Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 87–92
DOI: 10.15439/2025F0373
ISSN 2300-5963 ACSIS, Vol. 44

Towards Human-Robot Interaction in Agriculture Using Large Language Models

Lavanyan Rathy, Håvard Pedersen Brandal, Weria Khaksar
Norwegian University of Life Sciences, Ås, Norway
Email: lavanyan.rathy@nmbu.no; havard.pedersen.brandal@nmbu.no, weria.khaksar@nmbu.no

***Abstract*—Labor shortages and usability challenges limit the adoption of robotics in agriculture. This work explores how Large Language Models (LLMs) and Vision-Language Models (VLMs) can bridge this gap by enabling non-expert users to command robots using natural language. A modular system was developed to interpret instructions, execute tasks, and generate visual field reports. Evaluations in a simulated field showed that hybrid prompting strategies yielded reliable plans, while VLMs supported effective object detection and contextual reporting. This approach reduces entry barriers to robotics and promotes accessible, intelligent agricultural automation.**

Keywords: Large Language Models, AI, HRI, NLP, Precision farming, digital agriculture

I. Introduction

A. Motivation and Background

ROBOTICS is a rapidly evolving field with the potential to address pressing global challenges, particularly in sectors like agriculture [7]. However, deploying robotic systems in practice often demands high technical expertise, limiting accessibility for non-experts.

Norwegian agriculture, for example, faces critical challenges such as labor shortages, food waste, and reduced productivity [6], [4], [2]. Robotic solutions could address these issues by automating labor-intensive tasks. However, the complexity of current systems often discourages adoption, especially among farmers unfamiliar with robotics or programming [5].

Recent advances in artificial intelligence, particularly large language models (LLMs), present an opportunity to close this usability gap. LLMs can interpret and respond to natural language instructions, enabling intuitive, conversational interfaces. This could significantly lower barriers to adoption, allowing farmers to operate advanced robotic systems through simple, everyday language [8].

B. Problem Statement and Objectives

Despite the potential of robotics to transform agriculture, usability remains a core barrier. Most current systems are not designed for non-technical users, limiting their impact on productivity and sustainability [6].

This work addresses that challenge by exploring how LLMs and vision-language models (VLMs) can make human-robot interaction (HRI) more natural and accessible. Specifically, the system interprets written instructions, plans and executes robotic actions, and processes visual data to generate human-readable field reports.

The main objectives of this study are to:

- **Develop a multimodal LLM/VLM system** that translates natural language and visual input into ROS2-compatible robot actions.
- **Evaluate the accuracy and reliability of LLM-generated action plans**, including the impact of robotic hardware limitations.
- **Analyze how different prompt engineering strategies** affect command quality and consistency.
- **Assess VLM capabilities for object detection and spatial reasoning** in agricultural environments.
- **Demonstrate VLM-based visual reporting**, including structured outputs that enhance transparency and oversight.

C. Research Questions

To evaluate the proposed approach, this research is guided by the following questions:

- How accurately can an LLM generate executable ROS2 action plans from natural language instructions, and how do hardware limitations affect execution?
- How do different prompt engineering strategies influence output quality and consistency?
- How effectively can a VLM identify and localize agricultural objects, and what are its spatial limitations?
- Can VLMs produce interpretable, natural-language field reports from visual input that support human-robot collaboration?

II. Background and Related Work

Recent advances in LLMs and VLMs have enabled more intuitive human-robot interaction, particularly in contexts requiring high-level reasoning and accessibility for non-experts. LLMs such as GPT-4 exhibit strong generalization capabilities across tasks like planning, summarization, and code generation without retraining. Their ability to interpret natural language and produce structured outputs makes them a compelling option for high-level robotic control [8].

Prompt engineering has emerged as a key factor in improving the consistency and accuracy of LLM outputs. Direct prompting involves single-shot commands but often lacks reliability. Chain-of-thought (CoT) prompting helps by introducing intermediate reasoning steps, while few-shot prompting provides examples to anchor the model's behavior. Hybrid strategies, combining CoT and few-shot, can further enhance both interpretability and execution success in planning tasks [1].

VLMs extend this capability by jointly processing image and text inputs. Trained on large-scale image-caption datasets, models like CLIP and BLIP can identify and describe visual content, perform spatial reasoning, and generate contextual reports. This is particularly valuable in agriculture, where visual cues, such as detecting obstacles or crop conditions, play a vital role in robot operation [3].

