

Foundation Models for Robotics Study

Executive Summary: Insights & Industrial Applications

Authors:

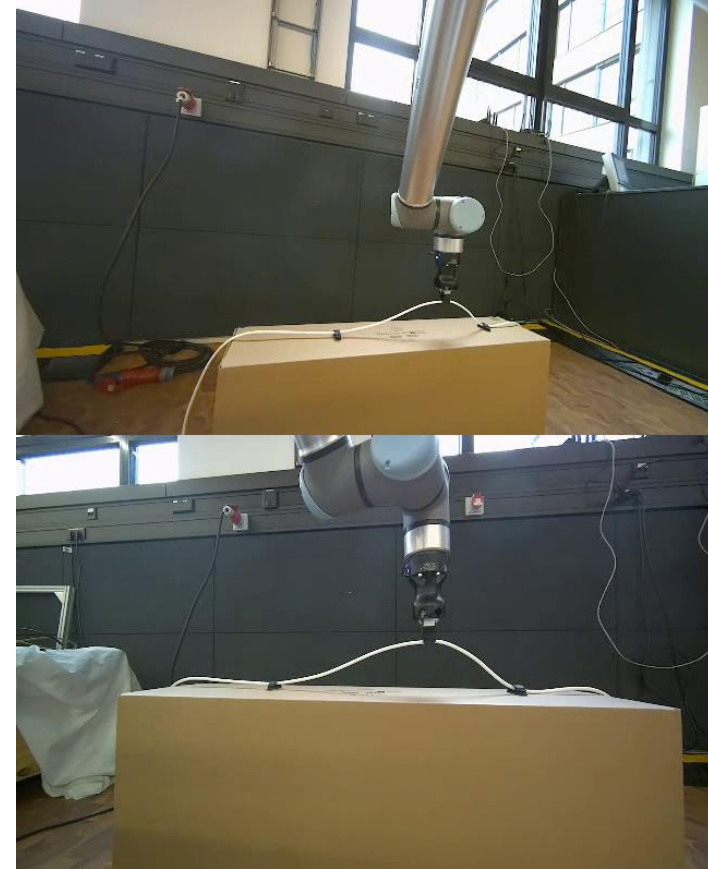
Çağatay Odabaşı, Florian Jordan, Jochen Lindermayr

Motivation

This comprehensive Foundation Models Study provides:

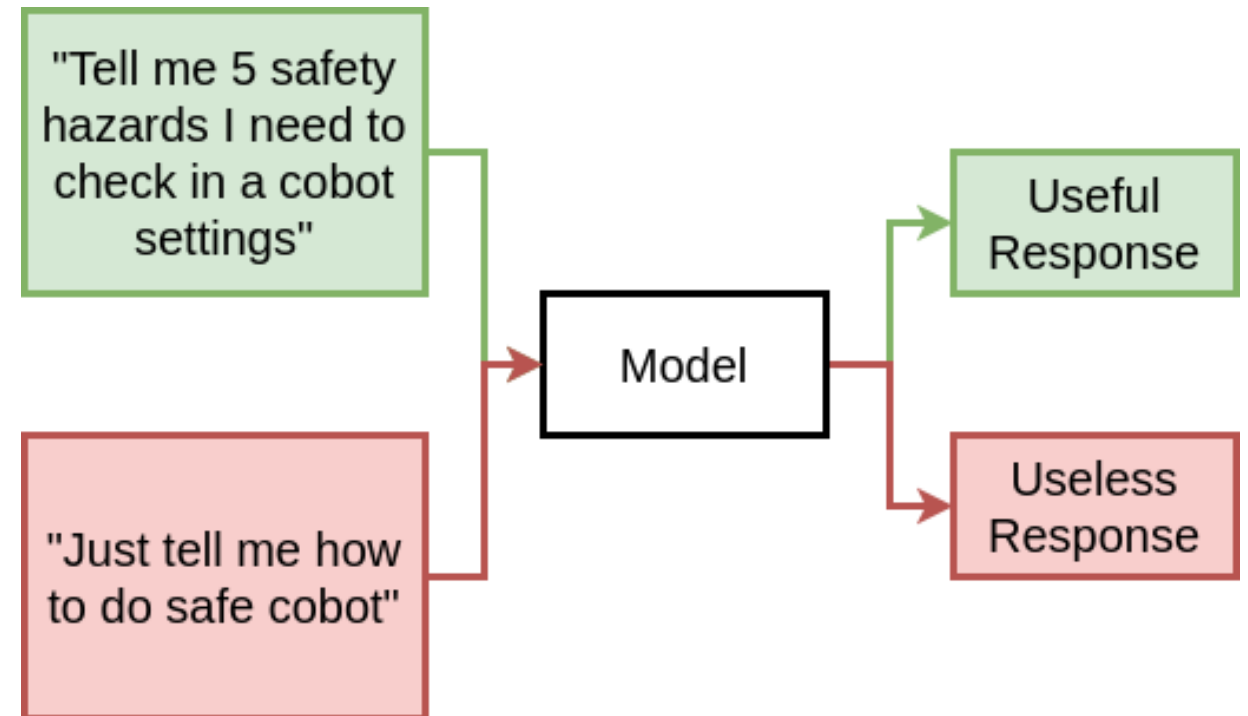
- **guidance** for those who want to integrate foundation models into their pipelines,
- **information** for those who want to explore foundation models with a strong focus on robotics.

