

Exploring the Use of LLMs for Agent Planning Strengths and Weaknesses

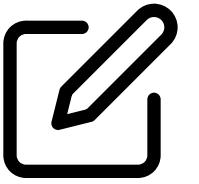
DAVIDE MODOLO
20/03/2025
Supervisor PAOLO GIORGINI



UNIVERSITÀ
DI TRENTO

Context

Large Language Models' Capabilities

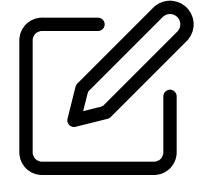


Text Generation
Their Scope

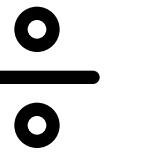
1. Context



Large Language Models' Capabilities



Text Generation
Their Scope



Math Reasoning
Emerging Behavior

1. Context

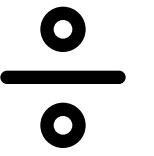


UNIVERSITÀ
DI TRENTO

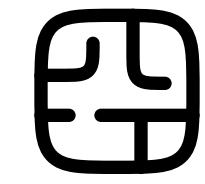
Large Language Models' Capabilities



Text Generation
Their Scope



Math Reasoning
Emerging Behavior



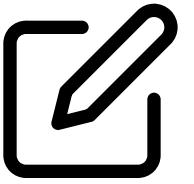
Planning Abilities
Emerging Behavior

1. Context

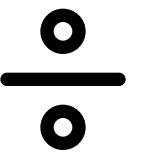


UNIVERSITÀ
DI TRENTO

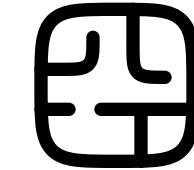
Large Language Models' Capabilities



Text Generation
Their Scope



Math Reasoning
Emerging Behavior



Planning Abilities
Emerging Behavior

...

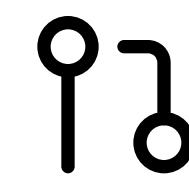
More

1. Context



UNIVERSITÀ
DI TRENTO

LLM-based Planning

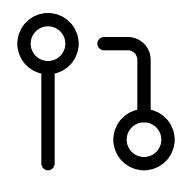


**Chain of
Thought**
Reasoning¹

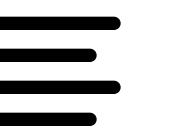
¹ *Chain-of-thought prompting elicits reasoning in large language models* - Wei et al., 2022



LLM-based Planning



**Chain of
Thought**
Reasoning¹

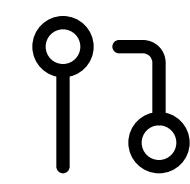


Few-Shots
Prompting²

¹ *Chain-of-thought prompting elicits reasoning in large language models* - Wei et al., 2022

² *PDDL planning with pretrained large language models* - Silver et al., 2022

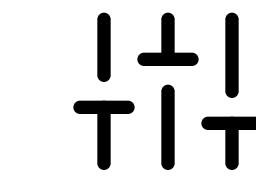
LLM-based Planning



Chain of Thought
Reasoning¹



Few-Shots
Prompting²



Fine-Tuning
Models³

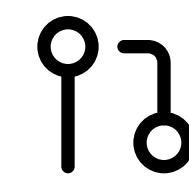
¹ *Chain-of-thought prompting elicits reasoning in large language models* - Wei et al., 2022

² *PDDL planning with pretrained large language models* - Silver et al., 2022

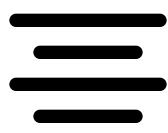
³ *Unlocking Large Language Model's Planning Capabilities with Maximum Diversity Fine-tuning* - Wenjun Li et al., 2024



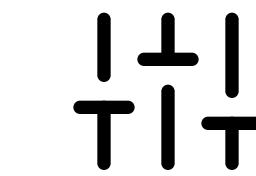
LLM-based Planning



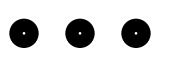
Chain of Thought
Reasoning¹



Few-Shots
Prompting²



Fine-Tuning
Models³



More

¹ *Chain-of-thought prompting elicits reasoning in large language models* - Wei et al., 2022

² *PDDL planning with pretrained large language models* - Silver et al., 2022

³ *Unlocking Large Language Model's Planning Capabilities with Maximum Diversity Fine-tuning* - Wenjun Li et al., 2024

Research Questions



What happens if we strip everything prior away?

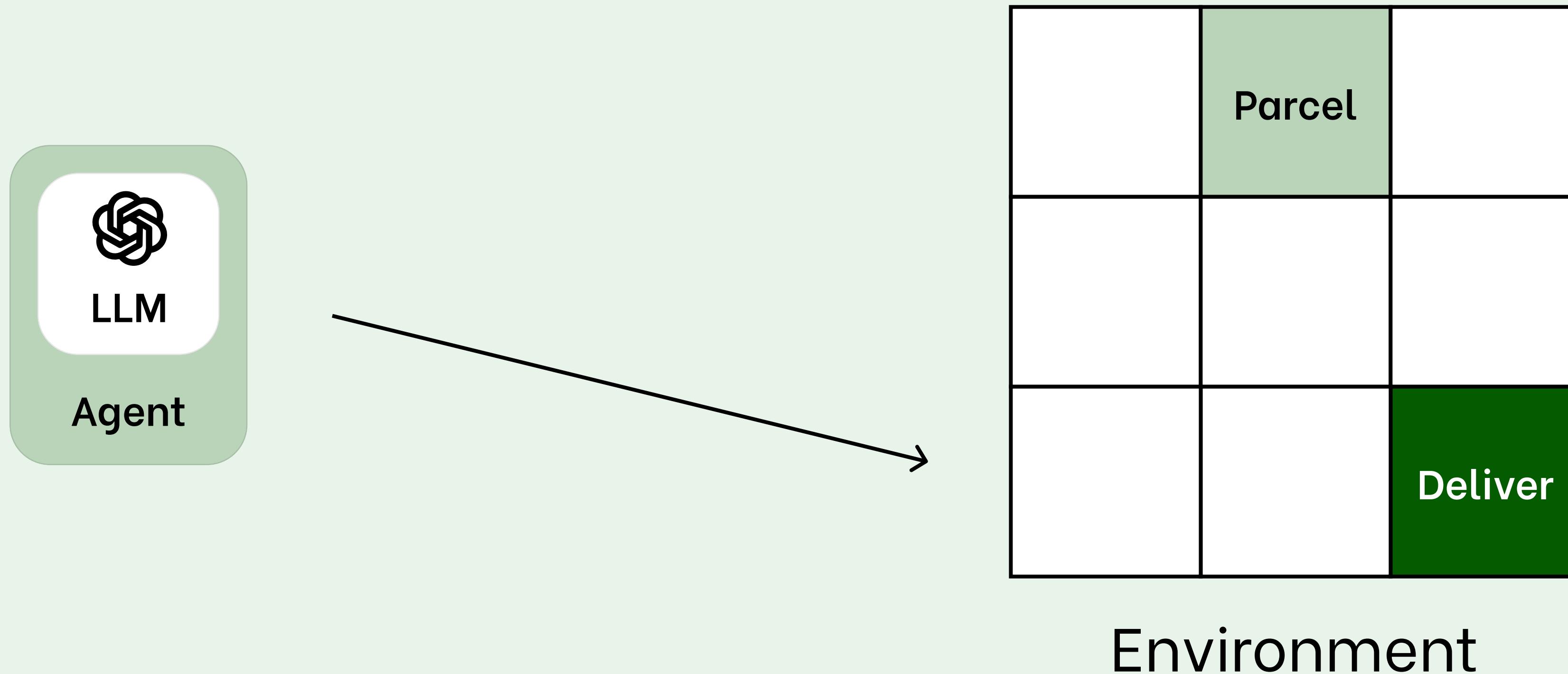


Can an LLM, without additional training or frameworks, effectively plan and navigate in an unknown environment?



How well can LLMs's generative capabilites make sequential decisions in such environments?

Idea



1. Context



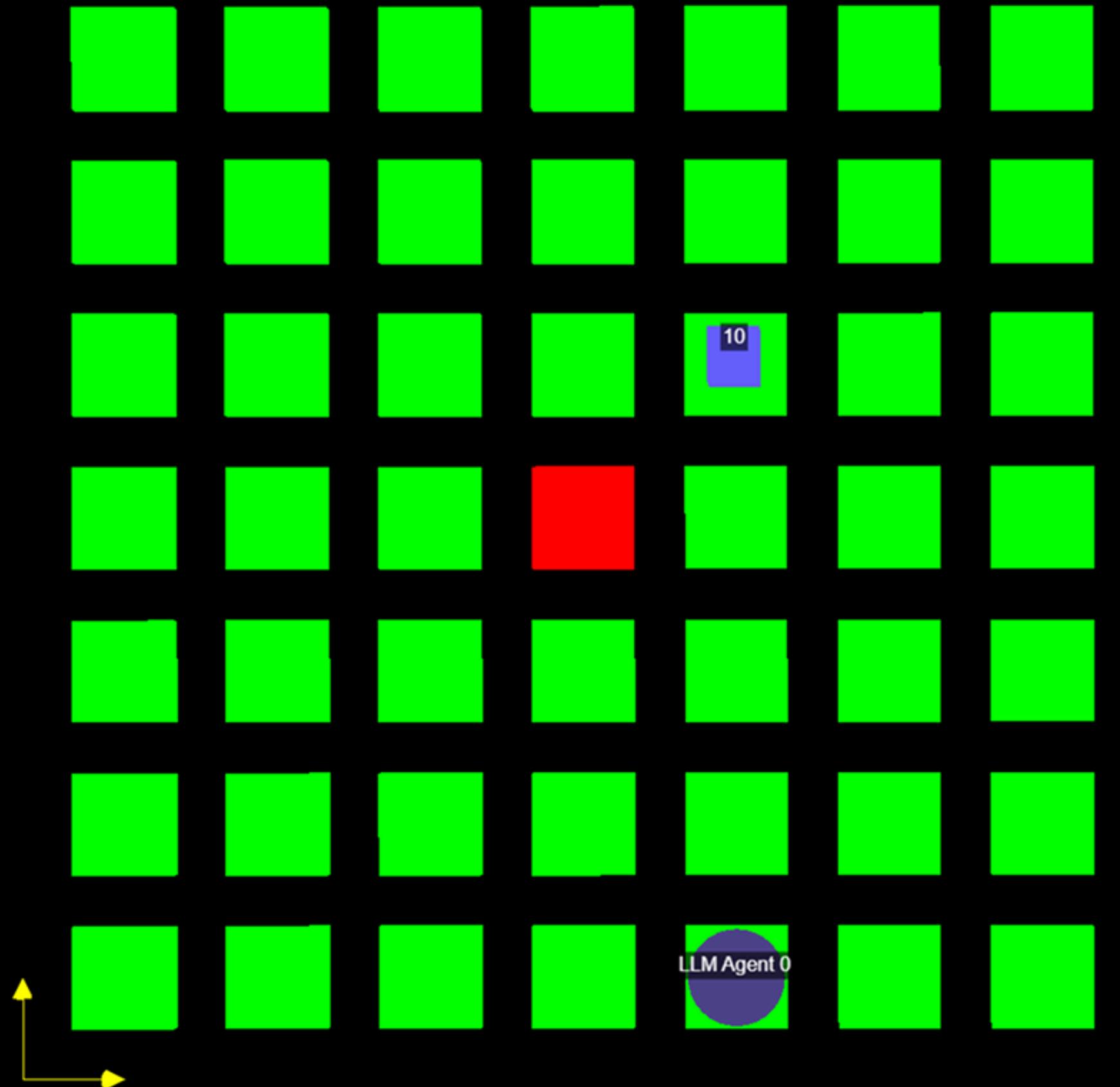
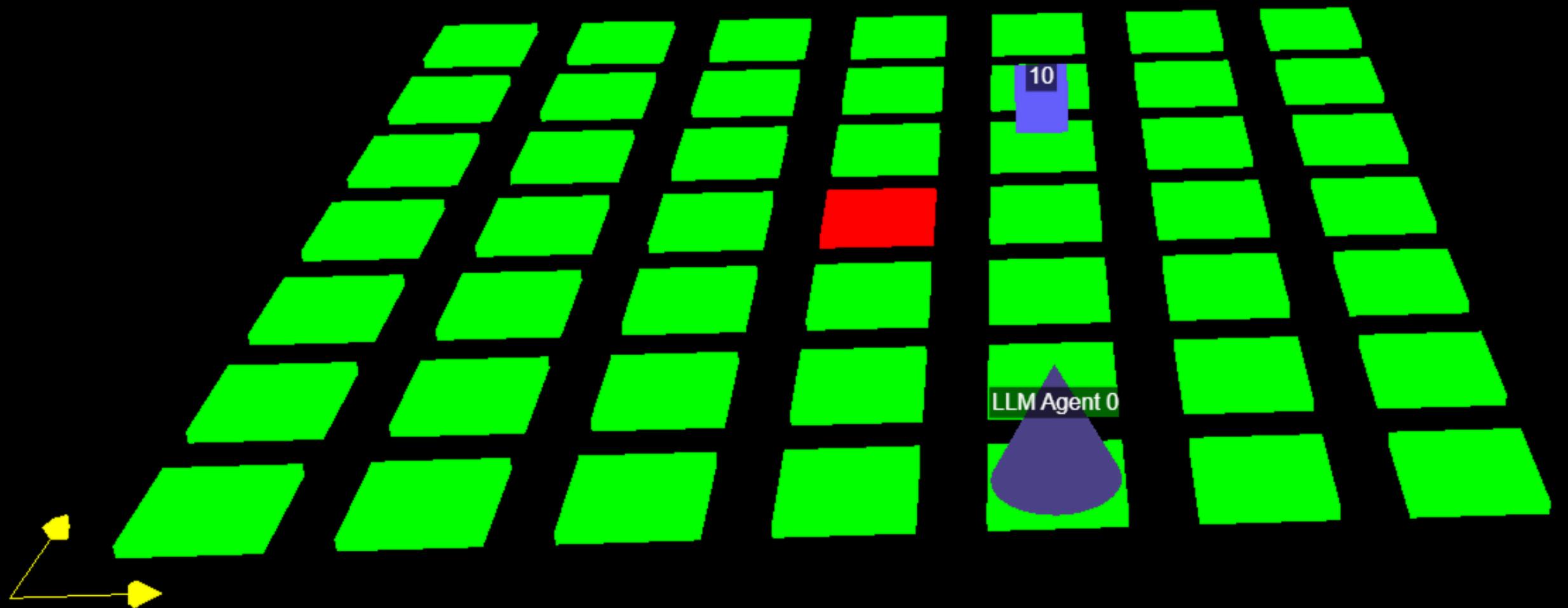
UNIVERSITÀ
DI TRENTO

Setting

Deliveroo.js

Educational Game

Parcels spawn all around the map. The goal is to pickup and deliver them.

