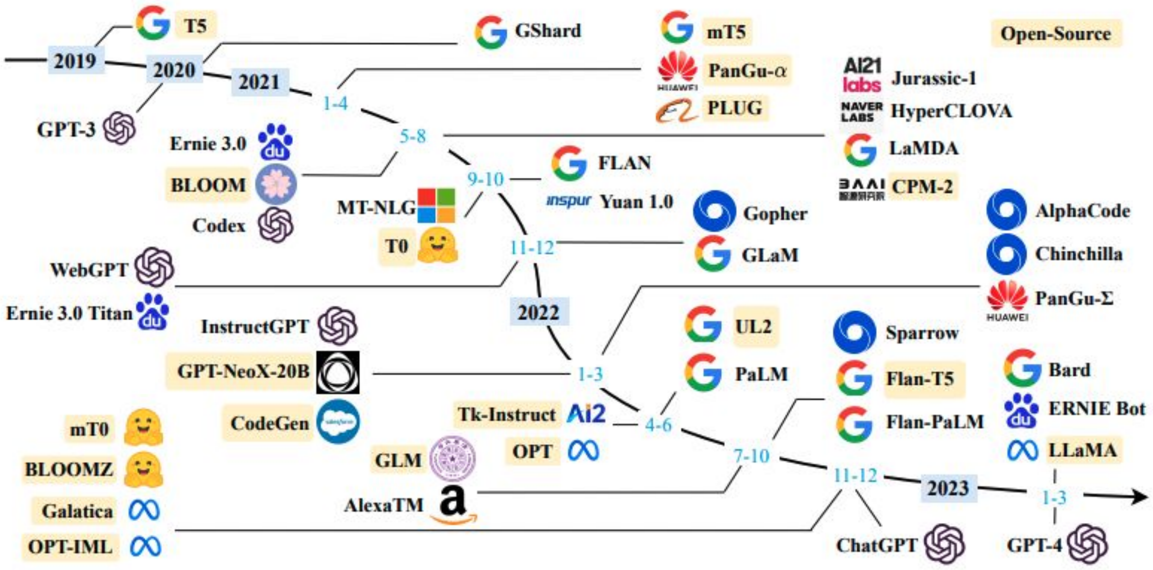
ICML International Conference On Machine Learning

Core Knowledge Deficits in Multi-Modal Language Models

Yijiang Li [1] Qingying Gao [*2] Tianwei Zhao [*2] Bingyang Wang [*3] Haoran Sun [2] Haiyun Lyu [4]
Robert D. Hawkins [5] Nuno Vasconcelos [1] Tal Golan [6] Dezhi Luo [7 8] Hokin Deng [9]

Large Foundation Models



Large Foundation Models

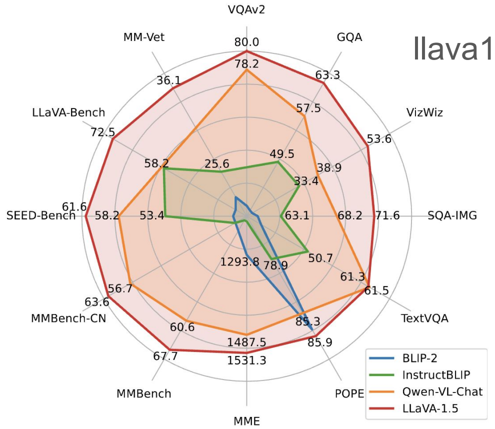
- Remarkable performance
- Reasoning & Perception (high-level)

	Gemini Ultra (pixel only)	Gemini Pro (pixel only)	Gemini Nano 2 (pixel only)	Gemini Nano 1 (pixel only)	GPT-4V	Prior SOTA
MMMU (val) Multi-discipline college-level problems (Yue et al., 2023)	**59.4%** pass@1 **62.4%** Maj1@32	47.9%	32.6%	26.3%	56.8%	56.8% GPT-4V, 0-shot
TextVQA (val) Text reading on natural images (Singh et al., 2019)	**82.3%**	74.6%	65.9%	62.5%	78.0%	79.5% Google PaLI-3, fine-tuned
DocVQA (test) Document understanding (Mathew et al., 2021)	**90.9%**	88.1%	74.3%	72.2%	88.4% (pixel only)	88.4% GPT-4V, 0-shot
ChartQA (test) Chart understanding (Masry et al., 2022)	**80.8%**	74.1%	51.9%	53.6%	78.5% (4-shot CoT)	79.3% Google DePlot, 1-shot PoT (Liu et al., 2023)
InfographicVQA (test) Infographic understanding (Mathew et al., 2022)	**80.3%**	75.2%	54.5%	51.1%	75.1% (pixel only)	75.1% GPT-4V, 0-shot
MathVista (testmini) Mathematical reasoning (Lu et al., 2023)	**53.0%**	45.2%	30.6%	27.3%	49.9%	49.9% GPT-4V, 0-shot
AI2D (test) Science diagrams (Kembhavi et al., 2016)	**79.5%**	73.9%	51.0%	37.9%	78.2%	81.4% Google PaLI-X, fine-tuned
VQAv2 (test-dev) Natural image understanding (Goyal et al., 2017)	**77.8%**	71.2%	67.5%	62.7%	77.2%	86.1% Google PaLI-X, fine-tuned

