Core Knowledge Deficits in Multi-Modal Language Models

ICML 2025

Yijiang Li, Qingying Gao*, Tianwei Zhao*, Bingyang Wang*, Haoran Sun, Haiyun Lyu,
Robert D. Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, Hokin Deng

*Equal Contribution

Core Knowledge Hypothesis

Pitfalls in MLLMs

- **Moravec's paradox**: Tasks that are easy to humans can be difficult to machines and vice versa
- Two types of linguistic competency (Mahowald et al., 2024)
  - LLMs excel in generating fluent language (formal)
  - But may **lack real-world understanding (functional)**
- Difficulties in **out-of-distribution (OOD)** tasks / **in-the-wild** generation (Zhang et al., 2025)

Whereas In Human—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)
- **Piaget**: Stage Theories of Cognitive Development (Piaget, 1976)

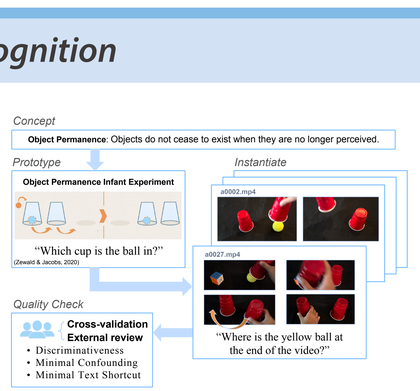
Core Knowledge Hypothesis

Above deficiencies stem from the absence of core knowledge—rudimentary cognitive abilities innate to humans from early childhood

CoreCognition

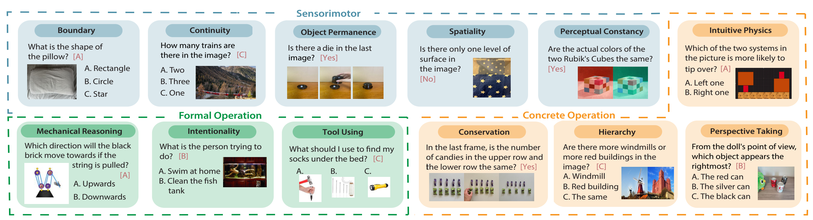
Dataset Curation

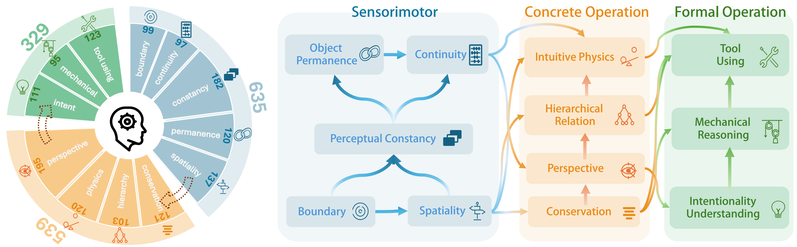
1. Choose **5–10 prototypes** that abstractly exemplify situations suited to evaluate **each core ability**
2. **Instantiate** prototypes using diverse visual media to form VQAs
3. Manual data **quality check** by cross-validating between annotators and crowd-sourcing



Concept	Definition	Concept	Definition	Concept	Definition
Boundary	The transition from existence to non-existence of objects.	Continuity	Physical properties of objects tend to exist in the same way.	Permanence	Things continue to exist when they are not in sight.
Spatiality	The a priori understanding of the Euclidean properties of our 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 ofinclusion 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.

Table 1. Abbreviated definitions of the 12 cognitive abilities assessed.

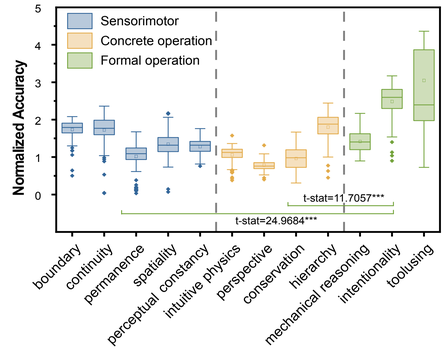




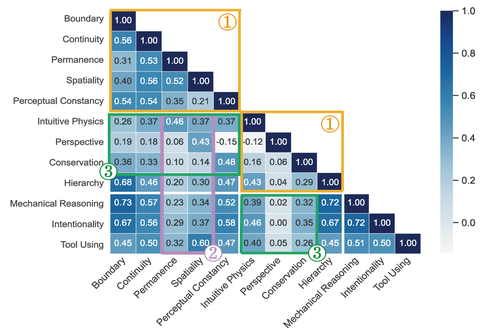
Core Knowledge Deficits

*We evaluated **230 models across 11 prompt types** (2,530 data points) for analysis*

