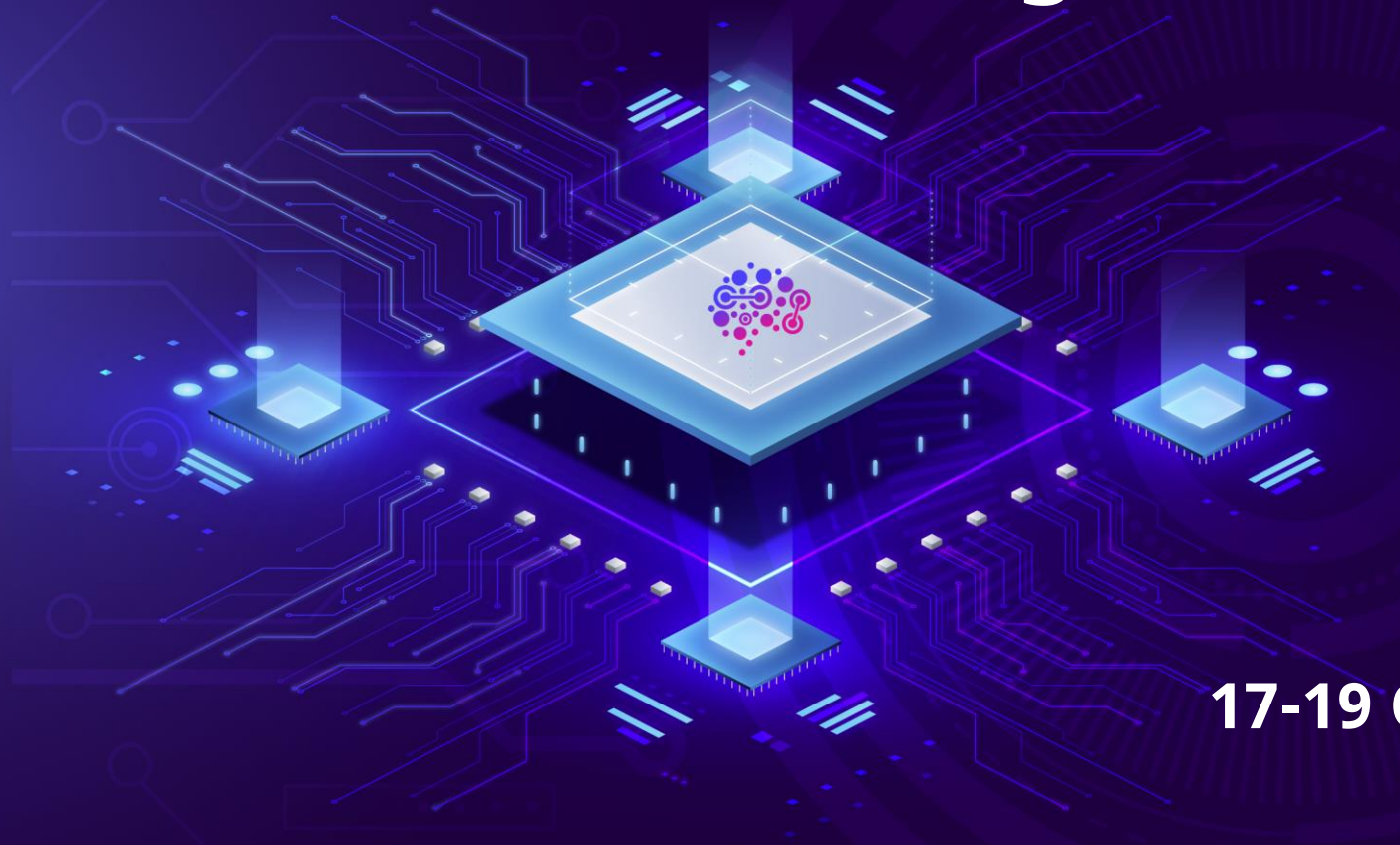
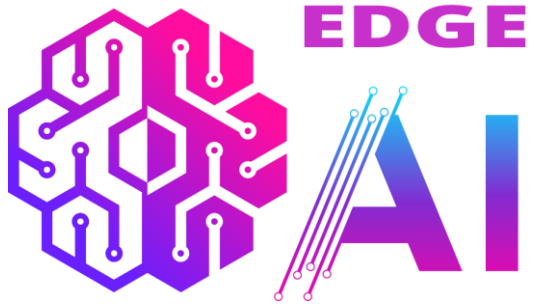
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



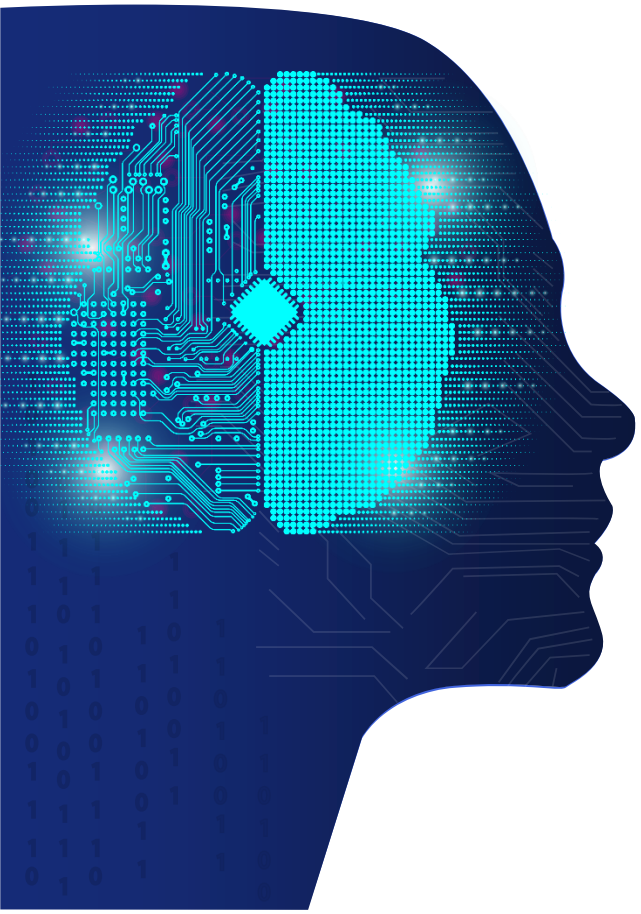
Natural Language Conditioned Planning of Complex Robotics Tasks

