TEACHING ROBOTS AND NAVIGATION

LLMs and Robotics

From embodied agency to vision-language-action models, robot-use agents and world models

Will an embodied LLM have agency?

If we embody LLMs in a robot, will they have agency?

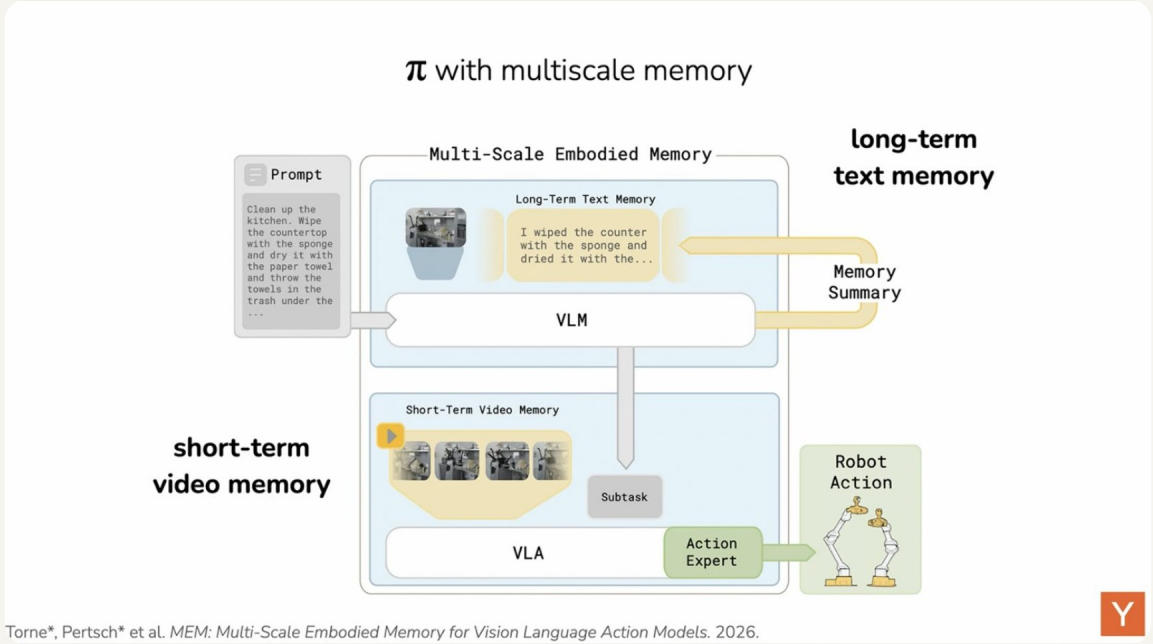
Alter3, a humanoid driven by an LLM: does it show a sense of agency?

Can Alter3 recognize itself when looking in a mirror?

Reading: Ikegami et al., Minimal Self in Humanoid Robot Alter3 Driven by a Large Language Model (arXiv 2406.11420)