Integrating LLMs and VLMs in robotic applications introduces a multimodal reasoning layer, enabling systems to move beyond hard-coded control toward flexible, adaptive interaction. Although prior work has demonstrated the potential of these models in lab settings, their deployment in field robotics, especially under agricultural constraints, remains underexplored. This research addresses that gap by combining LLM and VLM modules in a ROS2-based system that translates natural language commands and visual input into executable robot actions and structured field reports.

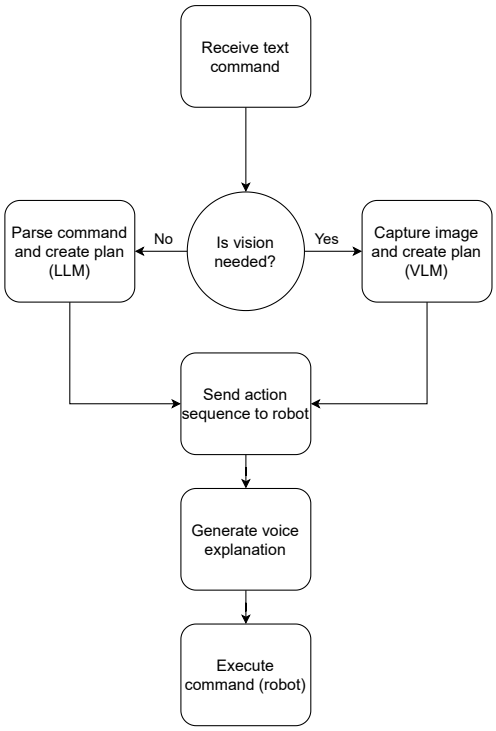

Fig. 1. High-Level Architecture for LLM-Based Robotic System

III. Methodology

A. System Architecture

The system follows a modular architecture combining language and vision models for robotic control. As shown in Figure 1, it processes natural language commands through an LLM to generate ROS2-compatible action plans. If visual input is required, a VLM interprets camera images to support perception and reporting. The robot then receives executable commands and a spoken summary of intent for transparent interaction.

The natural language command is processed through a Langchain pipeline using a `FewShotPromptTemplate`, which embeds dynamic user input and curated examples to shape the model's interpretation. The prompt structure includes a task description, spatial constraints, and example command formats. The LLM response contains a natural-language summary and a structured plan expressed in pseudo-code or action-like instructions. These are then parsed and verified using a YAML schema to ensure semantic and syntactic validity.
An example output may resemble:

```
Plan:
  - drive(2)
  - turn(90)
  - drive(2)
```

This intermediate representation allows modular validation and easier debugging. Internally, each action string is mapped to a corresponding ROS2-compatible function. For instance, `drive(2)` translates to a call to the navigation stack or a custom publisher on the `/cmd_vel` topic with linear velocity commands for a specified duration. Angle commands like `turn(90)` trigger a PID-regulated angular velocity loop with quaternion goals defined in radians. All interpreted commands are time-stamped and executed via a ROS2 executor, ensuring synchronization and feedback integration. For planning errors or misinterpretation, fallback handlers can re-query the LLM using augmented prompts that include failure context.

B. Simulation Environment

Development and validation were conducted in Gazebo Classic using the Peik robot, modeled in URDF/Xacro to replicate real-world geometry and sensor layout (Figure 2). ROS2 middleware facilitated communication across components. Peik's simulated sensors include a front-mounted RGB-D camera and an IMU. The robot was simulated in a maize field using Gazebo, with onboard RGB-D sensing and inertial measurement to support planning, perception, and trajectory tracking. A modular ROS2 architecture handled action execution and data flow between the LLM, VLM, and navigation stack.

The robot base is configured with a `'base_link'` and `'camera_link'` transform, aligned using static TF publishers. The URDF includes a ZED-like camera plugin with near-true RGB-D behavior. Odometry is simulated using differential drive parameters in Gazebo, allowing accurate benchmarking of LLM trajectory plans versus actual ground truth paths. The robot's rotational behavior is tuned with angular velocity limits of ± 1.5 rad/s and a max forward speed of 0.5 m/s, constrained for safety in narrow-field crop paths.