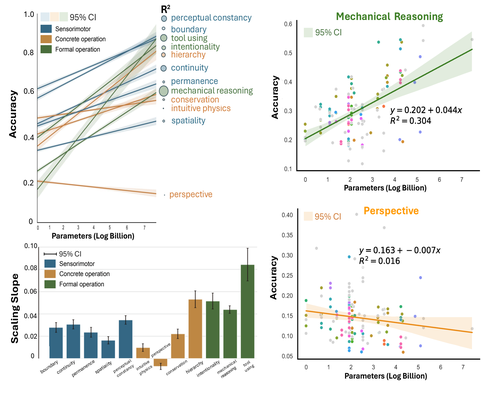
Finding 1: Core knowledge deficits



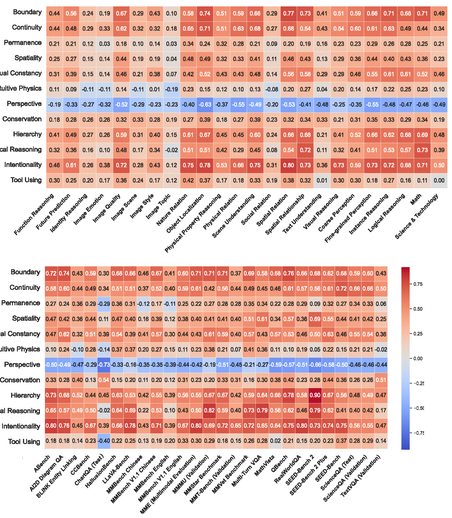
Finding 2: Misaligned dependency of core knowledge



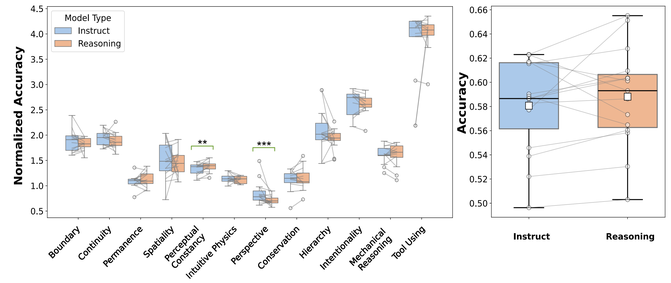
Finding 3: Core knowledge scaling deficits



Finding 4: Core knowledge is predictive of higher-level abilities

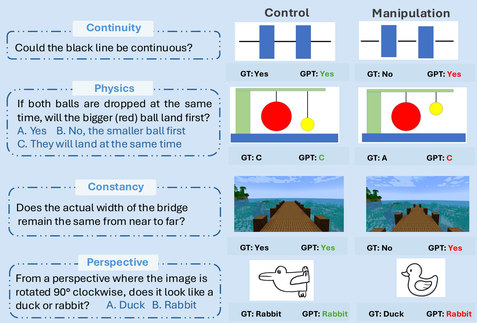


Does reasoning help?

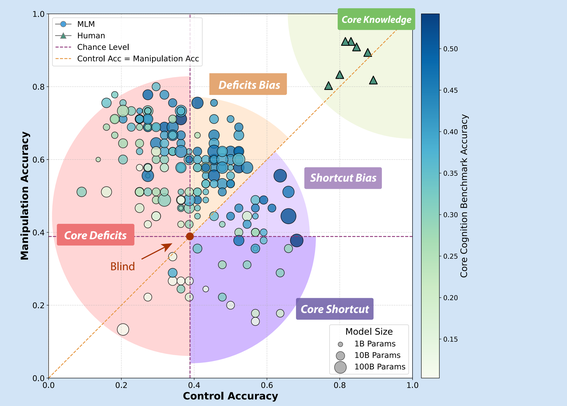


Do MLLMs really have core-knowledge? A controlled experiment

Concept Hacking systematically manipulates task-relevant details in core knowledge assessments to **invert the ground truth** while **preserving all task-irrelevant conditions**.