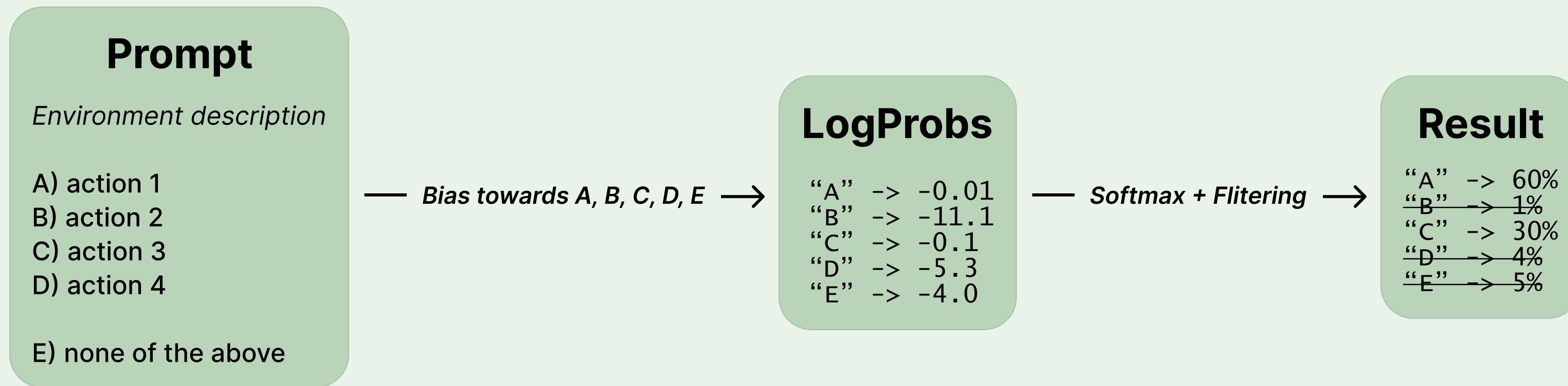


2. Setting



UNIVERSITÀ
DI TRENTO

KnowNo Uncertainty Framework

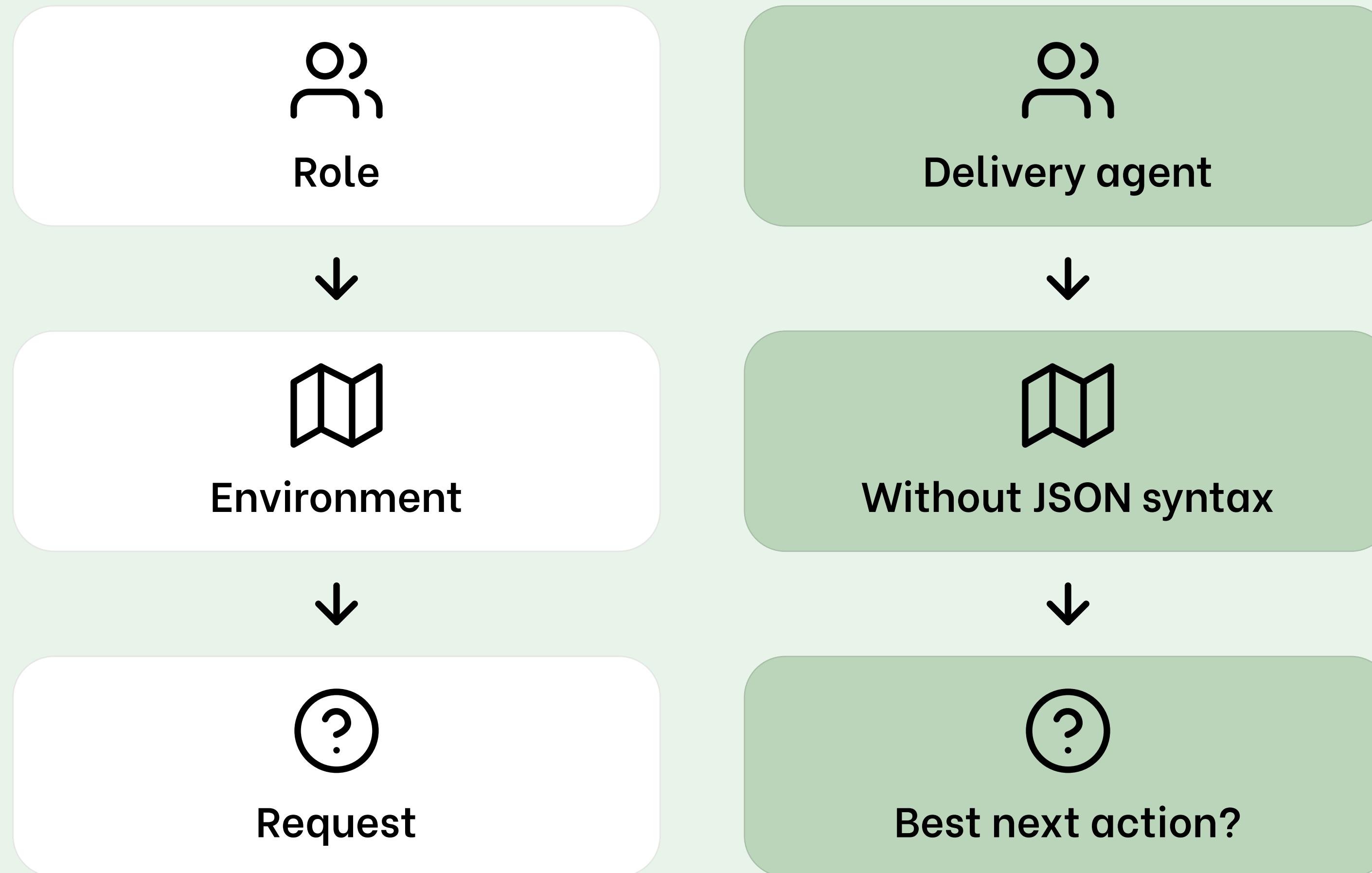


2. Setting



UNIVERSITÀ
DI TRENTO

Prompting Strategy



Model

	GPT-4o	GPT-4o-mini
top1%	77%	84%
top2%	95%	91%
top3%	96%	92%

- OpenAI models
- GPT-4o was the best
- GPT-4o-mini selected for price/performance

2. Setting



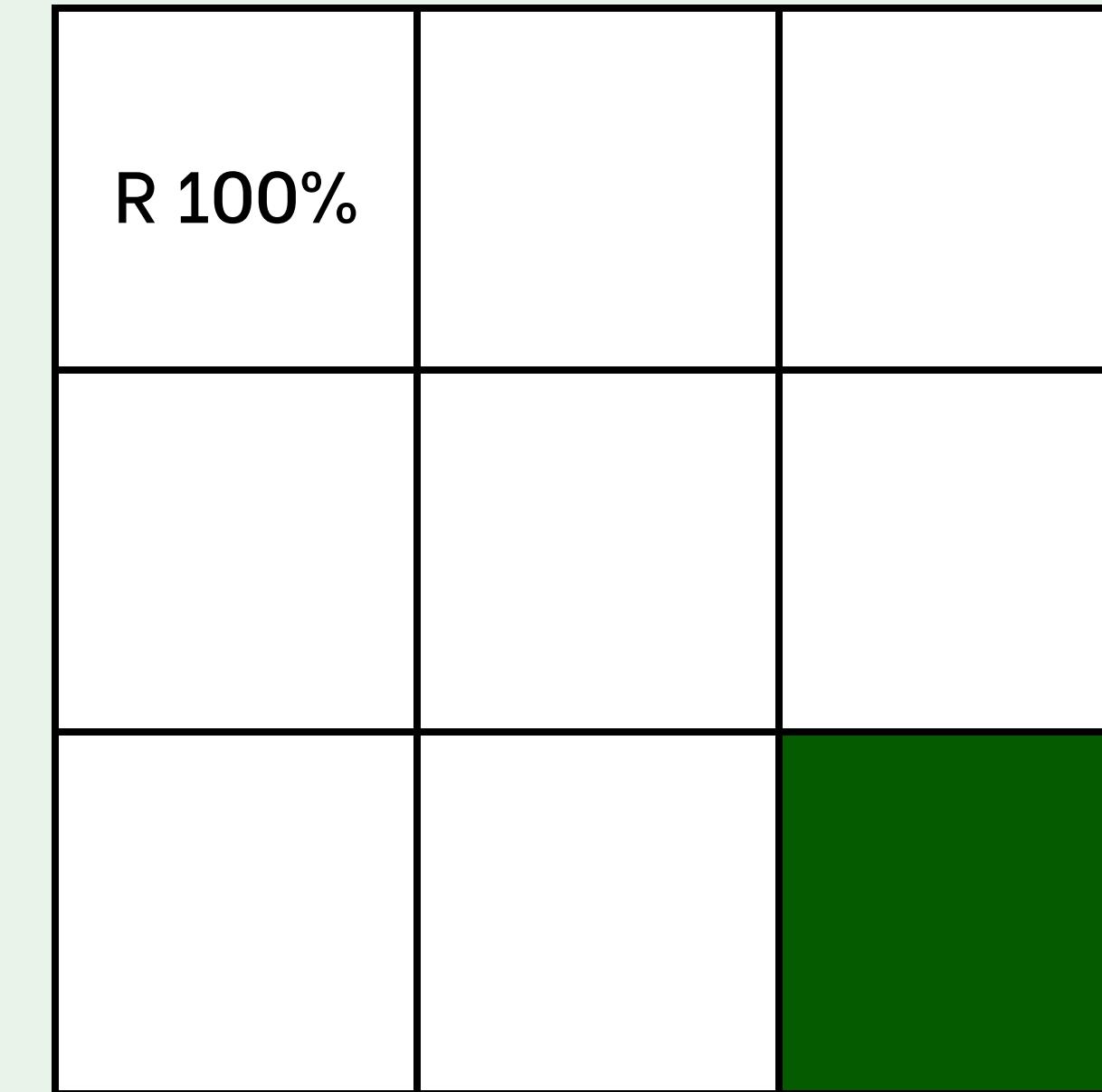
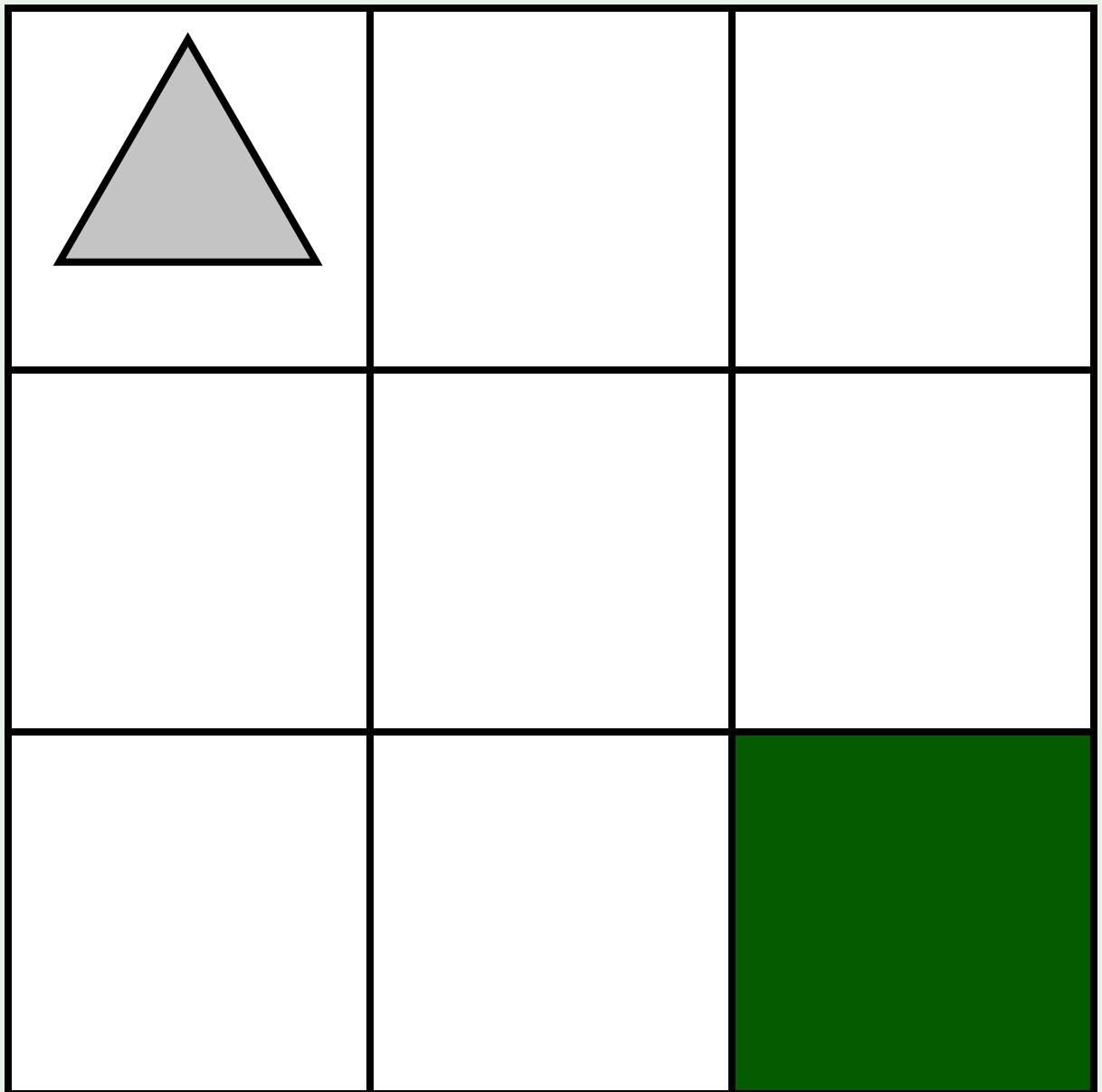
Data Collection