Fig. 2. Peik operating in a simulated maize field

C. LLM-Based Command Interpretation

User instructions are sent via a ROS2 topic and processed by an OpenAI-powered LLM using the Langchain framework. Prompts are dynamically constructed to include reasoning and explicit robot actions. Responses are parsed into a human-readable explanation (spoken aloud) and a command list (e.g., `drive(2)`, `turn(90)`), which is executed by the robot. The system triggers visual processing if the response contains the keyword `[CAMERA_REQUIRED]`.

D. VLM Integration for Perception

For visual reasoning, the system captures a JPEG image from the robot's camera, encodes it in base64, and sends it with a text prompt (e.g., `"What's in this image?"`) to a GPT-4-based VLM. The model returns a natural-language description of the scene, including obstacle presence or task-relevant objects. This output is both published and spoken by the robot for transparency.

IV. Results

A. Trajectory Execution

The system was evaluated using a square-pattern navigation task, where the LLM generated a plan from the command: *"Move in a square pattern, each side one meter long."*. The robot successfully executed the plan with minor trajectory drift. A PID controller improved tracking accuracy compared to open-loop control. Figure 3 shows the odometry trace before and after adjustment.

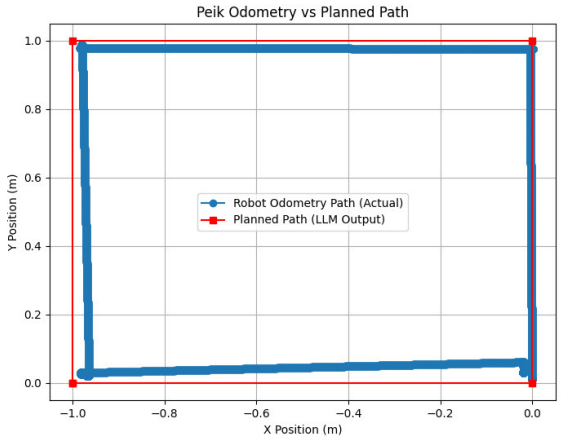

Fig. 3. Robot trajectory: Open-loop vs PID control

B. Prompting Strategy Comparison

Four prompting strategies were compared: Direct, Chain-of-Thought (CoT), Few-Shot, and Hybrid. Each strategy was tested using the same navigation task in the simulation. Figure 4, 5, 6 and 7 shows one example of each run.