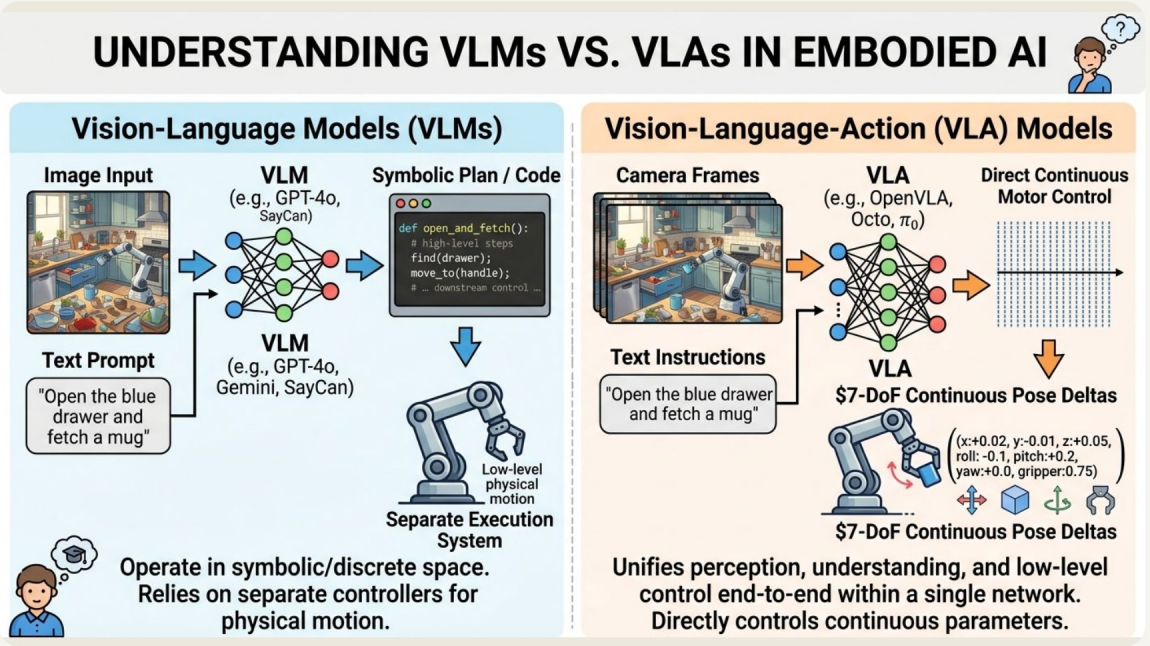
Inside a VLA model

- **State tokens**: internal state and sensor data (e.g. LIDAR)
- **Action tokens**: generated autoregressively, each one a low-level control signal such as a joint angle update or torque
- **Memory** carries context across a long task, such as drying the counter with a sponge



VLMs versus VLAs

- **VLM** (GPT-4o, Gemini, SayCan): image plus text in, text, plans or code out; needs downstream controllers
- **VLA** (OpenVLA, Octo, π0): camera frames plus text in, continuous 7-DoF deltas out (x, y, z, roll, pitch, yaw, gripper)
- One network unifies perception, instruction understanding and control



Safety of VLA models

- Visual input and language go straight to motor commands
- Open-world use needs robust behavioral safety guards
- Current VLAs lack them: adversarial images, jailbreaks and unsafe commands
- **Discuss:** what are the safety implications?

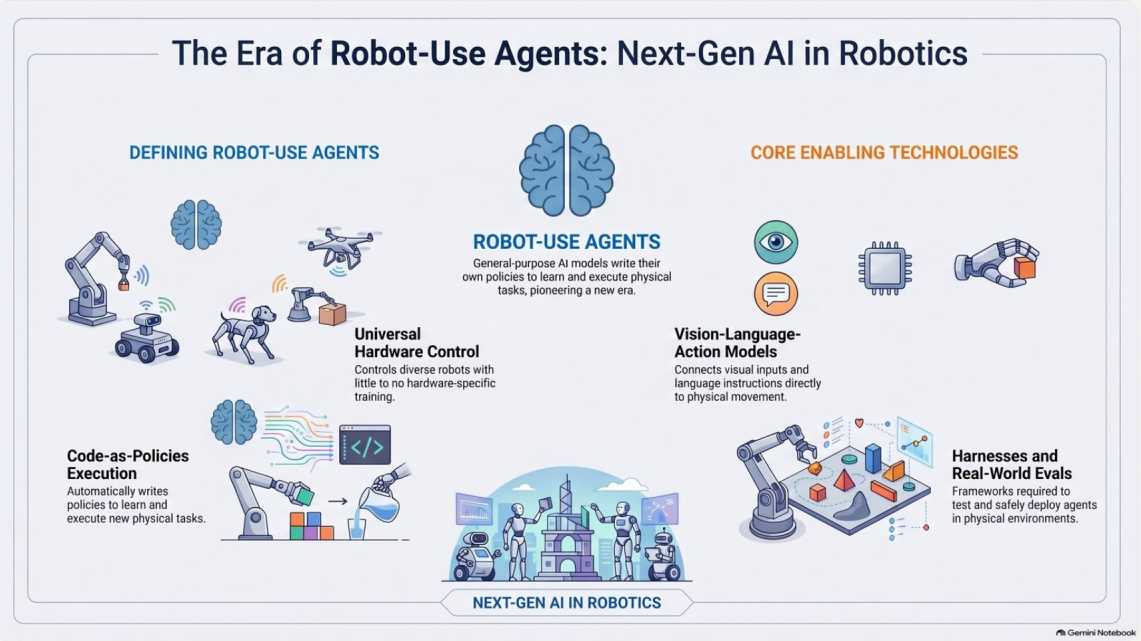


The era of robot-use agents

General-purpose models that control different robots, write policies and learn new physical tasks with little robot-specific training.