We use this knowledge base to guide our industry partners. Reach out to schedule a deep-dive session into the full report.



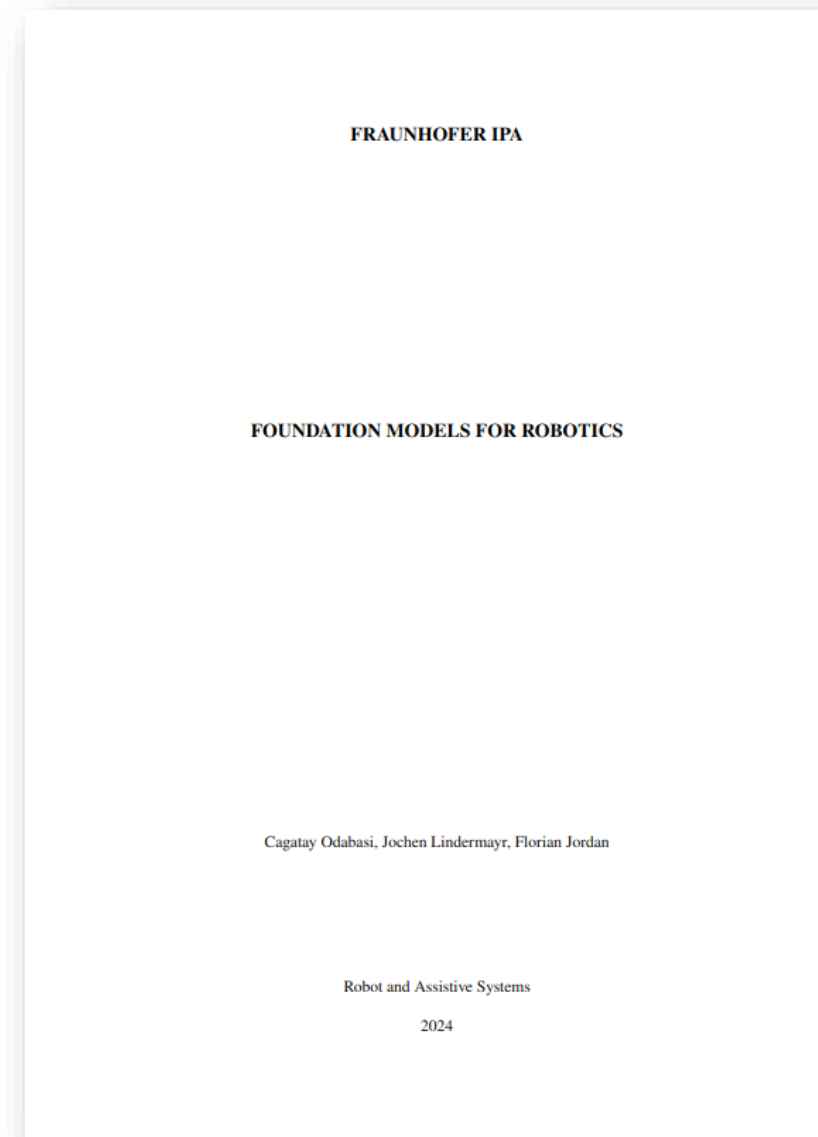
What is a Foundation Model?

- Foundation model:
 - serves as a flexible base model capable of performing various tasks due to its training on a diverse and extensive dataset.
 - flexible
 - trained with large scale data
 - zero shot capabilities
 - low alignment
 - unpredictable without safety alignment/guardrails
 - often requires instruction-tuning or strict prompting to execute specific tasks reliably.
- Examples for foundation models
 - LLMs, e.g. GPT-2
 - VLMs, e.g. Gemini
 - Foundation models for robotics, e.g. VLA (Vision Language Action) like Pi0.6

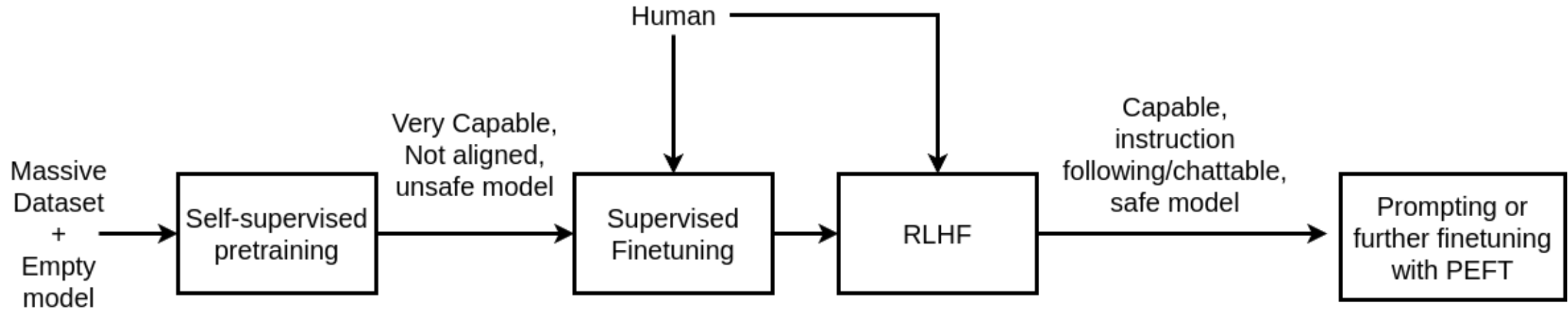


What is inside the full Study?

