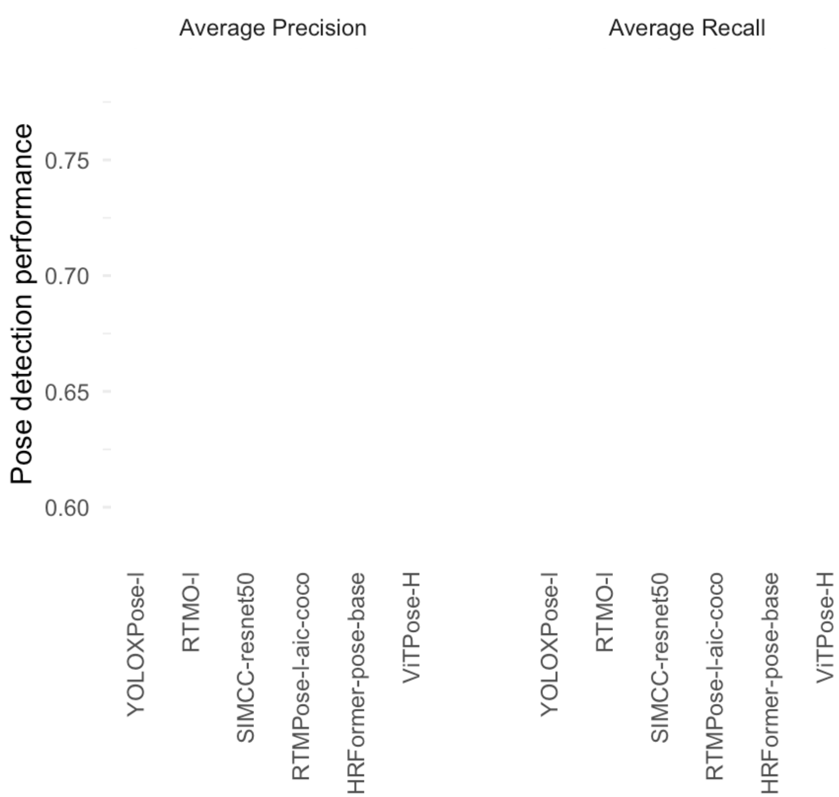
- ~25 families contributing to main (home recording) dataset
  - Mostly highly educated families
  - White, Asian, and mixed race
- All recordings at home with all adults consented
- Parents given multiple opportunities to remove videos
- Release through Databrary.org
  - Access for not-for-profits with IRB approval
  - Predicated on no reidentification & redistribution

Databrary



The BabyView dataset: Pose

Higher resolution videos & bigger models: improved pose detections (all keypoints)

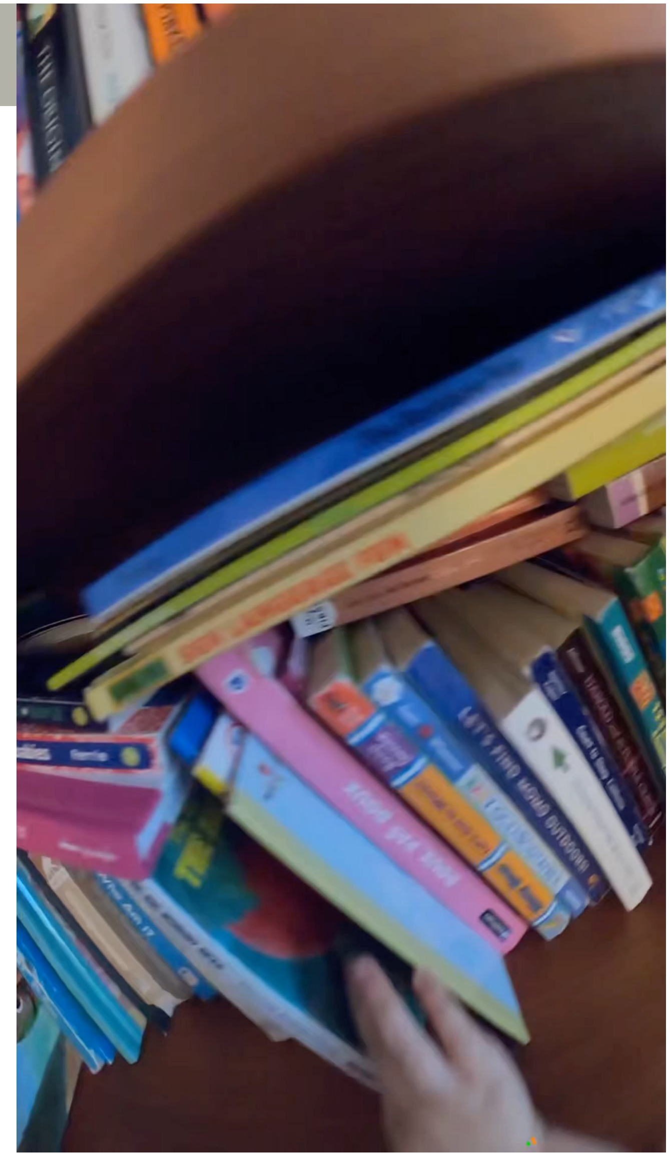


COCO Validation

BabyView Validation

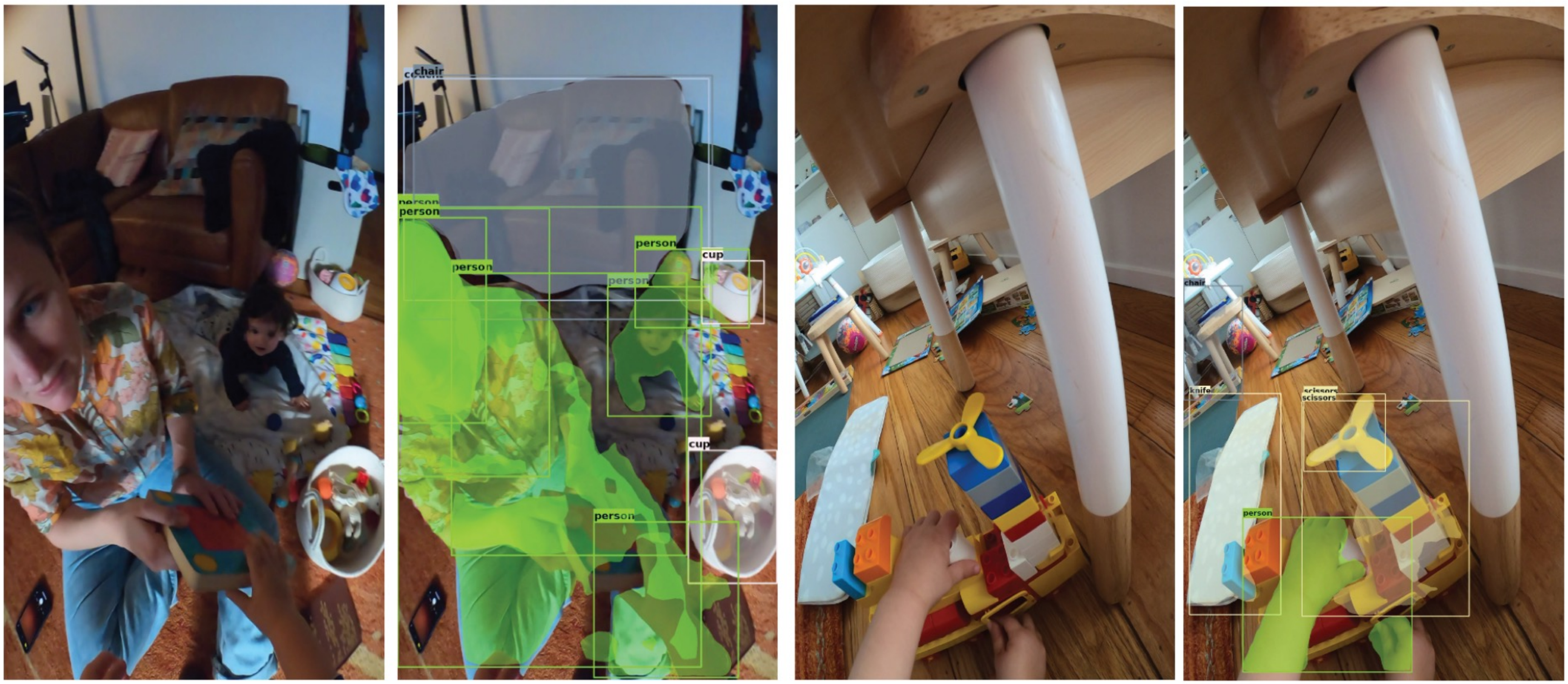
Dot size = #parameters

Long, Xiang, Stojanov et al. (2024), *arxiv*



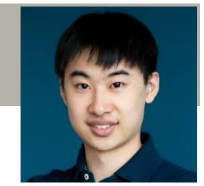
The BabyView dataset: off-the-shelf models

Better resolution: better (though imperfect) off-the-shelf object segmentations



Example Mask R-CNN segmentations (confidence > .3)

