Appendix 2 On Knowledge

World and instrumental knowledge.

can language be an instrument to build World knowledge

The clearest test is what counts as success. World knowledge is about how things are. It succeeds by being true, and it doesn't depend on any goal: knowing that water boils at a lower temperature at altitude serves a cook, a climber and a physicist, each for a different purpose. Instrumental knowledge is about how to get something done. It succeeds by working, and it is always tied to a purpose. This roughly matches Ryle's "knowing-that" versus "knowing-how."
A few markers help tell them apart:

- **How it fails:** world knowledge fails by being wrong. Instrumental knowledge fails by not working. The two can separate: a false belief can still support a recipe that works, as with some traditional remedies or a model that predicts well without capturing the real causes.
- **How far it travels:** world knowledge can be put to new goals. Instrumental knowledge tends to stay tied to its original task.
- **What prompts revision:** you revise world knowledge when evidence contradicts it. You revise instrumental knowledge when results disappoint, even if nobody knows why.

Pragmatists like Dewey argued the line is softer than this, since we test beliefs by acting on them. The distinction is still useful, though, because you can have reliable know-how without understanding, or understanding without the skill to act on it.
On language: yes, but it's an unusual instrument, because what it produces is itself world knowledge. It does at least three jobs:

- **Transmission:** most of what we know about the world we never observed ourselves. It reached us through what others told us.
- **Carving:** concepts make distinctions visible. Once you have a term like "confirmation bias," you start noticing cases of it. Vygotsky's point was that speech, once internalized, becomes a tool for thinking.
- **Stabilizing:** language fixes knowledge in a form others can check, dispute and correct. That is how private impressions become shared knowledge.

The catch is that language can also stand in for world knowledge instead of building it. Being fluent in a field's vocabulary can feel like understanding the field. Feynman's father made this point: knowing a bird's name in every language tells you nothing about the bird. Large language models make the issue concrete. Trained only on text, they show a lot of world-like competence, which suggests language carries a compressed map of the world. But it is a map learned without visiting the territory, and it inherits the blind spots of the texts it came from.
So language builds world knowledge, but its products have to be tested against contact with the world (observation, action, consequences). Otherwise they risk becoming knowledge about words.

- interesting, hotelera tiene is an artillería from yildrim and Paul frontera 2024 calle frontera task seductores to World models. Terminología is diferente. Could YouTube analiza and comentarios it, algo giving a link of where to get it and any follów ups? ..

Searched the web, read a page

The paper is Ilker Yildirim and L.A. Paul, *From task structures to world models: what do LLMs know?* The free preprint is arXiv 2310.04276, and the published version is in Trends in Cognitive Sciences, DOI 10.1016/j.tics.2024.02.008. Yale's lab also hosts the published PDF (links at the end). arxiv (https://arxiv.org/abs/2310.04276)

How their terms differ from the distinction I drew earlier

Their split is not "true vs. works." In their glossary, instrumental knowledge is knowledge of true propositions that a system has because of its set of abilities. Worldly knowledge requires that P has real-world content and that the subject grasps that content using a world model. A world model, for them, is a structure-preserving causal abstraction of real-world processes, the way a physics engine simulates how objects move. So both kinds concern true propositions. The difference is what the knowledge rests on: competence alone, or an internal model of the world. From task structures to world models: What do LLMs know? +2 (https://arxiv.org/pdf/2310.04276)

The argument in brief

- **Where instrumental knowledge comes from:** to predict the next word well, a model has to infer the task implied by the text (summarizing, translating, and so on) and condition its output on it. Machine translation is their example of a capability that may need very little shared knowledge of the physical world. arxiv (https://arxiv.org/pdf/2310.04276)arxiv (https://arxiv.org/pdf/2310.04276)
- **The leap:** compressing text might recover the process that generated it, and that process includes the world as people perceive and describe it. Their evidence is Othello-GPT: the board state could be read out of the model's internal layers, and editing it changed which moves the model predicted. But for color, the match between model and perceptual space was weak, a correlation of roughly 0.2. From task structures to world models: What do LLMs know? +2 (https://arxiv.org/pdf/2310.04276)
- **The resource-rational tradeoff:** worldly knowledge is costly, so a model acquires only as much as its frequent tasks require. That works when a coarse simulation is enough and breaks down when a finer one is needed. Their example: GPT-4 consistently answered the ball-on-a-box balancing question correctly, but mostly got it wrong when "ball" and "box" became "sphere" and "cube." arxiv (https://arxiv.org/pdf/2310.04276)arxiv (https://arxiv.org/pdf/2310.04276)

My comments

- **Strongest idea:** treating world knowledge as a matter of degree, with the degree set by task demands. That turns a yes/no metaphysical dispute into an empirical one. It also applies to people: humans also build sparse, task-driven representations of the world. So the gap between humans and LLMs may be one of degree rather than kind. arxiv (https://arxiv.org/pdf/2310.04276)

- **"Instrumental knowledge" is generous.** It counts reliable true output as knowledge with no belief, justification or agent behind it. This is where their critics push hardest.
- **"Worldly" is narrowly physical.** Their world models cover objects, scenes and agents. Much of what people know, such as law, history, institutions and mathematics, has no physics engine behind it. In those domains, language arguably is the world model, and the framework doesn't say what worldly knowledge would look like there.
- **The examples are dated.** They come from the GPT-4 era. By the paper's own logic, training on harder tasks should push models toward finer-grained world models. The authors list the effect of fine-tuning on worldly knowledge as an open question. arxiv (https://arxiv.org/pdf/2310.04276)

Follow-ups

- **Direct reply:** Goddu, Noë and Thompson (TiCS 28(11), 2024) raise two objections: the paper treats LLMs as agents rather than as models, and it assumes causal understanding can come from prediction alone. I didn't find a published rejoinder from Yildirim and Paul. philarchive (https://philarchive.org/rec/GODLDK)
- **Empirical test:** Vafa et al. (NeurIPS 2024) found that models passing standard world-model checks have far less coherent internal models than they appear to, which makes them fragile on slightly different tasks. A model trained on New York taxi routes produced valid routes about 99% of the time, yet its implied map of Manhattan was physically impossible. neurips (https://papers.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html)emergentmind (https://www.emergentmind.com/videos/evaluating-implicit-world-models-in-llms-996e4f05)
- **Counter-finding (2026):** a mechanistic study of that same taxi model found a highly faithful internal map. It traced the failures to how the model locates itself on the map, and argued the field should study how models use world models, not just whether they have one. arxiv (https://arxiv.org/pdf/2609.21748)
- **Philosophical follow-up:** a 2025 paper asks whether structural correspondence, the core of Yildirim and Paul's world-model criterion, is enough to give LLM representations real-world content. arxiv (https://arxiv.org/pdf/2506.16370)
- **The hybrid route they recommend:** Wong et al., "From Word Models to World Models," which translates natural language into a probabilistic language of thought. arxiv (https://arxiv.org/pdf/2310.04276)

Sources:

- arXiv preprint (Yildirim & Paul) (https://arxiv.org/abs/2310.04276)
- Published version, Yale lab PDF (https://cncl.yale.edu/assets/pdf/2024_YildirimPaul.pdf)
- Goddu, Noë & Thompson, "LLMs don't know anything" (PhilArchive) (https://philarchive.org/rec/GODLDK)
- Vafa et al., NeurIPS 2024 (https://papers.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html)

- World Modeling in Transformers (2026) (https://arxiv.org/pdf/2609.21748)
- Can structural correspondences ground real world representational content in LLMs? (https://arxiv.org/pdf/2506.16370)