Toms Eduards Zinars, Oskars Vismanis, Peteris Racinskis, **Janis Arents**, and Modris Greitans, **EDI, Latvia**

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline

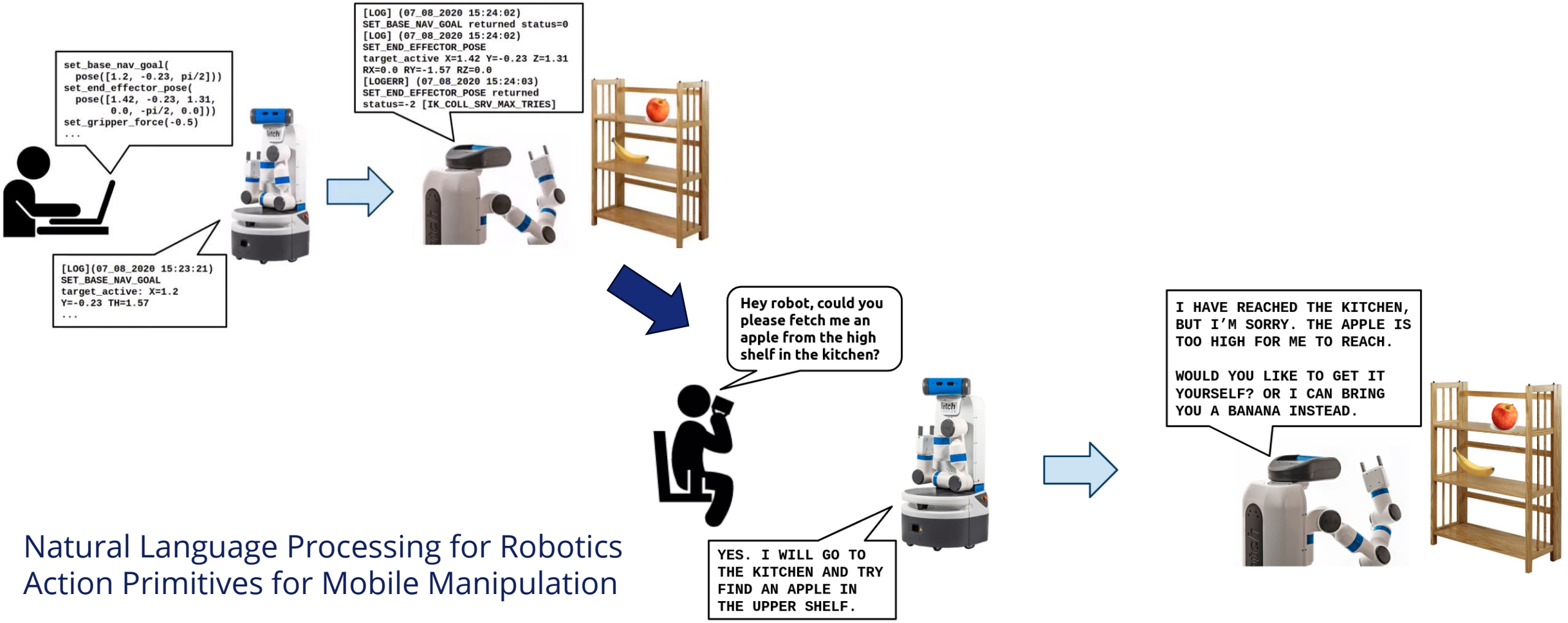


- Introduction
- Natural Language Processing for Robotics
- Action Primitives for Mobile Manipulation
- Identified Challenges
- Conceptual Architecture
- Conclusions and Outlook

Copyright © 2023

Introduction

- Natural Language Conditioned Planning Of Complex Robotics Tasks



- Natural Language Processing for Robotics
- Action Primitives for Mobile Manipulation

