From An LLM Swarm To A PDDL-Empowered HIVE: Planning Self-Executed Instructions In A Multi-Modal Jungle

Kaustubh Vyas, Damien Graux, Yijun Yang, Sébastien Montella, Chenxin Diao
Wendi Zhou, Pavlos Vougiouklis, Ruofei Lai, Yang Ren, Keshuang Li, Jeff Z. Pan

Huawei Technologies Ltd., UK University of Edinburgh, UK

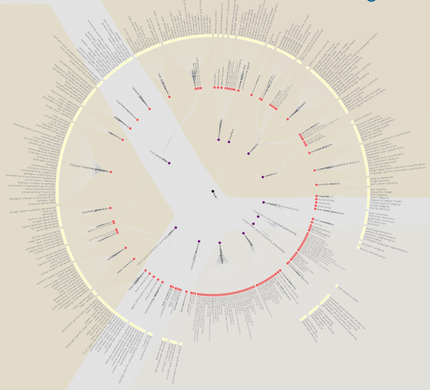
Context & Challenges

Making use of the plethora of available models!

Challenges – Existing approaches lack:

1. Formal Representation in Profiles for Capability Acquisition
2. Precise and Interpretable Planning and Memory Alignment

Taxonomy of NLP Tasks



Comprehensive taxonomy to control the domain classifier

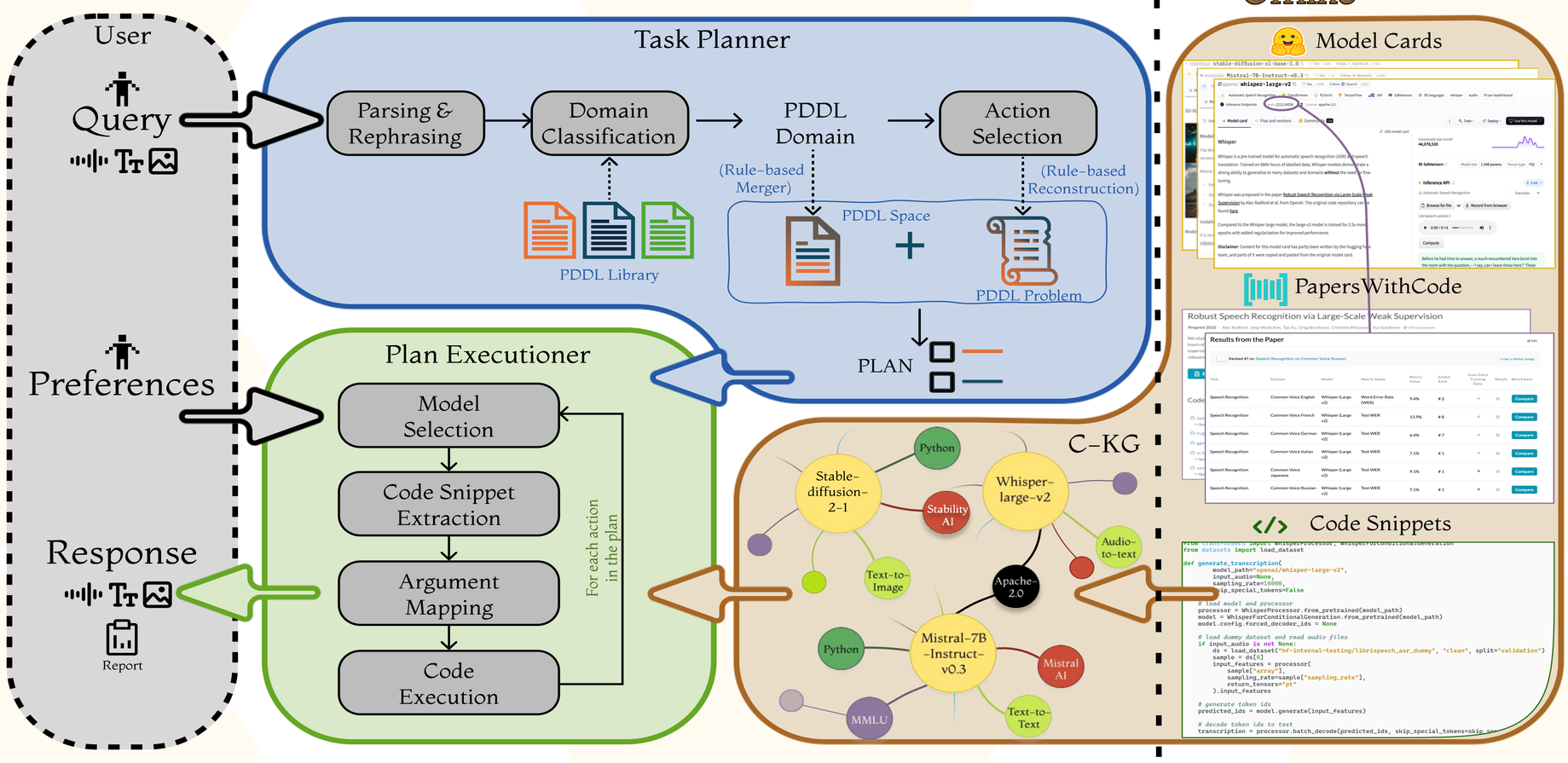
- 420 concepts
- 3 hierarchical levels

"Generate a summary for this audio track file" → PDDL Plan

```
(define (domain audio)
  (:requirements :strips :typing)
  (:types text audio)
  (:predicates
    (IsAudio ?input_audio - audio)
    (GenerateAudioFromText ?input_text - text)
    (IsText ?input_text - text)
    (SpeechRecognition ?input_audio - audio)
  )
  (:action text_to_audio ; given a piece of text,
                         convert it to audio
    :parameters (?input_text - text)
    :precondition (IsText ?input_text)
    :effect (GenerateAudioFromText ?input_text)
  )
  (:action audio_to_text ; given an audio,
                         transcribe it to text
    :parameters (?input_audio - audio)
    :precondition (IsAudio ?input_audio)
    :effect (SpeechRecognition ?input_audio)
  )
)
```

```
(define (domain summarisation)
  (:requirements :typing :strips)
  (:types text)
  (:predicates
    (IsText ?input_text - text)
    (InstructionBasedSummary ?instruction - text ?input_text -
text)
  )
  (:action get_instruction_based_summary ; instruction based
                                           summarisation