- **Curious about the details?**
 - **Contact us for a private walkthrough of our foundational 2024 report and our latest 2026 updates!**
- Literature survey
 - Introduced more than 300 references about LLMs, VLMs, and Robot Models
 - Identified research gaps
- Training techniques and costs
 - Detailed explanations of trainings and fine-tunings
 - Costs & tips to reduce these costs coming with the foundation models
- Pilot tests with foundation models
 - Result replications via official codes
 - Fine-tunings on different datasets
 - Testing LLMs and VLMs with different prompts



Training Overview of LLMs



Legend:

- RLHF
 - Reinforcement Learning from Human Feedback
- PEFT
 - Parameter Efficient Fine Tuning

LLM – Large Language Model

- LLMs:
 - process texts,
 - are the backbone of many different models,
 - come in many different sizes,
 - are very slow for robots, because token generation is not fast,
 - can hallucinate,
 - are trained to follow instructions,
 - native arithmetic reliability is limited without external tool-use,
 - useful for planning, user interaction, code generation, information retrieval.

USER 5 things to check in a cobot safety assessment

ASSISTANT : Here are 5 things to check in a cobot safety assessment:

1. **Risk Assessment:** Identify all potential hazards associated with the cobot's operation. This involves understanding the cobot's movement, speed, range of motion, and how it interacts with its surroundings.
2. **Safe Operating Environment:** Ensure the work area is free from clutter, obstacles, and other hazards that could interfere with the cobot's operation. Adequate space is crucial for safe movement and proper workspace design.
3. **Emergency Stop Functionality:** The cobot must have a clear and reliable emergency stop mechanism that can be easily activated in case of an emergency. This includes clear visual and tactile signals for operator awareness.
4. **Safety Features:** Assess the cobot's built-in safety features, including emergency stop buttons, sensors, and safety meshes. These features should be robust, properly calibrated, and functioning correctly.
5. **Task-Specific Considerations:** Beyond general safety measures, consider the specific task requirements and potential hazards associated with the cobot's operation. This may involve evaluating the cobot's handling capabilities, collision avoidance mechanisms, and interaction protocols.

Let me know if you have any other questions!

Say Something ... (Shift + Enter for new line)

Back

Upload Image

Stop

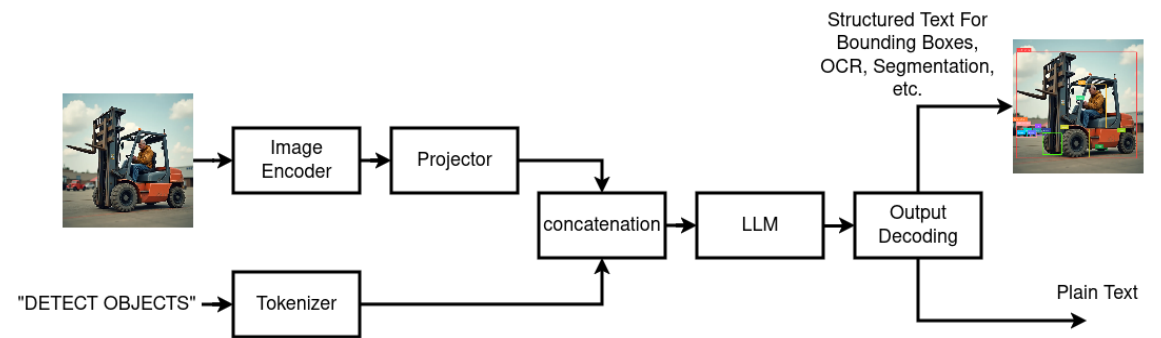
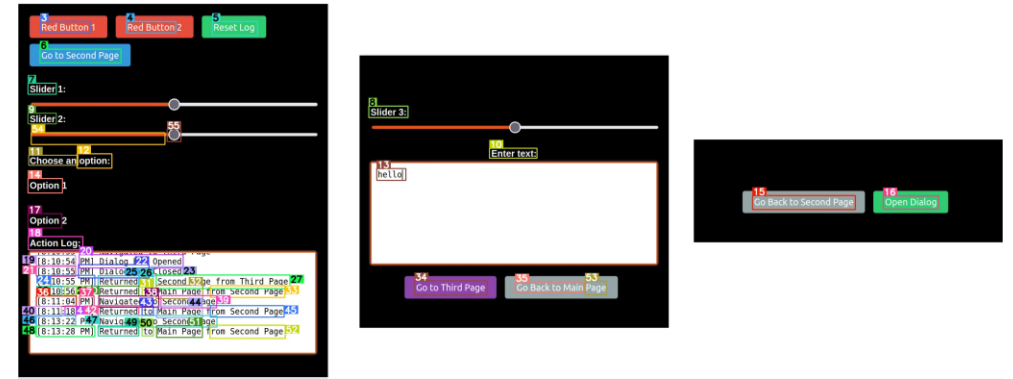
Submit