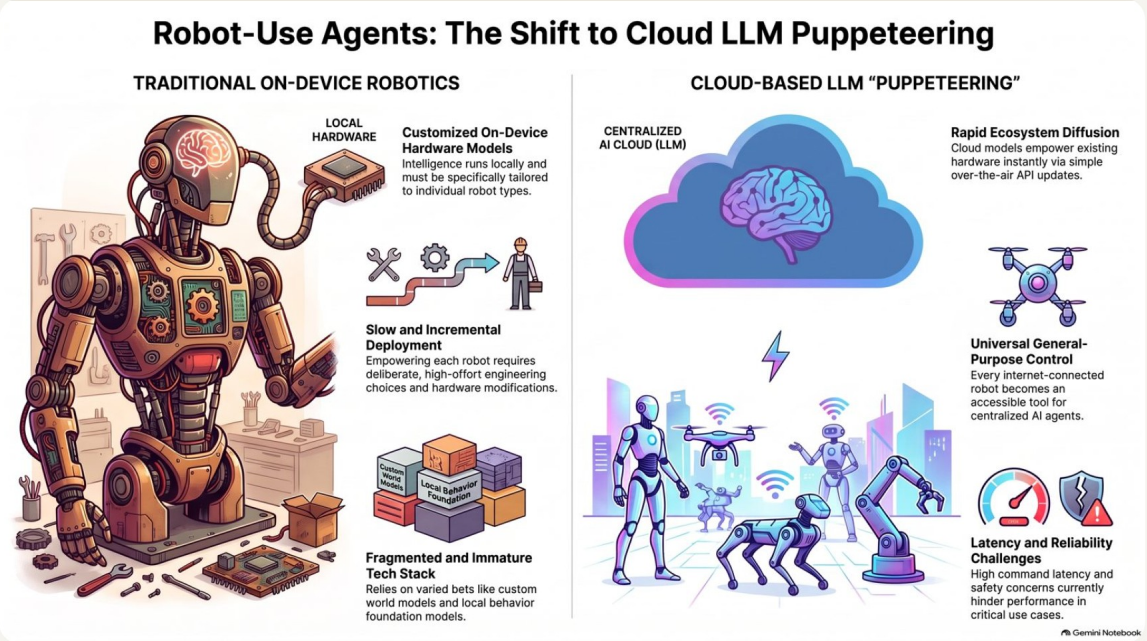
Philip Isola, MIT

Bitter Lesson (Sutton): leverage data and the LLM ability to write code and transfer it across domains.



LLM puppeteering from the cloud

- A shift from bespoke hardware to LLM agents with hardware in the loop
- Intelligence runs in the cloud and is not tuned to one robot
- A non-intelligent device could become AI-enabled with a software update
- **Caution:** puppeteering may be less reliable than proven robotic systems in safety-critical use