Qwen2.5VL

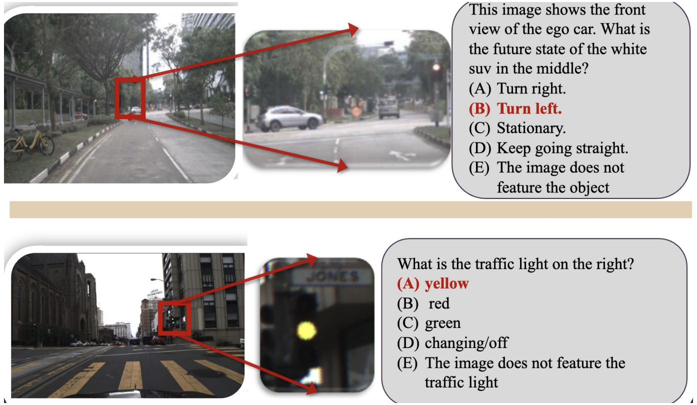
Gemini

llava1.6

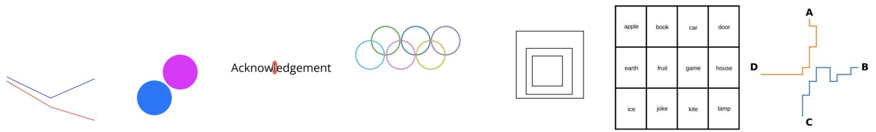


Pitfalls of Large Foundation Models

- Moravec's paradox (Tasks that are easy to humans could be difficult to machines and vice versa)
- OOD / in-the-wild Generalization
- Hallucinations



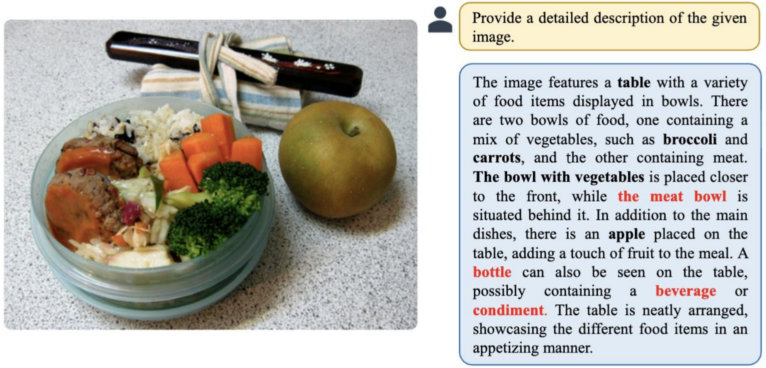
MLLM in-the-wild (Zhang et al., 2025)



	P1		P2		P3		P4		P5		P6		P7	
	1	✗	Yes	✗	o	✗	6	✓	5	✗	3×4	✓	1	✓
	1	✗	No	✓	w	✗	5	✗	3	✓	3×4	✓	2	✗
	1	✗	Yes	✗	o	✗	5	✗	4	✗	4×4	✗	2	✗
	0	✓	No	✓	l	✓	6	✓	3	✓	3×4	✓	1	✓

GPT-4o Gemini-1.5 Sonnet-3 Sonnet-3.5

BlindTest (Rahmanzadehgervi et al., 2024)



Visual hallucination (Li et al., 2023)

Why?

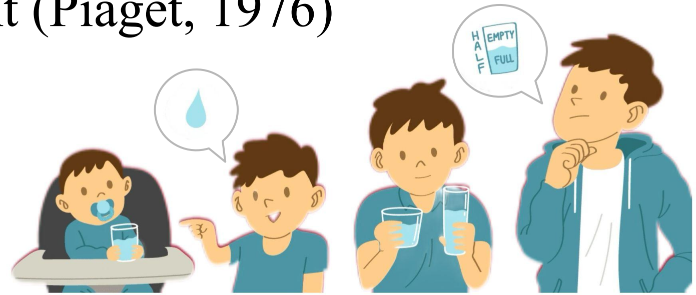
- Formal vs. Functional Linguistic Competence (Mahowald et al., 2024):
    - LLMs excel in generating fluent language (formal)
    - But may lack real-world understanding (functional)
- 🦜?