262 predicted, 279 cached, 4.06 tokens per second

Powered By llama.cpp and ggml.ai

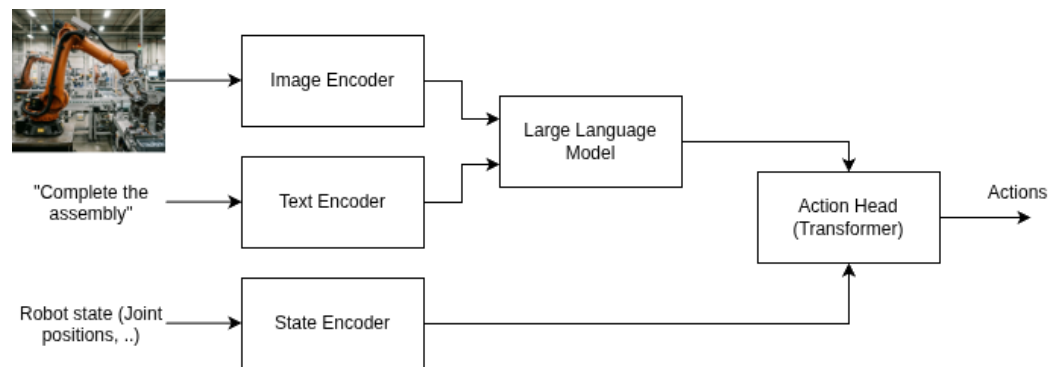
VLM – Vision Language Model

- VLMs:
 - can process text, image, multiple images and videos,
 - are generally built on top of a pretrained LLM, but they can be also natively multimodal,
 - are highly relevant to robotics,
 - are highly beneficial for document retrieval and classical computer vision,
 - token generation latency can bottleneck high-frequency real-time control loops.
- Precise millimeter-level 3D spatial reasoning is still evolving.



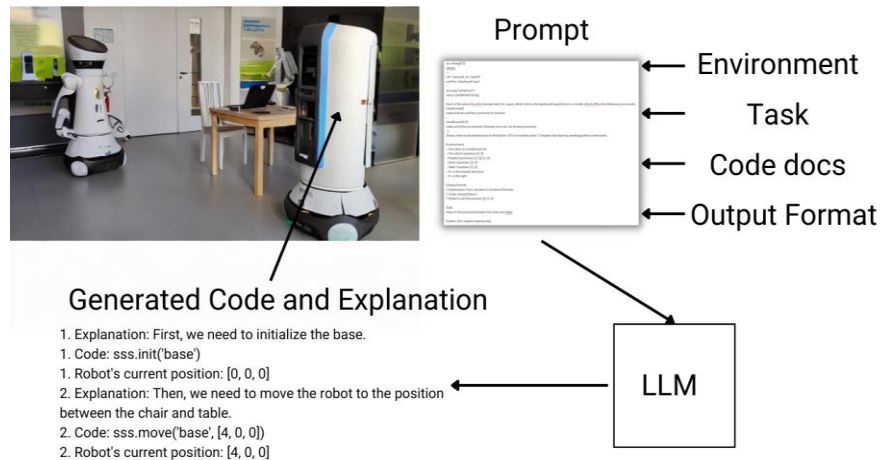
What is a Robot Foundation Model?

- Robot foundation model:
 - VLA's (Vision Language Action) are dominating the field but foundation models have many use cases in robotics beyond action generation
 - Generating actions to control the robots (still low-level controllers required)
 - Trained with diverse data from different robots and tasks
 - Currently poses safety risks for unconstrained public deployment
 - Able to control many different robots
 - Able to consume different input modalities (text, sensory data, vision, point cloud, ...)
 - Typically requires domain-specific fine-tuning for reliable execution.

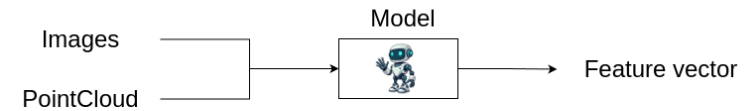


Types of Foundation Models for Robotics – Our Practically Oriented Segmentation

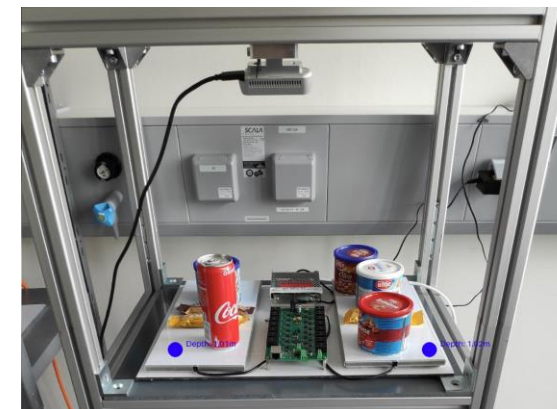
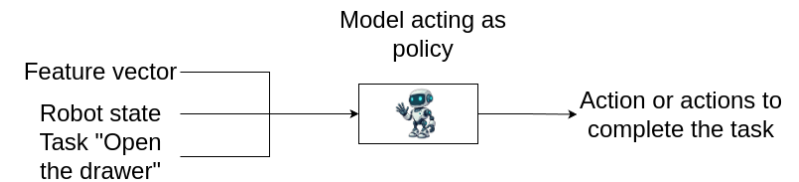
- VLA
- Visual Representation Models
- High-Level Planning
- LLM and VLM as Code Generators
- VLMs for Spatial Understanding