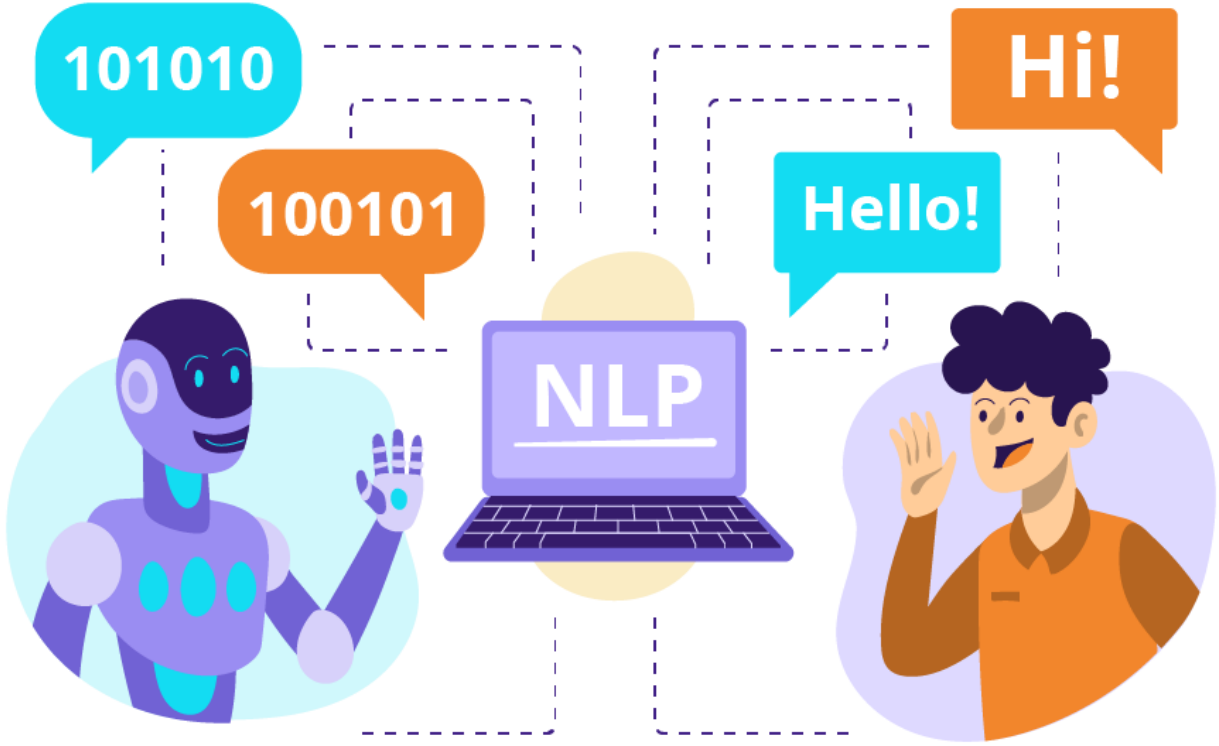
* https://roboticseabass.com/2020/07/28/introduction-to-natural-language-processing/

Copyright © 2023

Natural Language Processing for Robotics

- Natural language is vast and vague
  - NL command can consist of a request that only makes sense in the context of the situation and the environment
  - Robotic system must be able to ground this information in the environment it finds itself and act upon through a concise and safe plan of actions

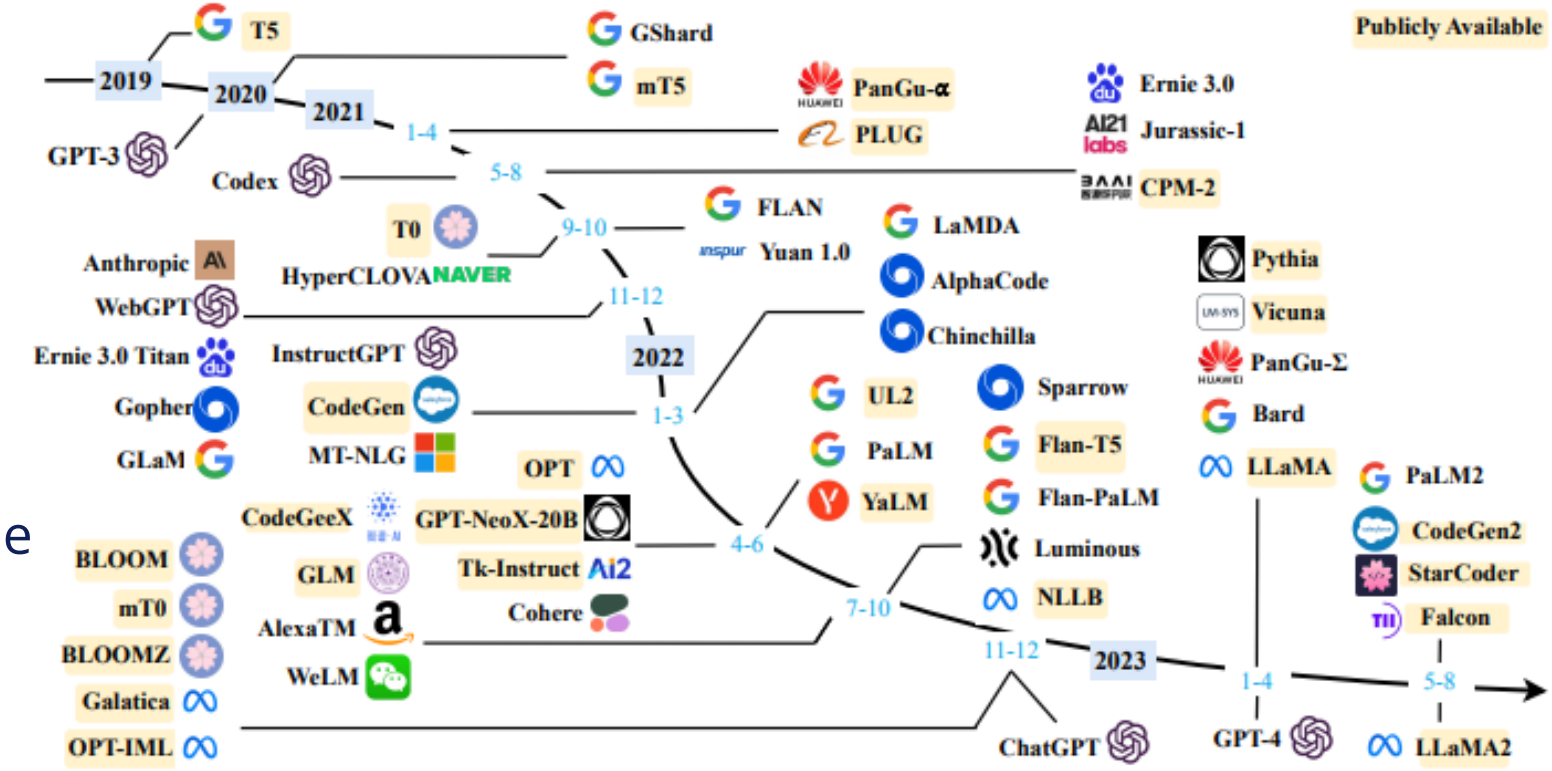
- Large language models (LLM)
- Multi-modal embeddings
- Recent implementations and uses in planning for robotic systems



Copyright © 2023

Natural Language Processing for Robotics
- Large Language Models

- Powerful natural language interpreters
- Mostly based on transformer architecture with high parameter count >10^9
- Trained on vast and diverse datasets of text from the internet (books, articles, social media, forums etc.)
- Broad knowledge base that can be leveraged to create a human-machine interface
- Capable of decoding obscure natural language requests into actions understandable to a robotic system.



*Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

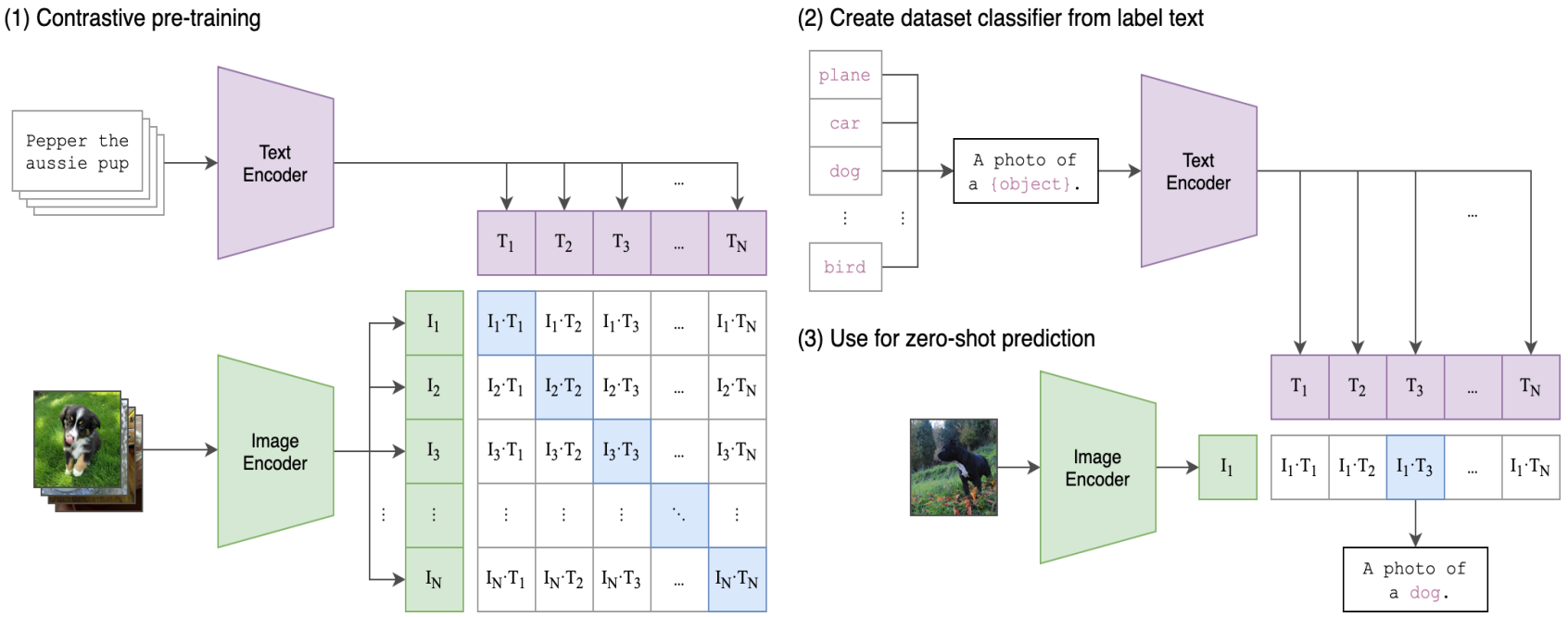
Copyright © 2023

Natural Language Processing for Robotics
- Multi-modal embeddings

The abstract vector nature of the tokens being processed by transformers, exploited to create mappings between radically different data.

Vision-Language Models:

- CLIP (Contrastive Language-Image Pre-training):
  - Jointly trains an image classifier and text encoder on image-caption pairs.
  - Generates vectors in a shared latent space.
  - Enables mapping dissimilar image and text inputs to a common latent "concept" space.



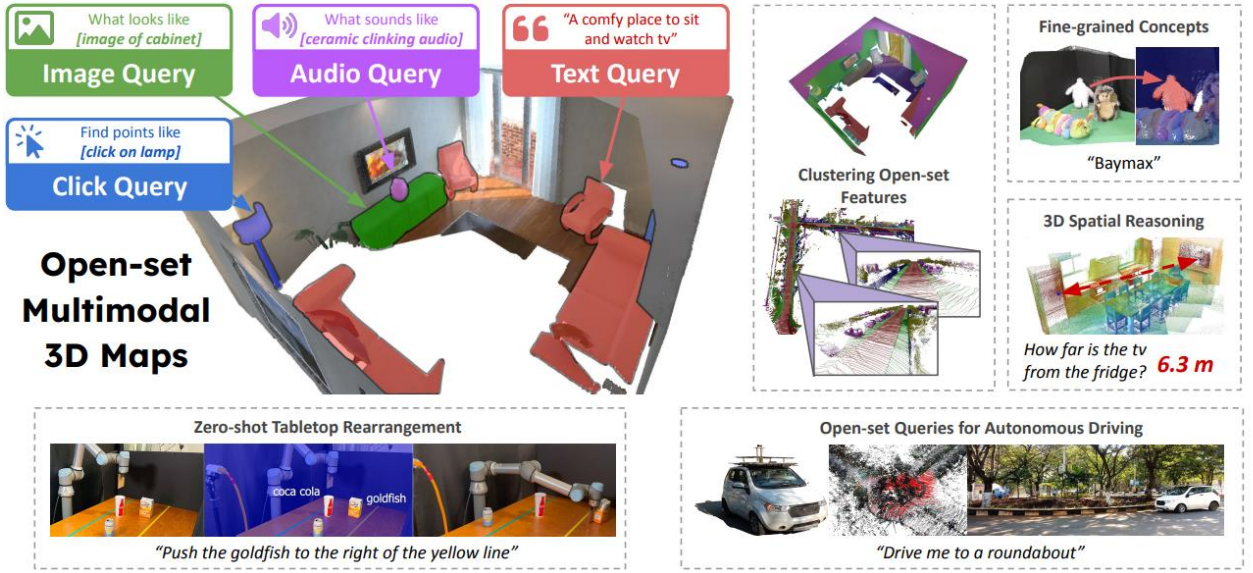
Radford, Alec, et al. “Learning Transferable Visual Models From Natural Language Supervision.” International Conference on Machine Learning (2021).