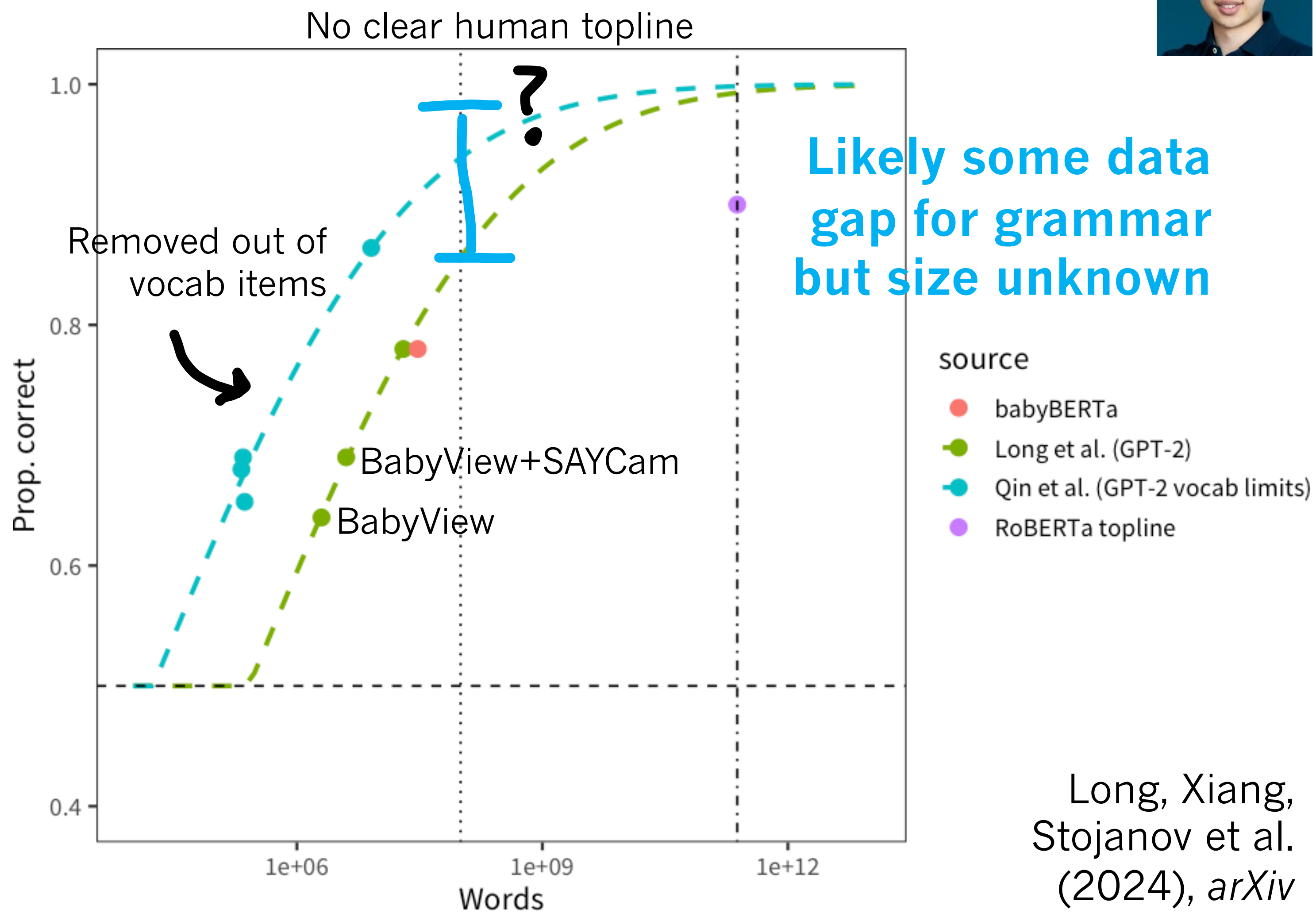
Training language models on BabyView+SayCam



Speech processing for SAYCam and Babyview:

Whisper v3 transcripts
All data WER = .38
Parents WER = ~.34
<18mo WER = ~1
18-30mo WER = .56

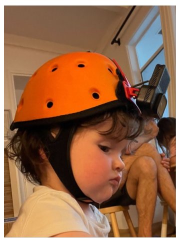
Also automated diarization (F around .6-.7; Lavechin et al. 2020)



Long, Xiang, Stojanov et al. (2024), *arXiv*

Training CV models on BabyView+SAYCam

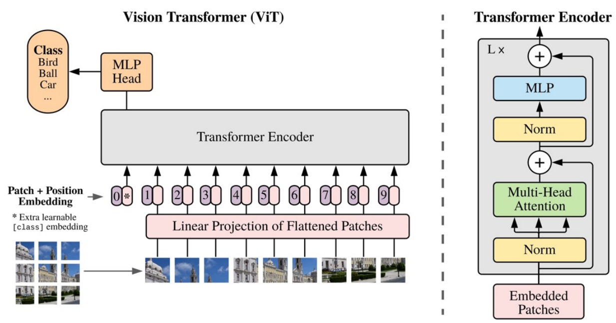

How well do self-supervised vision models perform as they learn from increasing amounts of developmental egocentric video data?




~900 hours total:
SAYCam + 433 hours BabyView
Home recordings

Training on 1%, 5%, 10% 25%, 50% & 100% of combined datasets

DINOv2
Self-supervised
transformer ViT-B/14



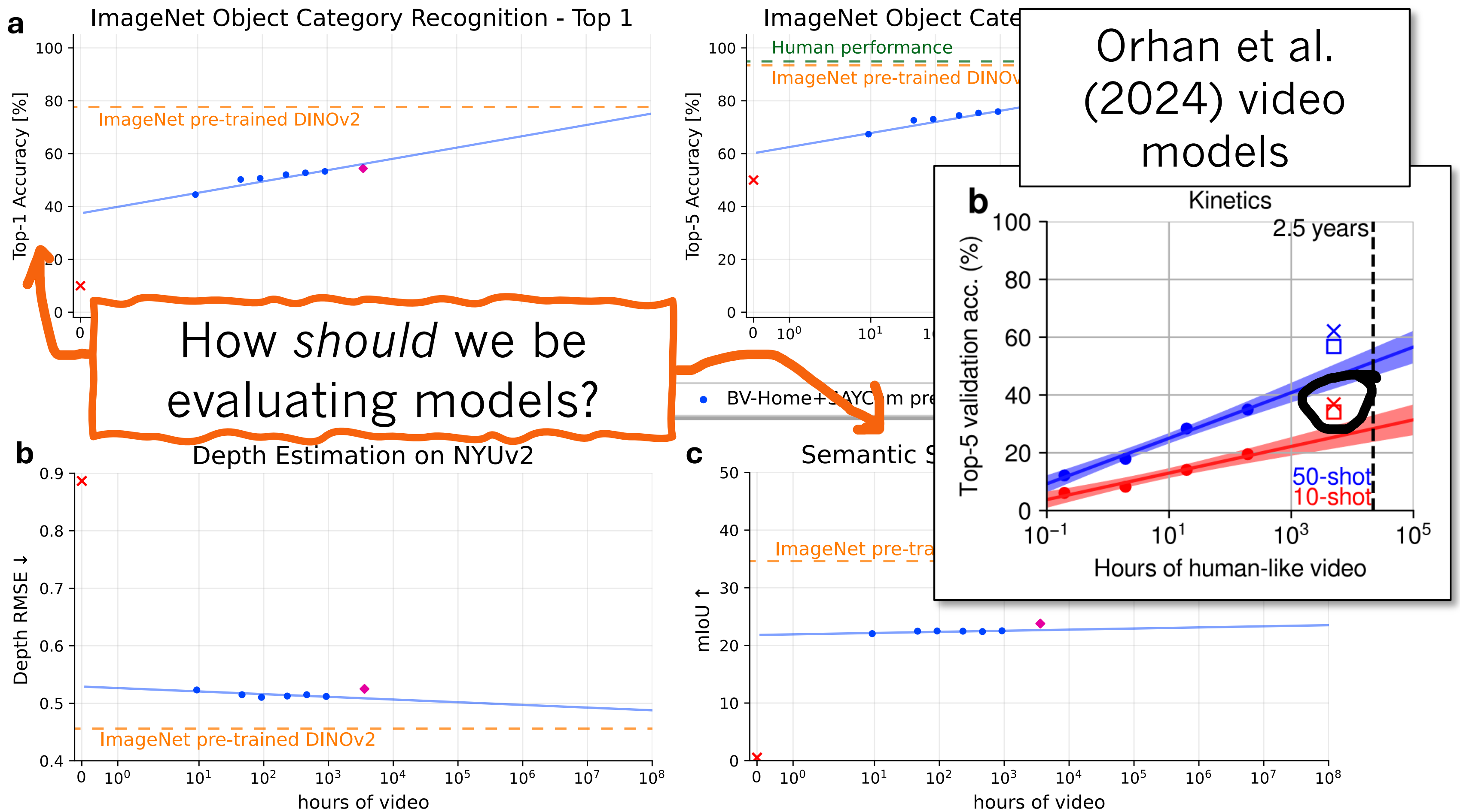
Oquab et al., 2023

Meta

Downstream tasks:
1. ImageNet Classification
2. Depth Estimation
3. Semantic Segmentation