Visual Representation Models



Feature Vector Usage



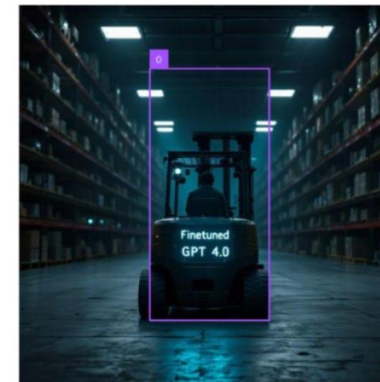
Our Strategic Recommendations for Industry to use Foundation Models in Robotics

- Use RAG (Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Lewis et al. 2021; see also next slide)
 - Leveraging documents in a database to ground answers of an LLM
- Use LoRA (Low-Rank Adaptation of Large Language Models, Hu et al. 2021; see also slide 15)
 - Finetuning only a fraction of model parameters to achieve near full finetuning performance
- Collect domain-specific, diverse robot data
- Leverage VLMs for fast PoC vision projects
- Use cloud GPUs – (Buying: 8x H100 – 400K, Renting: 8x H100 – 120K/year)
- Leverage cheaper LLMs
- Invest effort in foundation model peripherals
- Invest in simulation technologies
- Invest in agentic frameworks

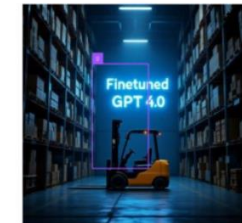
Findings on fine-tuning GPT-4 on a forklift detection dataset in a Proof of Concept (PoC)

Training 350 images at 640x640 resolution for 2 epochs used 590K training tokens.

☞ Trained tokens	589,650
☞ Epochs	2
☞ Batch size	1
☞ LR multiplier	2



It can learn a detection problem; however, it requires at least a few hundred images.



The dataset rules are strict: if a person appears in the data, those data points are ignored.

585 examples with images that were skipped for the following reasons: contains faces, contains people.

Retrieval-Augmented Generation RAG – What is it?

■ The Concept:

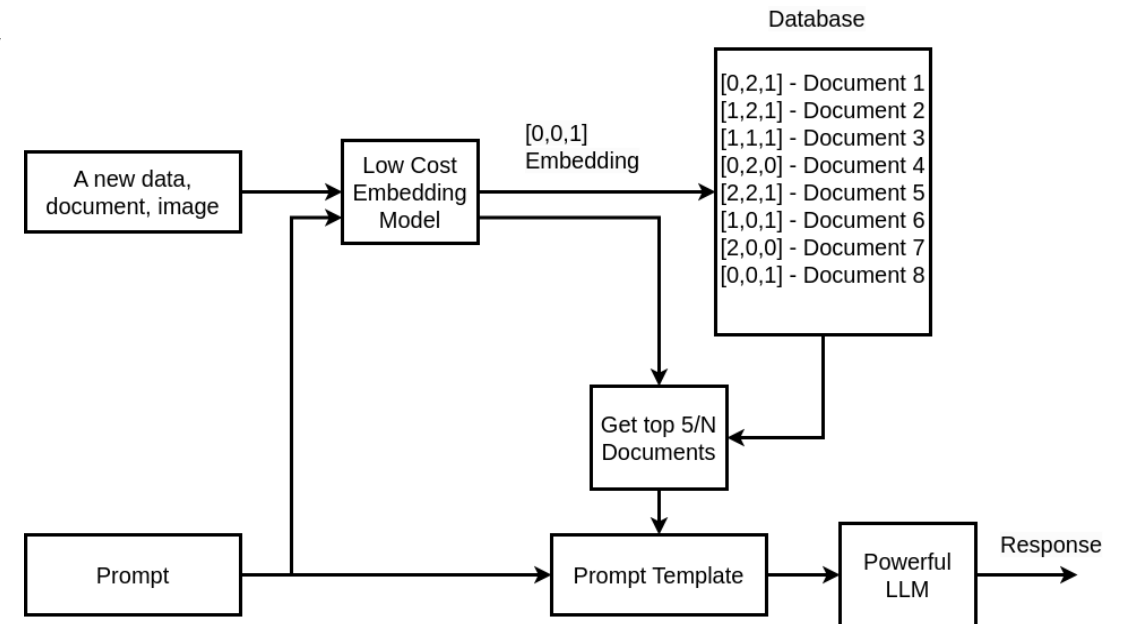
- Foundation models are capable, but their knowledge is frozen in time because of their fixed dataset. That's why retraining the LLM for every new data is not feasible. RAG is a concept of enrichment of LLM's memory with an intelligent search.