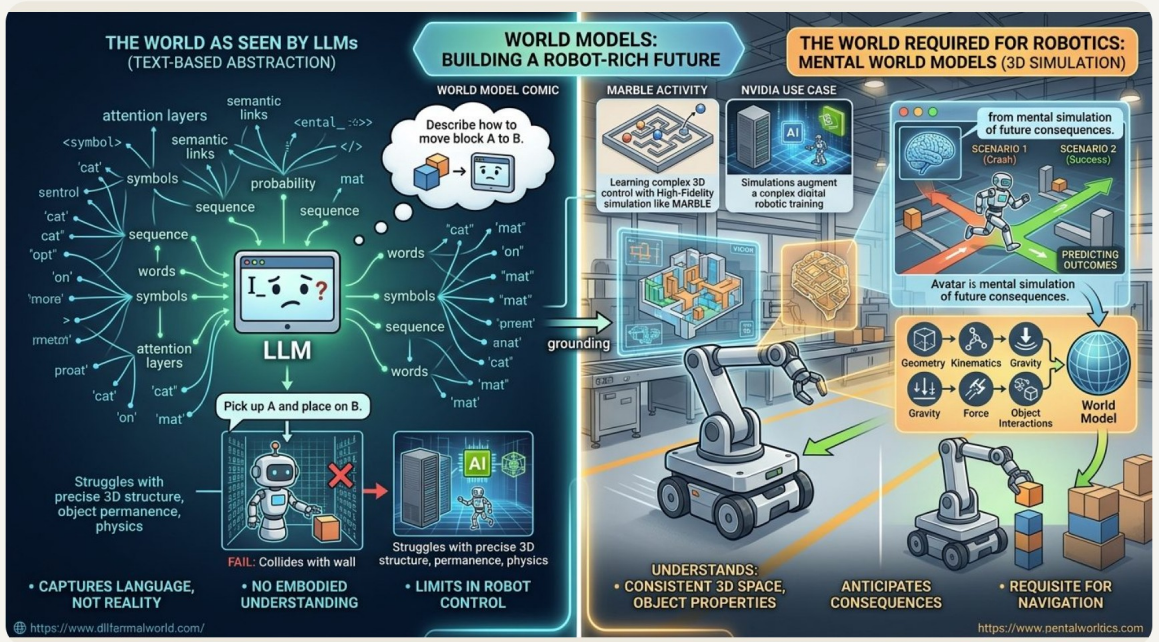


Claude plays robotics: 11 locomotion tasks

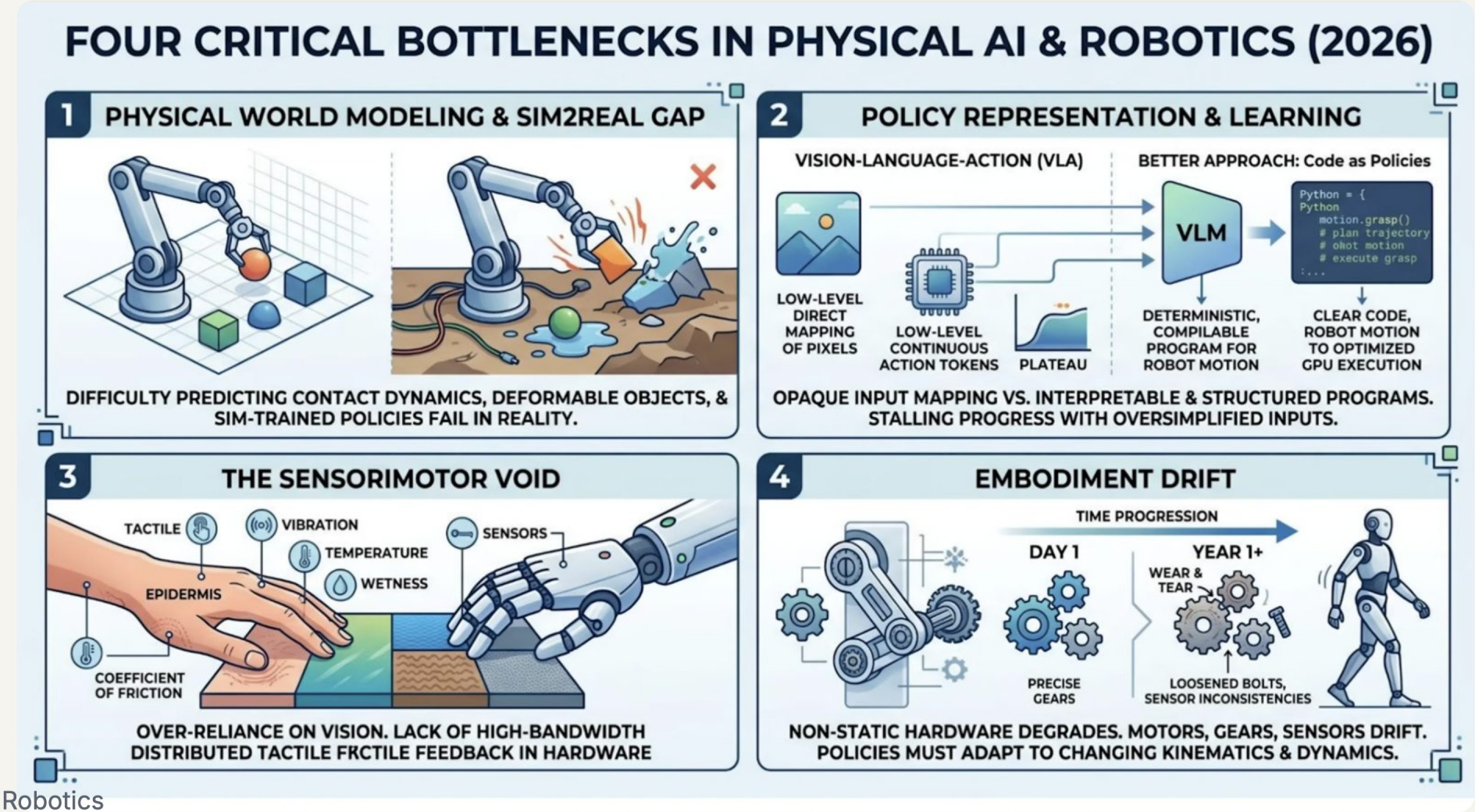
Task	What it tests
find_x	Walk to a table marked with a blue X, 25 ft away
visual_search	Find a red sphere hidden behind occluders in a 12 × 12 m arena
color_sequence	Visit colored circles in order; working memory
return_home	Follow waypoints, then return to origin from memory
procedural_maze	Navigate a maze with the forward camera only
invisible_walls	Replan around walls that cannot be seen
obstacle_course	Pass gaps; know the robot's dimensions
oneshot_course	Plan the whole command sequence from a top-down map
drift_detection	Notice silently corrupted commands and compensate
turn_correction	See that a turn was incomplete and correct it
explore_report	Explore, then answer layout questions from memory