Long, Xiang, Stojanov et al. (2024), *arXiv*

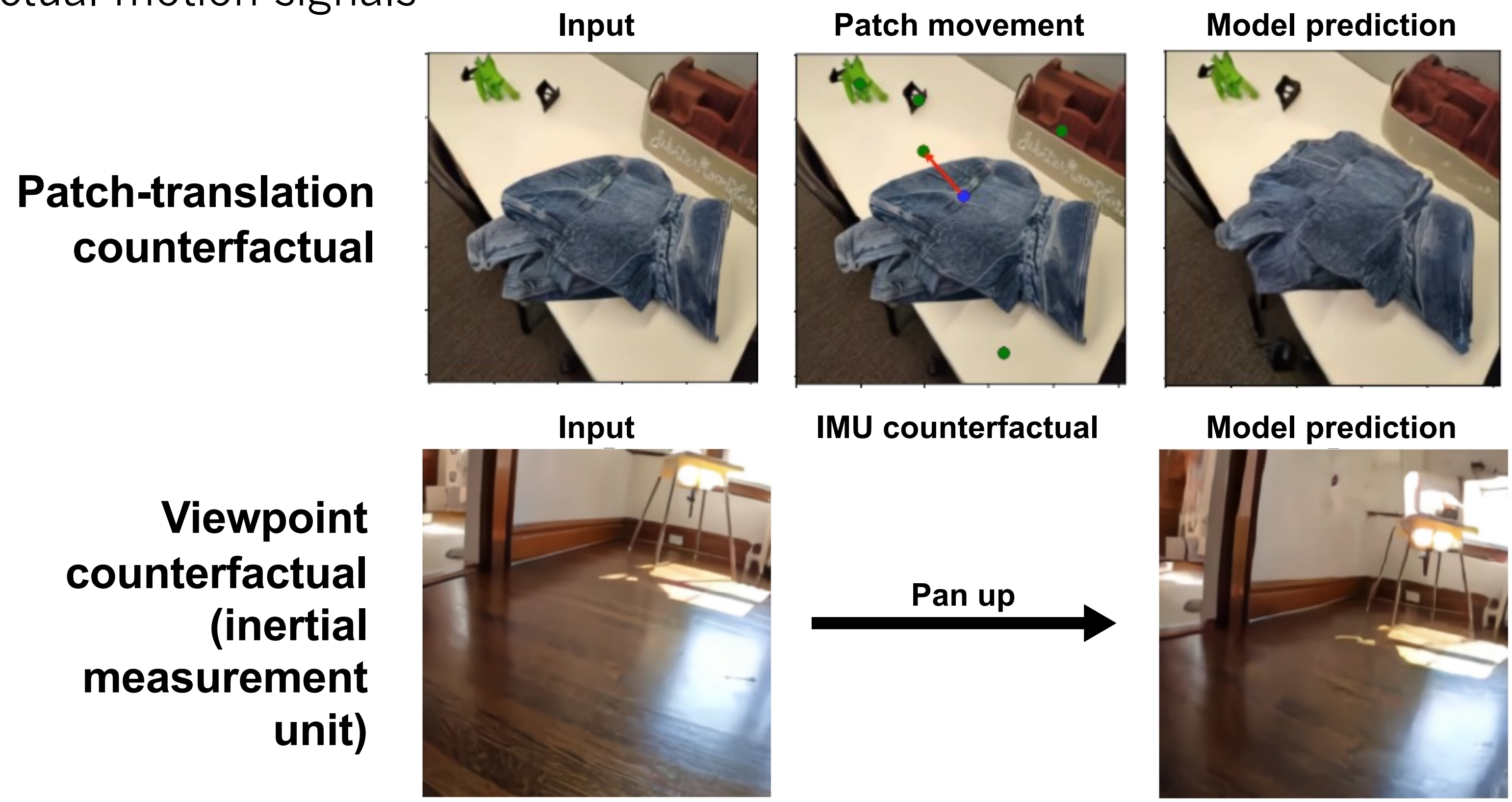
A data gap: Lack of efficient learning



Long, Xiang, Stojanov et al. (2024), *arXiv*

Training Causal Counterfactual World Model (CCWM)

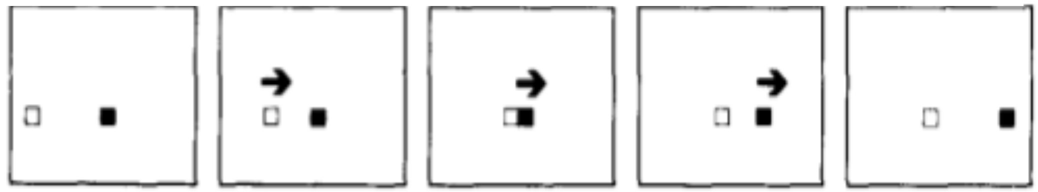
- CWM model uses masked prediction to ask about counterfactuals (Bear et al., 2023)
- CCWM advantage of IMU (accelerometer & gyro) to train models to predict based on actual motion signals

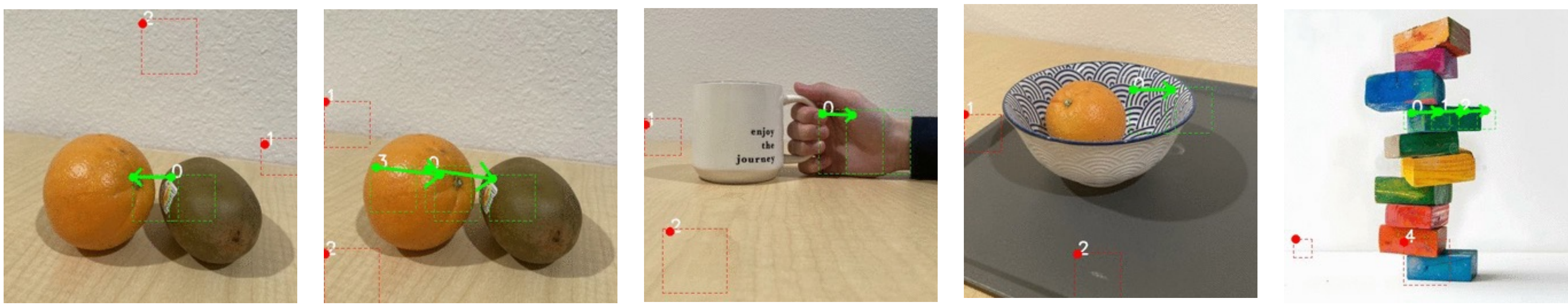


Khai Loonh Aw & Dan Yamins

CCWM as a cognitive model

Infant causal perception (Leslie & Keeble, 1987)





Khai Loonh Aw & Dan Yamins

Outline

1. Developmental training data (x axis)

2. **Developmental evaluation** (y axis)


AI Influencer 2000
@aibro

I let GPT-6 watch Disney's Moana and it cried at the end. I'm sure it has true human emotions!

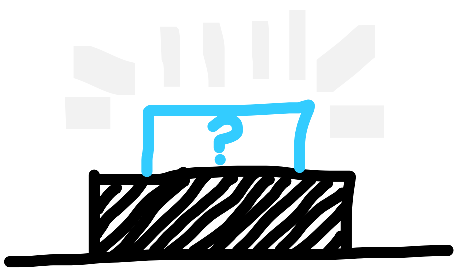
12:03 PM · Jun 1, 2023

1.1K Retweets **503** Quote Tweets **23.2K** Likes

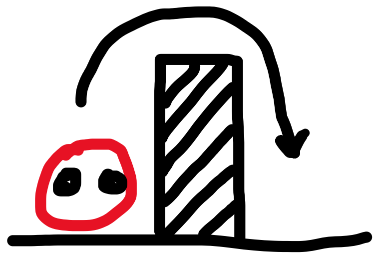
(not a real tweet)

"Baby steps" towards psychometric model evals

"it's a blicket!"



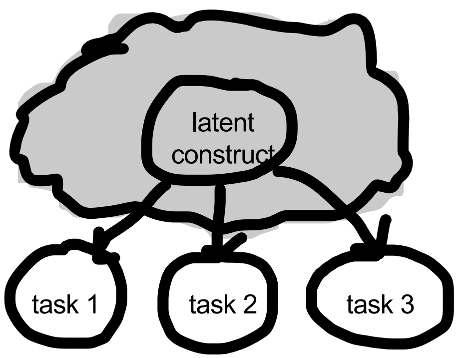
Novel materials
(cf. "data contamination"



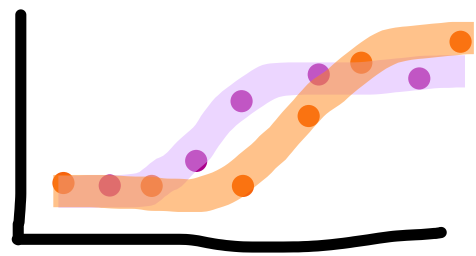
Schematic stimuli

	golabupadoti	golabu
1	tupirobidaku	
	golabutupiro	bupado
	rogolabupado	golabu
2	titupirobida	
	bupadotitupi	bupado