■ How it Works:

- RAG is starting with your question. The system finds most relevant texts in your documents and uses this up-to-date information to answer you.

■ Meaning for Robotics:

- Robot brains can search and reach millions of documents without requiring massive GPUs.



Retrieval-Augmented Generation RAG – Example

The screenshot displays a web-based chat interface for Retrieval-Augmented Generation (RAG). The top navigation bar includes 'Chat', 'Files', 'Resources', 'Settings', and 'Help'. The main content area is divided into three vertical sections:

- Conversations:** A sidebar on the left showing a list of chat sessions. The active session is titled 'Denoising Neural Networks in Deep Learning'. Below this are options for 'File Collection' (Search All, Search In File(s)) and 'Quick Upload'.
- Chat Area:** The central part of the interface. It shows a user query: 'What kind of network do they use for denoising part?'. The AI response is: 'They use two types of neural network architectures for the denoising part: 1. Convolutional Neural Networks (CNNs): They adopt the 1D temporal CNN from Janner et al. with some modifications. 2. Transformers: They introduce a novel transformer-based DDPM which adopts the transformer architecture from minGPT for action prediction.' Below the response are icons for copy, quote, and share.
- Retrieved Document:** A panel on the right showing a snippet from a document. The text discusses gradient descent steps and noise prediction networks. A citation is provided: 'Image from [Page 2] 2303.04137v4.pdf [score: 0.8] [Preview]'. Below the text is a small image of a robotic arm and a 'Relevance score: 0.8' indicator.

At the bottom of the chat area, there is a 'Chat Input' field, a 'Send' button, and a 'Regen' button. A 'Chat settings' dropdown is also visible.

<https://github.com/Cinnamon/kotaemon>

Use Low-Rank Adaptation of Large Language Models (LoRA) When Possible

■ The Concept:

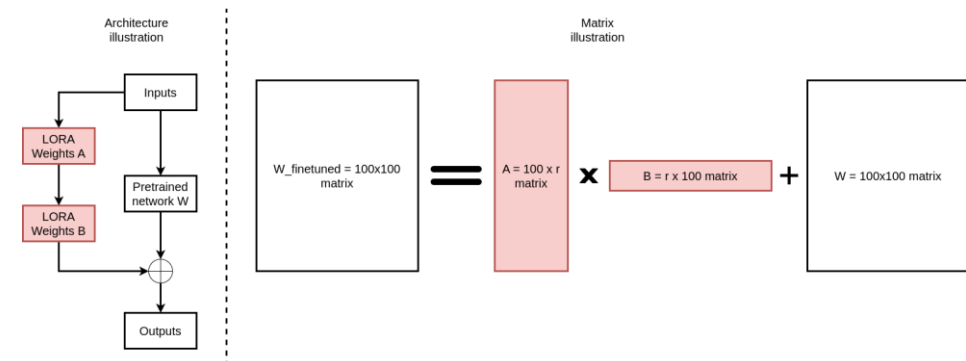
- Even fine-tuning a big LLM requires a massive training cost. That's why LoRA allows us to train a massive AI model in a very short time at a very low cost.

■ How it Works:

- LoRA intentionally introduces a minuscule number of parameters, i.e., 1% of the total network, and trains only those. It then merges them with the main network. Also, as it is easily pluggable and unpluggable, it is used to change the behavior of the AI model. For example, you can make it an expert in robot arms, or it can be an expert in humanoids, with different LoRA modules.

■ Meaning for Robotics

- Adapting a robot cell to a new application or to a new object is much less expensive.



Model Name	LoRA Rank	LoRA Alpha	Trainable Params to Total Param Ratio	GPU	Time	# of Images	MAP 50_95
Florence 2 Large FT	16	8	1%	4090	35 min.	882	0.55
Florence 2 Large	16	8	1%	4090	35 min.	882	0.55
Florence 2 Base FT	16	8	1.4%	4090	15 min.	882	0.61
Florence 2 Base	16	8	1.4%	4090	15 min.	882	0.60
Florence 2 Large	Full Fine Tune	Full Fine Tune	100%	A100	60 min.	882	0.68
Florence 2 Base	Full Fine Tune	Full Fine Tune	100%	A100	25 min.	882	0.66
Florence 2 Base	128	1	10%	4090	15 min.	882	0.57
Florence 2 Base	8	1	0.7%	4090	15 min.	882	0.49
Florence 2 Base	32	8	2.7%	4090	15 min.	882	0.61
Florence 2 Base	8	8	0.7%	4090	15 min.	882	0.58