Core-knowledge hypothesis

What about humans? – the human path

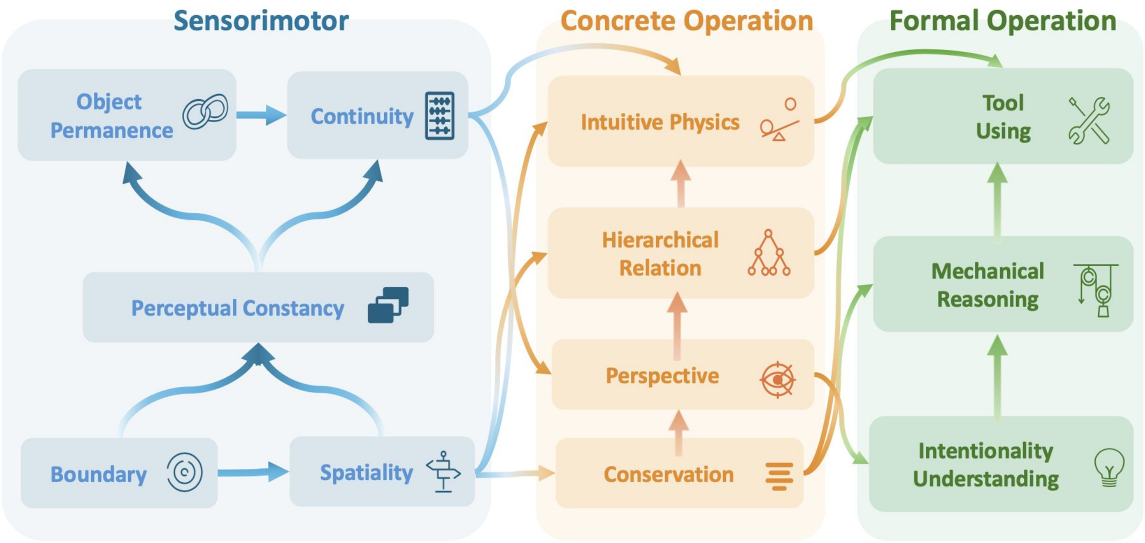
Innateness

- Plato’s Meno: everything we know is innate
- Leibniz: something in the mind must be innate, if it is only the mechanisms that do the learning (Pinker, 2002)
- Stage Theories of Cognitive Development (Piaget, 1976)



Growing up

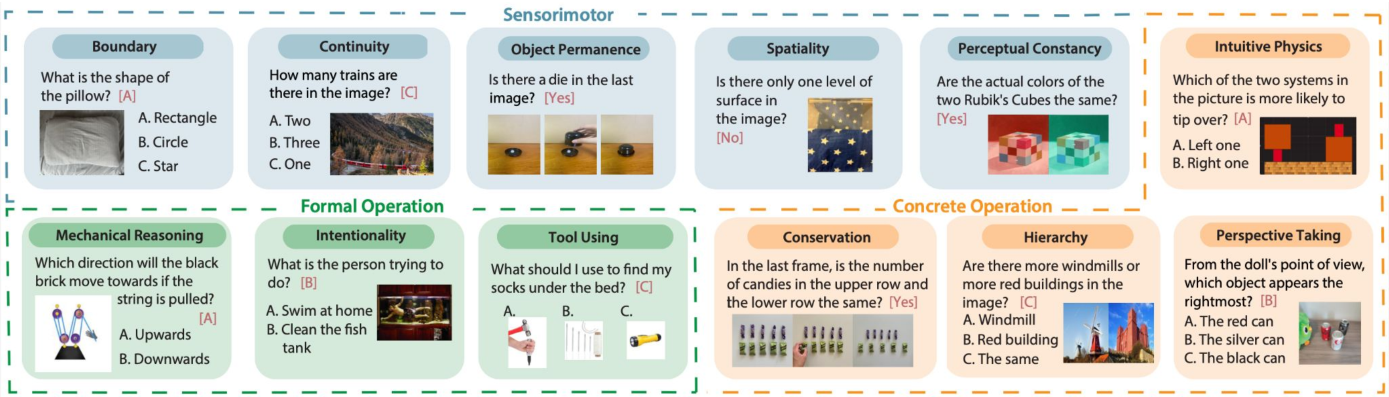
- Children develops along distinct stages of conceptualizing the world, each stage is marked by previously inaccessible abilities
- Early, simpler abilities serve as the basis for later, complex abilities (“**grounding**”)