Closely matched controls



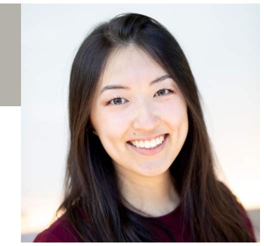
Multiple tasks and measures

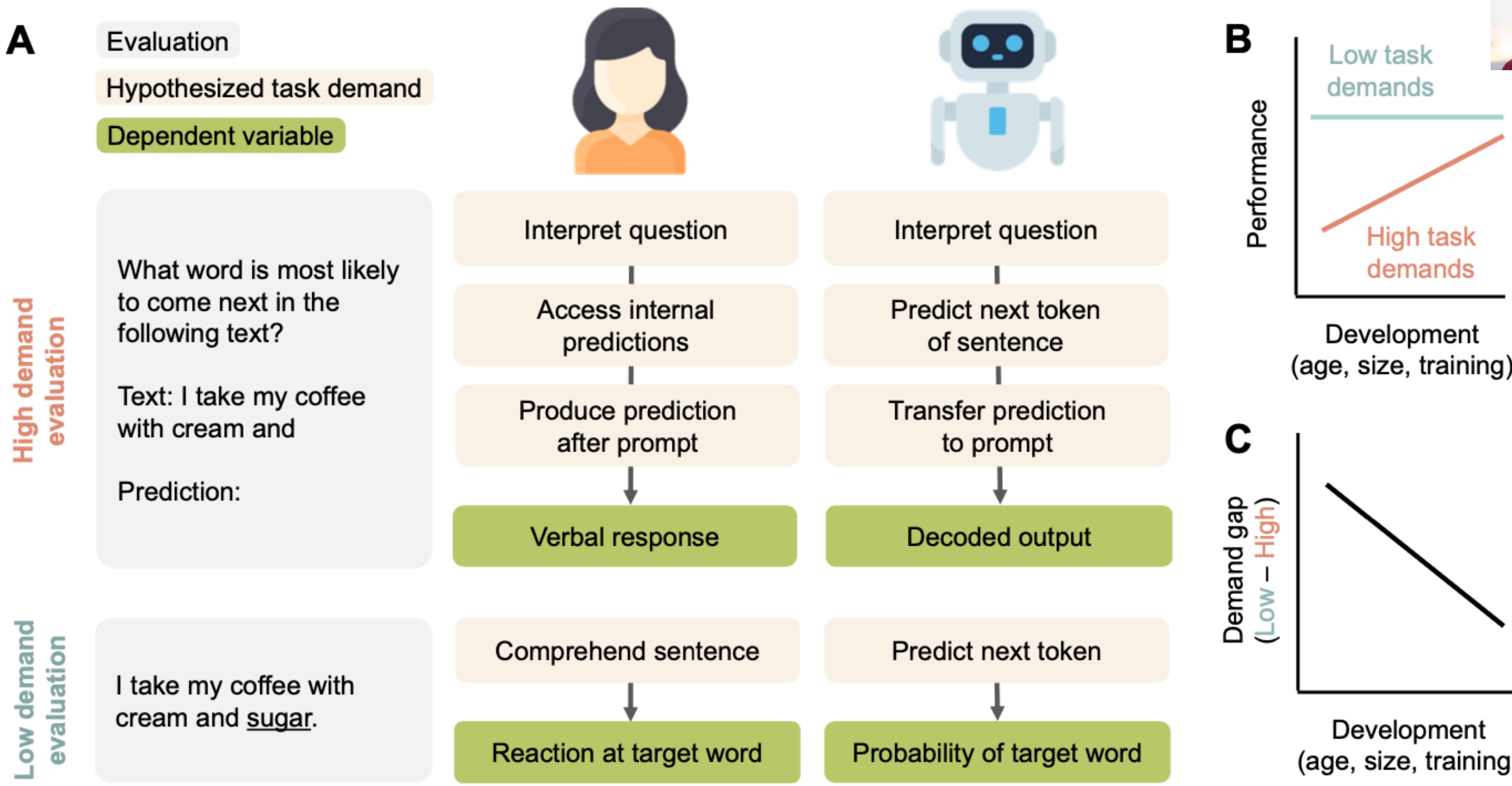


Measuring dose-response function
(in both learning input and treatment)

Frank (2023), *Nature Rev. Psych*

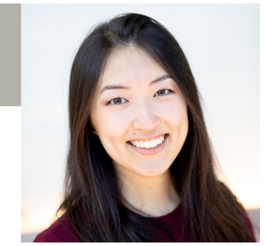
Evaluating models like kids – task demands?