Invest in Agents

Generation of a carehome environment with different personas for a drink serving robot

```
factory = TinyPersonFactory("A carehome")
manager = factory.generate_person(
    """
    Care home manager. Her name is Lena.
    """)
)
level_3 = factory.generate_person(
    """
    careworker who is taking care of people at level 3. Her name is Jenny.
    """)
)
level_2 = factory.generate_person(
    """
    careworker who is taking care of people at level 2. His name is Daniel.
    """)
)
chef = factory.generate_person(
    """
    Chef of the carehome. He is responsible of drinks and meals for the carehome. His name is Mark.
    """)
)
brian = factory.generate_person(
    """
    A resident of Carehome. His name is Brian. He is interested in Robots. Reads physical newspaper. Lives at level 2.
    """)
)
may = factory.generate_person(
    """
    A resident of Carehome. Her name is May. She is not interested in tech. She has no cellphone. Lives at level 3.
    """)
)
world = TinyWorld("Focus group", [manager, level_3, level_2, chef, brian, may])
world.broadcast("""
A drink serving robot is deployed at level 3. Careworkers should fill it in with drinks
and it serves these drinks to the residents.
We need a feedback from you all about the 2 weeks of experient.
""")
```

Feedback from some agents. Interestingly they came up with the same problems and benefits that we found during the real deployment without even seeing the robot

```
Brian --> Lena: [CONVERSATION]
> The drink serving robot has been a fascinating addition to our care home. It offers a
> unique experience for the residents and allows careworkers to focus more on direct
> care. However, like others have mentioned, it struggles with navigation in crowded or
> tight spaces and requires frequent restocking. Improving these aspects could
> significantly enhance its efficiency and reliability.
```

```
Lena --> Jenny: [CONVERSATION]
> Based on the collective feedback, it's clear that while the drink serving robot has been
> a positive addition, offering a unique experience for residents and allowing careworkers
> to focus more on direct care, there are consistent challenges with its navigation in
> tight spaces and the need for frequent restocking. Addressing these issues could
> significantly enhance its efficiency and reliability.
```

Images from real experiment



Limitations of Existing Models in Robotics

- Data and Training
 - Domain Knowledge Gaps
 - Training Challenges
 - Multimodal Data Requirements
 - Sim-to-Real Gap
 - Data Transparency Issues
- Requirement Gap
 - Spatial Reasoning Limitations
 - Inference Time and Latency
 - Long-Running Operations
- Robustness and Safety
 - Interpretability Issues
 - Reliability Concerns
 - Safety Measures

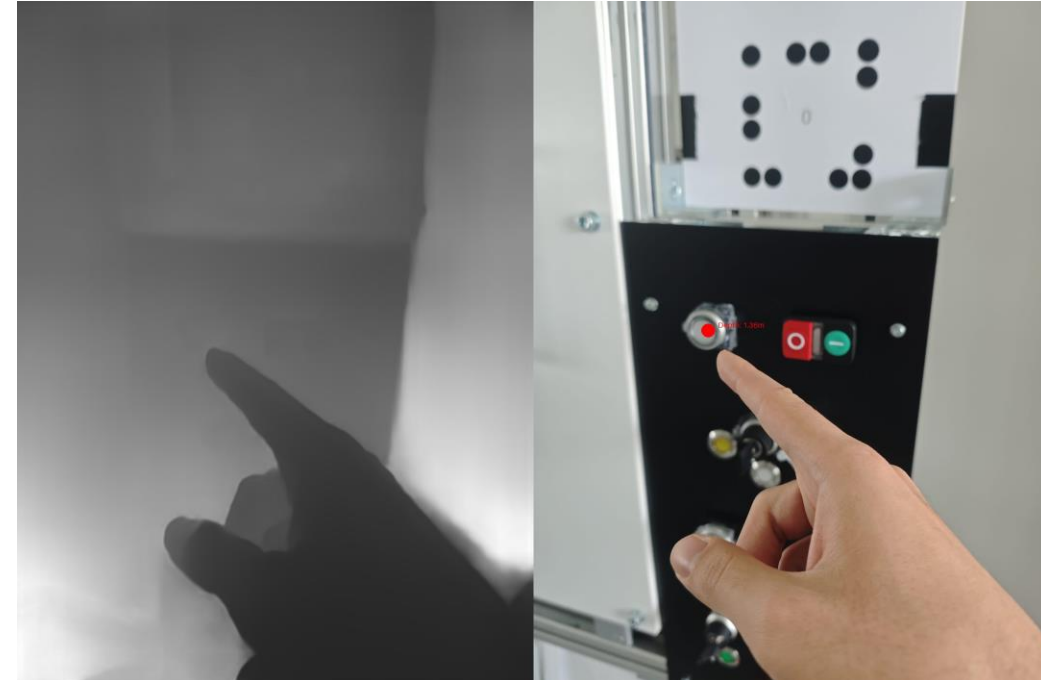
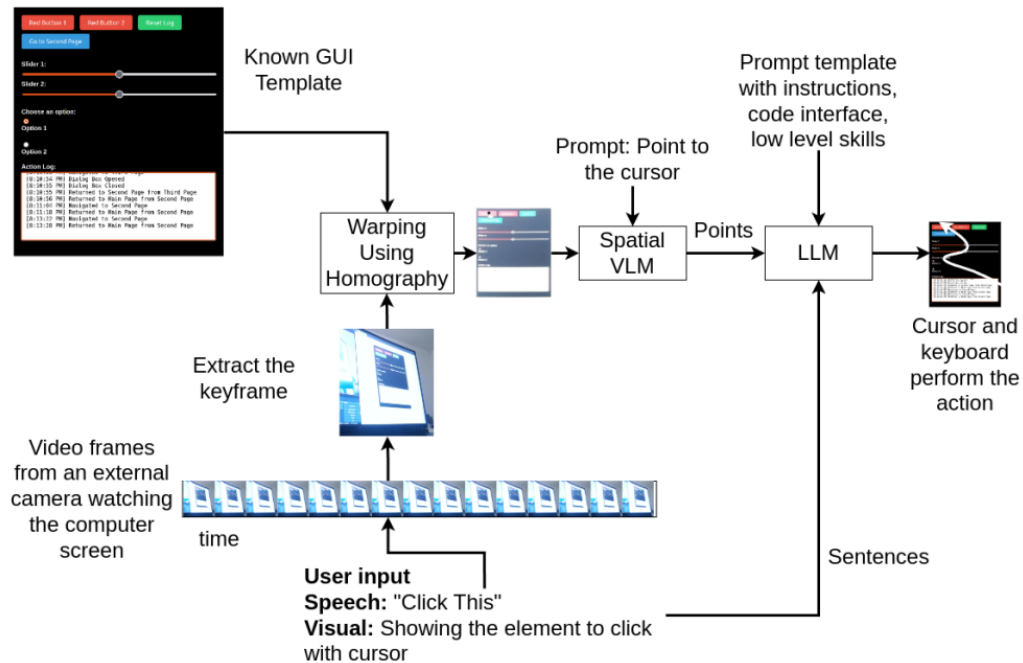
Joint Development Opportunities with You