DAVIDE MODOLO
20/03/2025
Supervisor PAOLO GIORGINI

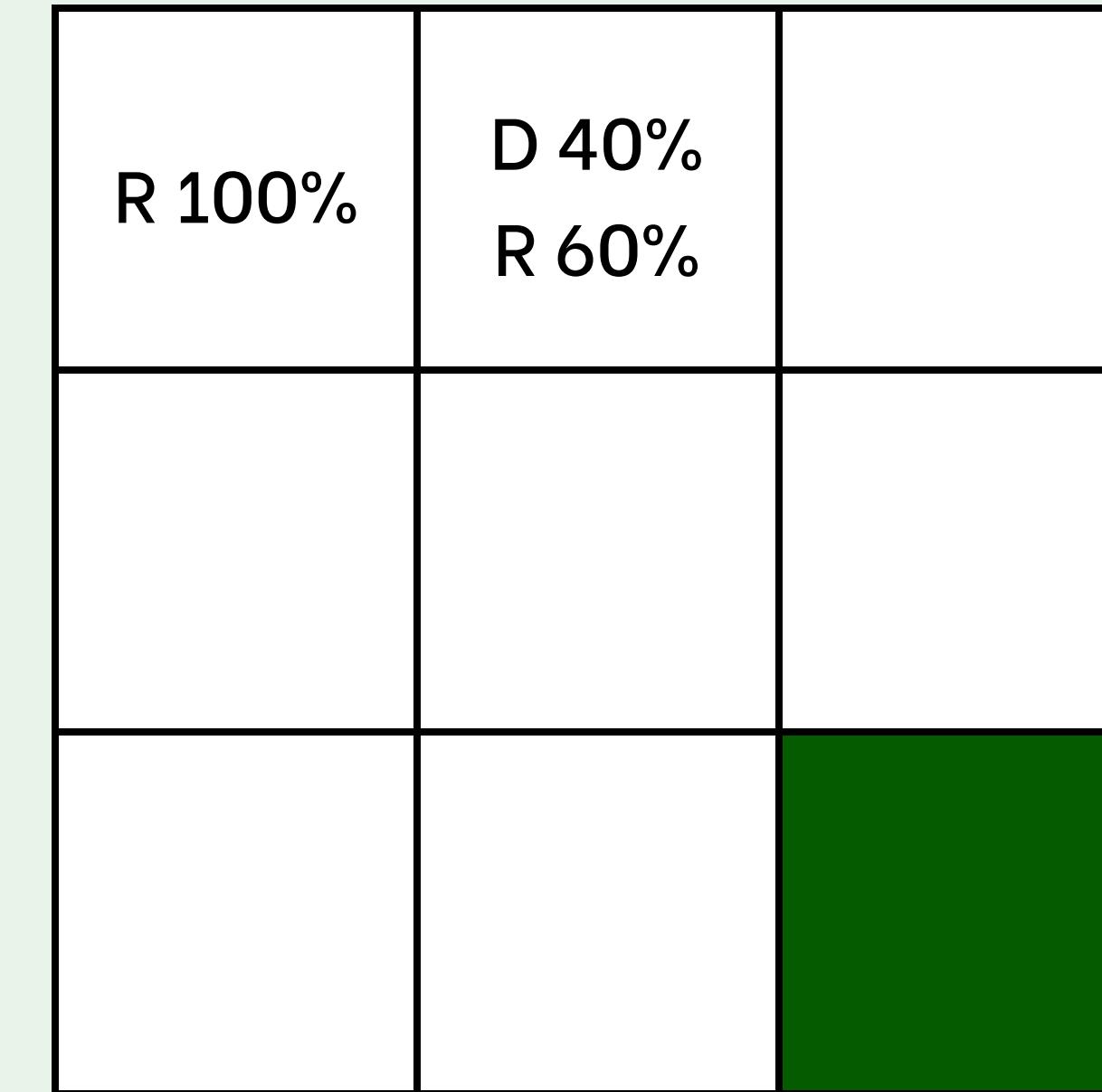
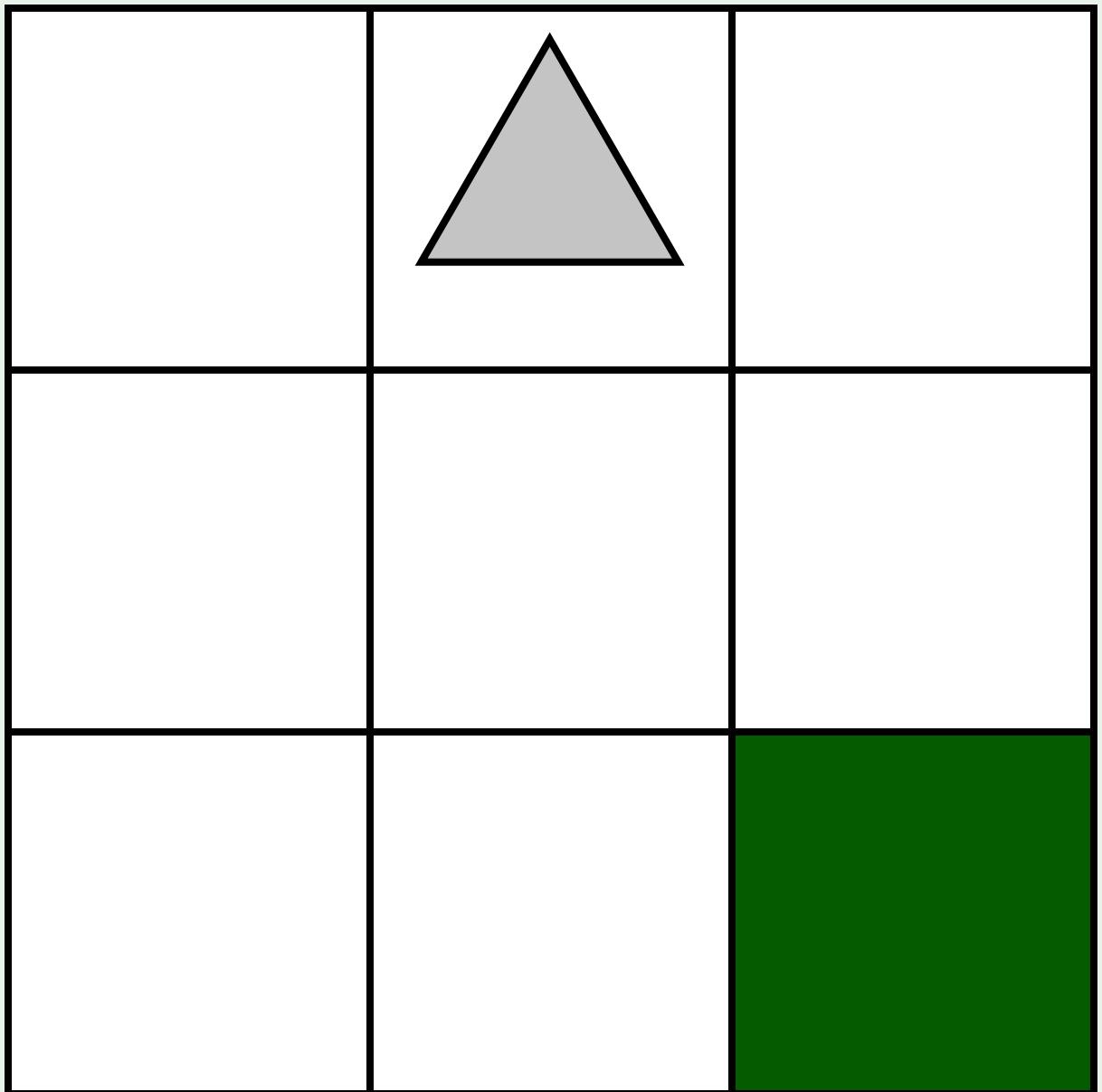