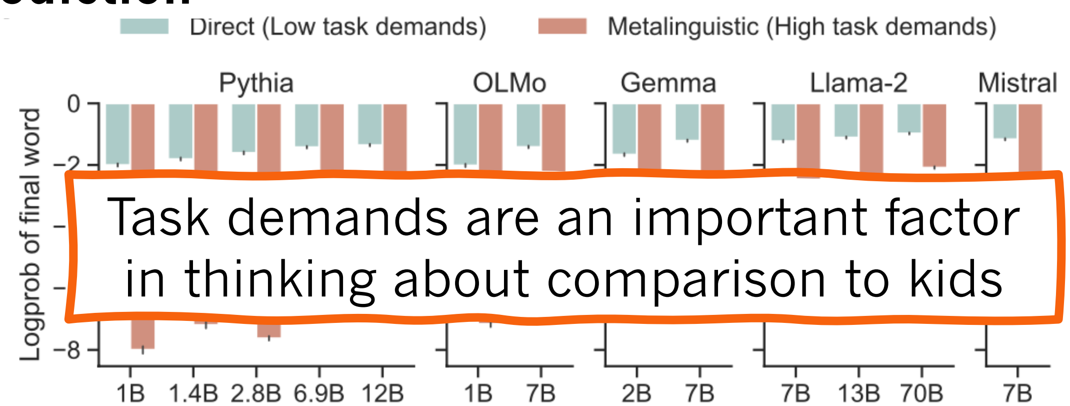


Hu & Frank (2024), *CoLM*

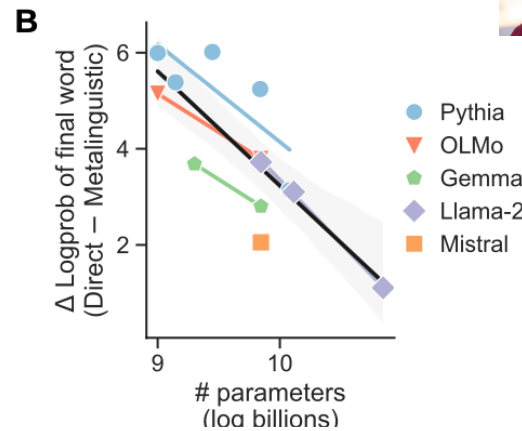
Task demands affect smaller models



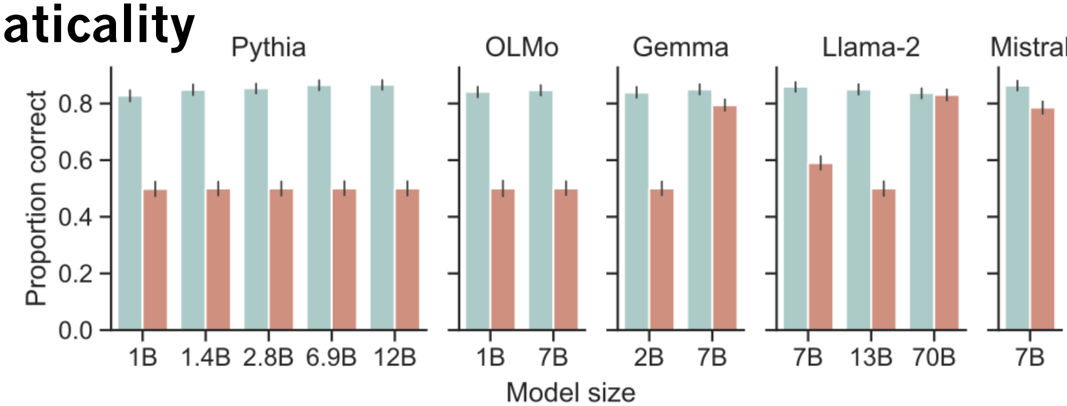
Word prediction

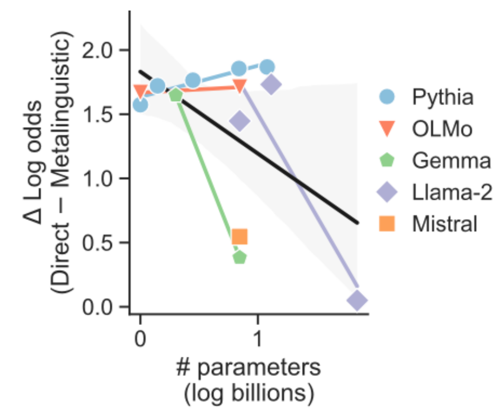


Task demands are an important factor in thinking about comparison to kids



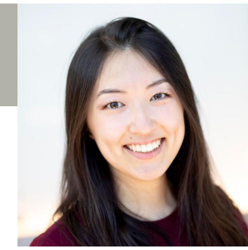
Grammaticality



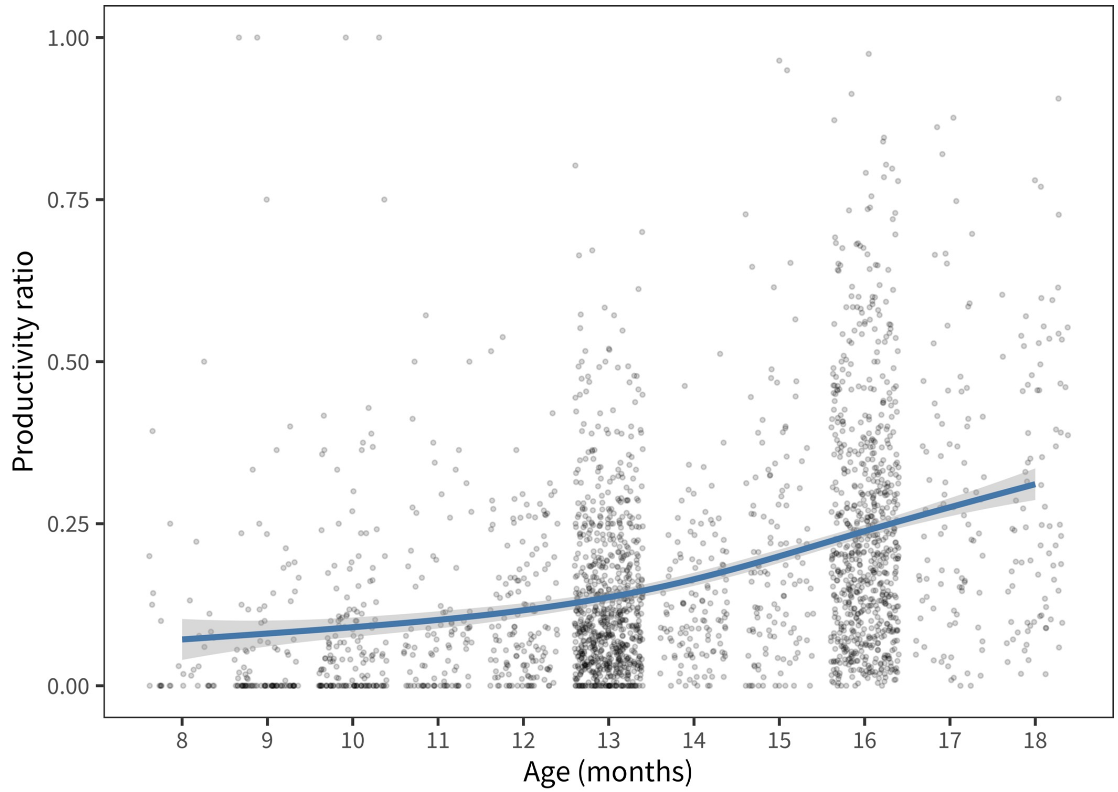


Hu & Frank (2024), *CoLM*

Machine comprehension / production

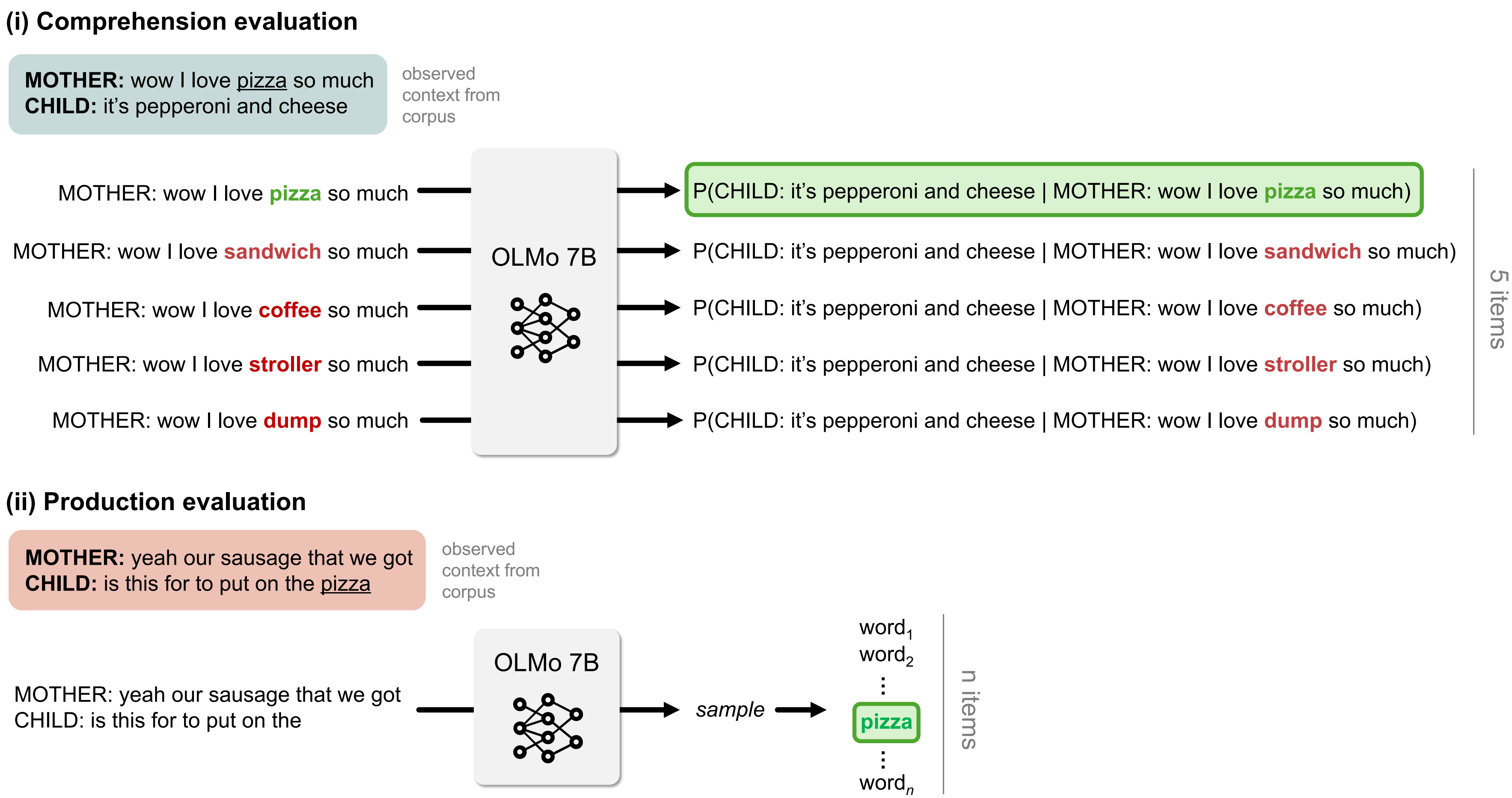


Kids understand more than they produce: do models also?



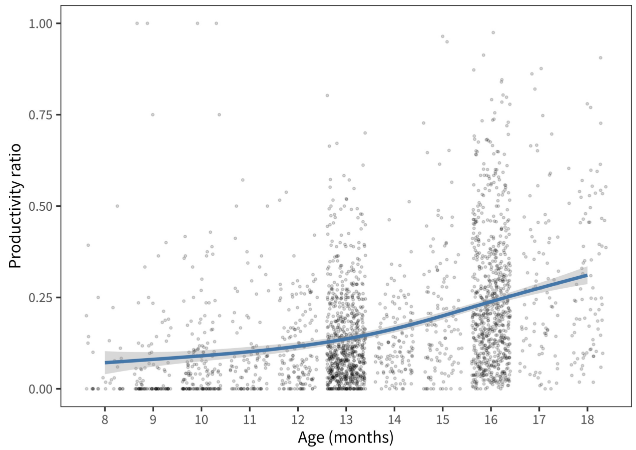
Hu et al. (under review)

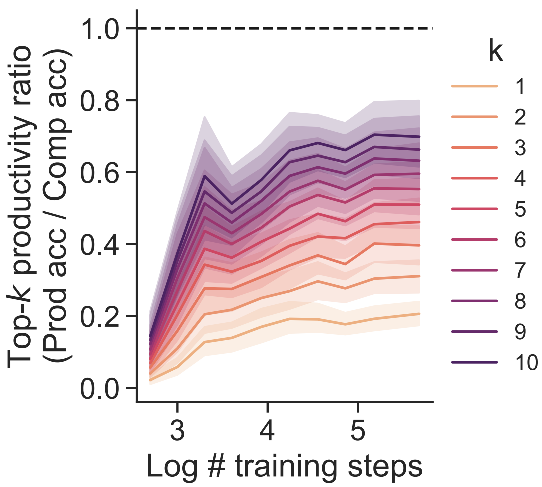
CDI production / comprehension

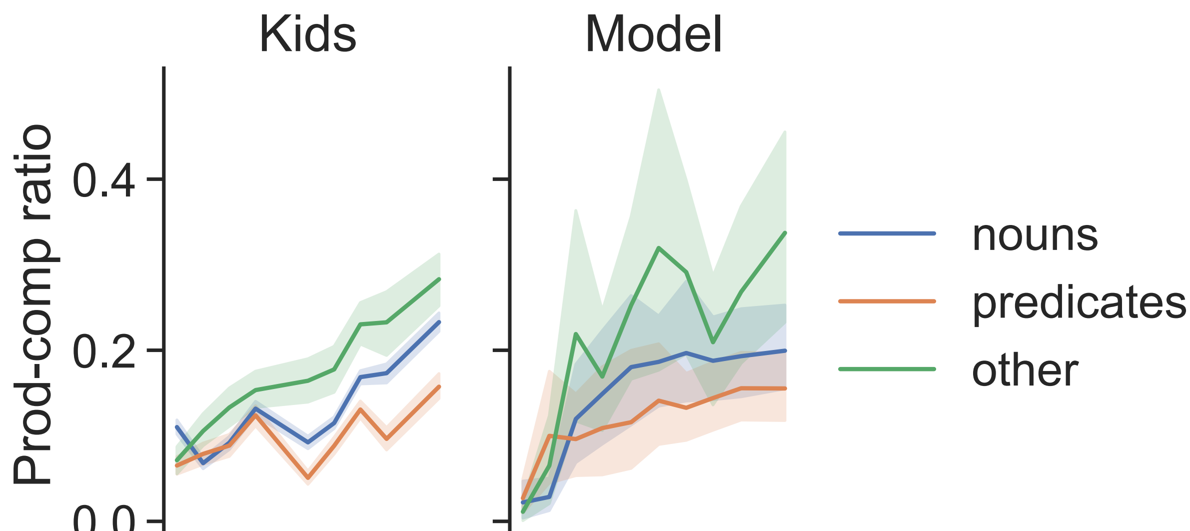


Hu et al. (under review)

CDI results







How can we hold task demands somewhat constant and examine the development of children's abilities?

Hu et al. (under review)

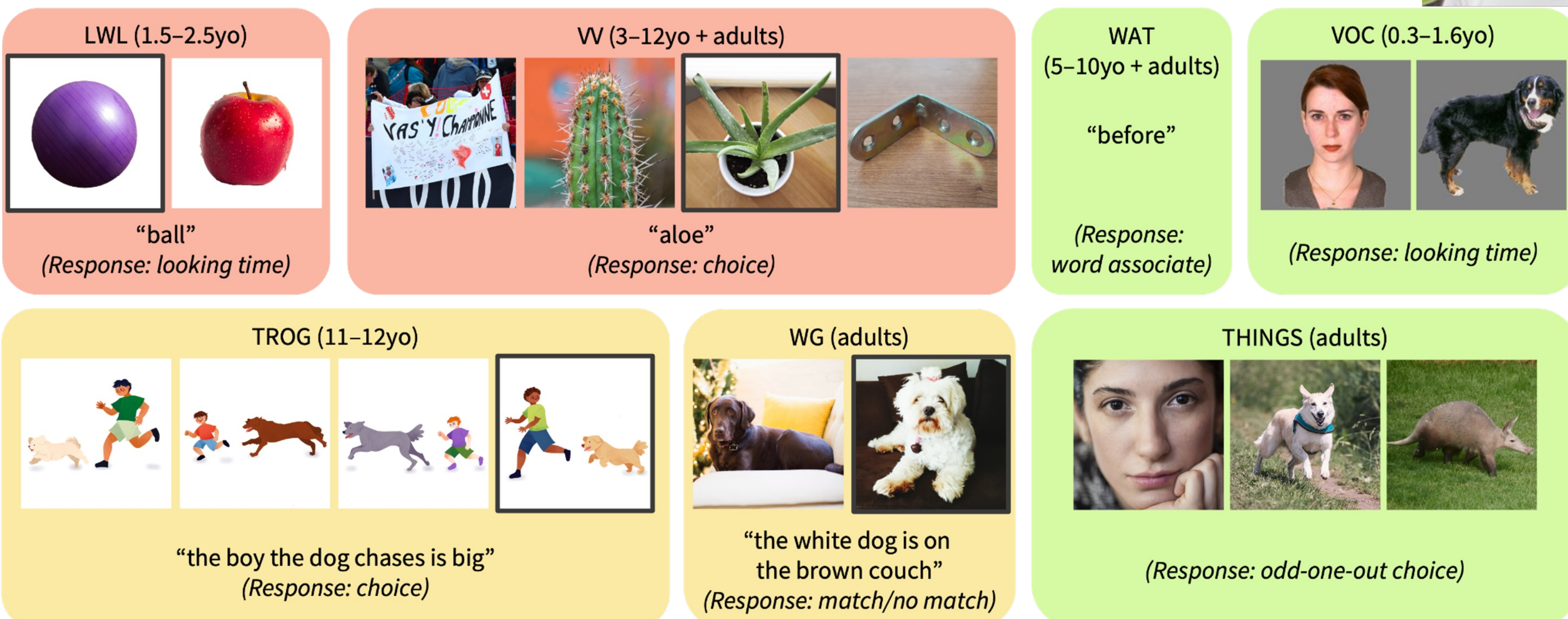
Do we really need another benchmark?