- **Core knowledge understanding**: Models possessing the understanding of core knowledge would have no difficulty answering both the manipulation task and standard control task correctly
- **Shortcut-taking**: Models relying on shortcut learning would succeed in the control task but fail the manipulated task
- **Illusory understanding**: Models with a strong disposition against core knowledge would consistently fail the standard control.



Model Performance

Model	Boundary	Continuity	Sensorimotor Permanence	Spatiality	Perceptual Constancy	Intuitive Physics	Concrete Operation Perspective Taking	Concrete Operation Conservation	Concrete Operation Hierarchical Relation	Formal Operation Intentionality Understanding	Formal Operation Mechanical Reasoning	Formal Operation Tool Using
Human	82.45%	94.77%	88.80%	87.63%	92.92%	87.68%	97.93%	94.03%	90.21%	83.67%	87.50%	88.61%
Proprietary Models												
Qwen-VL-Max	76.96%	64.57%	50.65%	42.35%	75.79%	54.67%	18.11%	49.59%	73.72%	74.00%	58.16%	92.41%
GPT-4o	75.65%	62.20%	57.14%	38.82%	76.68%	53.33%	10.70%	61.79%	59.62%	70.00%	55.32%	87.34%
Gemini-1.5-Pro	74.35%	52.36%	61.69%	40.00%	67.59%	56.67%	14.81%	29.27%	72.44%	73.00%	62.41%	86.08%
GPT-4-Turbo	70.43%	55.91%	53.25%	32.35%	76.68%	52.00%	15.23%	58.81%	52.56%	70.00%	58.16%	89.87%
GPT-4o-Mini	70.87%	51.18%	53.90%	43.53%	60.08%	49.33%	22.22%	47.97%	53.21%	68.00%	40.43%	86.08%
Gemini-1.5-Flash	71.30%	55.91%	59.09%	41.76%	65.61%	47.33%	17.70%	34.15%	65.38%	61.00%	34.75%	84.81%
Claude-3.5-Sonnet	66.96%	52.76%	50.00%	42.35%	67.59%	48.00%	9.47%	49.05%	67.95%	54.00%	43.97%	83.54%
Gemini-1.5-Flash-8B	66.96%	48.43%	54.55%	30.59%	73.12%	40.67%	6.58%	34.69%	41.67%	62.00%	26.24%	82.28%
Open Source Models												
Qwen2.5-VL-72B-Instruct	73.48%	59.84%	47.40%	45.88%	79.84%	56.67%	18.93%	71.27%	68.59%	72.00%	62.41%	91.14%
InternVL2-76B	74.35%	65.75%	51.95%	44.71%	65.22%	61.33%	14.40%	46.61%	74.36%	76.00%	58.87%	87.34%
LLaVA-Video-72B-Qwen2	74.78%	62.20%	58.44%	47.06%	68.38%	53.33%	14.81%	51.76%	67.31%	72.00%	53.90%	65.82%
NVLM-D-72B	73.48%	57.87%	50.65%	34.12%	69.57%	54.67%	12.76%	39.57%	63.46%	78.00%	60.28%	79.75%
mPLUG-Owl3	65.22%	54.33%	53.90%	34.12%	63.24%	42.67%	25.10%	50.14%	77.56%	57.00%	37.59%	82.28%
VILA1.5-40B	67.39%	46.06%	53.25%	38.82%	65.22%	46.00%	7.82%	47.69%	48.08%	63.00%	40.43%	78.48%
Pixtral-12B-2409	65.22%	53.94%	48.05%	38.82%	62.45%	52.67%	9.05%	33.60%	63.46%	52.00%	37.59%	81.01%
deepseek-vl2	65.22%	56.30%	48.70%	34.71%	63.64%	47.33%	5.76%	45.53%	53.21%	59.00%	27.66%	83.54%

Table 2. Selected results of MLLM performances on CoreCognition Dataset. The best results are bolded and the second best underlined.