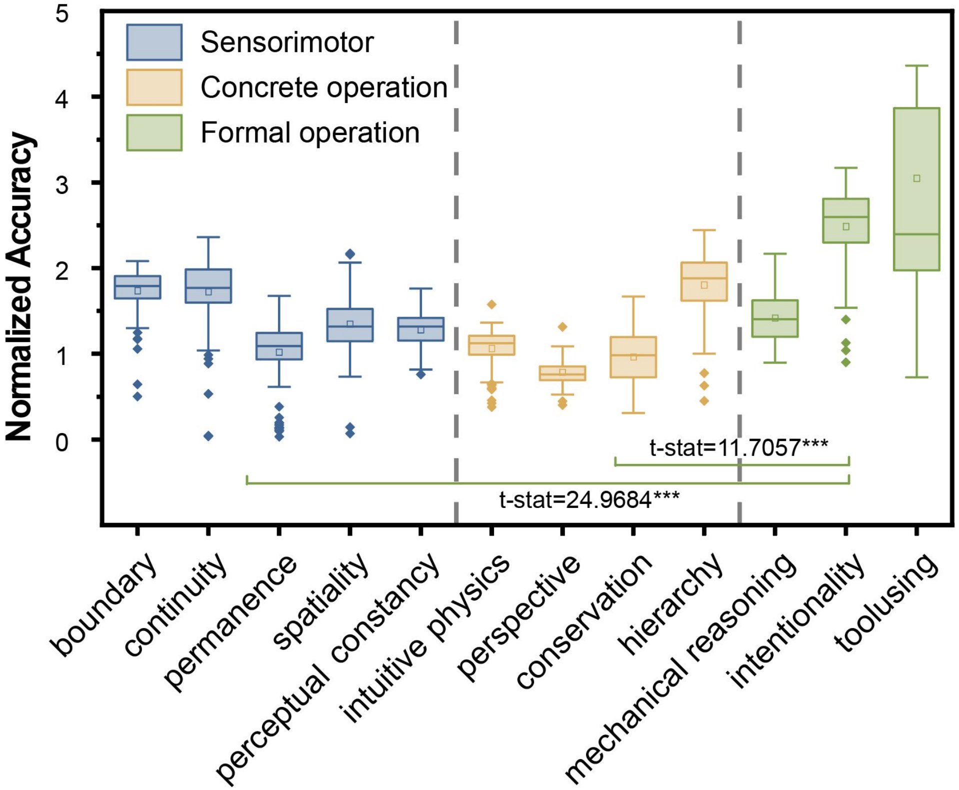
Core-knowledge in MLLMs

- Classifying taxonomy (grounded in cog-sci literature)
- 1500+ samples plus 200 + MLLMs

Concept	Definition	Concept	Definition	Concept	Definition
Boundary	The transition from one object to another.	**Continuity**	Objects persist as unified, cohesive entities across space and time.	**Permanence**	Objects do not cease to exist when they are no longer perceived.
Spatiality	The *a priori* understanding of the Euclidean properties of the world.	**Perceptual Constancy**	Changes in appearances don't mean changes in physical properties.	**Intuitive Physics**	Intuitions about the laws of how things interact in the physical world.
Perspective	To see what others see.	**Hierarchy**	Understanding of inclusion and exclusion of objects and categories.	**Conservation**	Invariances of properties despite transformations.
Tool Use	The capacity to manipulate specific objects to achieve goals.	**Intentionality**	To see what others want.	**Mechanical Reasoning**	Inferring actions from system states and vice versa.



Core knowledge deficits

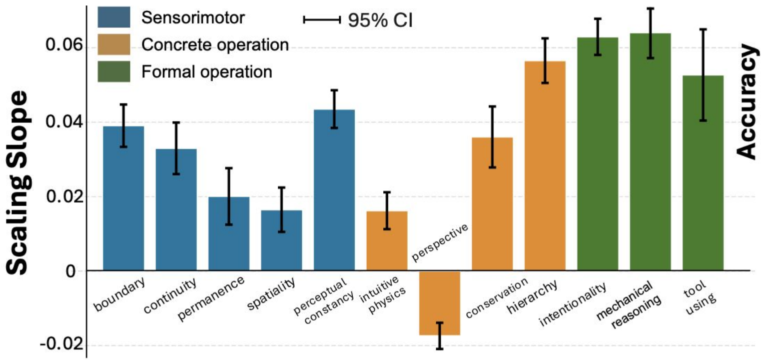


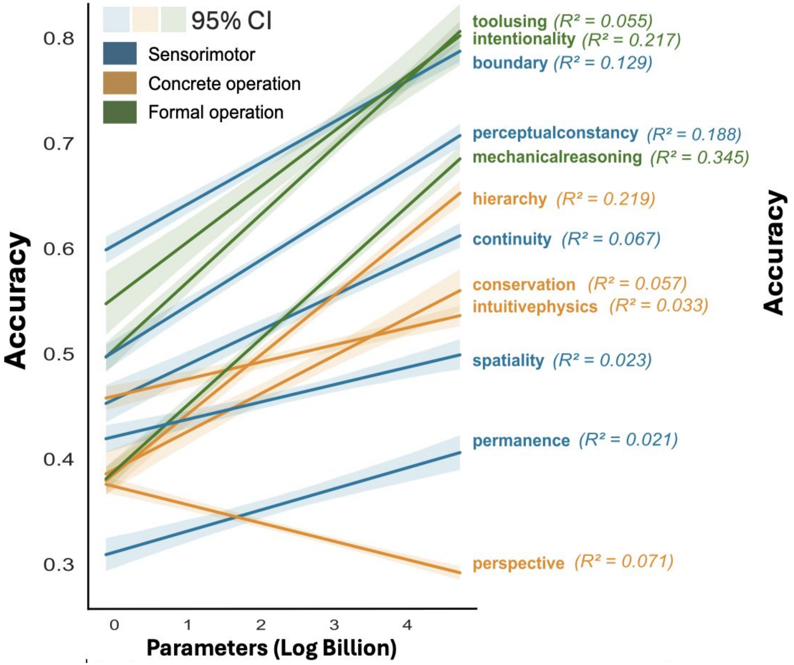
Key Finding 1 (Core Knowledge Deficits): MLLMs excel at higher-level abilities associated with later developmental stages but consistently struggle with lower-level abilities that typically emerge earlier in human cognition.