**UNIVERSITÀ
DI TRENTO**

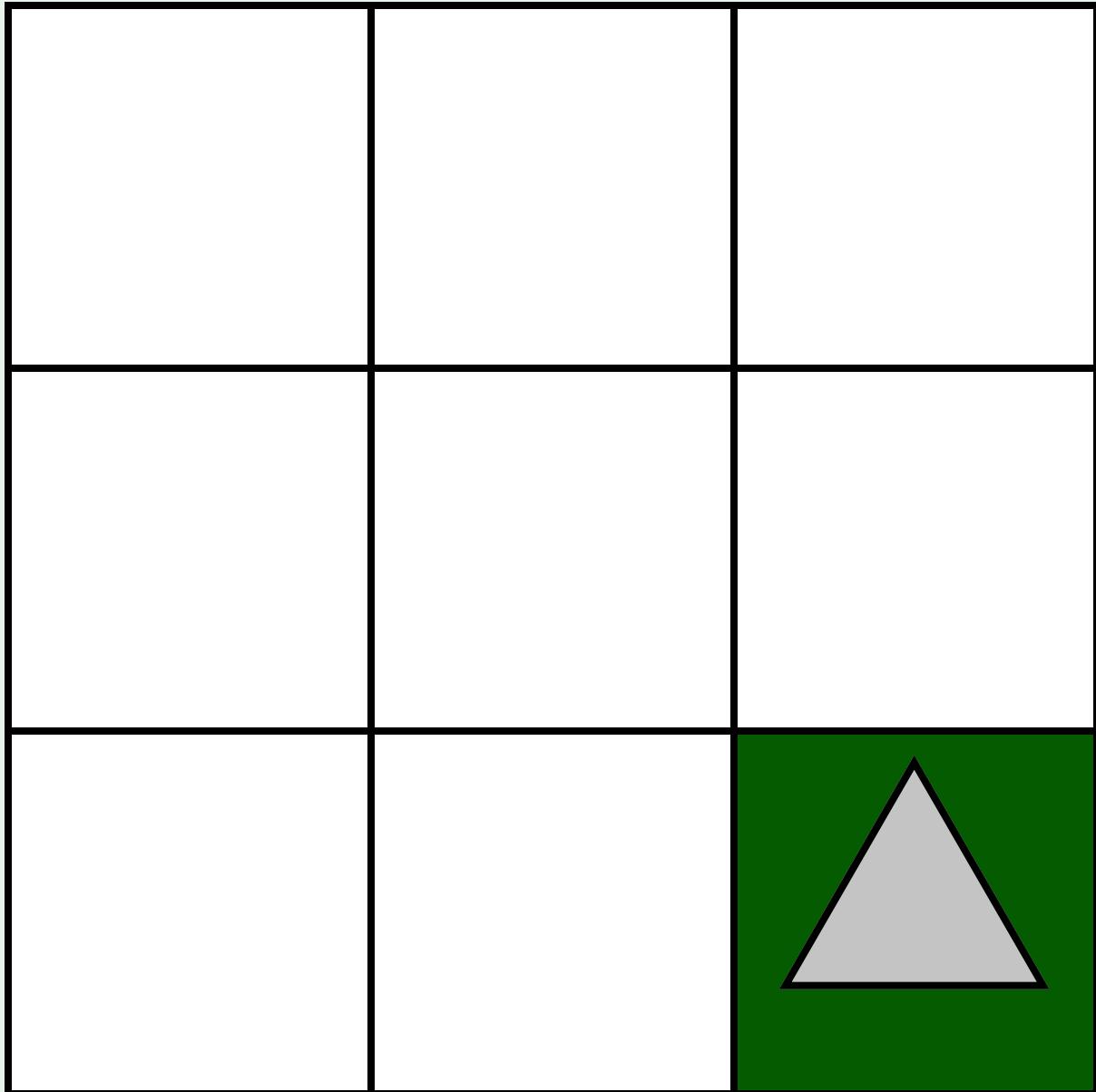
Heatmaps Creation #1



Heatmaps Creation #1



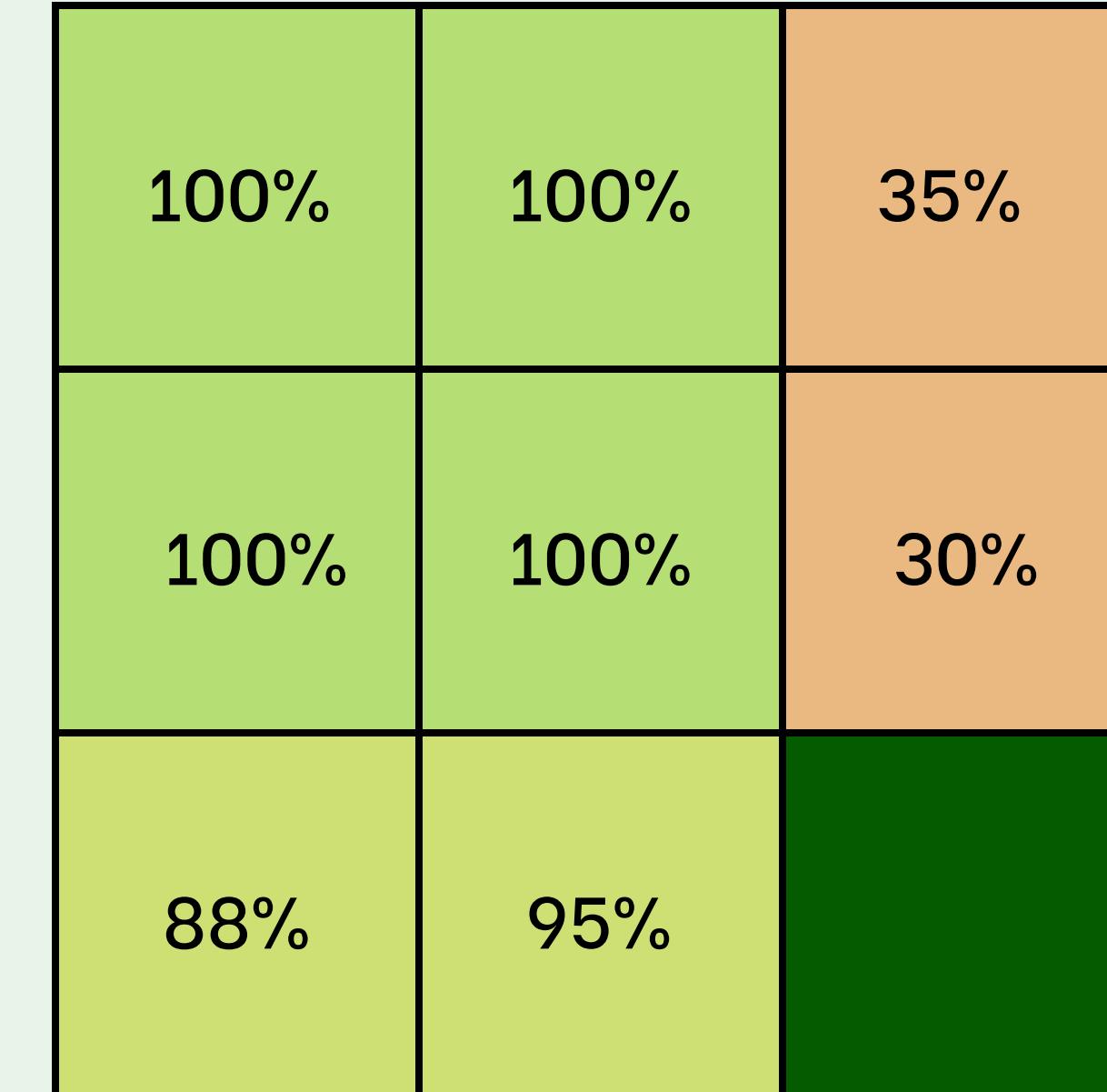
Heatmaps Creation #1



R 100%	D 40% R 60%	D 35% R 65%
D 5% R 95%	D 5% R 95%	D 30% R 70%
D 5% U 7% R 88%	D 5% R 95%	

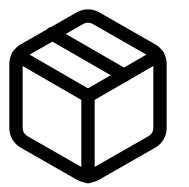
Heatmaps Creation #2

R 100%	D 40% R 60%	D 35% R 65%
D 5% R 95%	D 5% R 95%	D 30% R 70%
D 5% U 7% R 88%	D 5% R 95%	



Testing Strategy

Goals



Pickup

Goal tile explicit



Deliver

Goal tile identified in
map description

Average top1%
87%

Average top1%
81%

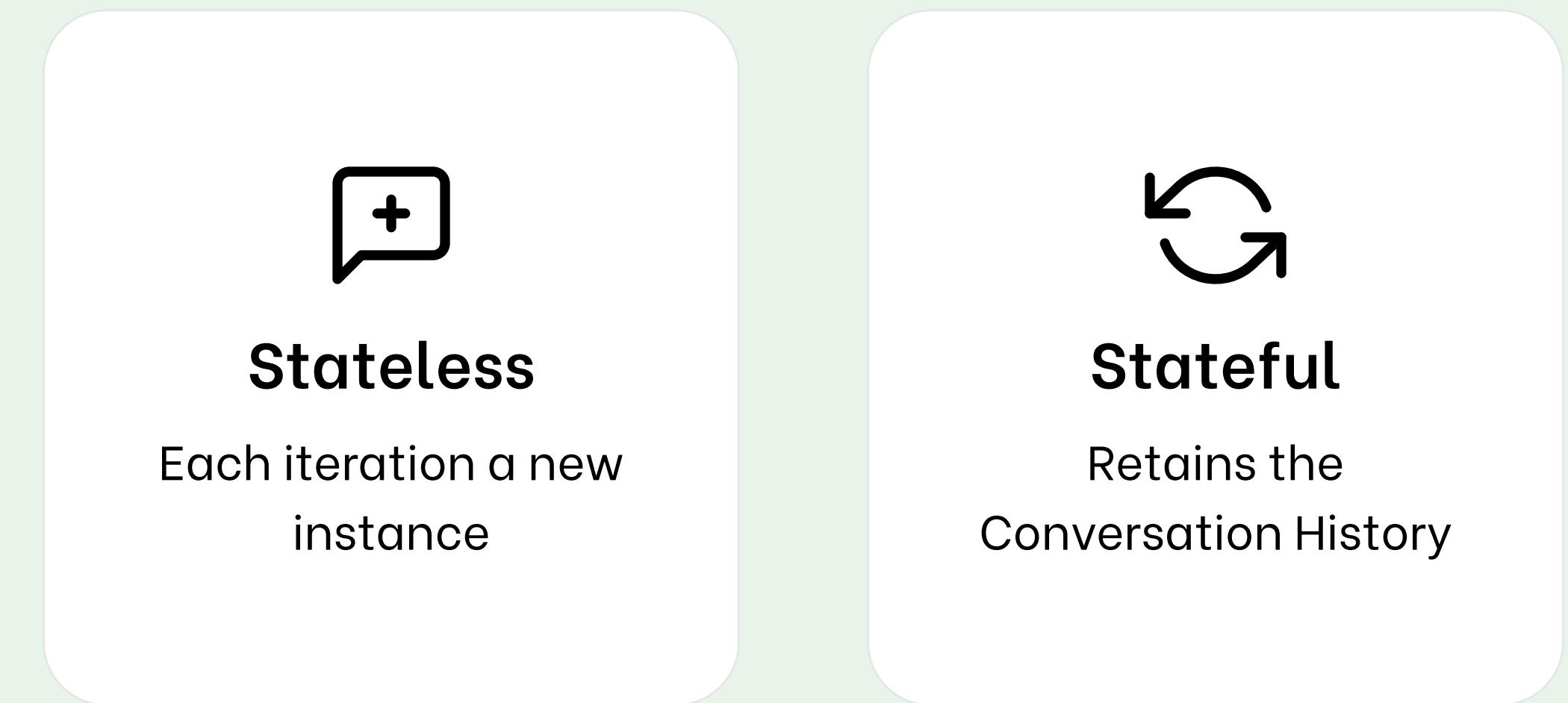
3. Data Collection



UNIVERSITÀ
DI TRENTO

Testing Strategy

Agents



3. Data Collection



UNIVERSITÀ
DI TRENTO

Our Findings

DAVIDE MODOLO
20/03/2025
Supervisor PAOLO GIORGINI



**UNIVERSITÀ
DI TRENTO**

Map Orientation

“Since we did not provide any info about the orientation,
how does the LLM perceive it?”