World models and spatial reasoning

- **Discuss:** will LLMs struggle with spatial reasoning?
- Fei-Fei Li: a fully consistent 3D world is required for navigation
- LeCun: robots need AI that anticipates the consequences of actions
- Text-only intelligence lacks the physical grounding; a world model lets an agent test plans safely in simulation

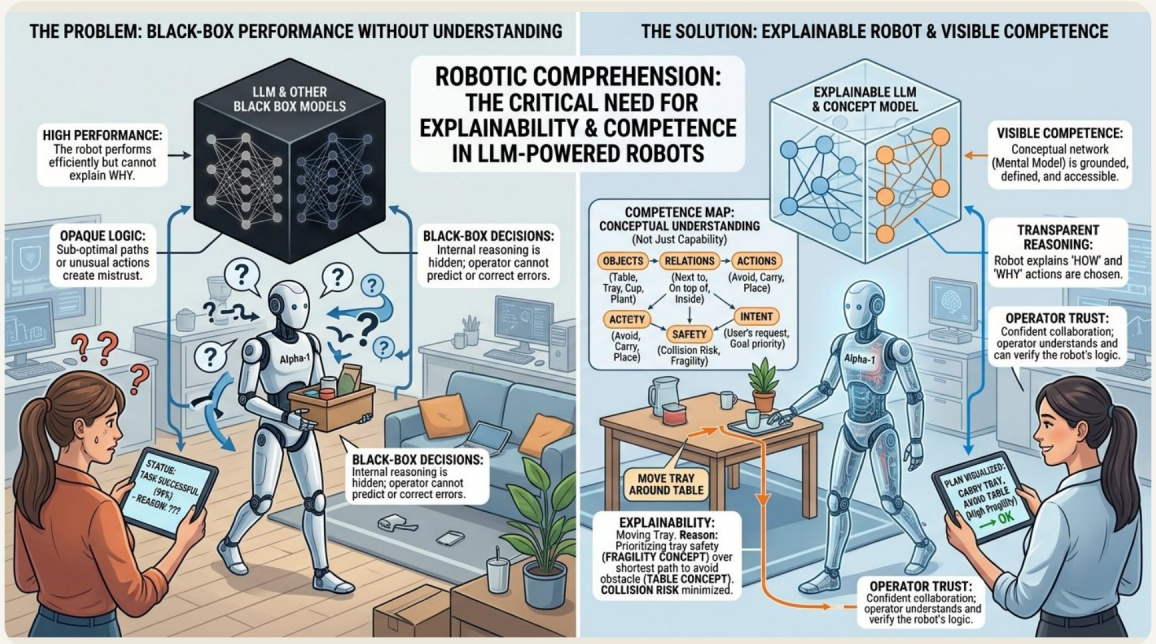


Bottlenecks for robotics (c. 2026)