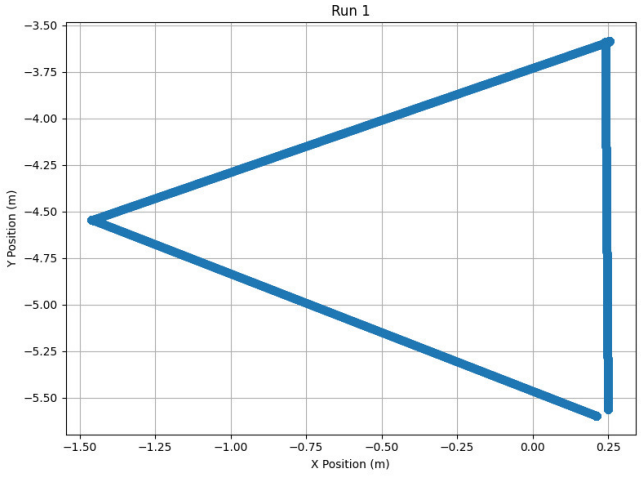

Fig. 4. Example of prompt strategy (CoT) run

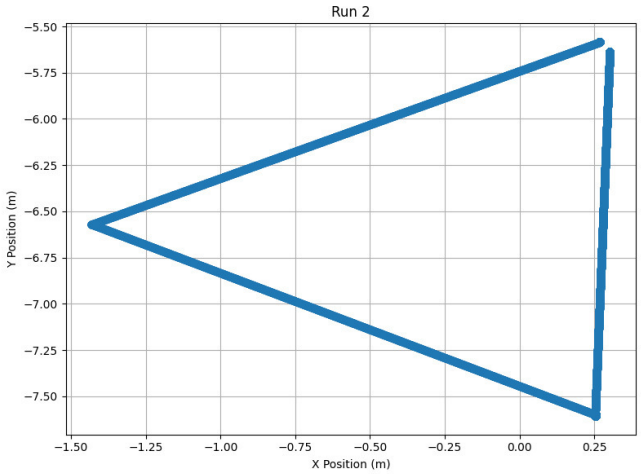

Fig. 5. Example of prompt strategy (Direct) run

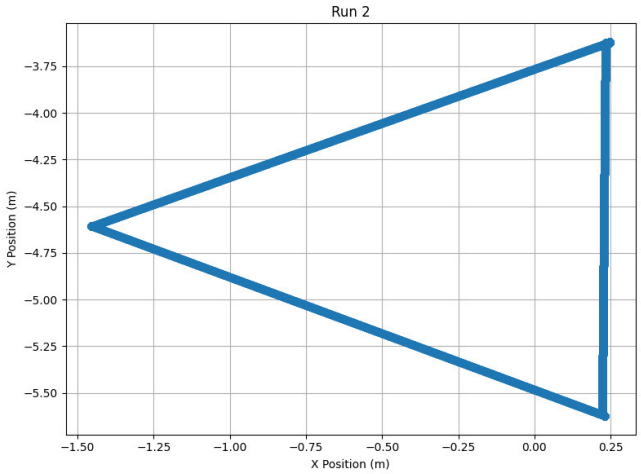

Fig. 7. Example of prompt strategy (Hybrid) run

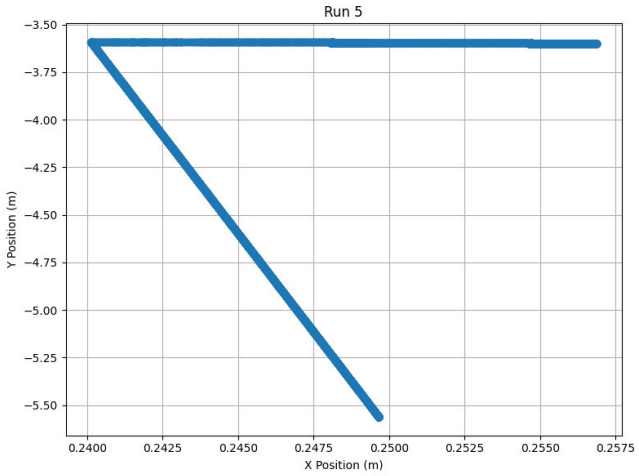

Fig. 6. Example of prompt strategy (Few-Shot) run

A quantitative comparison assessed each strategy's performance over five repetitions of a trajectory planning task. Table I summarizes the average task success rate and angular deviation across strategies.

TABLE I
PROMPT STRATEGY EVALUATION

Strategy	Success Rate
Direct Prompt	5/5
Chain-of-Thought	5/5
Few-Shot	1/5
Hybrid (CoT + FS)	3/5

C. Object Detection via VLM

The robot captured field images and passed them to GPT-4 with prompts like *"Describe what's in this image"*. The VLM consistently identified crops, tools, and obstacles like bottles or weeds.

D. Visual Field Reporting

In extended prompts (e.g., *"Generate a report of what you see in this field"*), the VLM produced coherent natural-language summaries highlighting plant health, potential obstructions, and environmental conditions. These reports were structured and human-readable, supporting autonomous decisions and remote operator review.

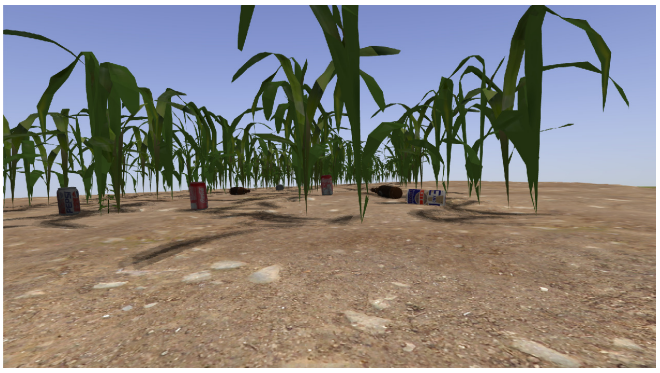

Fig. 8. Example of robot pov for GPT-4-generated field report

V. DISCUSSION

A. LLMs as Planners, Not Controllers

The findings validate the role of LLMs as high-level planners capable of translating abstract natural language instructions into executable robot behaviors. However, while LLMs can produce coherent and logically sound plans, their real-time execution fidelity is limited by hardware-level dynamics and environmental variance. As shown in the square-pattern task, deviations from expected paths were frequent in

open-loop mode, highlighting the importance of integrating traditional low-level control mechanisms like PID regulators. This reinforces the necessity of hybrid architectures, where symbolic reasoning from LLMs is grounded by deterministic feedback control.