Copyright © 2023

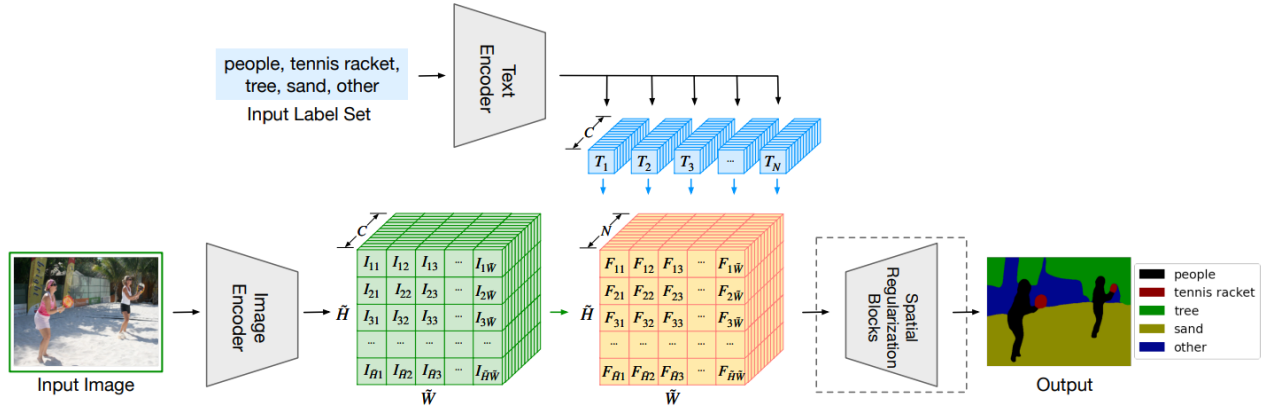
Natural Language Processing for Robotics
- Multi-modal embeddings

Vision-Language Models:

- ConceptFusion:
  - Focuses on producing 3D embedding-tagged point clouds.
  - Used for advanced mapping problems, creating complex 3D embeddings for robotic applications.
- LSeg:
  - Focuses on generating segmentation maps and embeddings for each image pixel.
  - Useful for creating maps amenable to natural language queries and navigation goals in robotic applications.



*Jatavallabhula, Krishna Murthy, et al. "Conceptfusion: Open-set multimodal 3d mapping." *arXiv preprint arXiv:2302.07241* (2023).

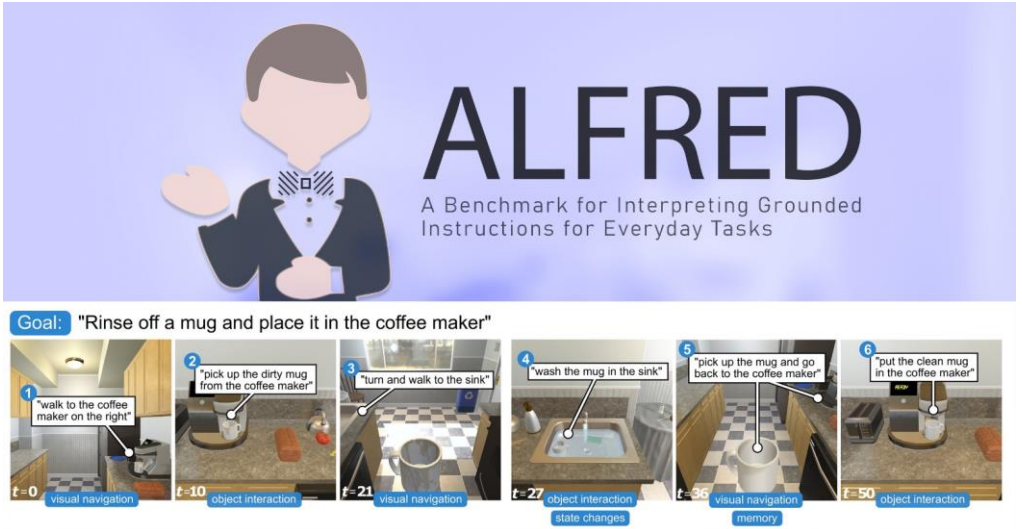


Li, Boyi, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun and Rene Ranftl. "Language-driven Semantic Segmentation." ArXivabs/2201.03546 (2022).

Copyright © 2023

Natural Language Processing for Robotics
- Recent implementation and uses in planning for robotic systems

- Implementation approaches vary, from feature extraction to high-level planning.
- Map integration and proper prompting techniques enhance performance.
  - ✓ ProgPrompt: Generate plans from Python code.
  - ✓ Lang2LTL: Create logical formulas.
  - ✓ Text2Motion: Plan and validate actions.
  - ✓ Language to Rewards: Generate reward functions.
  - ✓ Say-Can: Ground LLM planning in the real world.
  - ✓ Inner Monologue: Feedback-driven task completion.
- ALFRED benchmark
  - ✓ Test agent systems that use both natural language instructions and ego-centric vision.
  - ✓ Encourages the development of robust mobile manipulation systems.
  - ✓ Currently the best preforming systems reaches scores ~50% what human can.