- How do we assess these models?
  - *Currently:* (Explicit or implicit) adult-level evaluations
- What types of language ability are assessed?
  - *Currently:* Ad-hoc lexical, grammatical, or reasoning abilities
- What is the nature of these evaluations?
  - *Currently:* Limited task similarity with humans; limited corresponding human data; typically unimodal

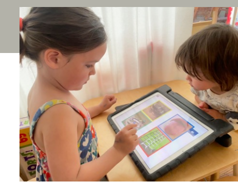
Constructing a developmental benchmark

+ Greater dynamic range (allows us to test smaller models or models trained on less data)
+ Allows for direct comparison with children (by using actual child tasks)
+ Permits analyses of development over training/maturation
+ Multiple levels of linguistic representation

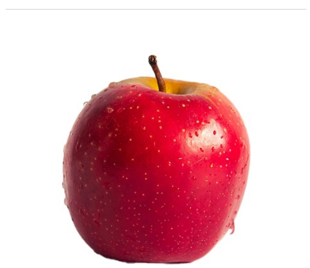
DevBench tasks



DevBench: Lexical tasks



Looking while listening

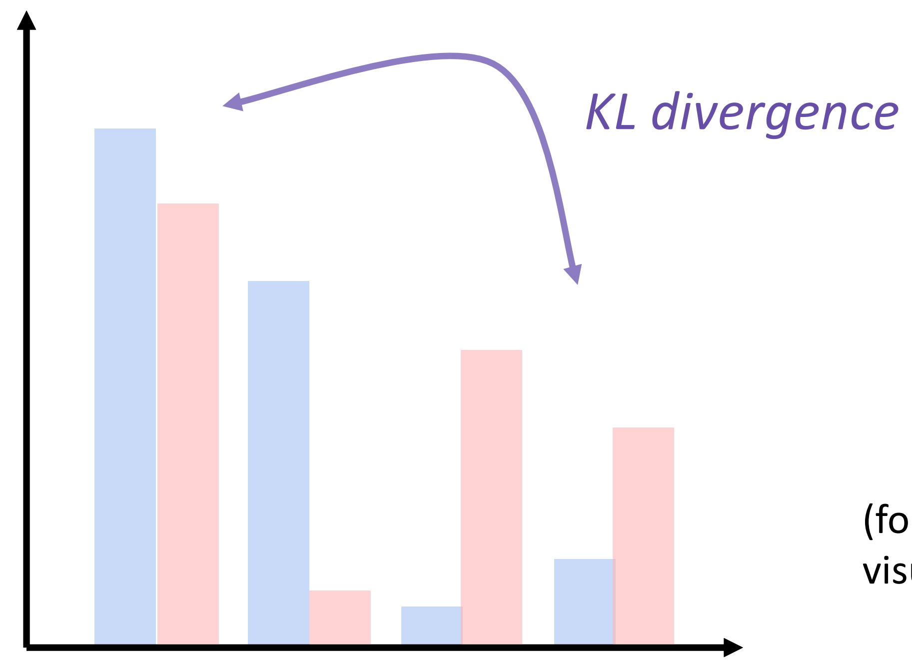


"look at the ball!"

Visual vocabulary

"aloe"

Measuring model–human similarity

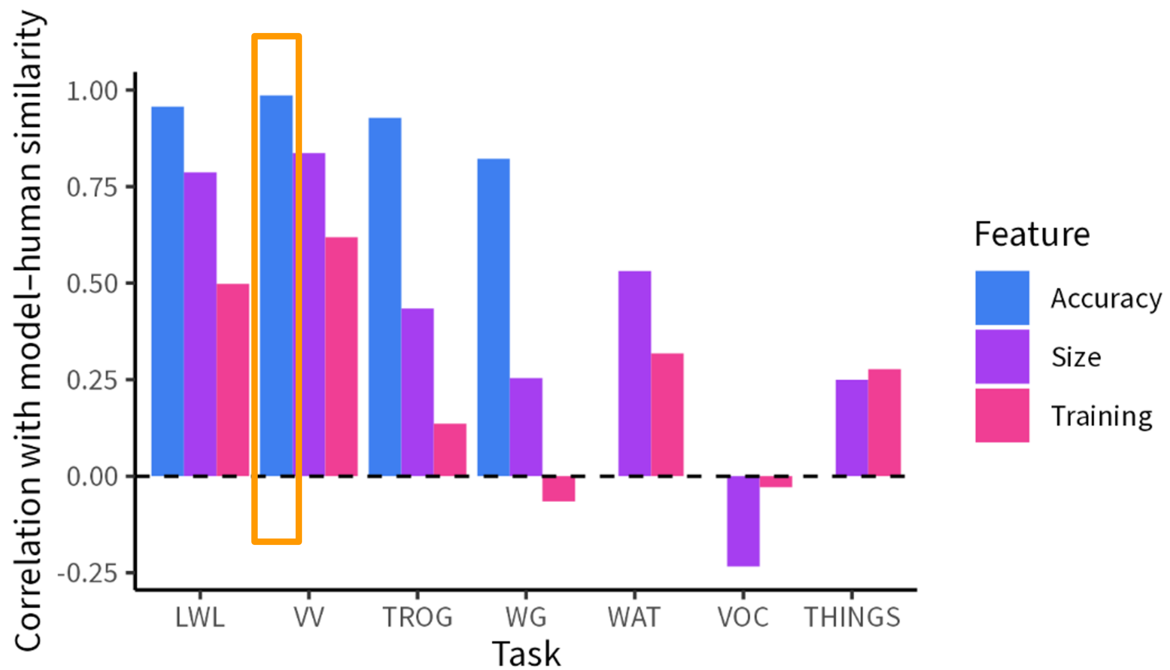


(for most tasks; RSA for visual semantic tasks)

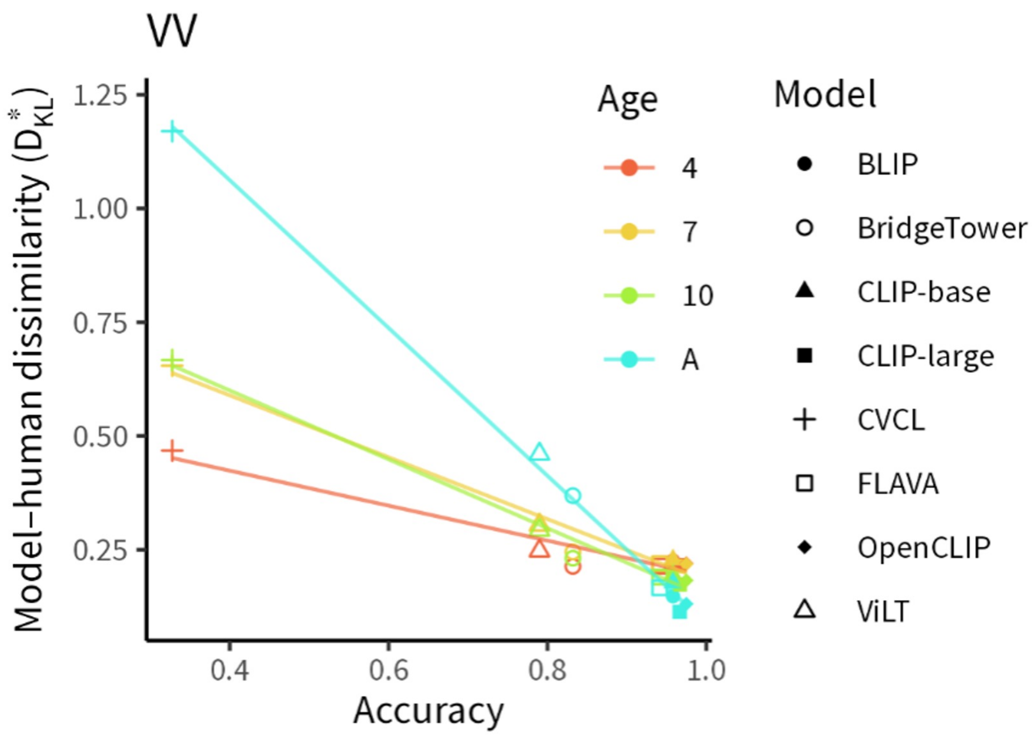
Benchmark results

			Lexicon		Syntax		Semantics		
Model	# params	# images	LWL (↓)	VV (↓)	TROG (↓)	WG (↓)	WAT (↓)	VOC (↑)	THINGS (↑)
CLIP-base [48]	149M	400M	0.014	0.205	0.732	0.256	0.495	-0.081	**0.397**
CLIP-large [48]	428M	400M	0.013	**0.179**	0.692	0.256	0.495	0.005	0.246
ViLT [49]	87M	4.1M	0.009	0.326	0.682	0.252	0.495	-0.053	0.127
FLAVA [50]	350M	70M	0.013	0.197	0.912	0.254	0.495	-0.042	0.189
BLIP [51]	252M	14M	0.010	0.193	**0.576**	0.259	0.495	-0.104	0.185
BridgeTower [52]	333M	4M	**0.008**	0.265	0.584	**0.250**	0.495	-0.095	0.345
OpenCLIP-H [53]	1.0B	32B	0.012	0.188	0.683	0.255	**0.495**	0.031	0.227
SigLIP [54]	800M	9B	0.067	0.612	0.888	0.258	0.495	-0.028	0.192
CVCL [4]	26M	600K	0.060	0.740	0.911	0.258	0.495	**0.138**	0.175
Human			0.010	0.091	0.028			0.251	
Random (OpenCLIP)	1.0B	0	0.087	0.740	0.908	0.258	0.495	0.246	0.054