4. Findings



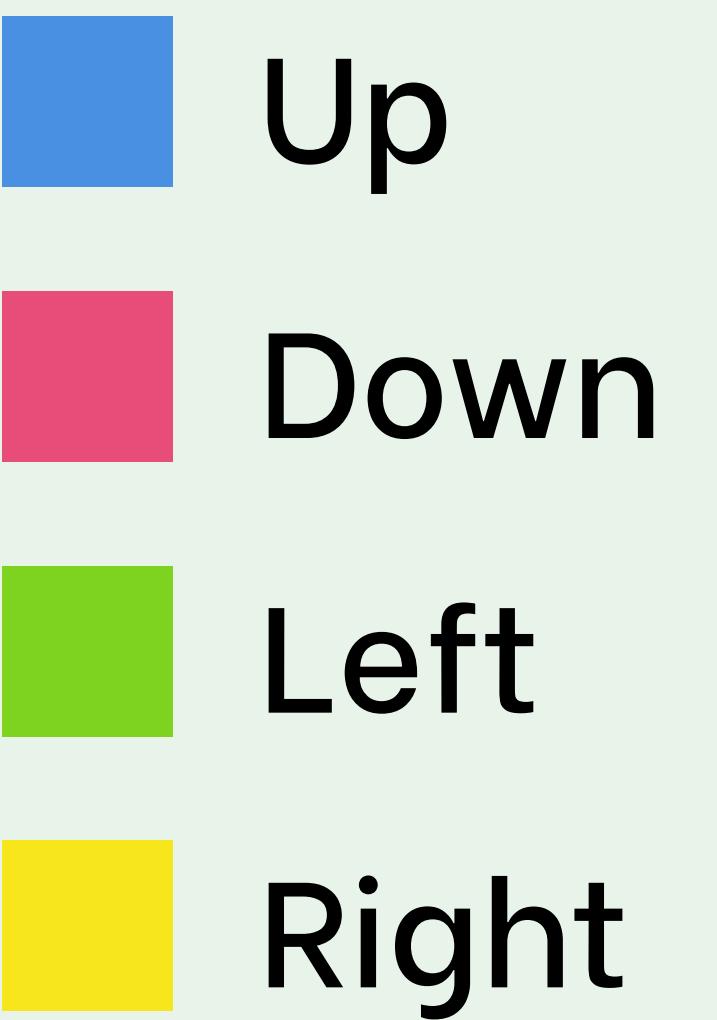
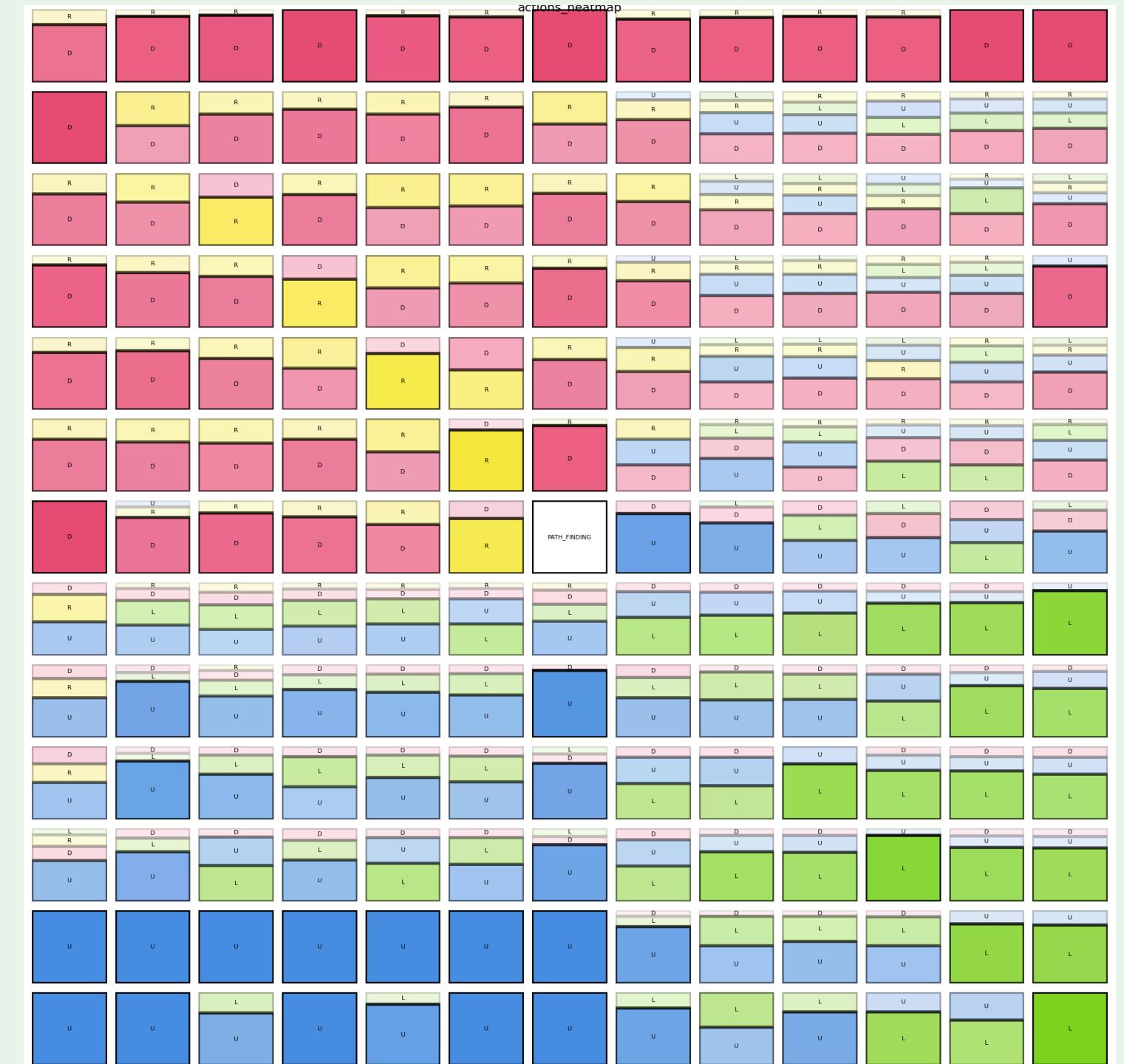
UNIVERSITÀ
DI TRENTO

Map Orientation

(0,0)



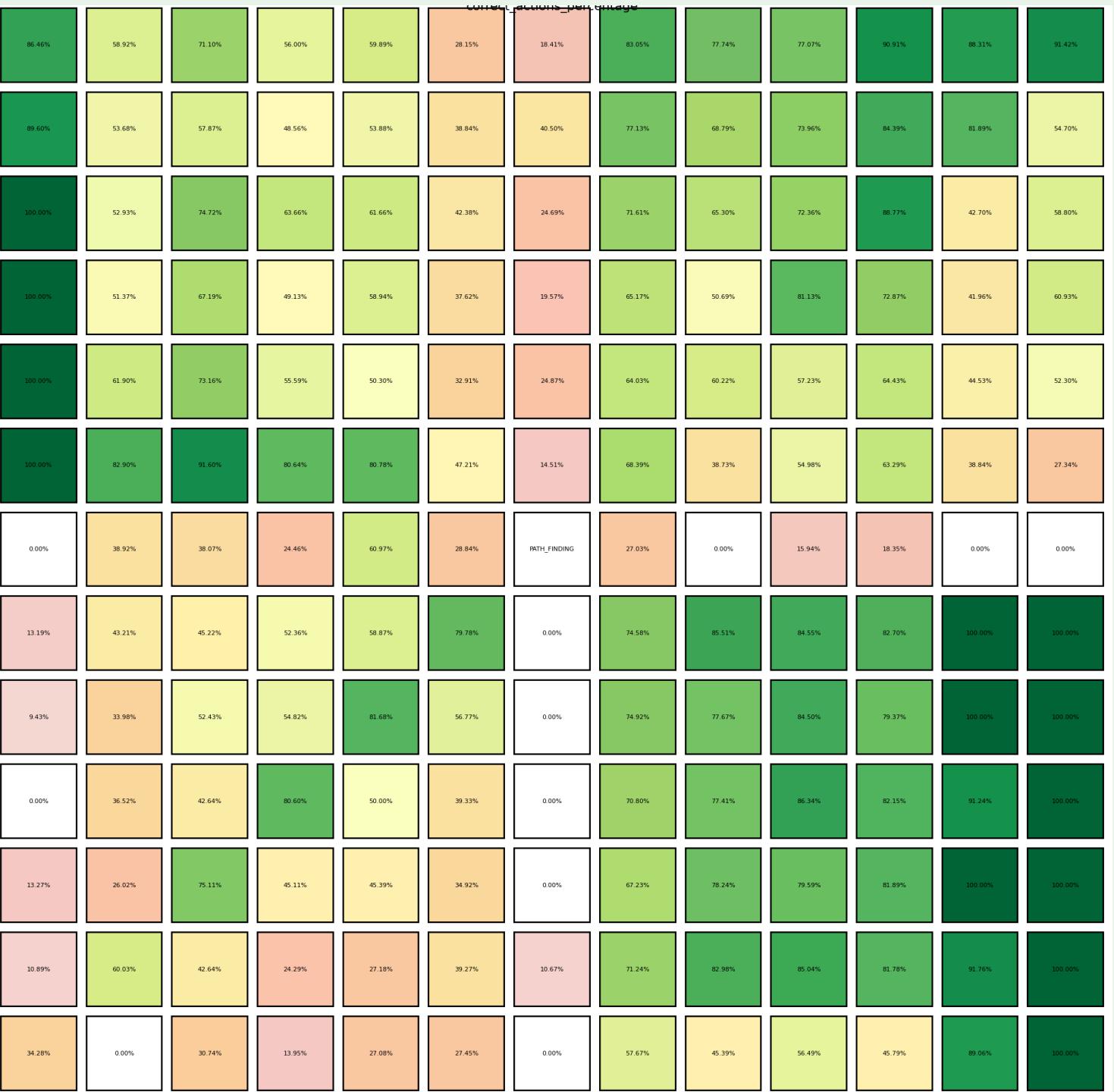
(0,0)



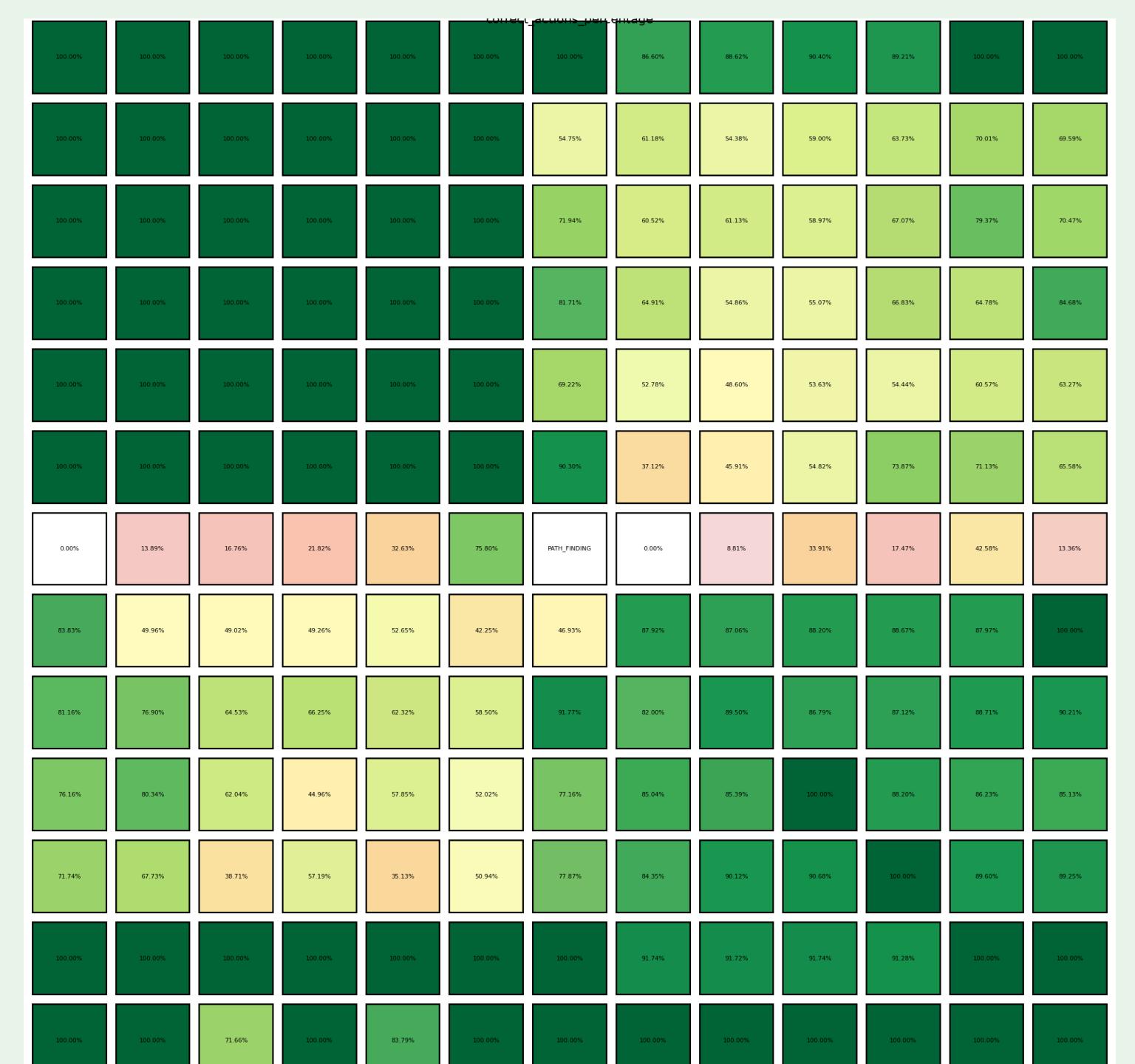
4. Findings

Map Orientation

(0,0)



(0,0)