Scalability

Scalability: performance on an ability improvement as model grows (slope of linear fitting)

High-level abilities in general shows much higher scalability.





Key Finding 4 (Not Scaling): MLLMs exhibit limited or no scalability on low-level abilities compared to high-level abilities

Core-knowledge is predictive of higher-level abilities

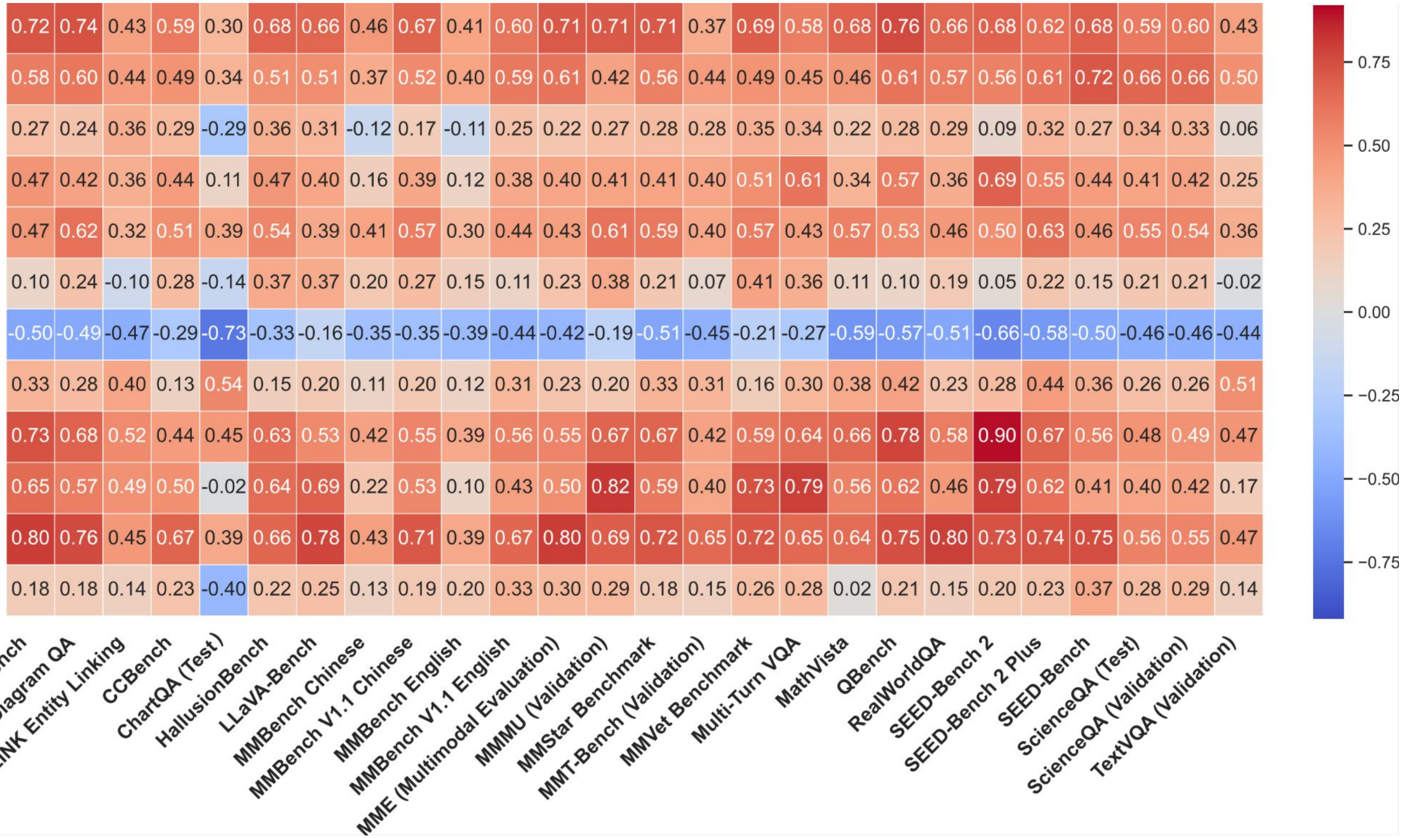
Key Finding 3 (Predictability)