B. Prompt Engineering Trade-offs

The prompt design significantly influenced output quality, with hybrid prompting (Few-shot + CoT) achieving the best balance of reliability and generalization. Direct prompts were quick to generate but tended to fail under ambiguity or complex task structures. Chain-of-thought prompting improved transparency by encouraging intermediate reasoning, sometimes resulting in verbose or over-engineered plans. Few-shot prompting offered stability by anchoring the model's output style with curated examples, but in practice, it did not generalize well to tasks requiring geometric adaptation. Hybrid prompting combined examples with reasoning, improving robustness in some cases but introducing inconsistency in others. This aligns with observations from the thesis, which showed that prompt selection directly affects the syntactic structure, interpretability, and trajectory adherence, especially in angle-sensitive instructions like turning 120° versus 90°.

The results of the triangle movement experiment further highlight the impact of system prompt design on LLM-driven control. Despite using the same user prompt ("Move in a triangle pattern"), the system's output and robot behavior varied significantly across prompting strategies.

1) Direct Prompting: Direct prompting achieved excellent performance, with 5 out of 5 successful runs and high consistency. This approach benefited from a system prompt instructing the LLM to generate concise, minimal step-by-step outputs without explicit reasoning. However, direct prompting is highly dependent on a well-phrased initial instruction. If user input is vague or lacks geometric precision, the model lacks mechanisms to infer missing context, potentially reducing robustness.

2) Chain-of-Thought (CoT) Prompting: CoT prompting also yielded strong performance, matching direct prompting with 5 out of 5 successful runs. In this case, the model was guided to reason that a triangle requires three sides of equal length and external angles of 120°. This explicit explanation helped the LLM generalize to the correct geometry.

3) Few-shot Prompting: Few-shot prompting demonstrated poor generalization, with only 1 out of 5 successful executions. Although the model was provided with examples (e.g., moving in a square), it frequently overfitted to these patterns and failed to extrapolate to triangles. Common errors included using 90° turns instead of 120° or stopping prematurely after one or two sides.

4) Hybrid Prompting (Few-shot + CoT): Hybrid prompting, which combines examples with structured reasoning, achieved 3 out of 5 successful runs. This method produced promising results when the examples and reasoning segments were well-aligned. While hybrid prompting offers strong potential, its effectiveness depends on carefully crafted prompt design to avoid interference between modes.

5) Overall Observations: Direct and chain-of-thought prompting emerged as the most reliable methods for producing executable, ROS2-compatible plans in geometric movement tasks. Few-shot prompting alone lacked adaptability, and hybrid prompting, while promising, introduced occasional inconsistencies. These findings underscore that prompting strategy plays a central role in shaping language output and real-world robot behavior.

For robotics applications, prompt clarity, structure, and internal logic are critical to minimize ambiguity and execution failure. Future research should explore combining prompt-based control with parameterized templates, explicit reasoning paths, or constrained decoding to improve interpretability and task repeatability.

C. VLM-Based Perception and Reporting

The VLM component effectively grounded visual input into human-readable outputs, such as object labels and structured reports. Agricultural scenes were typically parsed with high accuracy, though occlusions and low-contrast conditions introduced occasional misclassifications, especially in cluttered environments. This confirms the thesis's insight that VLMs can enhance field awareness but are sensitive to camera placement, field layout, and scene quality. Additionally, the ability to produce spoken reports supports explainability, which is crucial for human trust in robot decision-making.

The object detection experiments revealed that GPT-4-based VLMs consistently identified foreign objects such as bottles, soda cans, and weeds, and provided type-correct descriptions. Crucially, when no objects were present, the model did not hallucinate, correctly reporting empty scenes. This ability to maintain grounded, reality-consistent outputs suggests strong baseline reliability under normal field conditions. However, spatial localization, particularly left/right/center descriptions, showed inconsistencies, with subjective or frame-dependent language used to describe object position. More structured prompting (e.g., referencing rows or distance bands) could improve spatial clarity.