100%



0%

4. Findings



UNIVERSITÀ
DI TRENTO

Map Orientation

	Bottom-Left Origin	Top-Left Origin
top1%	62%	92%
top2%	92%	97%
top3%	93%	99%

4. Findings



Our Findings

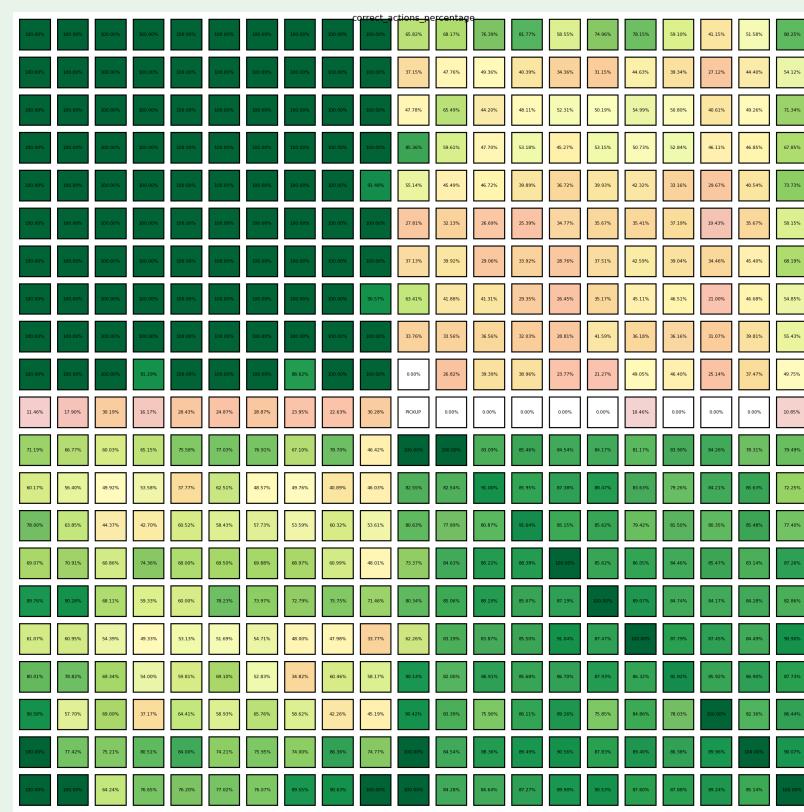
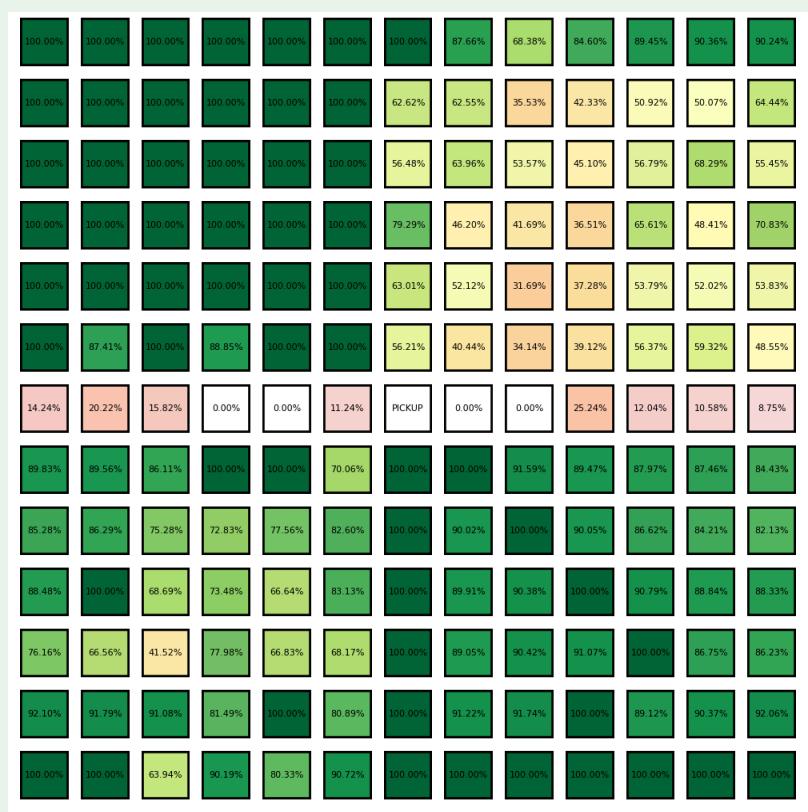
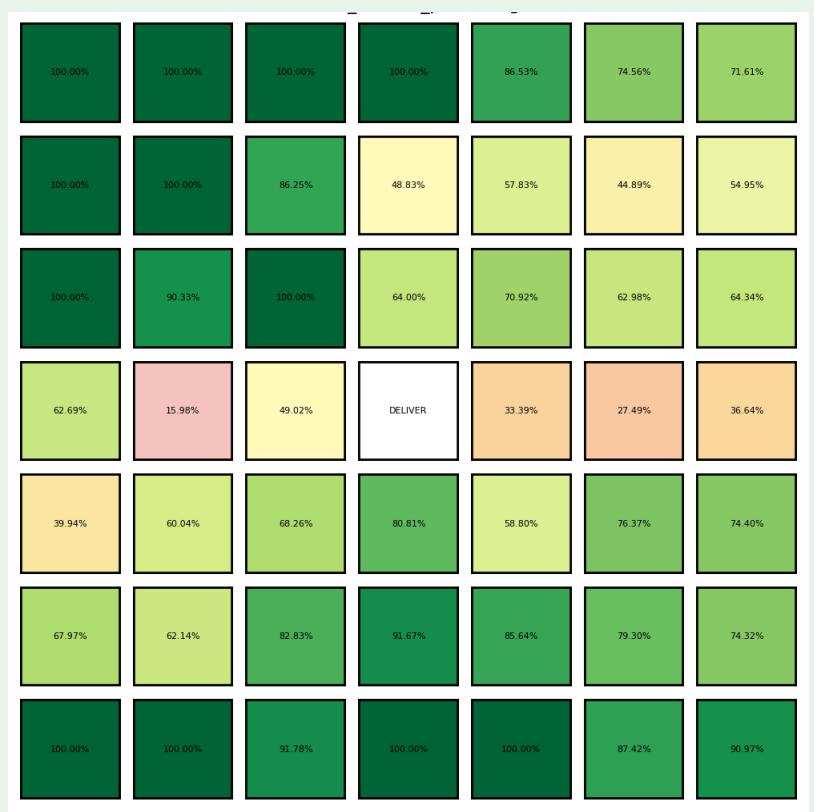
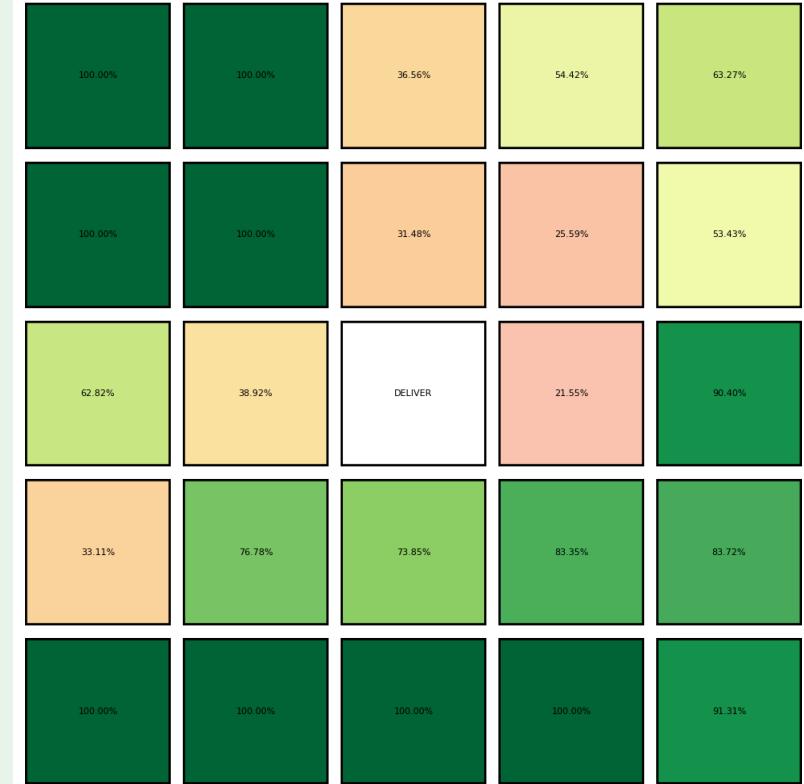
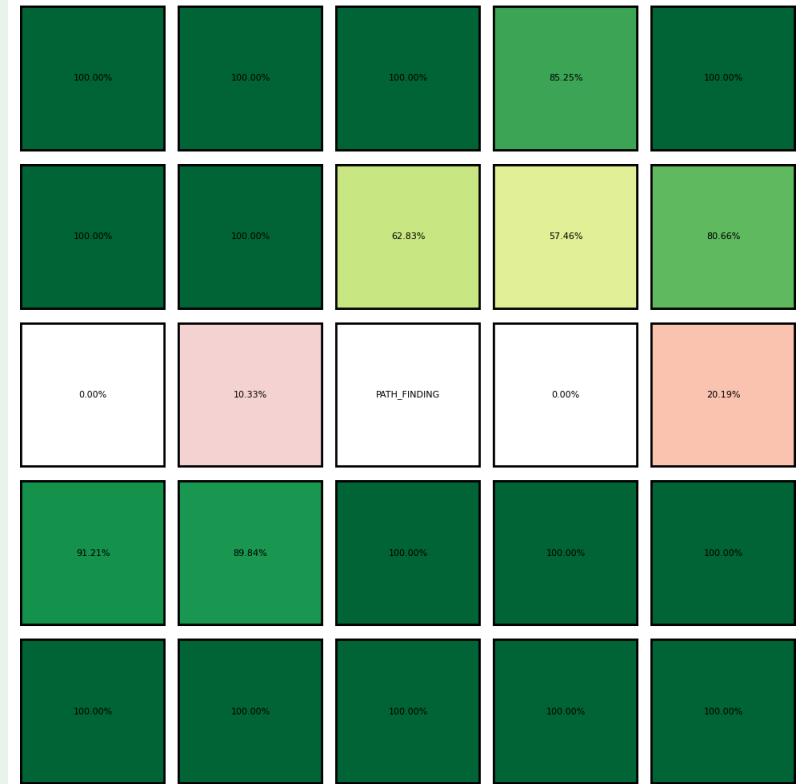
Common Uncertainty Patterns