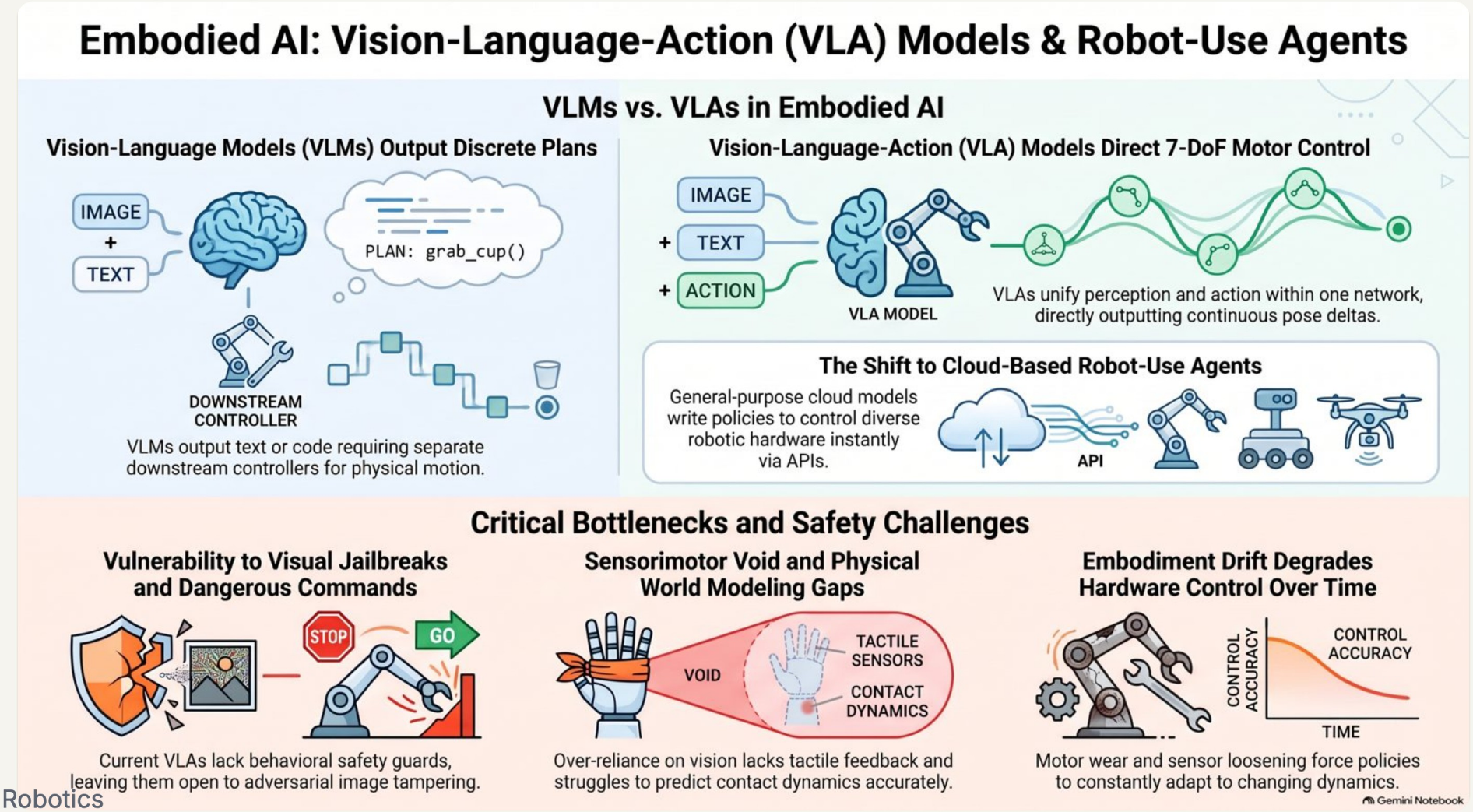


What about explainability?

- Is explainability important for these robots? **Discuss.**
- It links to competence versus performance
- LLMs may hallucinate their explanations



Summary: embodied AI models and challenges



Practicals, benchmarks and roadmap

Script A: a conceptual VLA in PyTorch, CPU only

Script B: OpenVLA-7B with 4-bit quantization on a Colab T4

Benchmarks: LIBERO, DROID, RoboCasa, BEHAVIOR-1K

Tools: LeRobot, Open X-Embodiment, inspect-robots

Vision → Language → Action → Data → Training → Evaluation → Deployment