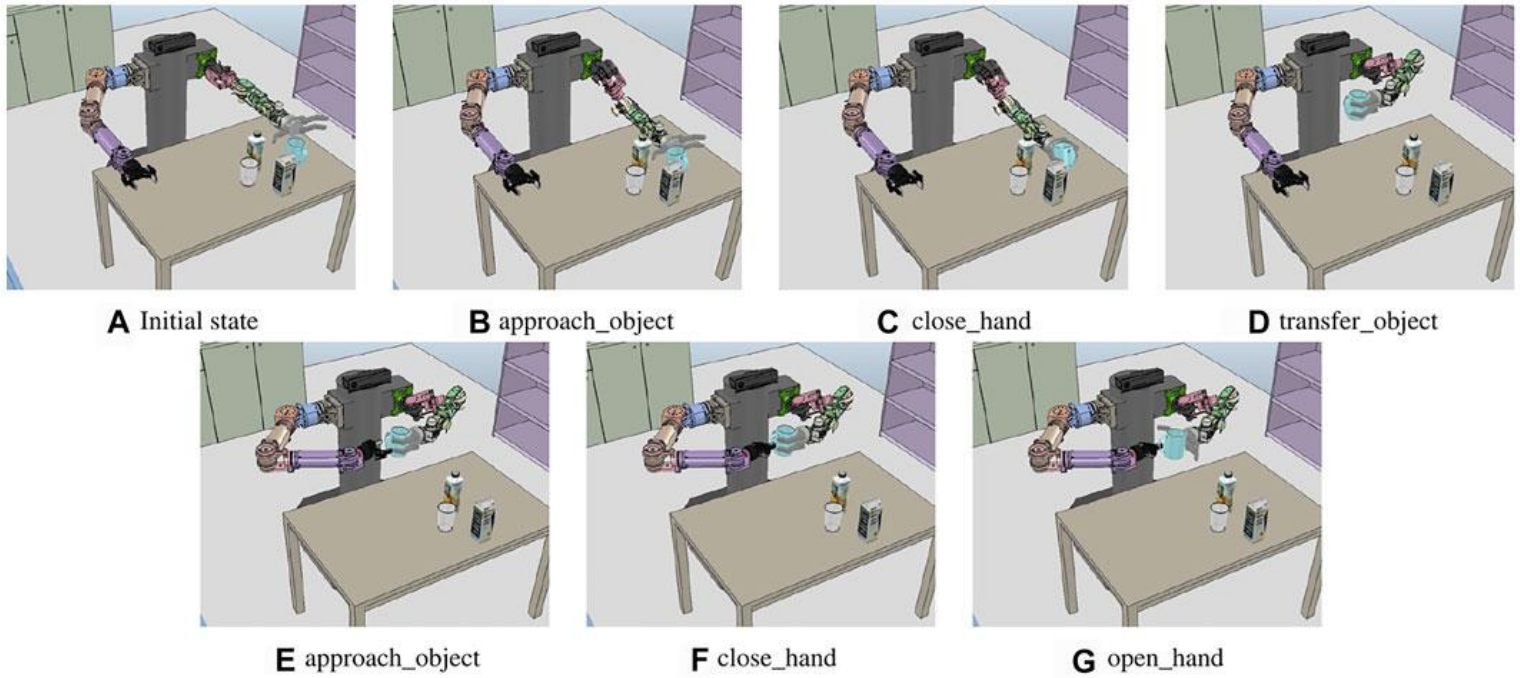


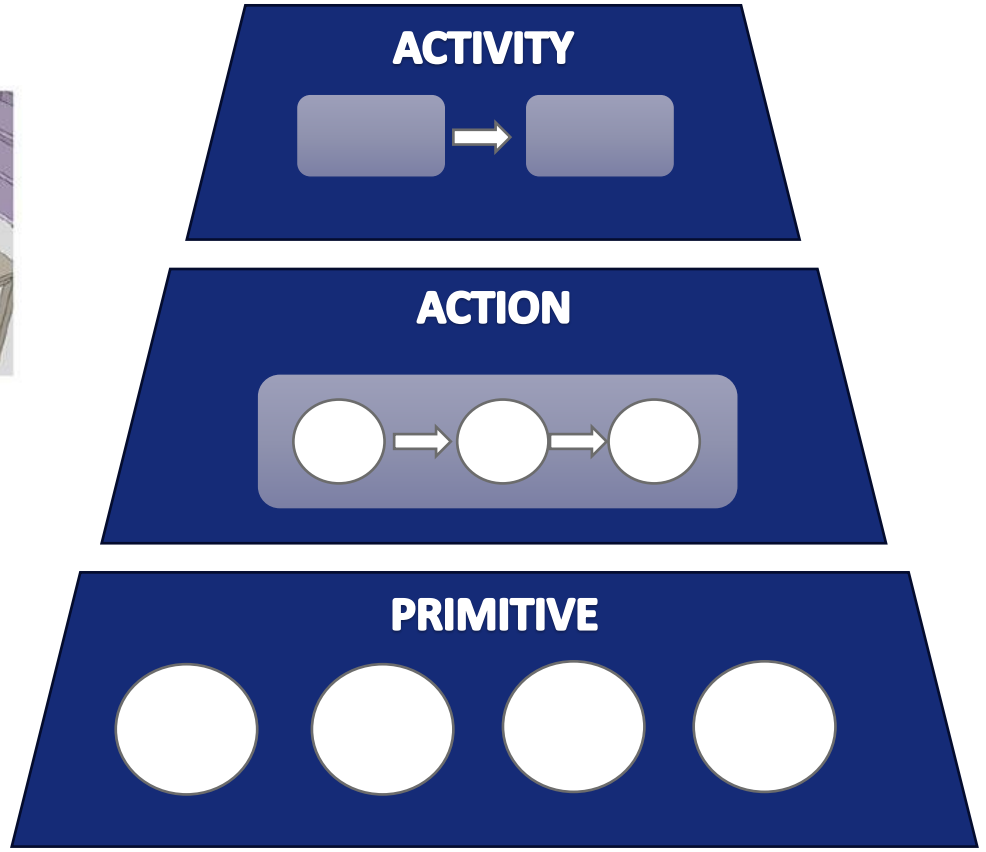
Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han,W., Mottaghi, R., Zettlemoyer, L., & Fox, D. (2019). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10737-10746.