DAVIDE MODOLO
20/03/2025
Supervisor PAOLO GIORGINI



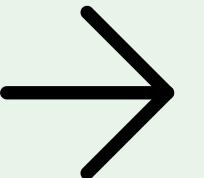
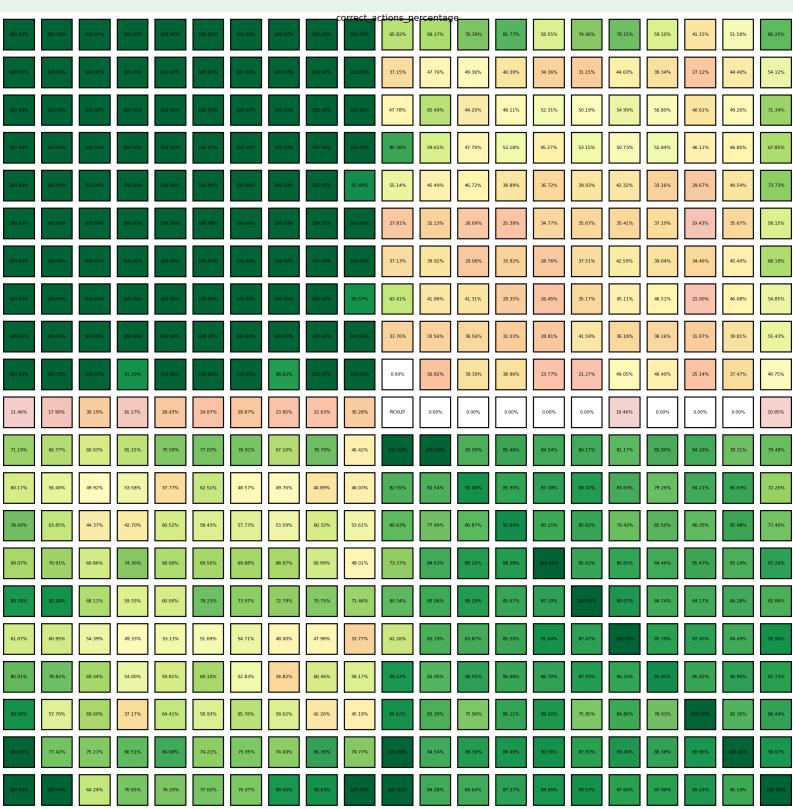
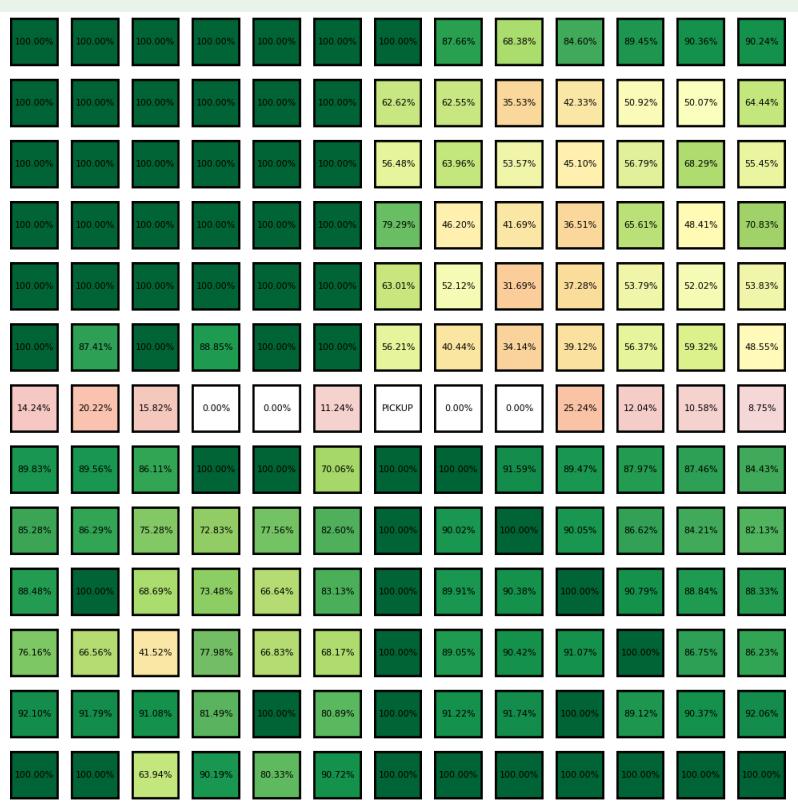
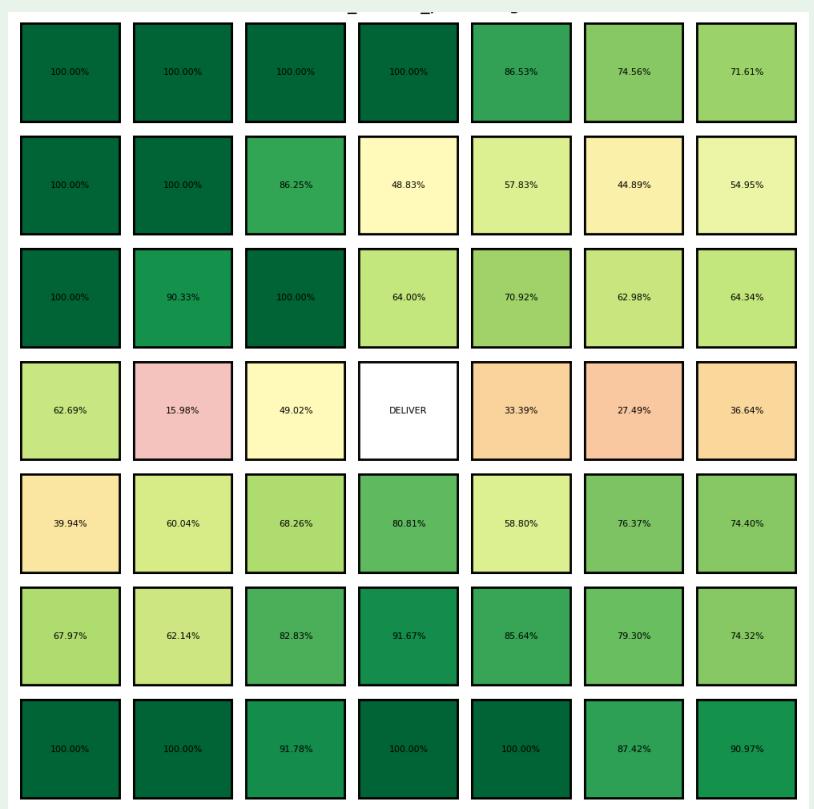
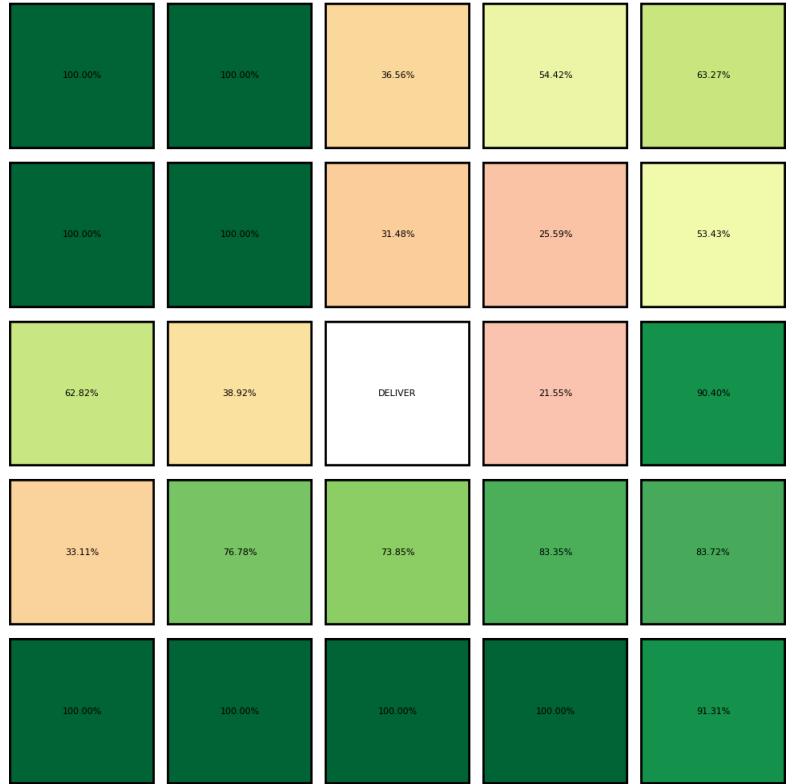
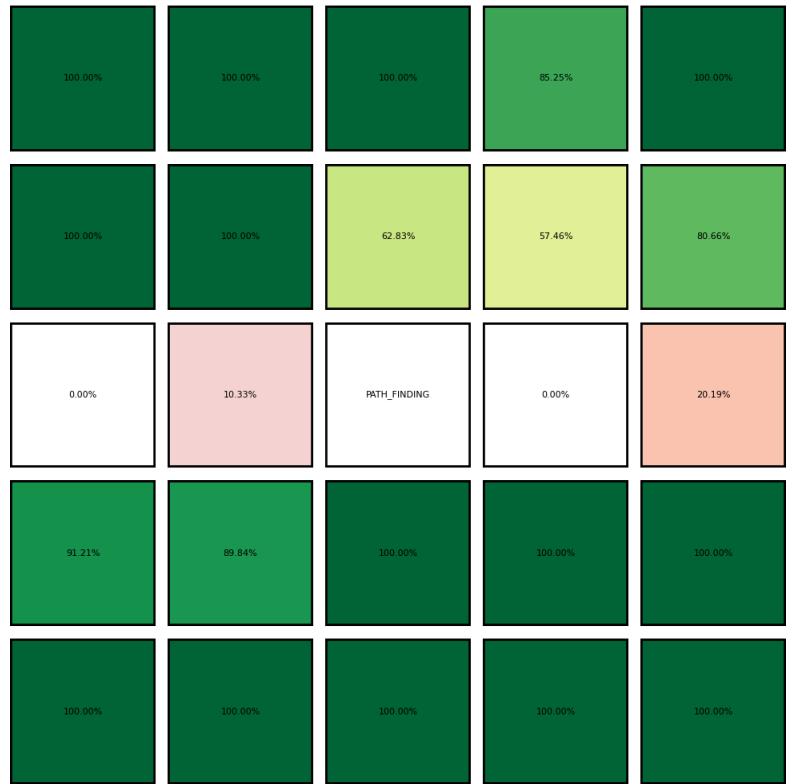
UNIVERSITÀ
DI TRENTO