- Simplify robot programming for complex tasks with LLMs (e.g., machine tending, maintenance)
- Enable better scene understanding for mobile robots in dynamic environments
- Integrate LLMs for advanced error handling and autonomous recovery
- Agentic Robotics implementations
- PoC tests with Teleoperation



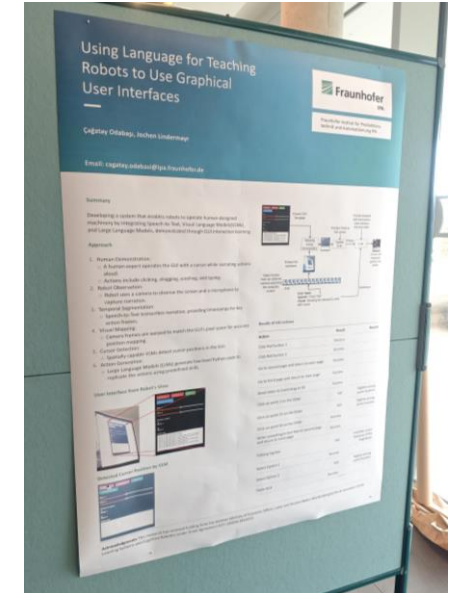
Our Paper at LangRob CoRL 2024 Workshop

Using Language for Teaching Robots to Use Graphical User Interfaces



What have we done during and after the study?

- **Publications:** Published a paper at [CASE 2025](#) and presented earlier works at the CoRL 2024 LangRob Workshop.
- **Media:** Published in a widely-read [Handelsblatt article](#) and contributed to various articles on Foundation models and Humanoids.
- **Industry & Demos:** Presented SOTA insights to major industry players at IPAI and executed highly successful live validation demos at Automatica.
- **Proposals:** Insights utilized to drive major EU/National proposals as well as industrial project acquisitions.



Contact

Çağatay Odabaşı

cagatay.odabasi@ipa.fraunhofer.de

Florian Jordan

florian.jordan@ipa.fraunhofer.de

Jochen Lindermayr

jochen.lindermayr@ipa.fraunhofer.de

Fraunhofer IPA
Nobelstrasse 12
70569 Stuttgart



Funded by  **Baden-Württemberg**
Ministerium für Wirtschaft,
Arbeit und Tourismus

The complete Foundation Models for Robotics Internal Report contains deep dives, extensive literature reviews, and benchmark data. Contact us to schedule a presentation of the full study or to explore joint collaborations.

Imprint

Editors

Prof. Dr.-Ing. Thomas Bauernhansl, Prof. Dr.-Ing. Marco Huber, Dr.-Ing. Werner Kraus

Authors

Cagatay Odabasi, Florian Jordan, Jochen Lindermayr

Funding Body

This study was made possible by funding from the AI Innovation Center "Learning Systems and Cognitive Robotics."
It is part of a series of studies focusing on the topics of AI and robotics. The authors are solely responsible for the content of this publication.

Contact address

Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Nobelstrasse 12, 70569 Stuttgart
Phone +49 711 970-3874
presse@ipa.fraunhofer.de
<https://www.ipa.fraunhofer.de>

License

The study "Foundation Models for Robotics" is licensed under the following Creative Commons license:
Attribution – Non-Commercial – No Derivatives
International 4.0 (CC BY-NC-ND 4.0).
Details about the license: <https://creativecommons.org/licenses/by-nc-nd/4.0/>

All rights reserved

© Fraunhofer Institute for Manufacturing Engineering and Automation IPA, April 2026



Funded
by



Baden-Württemberg
Ministerium für Wirtschaft,
Arbeit und Tourismus