	Function Reasoning	Future Prediction	Identity Reasoning	Image Emotion	Image Quality	Image Scene	Image Style	Image Topic	Nature Relation	Object Localization	Physical Property Reasoning	Physical Relation	Scene Understanding	Social Relation	Spatial Relation	Spatial Relationship	Text Understanding	Visual Reasoning	Coarse Perception	Fine-grained Perception	Instance Reasoning	Logical Reasoning	Math	Science & Technology
Boundary	0.44	0.56	0.24	0.19	0.67	0.29	0.43	0.10	0.58	0.74	0.51	0.59	0.66	0.29	0.77	0.73	0.41	0.51	0.59	0.66	0.71	0.68	0.71	0.49
Continuity	0.44	0.48	0.29	0.33	0.62	0.32	0.32	0.18	0.65	0.71	0.51	0.63	0.68	0.27	0.68	0.54	0.43	0.54	0.60	0.61	0.63	0.49	0.44	0.34
Permanence	0.21	0.21	0.12	0.03	0.18	0.10	0.14	0.03	0.34	0.24	0.32	0.28	0.21	0.09	0.20	0.19	0.16	0.23	0.23	0.29	0.26	0.28	0.25	0.21
Spatiality	0.25	0.27	0.15	0.14	0.44	0.19	0.19	0.04	0.48	0.49	0.32	0.33	0.41	0.11	0.46	0.43	0.13	0.29	0.36	0.44	0.40	0.43	0.36	0.23
Perceptual Constancy	0.31	0.39	0.15	0.14	0.46	0.21	0.38	0.07	0.42	0.52	0.43	0.43	0.48	0.14	0.56	0.58	0.29	0.29	0.46	0.55	0.61	0.61	0.52	0.46
Intuitive Physics	0.11	0.09	-0.11	-0.11	0.14	-0.11	0.01	-0.19	0.23	0.15	0.12	0.10	0.13	-0.08	0.20	0.27	0.04	0.20	0.14	0.17	0.22	0.25	0.23	0.10
Perspective	-0.19	-0.33	-0.27	-0.32	-0.52	-0.29	-0.23	-0.23	-0.40	-0.63	-0.37	-0.55	-0.49	-0.20	-0.53	-0.41	-0.48	-0.25	-0.35	-0.55	-0.48	-0.47	-0.46	-0.49
Conservation	0.18	0.28	0.26	0.26	0.32	0.33	0.28	0.19	0.27	0.39	0.18	0.27	0.39	0.23	0.32	0.34	0.33	0.21	0.31	0.35	0.33	0.29	0.34	0.19
Hierarchy	0.41	0.49	0.27	0.26	0.59	0.31	0.40	0.15	0.61	0.67	0.45	0.45	0.60	0.24	0.66	0.68	0.21	0.41	0.52	0.66	0.62	0.68	0.69	0.48
Mechanical Reasoning	0.32	0.36	0.16	0.10	0.48	0.17	0.34	-0.02	0.51	0.51	0.42	0.29	0.45	0.08	0.54	0.72	0.11	0.32	0.41	0.51	0.53	0.57	0.73	0.39
Intentionality	0.46	0.61	0.26	0.38	0.72	0.28	0.43	0.12	0.75	0.78	0.53	0.66	0.75	0.31	0.80	0.73	0.36	0.73	0.59	0.73	0.72	0.68	0.71	0.50
Tool Using	0.30	0.25	0.20	0.17	0.36	0.24	0.17	0.12	0.42	0.37	0.17	0.18	0.33	0.19	0.38	0.32	0.01	0.30	0.30	0.18	0.27	0.16	0.11	0.00

ICML International Conference On Machine Learning

Core-knowledge is predictive of higher-level abilities

Key Finding 3 (Predictability).