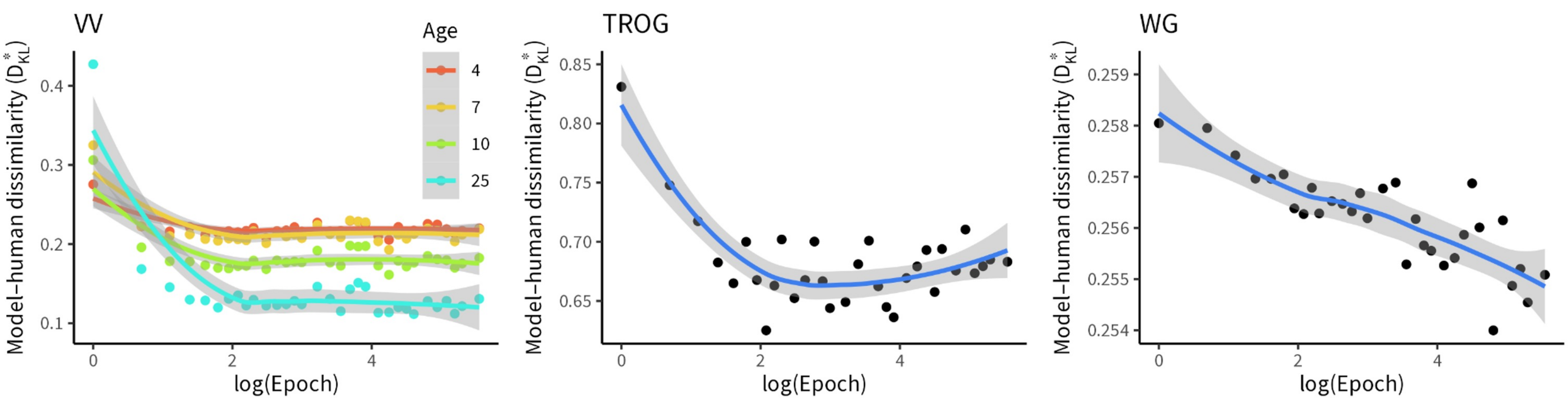
1. Better models are more human-like



1. Better models are more *adult*-like



2. OpenCLIP's trajectories are somewhat human-like

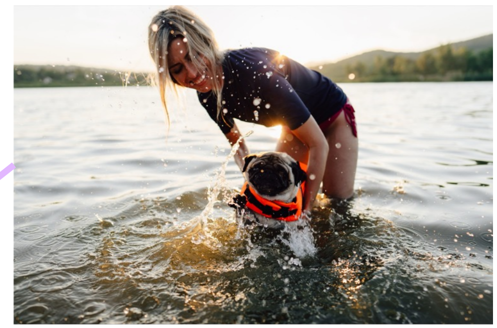


3. Models are most dissimilar to humans with ambiguity

polysemous targets

targets with other possible labels

hyponymy



VV

horn (distractors: bone, chin, ladybug)
hoe (distractors: peg, dustpan, beaker)
flan (distractors: fuse, amplifier, turnstile)
net (distractors: tee, domino, hydrant)
lollipop (distractors: candy, doorbell, crumb)

WG

a person whispering into a dog's ear / a dog whispering into a person's ear
there are more ladybugs than flowers / there are more flowers than ladybugs
the dog is swimming and the person is standing / the dog is standing and the person is swimming
blue pants and green top / green pants and blue top
a person sits and a dog stands / a person stands and a dog sits

Where next?

Model	# params	# images	Lexicon		Syntax		Semantics		
			LWL (↓)	VV (↓)	TROG (↓)	WG (↓)	WAT (↓)	VOC (↑)	THINGS (↑)
CLIP-base [48]	149M	400M	0.014	0.205	0.732	0.256	0.495	-0.081	**0.397**
CLIP-large [48]	428M	400M	0.013	**0.179**	0.692	0.256	0.495	0.005	0.246
ViLT [49]	87M	4.1M	0.009	0.326	0.682	0.252	0.495	-0.053	0.127
FLAVA [50]	350M	70M	0.013	0.197	0.912	0.254	0.495	-0.042	0.189
BLIP [51]	252M	14M	0.010	0.193	**0.576**	0.259	0.495	-0.104	0.185
BridgeTower [52]	333M	4M	**0.008**	0.265	0.584	**0.250**	0.495	-0.095	0.345
OpenCLIP-H [53]	1.0B	32B	0.012	0.188	0.683	0.255	**0.495**	0.031	0.227
SigLIP [54]	800M	9B	0.067	0.612	0.888	0.258	0.495	-0.028	0.192
CVCL [4]	26M	600K	0.060	0.740	0.911	0.258	0.495	**0.138**	0.175
Human			0.010	0.091	0.028			0.251	
Random (OpenCLIP)	1.0B	0	0.087	0.740	0.908	0.258	0.495	0.246	0.054

… DevBench 2.0?

LEVANTE

AN INITIATIVE OF THE
JACOBS FOUNDATION

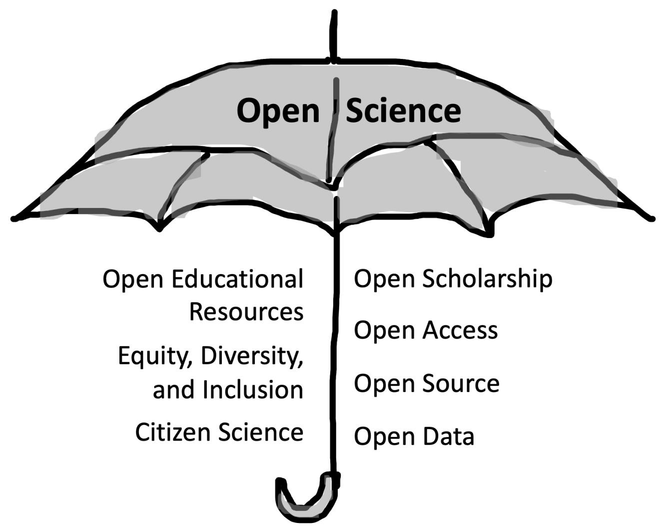
- **Framework**
  - Technical platform for data collection and sharing
  - Internationalized measures of learning and development
- **Network**
  - Teams use the framework to collect data in global contexts
  - Accelerated longitudinal data ages 2 – 12
  - Document children’s variability within and across individuals, groups, and cultures

(LEVANTE = Learning Variability Network Exchange)

Frank et al. (in revision)

Open science principles

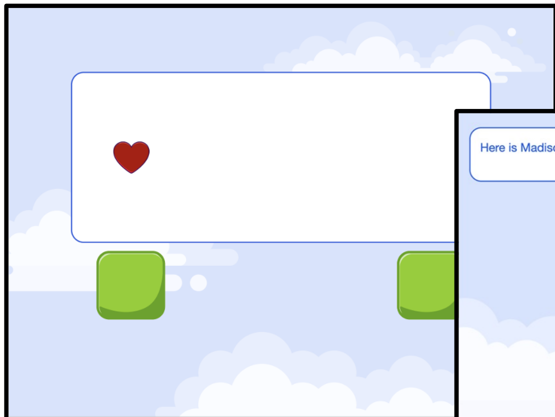
- All research products from LEVANTE will be made openly accessible under permissive licenses
- All tasks available free of charge
- All analysis and measure code will be developed as open source
- Core measure data released rapidly after collection



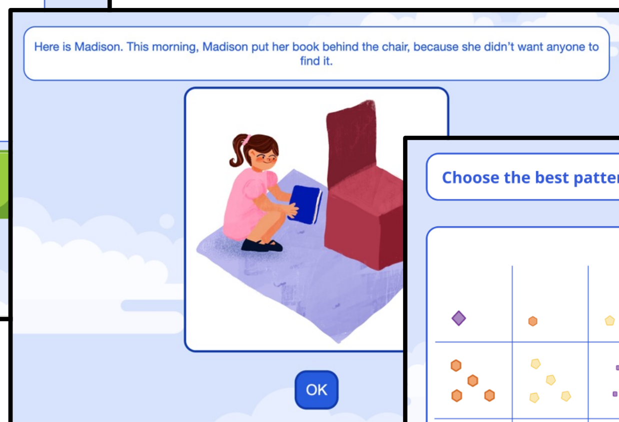
Frank et al. (2024)
experimentology.io

Task examples

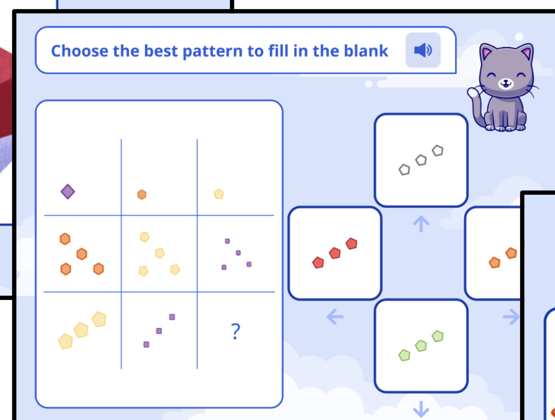
Multi-lingual AI-generated audio



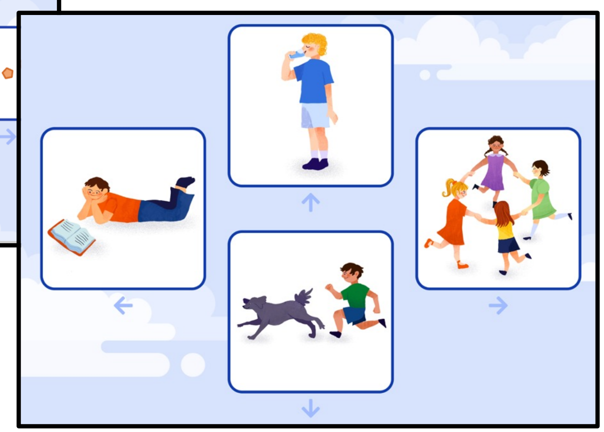
Inhibition (Hearts & Flowers)