Copyright © 2023

Action Primitives for Mobile Manipulation

- Different approaches to robot control: reactive control, deliberative control, hybrid control, behaviour-based control.
- From the concept of behaviors, a primitive behavior can be derived, which is a more general function made to be reusable in different scenarios, often referred to as the action primitives.





*Jeon, Jeongmin, et al. "Primitive Action Based Combined Task and Motion Planning for the Service Robot." *Frontiers in Robotics and AI* 9 (2022): 713470.

Copyright © 2023

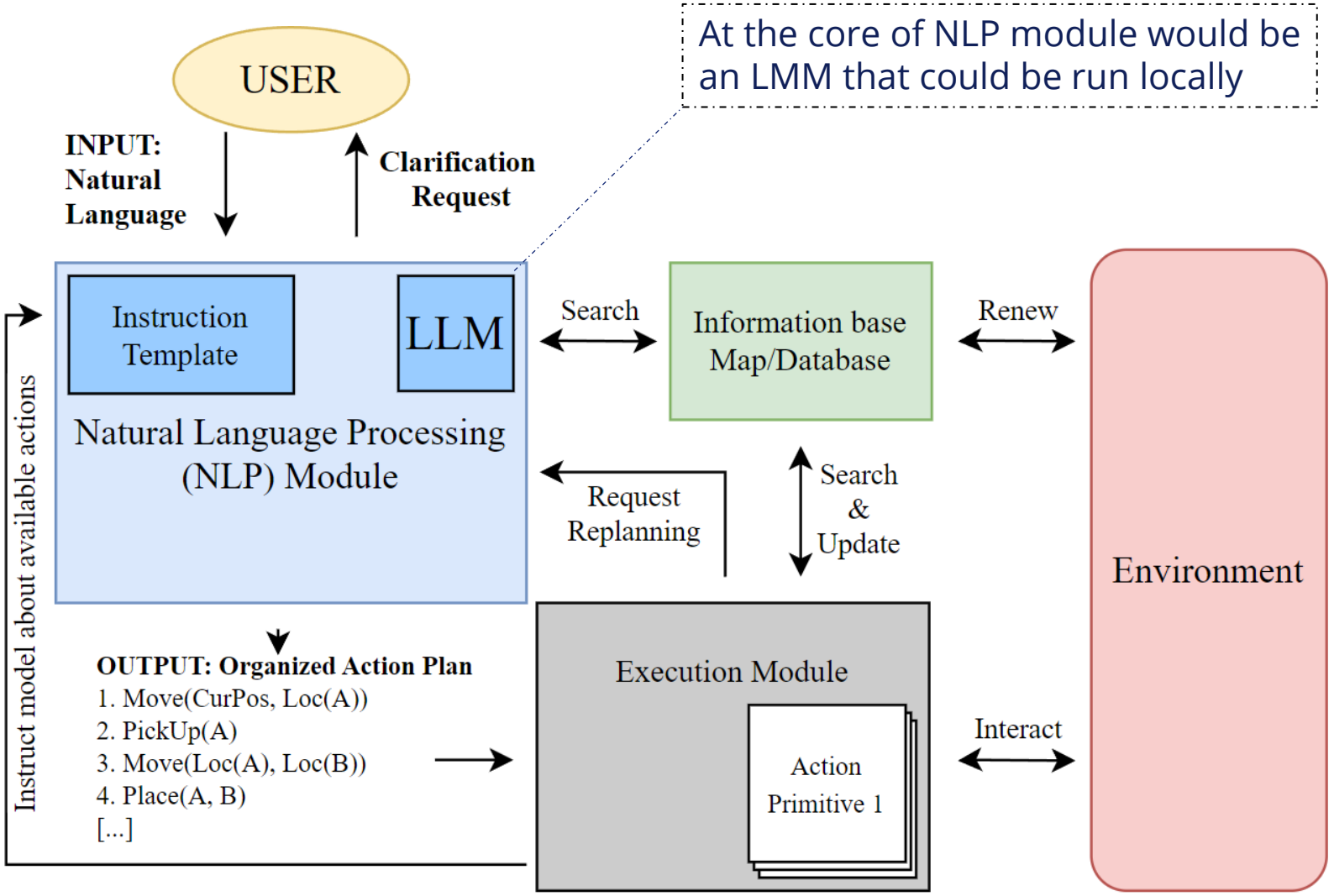
Identified Challenges

- Many implementations rely on using closed-source GPT models, requiring constant connection to cloud services:
  - ➢Lack of practical autonomy for mobile agent system.
- LLMs can contain 100B+ parameters
  - ➢Even with quantization, these larger models would require powerful GPUs that are impractical to be placed onto a physical agent to ensure autonomous operation.
- Range and capabilities of the skill mobile agents possess
  - ➢Even if the NL command can be translated to the system correctly, it cannot execute something it does not know how.
  - ➢Manipulation of wide variety of objects in dynamic environments.

Copyright © 2023

Conceptual Architecture

- Two main modules:
  - NLP module
  - Execution module
- NLP module's task is to interpret the user's input request in a way that the rest of the system may act upon it.
- NLP modules Interface with map, grounds the LLM in its current environment and permits NLP module access to task-relevant information.
- Execution module consists of an action library – action primitives that agent can execute to accomplish tasks.