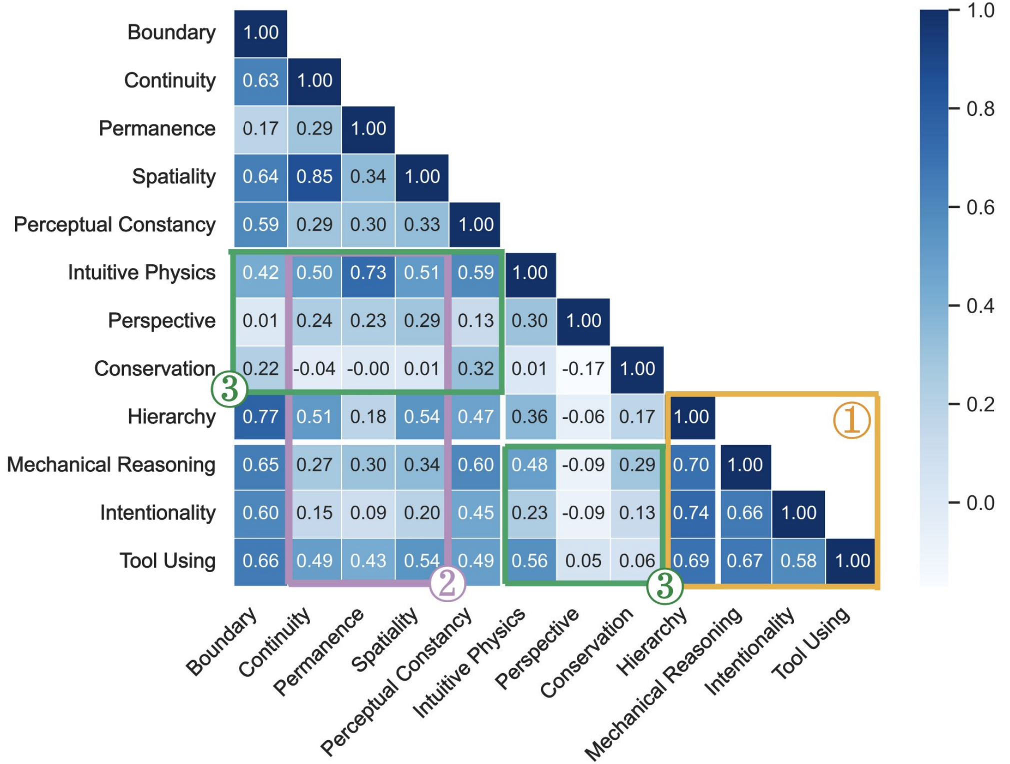


Dependency of core-abilities

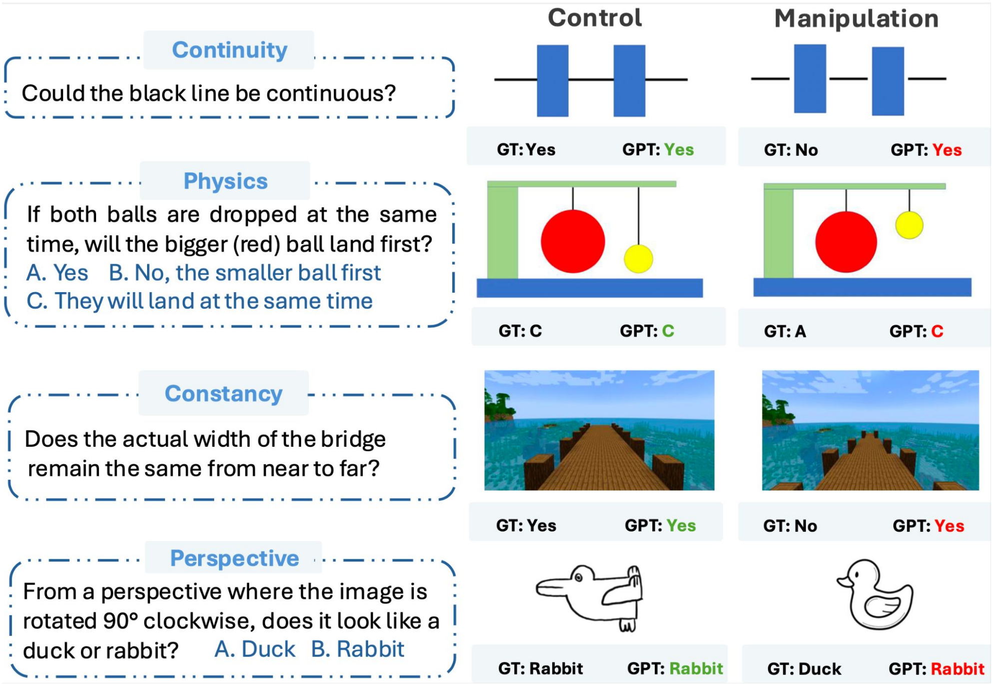
Key Finding 2 (Misaligned Dependency): Core abilities exhibit weak cross-stage correlations, indicating an absence of developmental scaffolding.

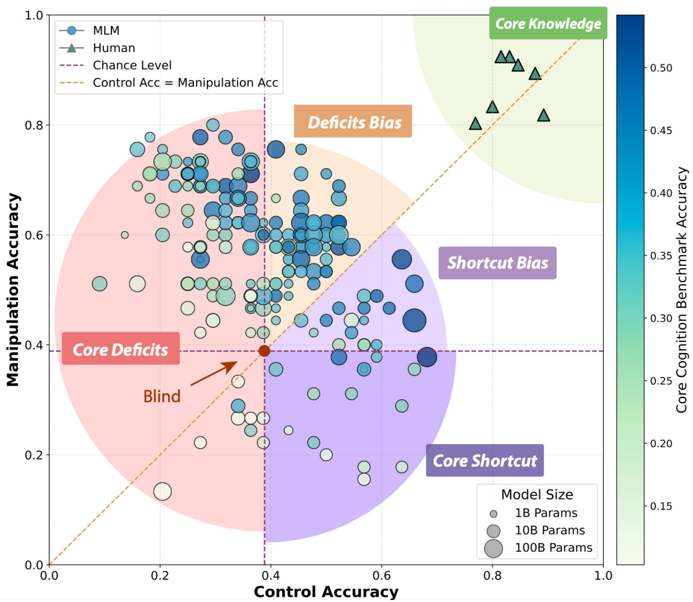
Human: high correlation within and across stages