Social Cognition



Matrix Reasoning

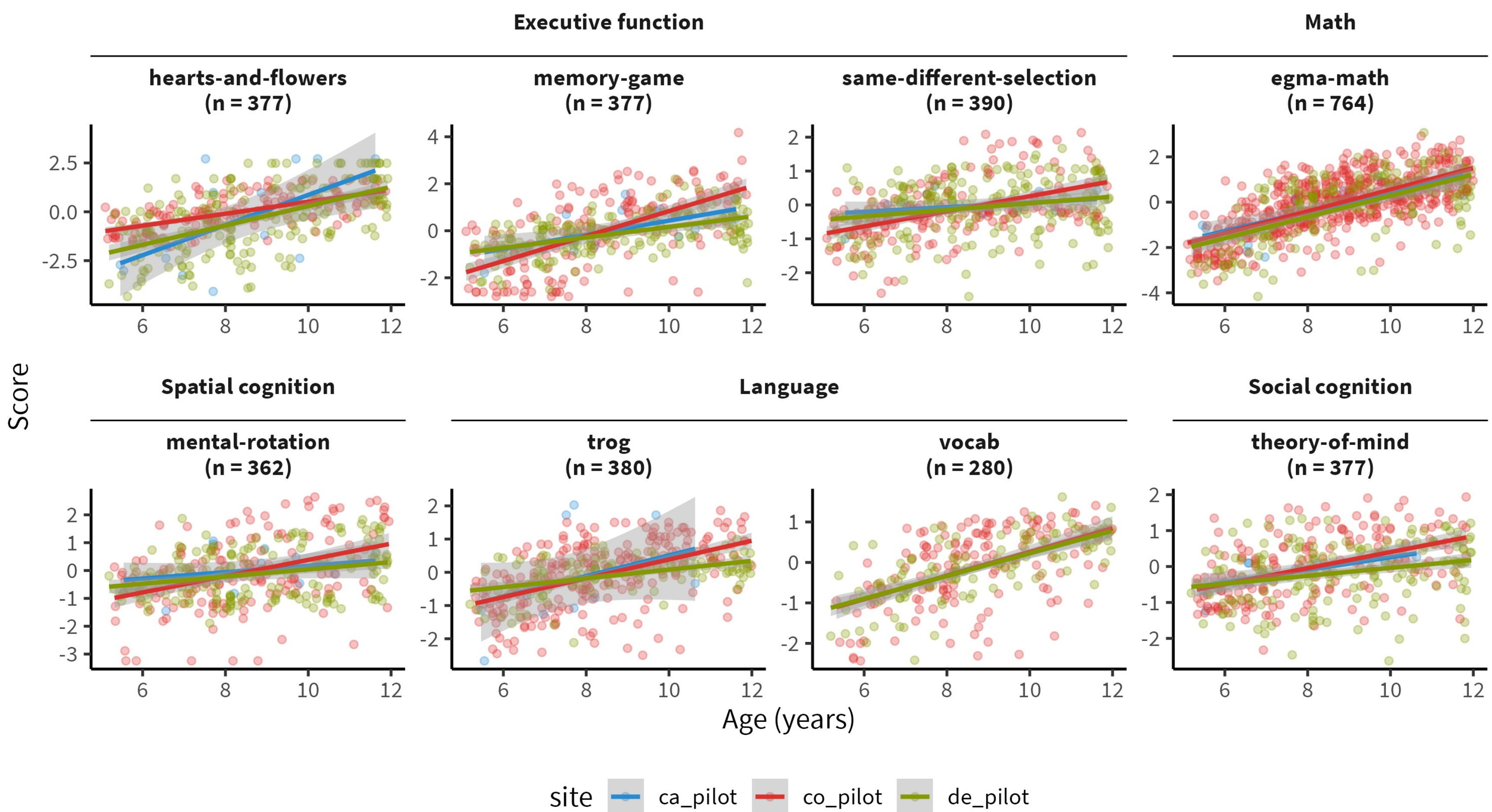


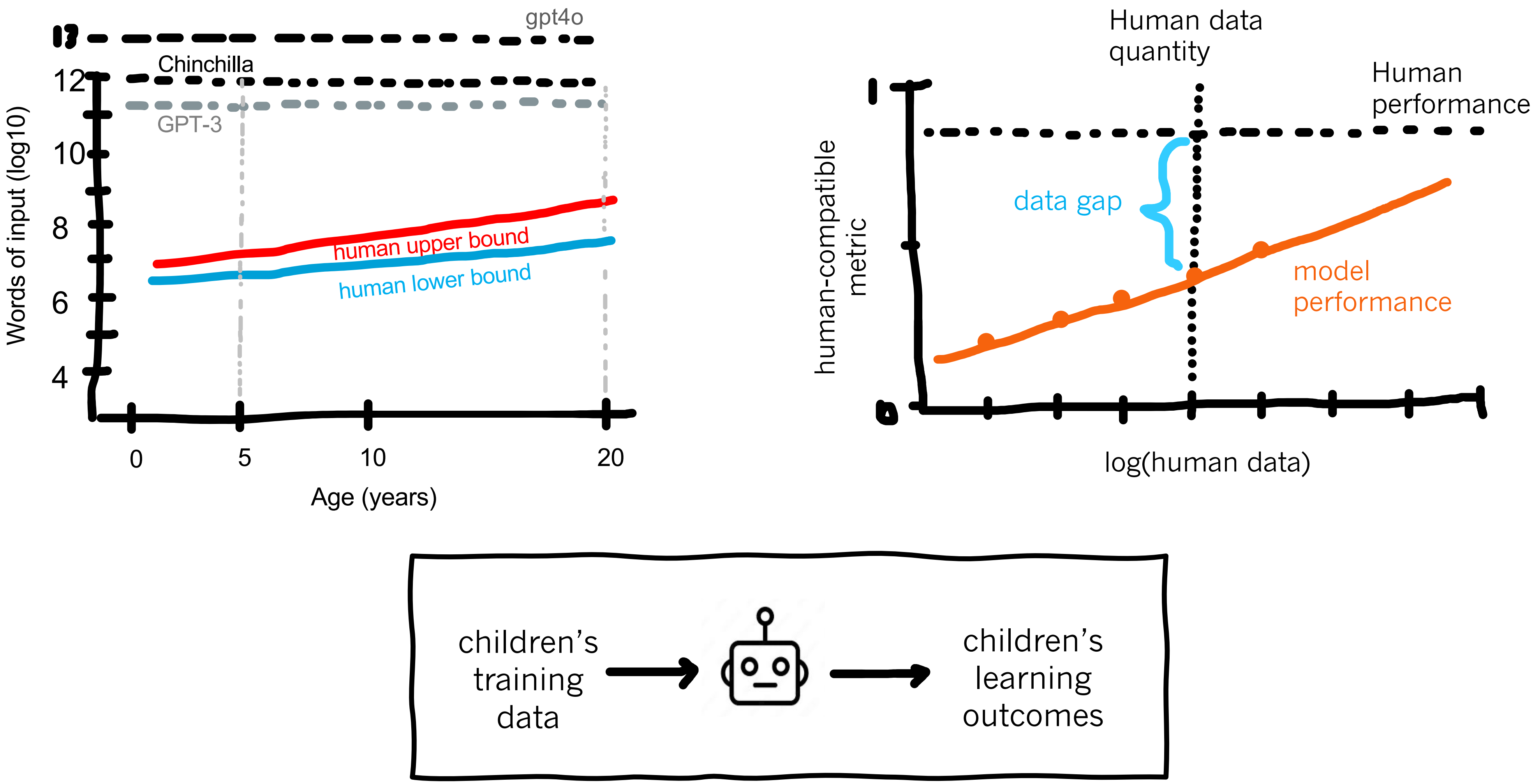
Grammar

Key features: **Internationalization** (English, Spanish, German, French), **cross-device capability**, **offline functionality**

LEVANTE pilot data

Data from Bogota, Colombia; Ontario, Canada; & Leipzig, Germany




gpt4o
Chinchilla
GPT-3
12
10
8
6
4
Words of input (log10)
human upper bound
human lower bound
0
5
10
20
Age (years)
Human data quantity
Human performance
1
0
human-compatible metric
data gap
model performance
log(human data)
children's training data
children's learning outcomes

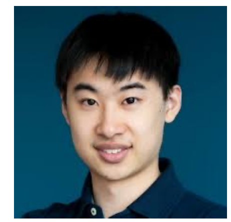
Bria Long

Bobby Sparks

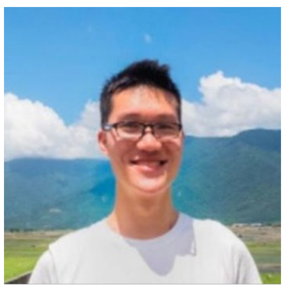
Language and Cognition
LAB
at Stanford University

Thank you!

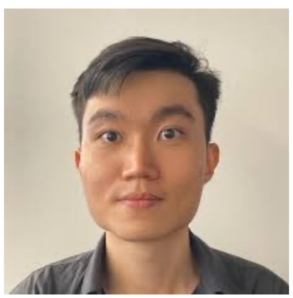
Jenn Hu



Steven Feng



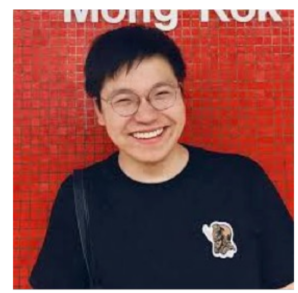
Alvin Tan



Khai Loong Aw

Stefan Stojanov

Violet Xiang



Zi Yin

Dan Yamins