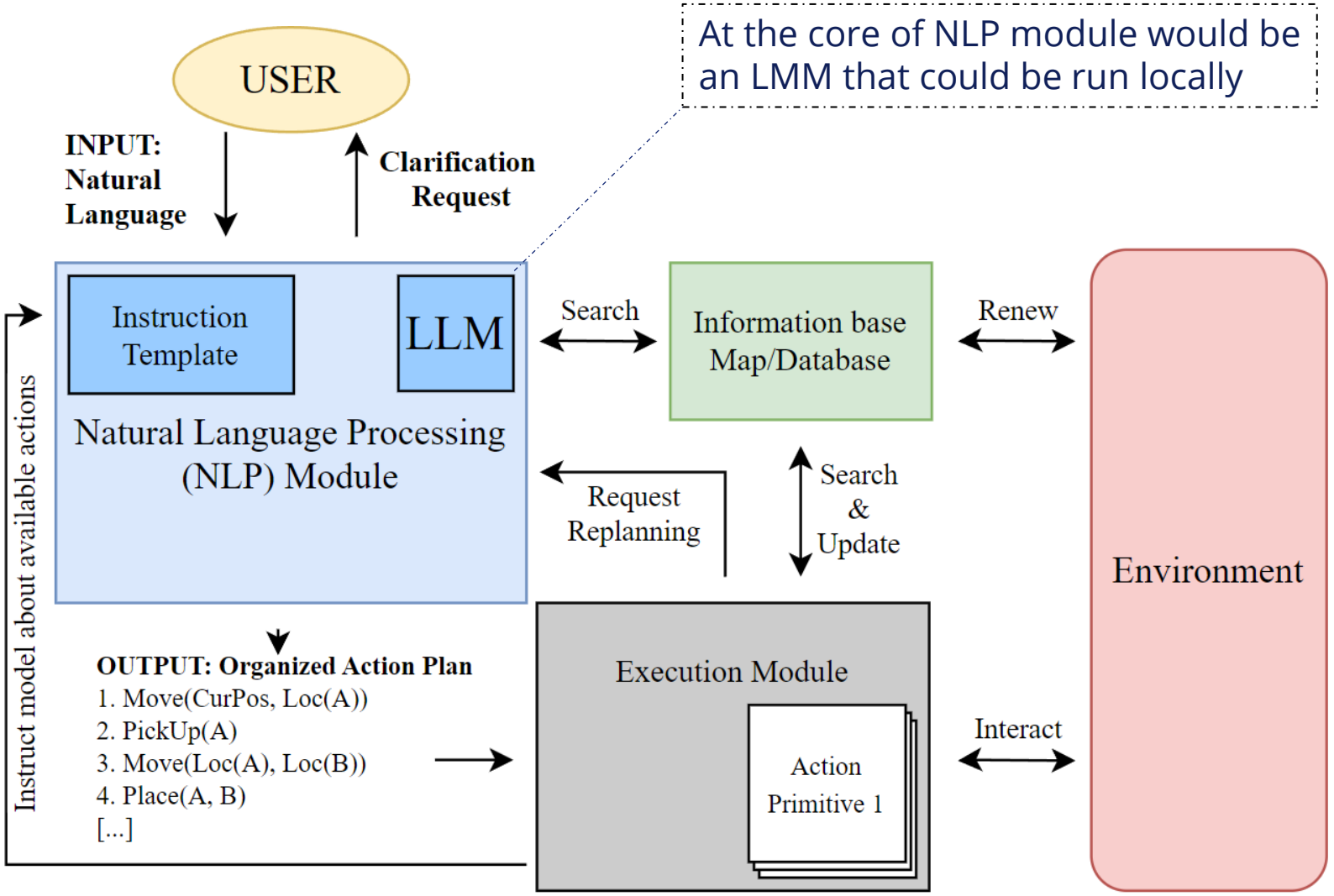
Copyright © 2023

Conceptual Architecture

Main benefits:

- Information about the environment is contained within the semantic map, while the LLM possesses general linguistic knowledge.
- The system is not hard-coded to a particular environment

By relying on a smaller and locally run LLM instance, the hope is to ensure the ability of the system to operate successfully in an edge AI use case.



Copyright © 2023

Conclusions

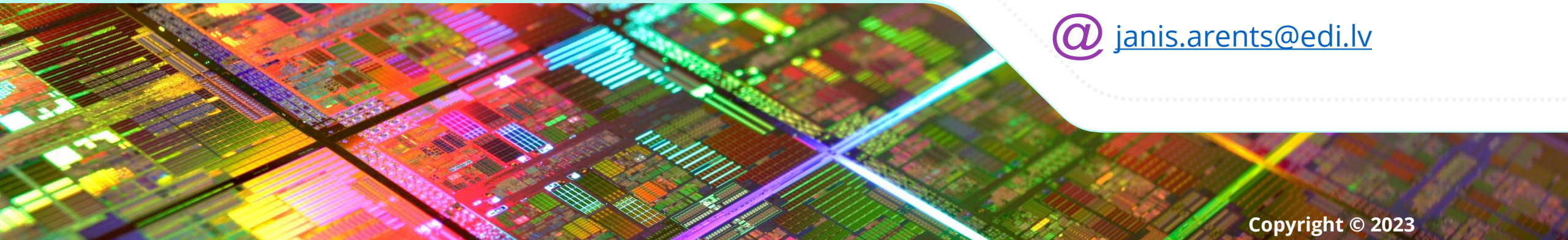
- The rise of LLMs has made applying general language understanding to computer systems seem deceptively trivial, but there is still much ambiguity to overcome.
- There is a need to bring these high-level abilities down to smaller and smaller model sizes to enable true edge AI applications.
- When working in complex environments, action primitives can be used as a powerful tool to generalize actions available to a robotic system.

Copyright © 2023

EDGE AI

Thank You

For your attention



@ janis.arents@edi.lv

Copyright © 2023