    :parameters (?input_text - text ?instruction - text)
    :precondition (and (IsText ?input_text) (IsText ?instruction))
    :effect (InstructionBasedSummary ?instruction ?input_text)
  )
)
```

PLAN: { get_instruction_based_summary (audio_to_text (*file*)) }



Evaluation with the MuSE* benchmark (new)

> Experts from diverse linguistic backgrounds create 100 heterogeneous, real-world queries

> Three types: Single-task, Two-task, and Three-task potentially multi-modal

> Cover various task domains, such as automatic speech recognition, question answering, and image generation, involving **15** models across **10** different modalities

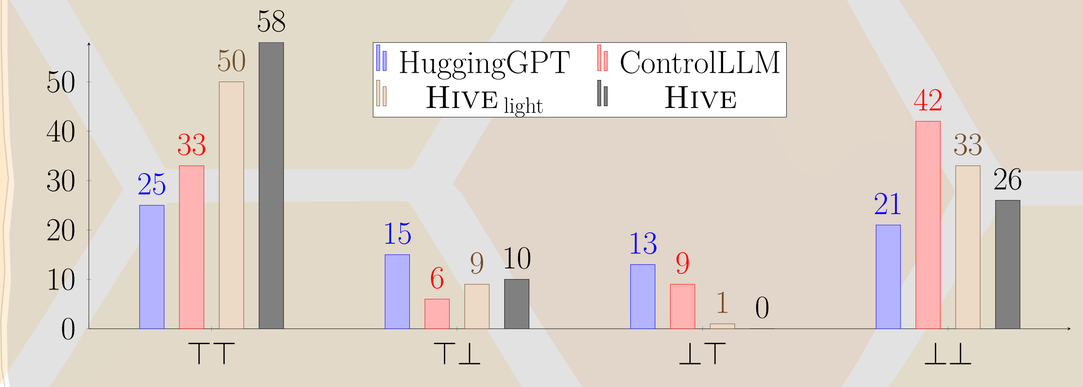
Metrics

- Task Selection (TS): identifying the required tasks from the query
- Flow of Thought (FoT): Evaluating the logical sequence and integration of the tasks
- Final Output (O): Assessing the correctness of the final result

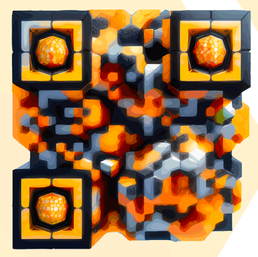
Query Types	HuggingGPT			ControlLLM			Hive light			Hive		
	TS	FoT	O	TS	FoT	O	TS	FoT	O	TS	FoT	O
Single Task	0.47	0.47	0.83	0.74	0.74	0.74	0.80	0.80	0.72	0.88	0.88	0.79
Two Tasks	0.64	0.55	0.44	0.33	0.33	0.38	0.67	0.62	0.52	0.71	0.69	0.58
Three Tasks	0.42	0.42	0.30	0.36	0.36	0.33	0.57	0.43	0.33	0.67	0.67	0.46
Overall	0.57	0.51	0.53	0.43	0.43	0.47	0.69	0.64	0.55	**0.74**	**0.73**	**0.62**

* MUlti-modal Sub-task Execution benchmark

Who can be trusted? *Justifications* (TS AND FoT) Vs *Outputs* (O)



Scan for demo ☞
or access through qrc-ai.com/WAeS8



Cross-Modal Performance

In↓ Out→	Text		
	HuggingGPT	ControlLLM	Hive light
Image	0.48	0.45	**0.52**
Audio	0.25	0.25	**0.67**

In↓ Out→	Image		
	HuggingGPT	ControlLLM	Hive light
Text	0.36	0.18	**0.75**
Audio	0.50	0.25	**1.00**

In↓ Out→	Audio		
	HuggingGPT	ControlLLM	Hive light
Text	0.33	0.14	**0.71**
Image	**0.80**	0.00	0.00