1. Alignment with human ($\rho > 0.65$) within high level abilities
2. Three Sensorimotor abilities *(Permanence, Spatiality, and Continuity)* exhibit weak correlations with most higher-stage abilities
3. Three concrete operational abilities (Perspective, Conservation, and Intuitive Physics) also show weak cross-stage correlations



Do MLLMs genuinely has core-knowledge? ⇒ A Controlled experiment





Key Finding 5 (Core Deficits v.s. Shortcut Taking): Models increasing in size exhibit deficits and shortcut-taking behaviors rather than progressing toward conceptual understanding of core knowledge.

But why is it important?

- **Paradigm agnostic**
  - **Core knowledge** may function as “developmental start-up software” (Lake, 2017)
  - **Shared Prerequisite** (e.g. computational/representational power) across intelligence
- Human Path
  - Inspiration from human
  - Alignment with human intelligence

Main take-away (What does it imply?)

- Misalignment from human (not a good sign)
  - Lack of core-knowledge
  - performance on high-level abilities does not correlate with the corresponding low-level abilities that ground them in humans.
- ⇒ shortcut ? parrot?
- (current) Scaling fails (at least not human-aligned)
- Do we need human aligned?

Future

- Scaling ⇒ core-abilities
  - Objective?
  - Data?
  - Architecture?
- shared **prerequisite** + “developmental start-up software” (Lake, 2017)
  - Learn core-knowledge first then pre-training
  - MoE to counterfact catastrophic forgetting
- More analysis
  - Causal instead of correlation for dependency
  - Training as causal Intervention
  - System-2 results (compared to system-1 counterpart)
  - …etc

Discussion

- Do AI need to be human-aligned?
  - Inspiration? Standard for AGI?
- ⇒ Argument: core-knowledge as shared **prerequisite** for all intelligence!
- Shortcut? Stochastic parrot?
  - In low-level core-abilities (at least)
- Distributed representation
  - Pre-training learn core-knowledge
  - But hard to retrieve due to distributiveness
  - System-2 thinking? RL?

Thanks to


ICLR
irteenth International Conference
On Learning Representations
Singapore Expo | April 24 - 28, 2025
Haoran Sun
Johns Hopkins
University

ICML
International Conference
On Machine Learning

GrowAI Community

Growing AI like a Child, at Scale

Humans never "learn" intelligence. Humans *develop* intelligence. Biological lives on this planet take heavy advantage of intelligent primitives embedded in their genes. Cats never "learn" to backflip. Birds never "learn" to fly. In the same way, humans never "learn" to *cognize*. Humans are born with a set of *core cognition*, that sets the foundation for our perception and action in the physical world.

Our *core cognition* unravels through a specific developmental trajectory as we grow into adulthood. Here, we seek to do the same for our machines, leveraging heavy cognitive literature in developmental psychology, e.g., Piagetian theory of cognitive development, to design our *growing up* curriculum. In addition, we also want to learn from the current success of machine intelligence, specifically the *scaling* law.

Instead of putting *growing up* and *scaling up* into opposite camps, we argue the next step towards human-like artificial general intelligence is to *grow AI like a child, at scale*. We come together as GrowAI, an open-source community uniting researchers from computer science, cognitive science, psychology, linguistics, philosophy, and beyond. Our ongoing research focuses on the following areas:

- Cognitive Competence: Investigating and evaluating the cognitive behaviors and limitations of pre-trained models beyond leaderboard chasing.
- Core Knowledge: Identifying and building fundamental knowledge and core cognitive scenarios for benchmarking and evaluating human-like intelligence.
- Developmental AI: Understanding the developmental trajectories and training dynamics of scalable systems toward human capabilities.

Updates

🎉 **Four of our papers accepted at ICLR 2025 Workshops!** 🎉

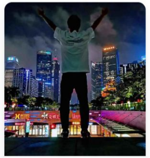
Vision Language Models See What You Want but not What You See BiAlign @ ICLR 2025

Vision Language Models Know Law of Conservation without Understanding More-or-Less Bidirectional Human-AI Alignment @ ICLR 2025

Probing Mechanical Reasoning in Large Vision Language Models Bidirectional Human-AI Alignment @ ICLR 2025

Active Members


Hokin Deng
Carnegie Mellon University


Yijiang Li
University of California, San Diego


Dezhi Luo
University of Michigan


Qingying Gao
Johns Hopkins University


Ziqiao Ma
University of Michigan


Emmy Liu
Carnegie Mellon University


Icy Wang
Emory University


Tianwei Zhao
Johns Hopkins University


Yixuan Wang
University of Florida


Pooyan Rahmanzadehgervi
Auburn University


Avi Bhattacharya
University of Michigan


Zory Zhang
Brown University

THANK YOU !