Contextual understanding was also demonstrated: the model inferred partial occlusion when overlapping objects were present and improved classification when similar items appeared at varying distances. For example, in one run, a far object was generically labeled as "debris," while a closer object in a similar class was correctly described as a "glass

bottle." This reflects a degree of contextual refinement, where object interpretation improves with better visual cues.

Field reporting experiments further validated the model's ability to assess environmental risks and suggest mitigation strategies. Detected objects were categorized by potential hazard (e.g., "the bottle might shatter and harm equipment"), with risk ratings inferred from visible features like size and material. In empty field scenarios, the VLM demonstrated conservative behavior, noting small rocks as minor concerns rather than hallucinating threats, showing an ability to scale its judgment based on visual evidence. However, it sometimes underestimated cumulative risks (e.g., multiple soda cans described without reference to quantity), highlighting a limitation in quantitative reasoning.

Finally, the model showed early signs of predictive reasoning: in occluded scenes, it inferred the likely presence of a second object based on partial shape overlap. Such capabilities could be valuable for hazard anticipation and proactive avoidance. Nonetheless, challenges remain in depth estimation, localization precision, and interpretability across varying field conditions. Structured prompts, confidence scoring, and hybrid visual reasoning modules could help mitigate these issues for real-world deployments.

D. Human-Robot Interaction Implications

The system enables a shift in human-robot interaction (HRI) toward natural-language-based collaboration. This reduces the cognitive and technical burden on end users, making robotics more accessible for domains like agriculture, where operators are often domain experts but not programmers. This positions conversational robotics as a tool for automation and augmenting field intelligence.

E. Limitations and Future Directions

While the results obtained in the simulation were promising, several limitations remain that must be addressed to enable real-world deployment:

- **Latency**: API-based model queries, especially those involving VLMs, introduced non-deterministic delays, which hinder real-time performance.
- **Robustness**: LLM behavior became less predictable during long or complex task sequences. Inconsistent internet connectivity in field environments further reduces system reliability.
- **Scalability**: The current modular architecture supports isolated tasks but lacks mechanisms for multi-step workflows, memory across sessions, and coordination between multiple agents.

To address these limitations, future work should explore deploying LLM and VLM inference directly on edge devices to reduce latency and improve autonomy. Visual capabilities could also include crop growth monitoring, disease detection, and environmental stress assessment. Additionally, integrating adaptive feedback loops, where the robot asks for clarification when uncertain, could significantly enhance task reliability and user trust in ambiguous situations.

VI. Conclusion

This work demonstrates a modular system integrating LLMs and VLMs to enable intuitive, explainable robot control for agricultural tasks. By translating natural language instructions into executable ROS2 actions and combining this with visual perception and reporting, the system allows non-expert users to interact with robots in accessible ways. Experimental results show that LLMs can generate high-level plans reliably when supported by classical control and that VLMs can effectively interpret agricultural scenes to produce structured field reports. This approach reduces the barrier to robotics adoption in farming and opens new opportunities for human-robot collaboration in semi-structured environments.

Future work will improve real-time robustness, deploy models locally for field use, and extend visual understanding to support crop-specific tasks such as growth analysis and anomaly detection.

Acknowledgment

This work is a part of DigiFoods SFI funded by the research council of Norway under the agreement 309259.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. arXiv:2005.14165 [cs].

[2] Magnus Skatvedt Iversen. Norges Bondelag vil gjøre det lettere å få tak i sesongarbeidere, June 2024. Section: dk.

[3] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From CLIP to DINO: Visual Encoders Shout in Multi-modal Large Language Models, March 2024. arXiv:2310.08825 [cs].

[4] OECD. *Policies for the Future of Farming and Food in Norway*. OECD Agriculture and Food Policy Reviews. OECD, March 2021.

[5] David Christian Rose and Jason Chilvers. Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming. *Frontiers in Sustainable Food Systems*, 2, December 2018. Publisher: Frontiers.

[6] Michael Ryan. Labour and skills shortages in the agro-food sector. *OECD*, January 2023.

[7] Bruno Siciliano and Oussama Khatib, editors. *Springer Handbook of Robotics*. Springer Handbooks. Springer International Publishing, Cham, 2016.

[8] Minghe Wang, Alexandra Kapp, Trever Schirmer, Tobias Pfandzelter, and David Bermbach. Exploring Influence Factors on LLM Suitability for No-Code Development of End User IoT Applications, May 2025. arXiv:2505.04710 [cs].