Common Uncertainty Patterns #1

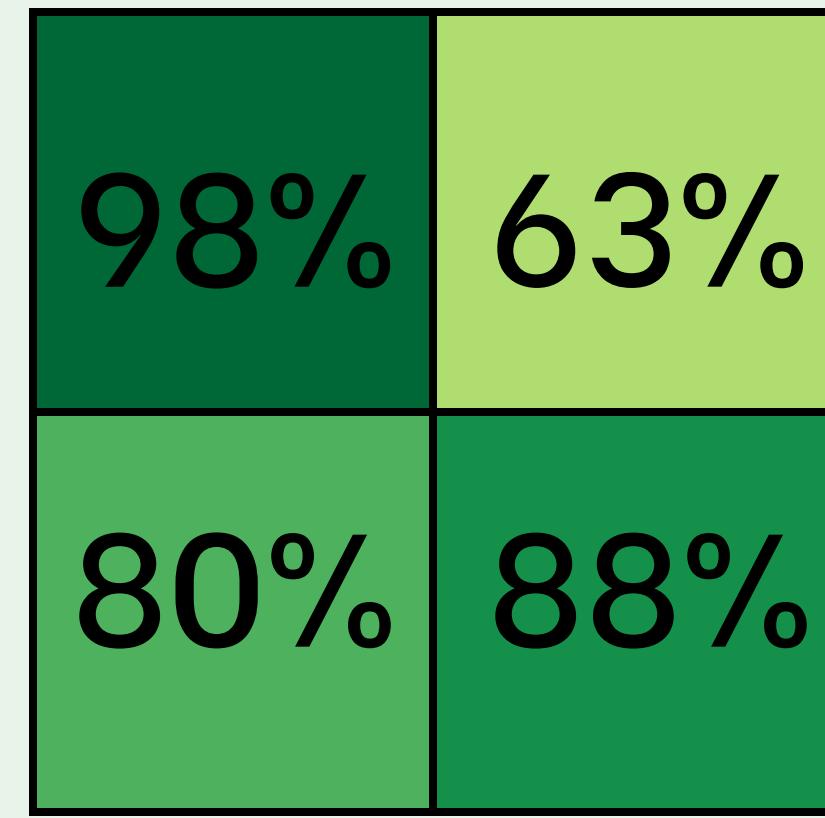


4. Findings

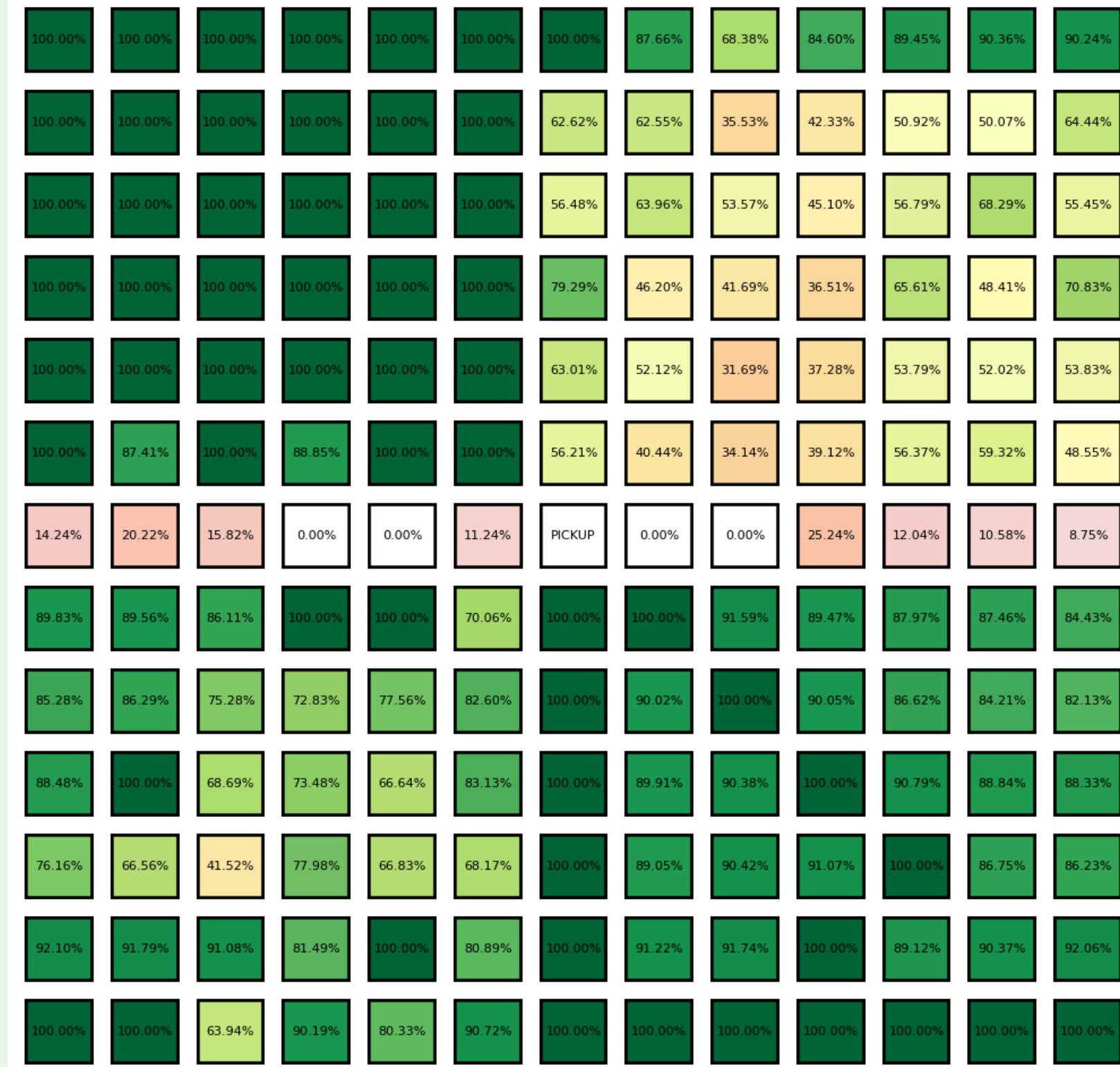
Common Uncertainty Patterns #1



Average

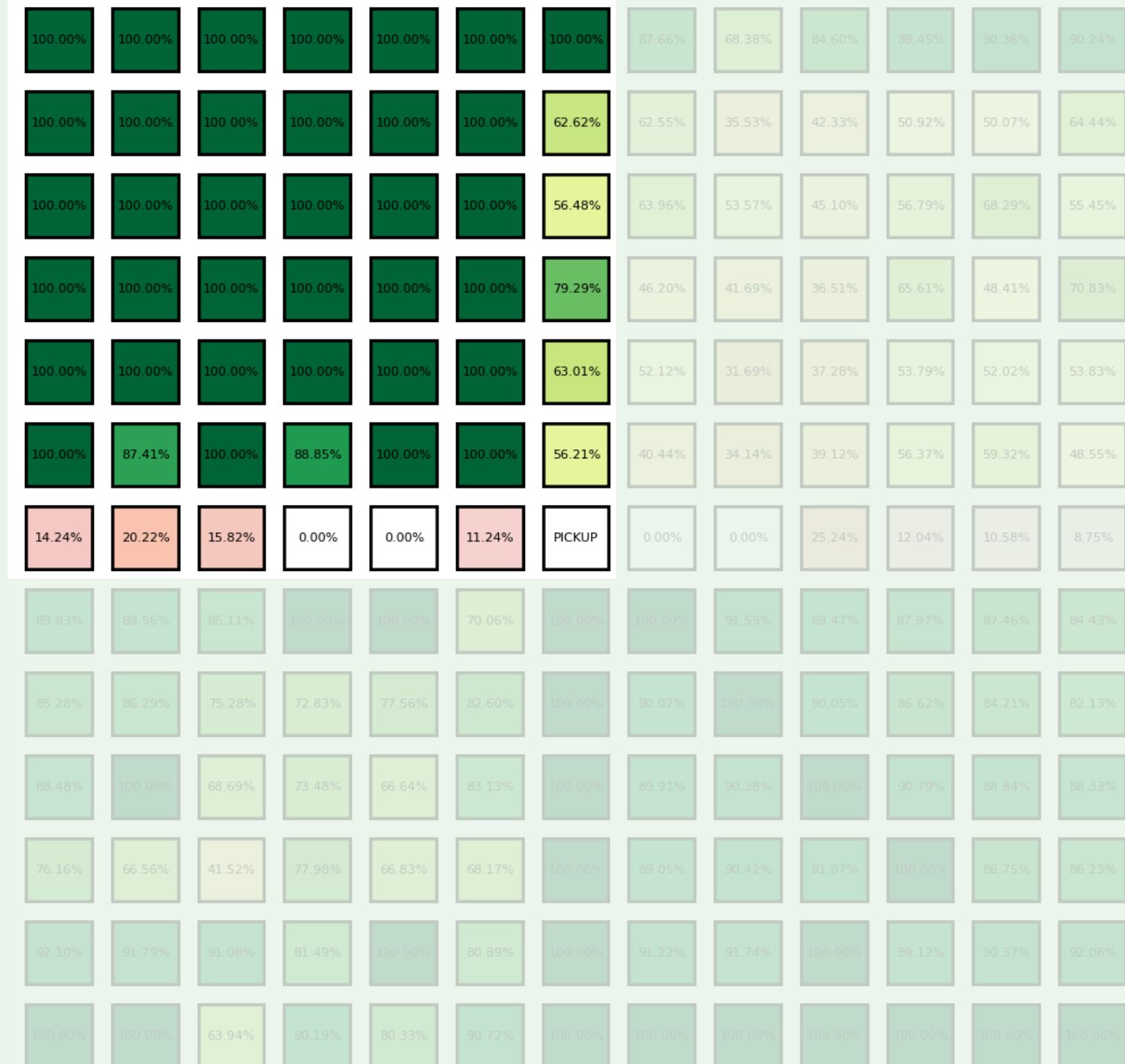


Common Uncertainty Patterns #1.1



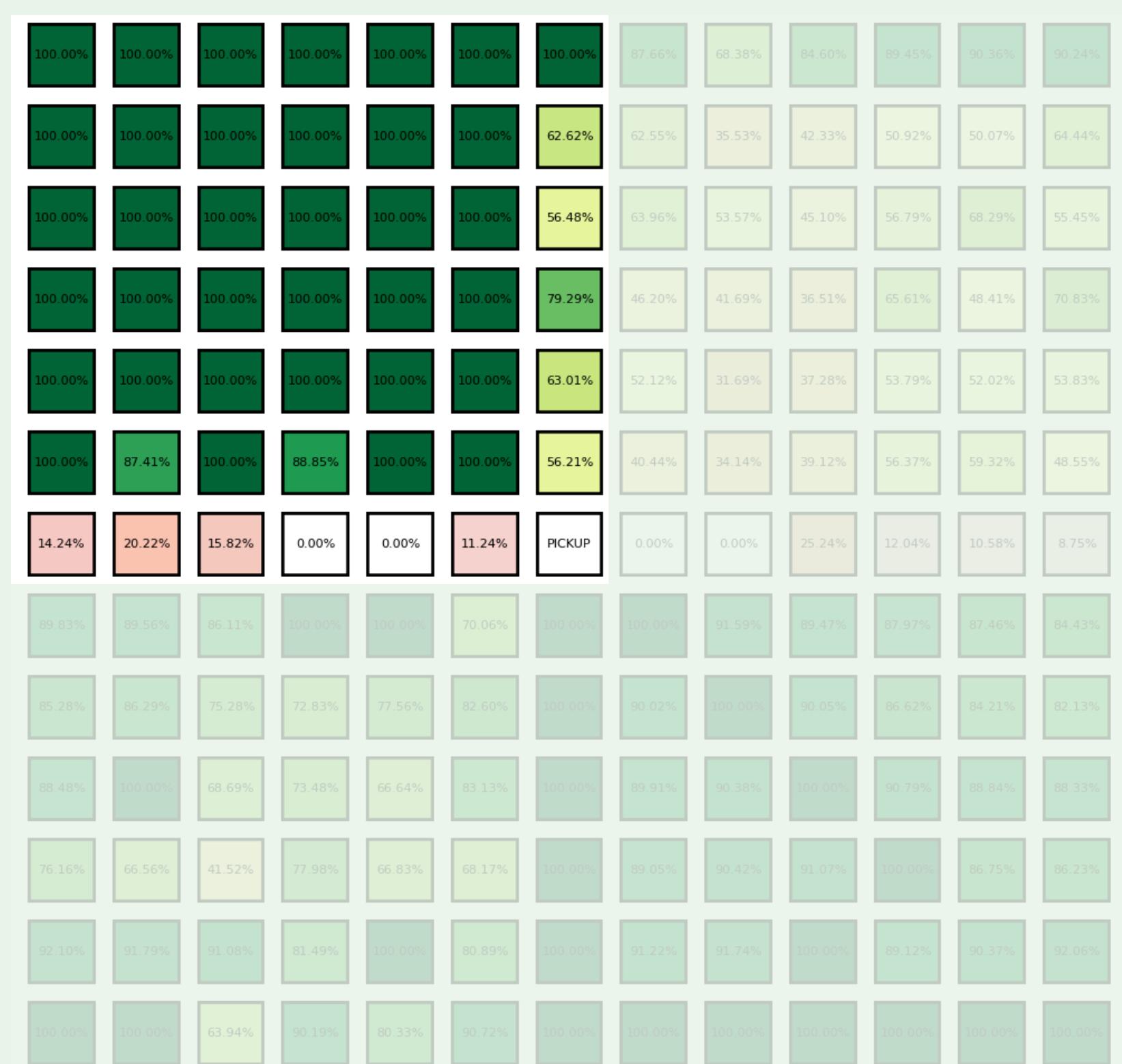
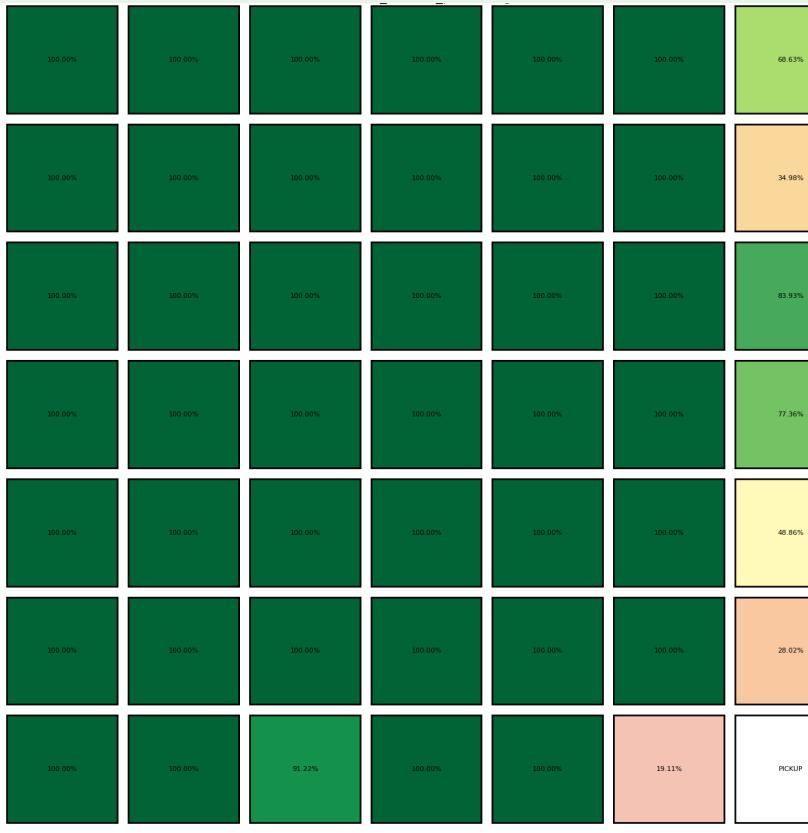
4. Findings

Common Uncertainty Patterns #1.1



4. Findings

Common Uncertainty Patterns #1.1

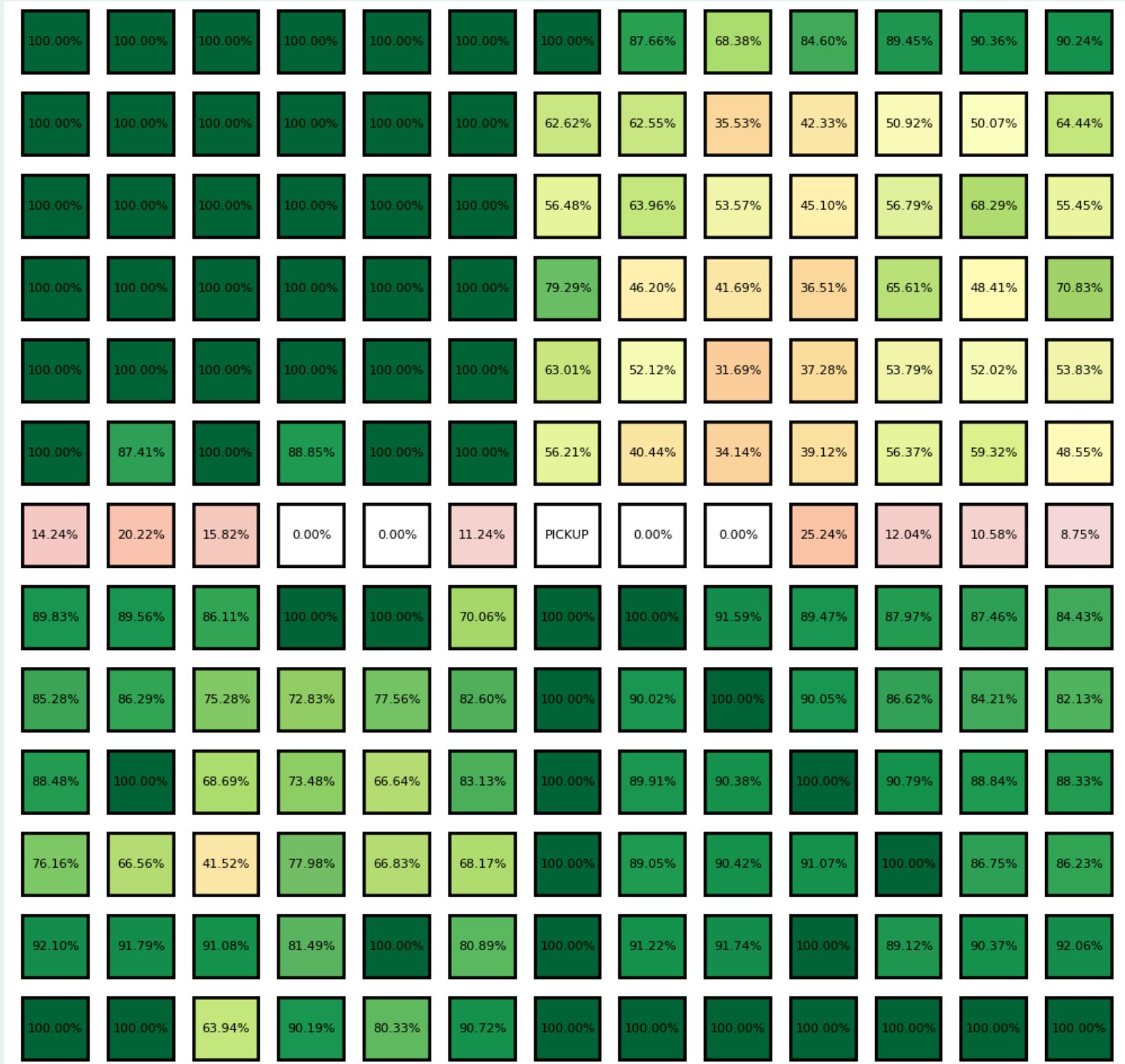


4. Findings



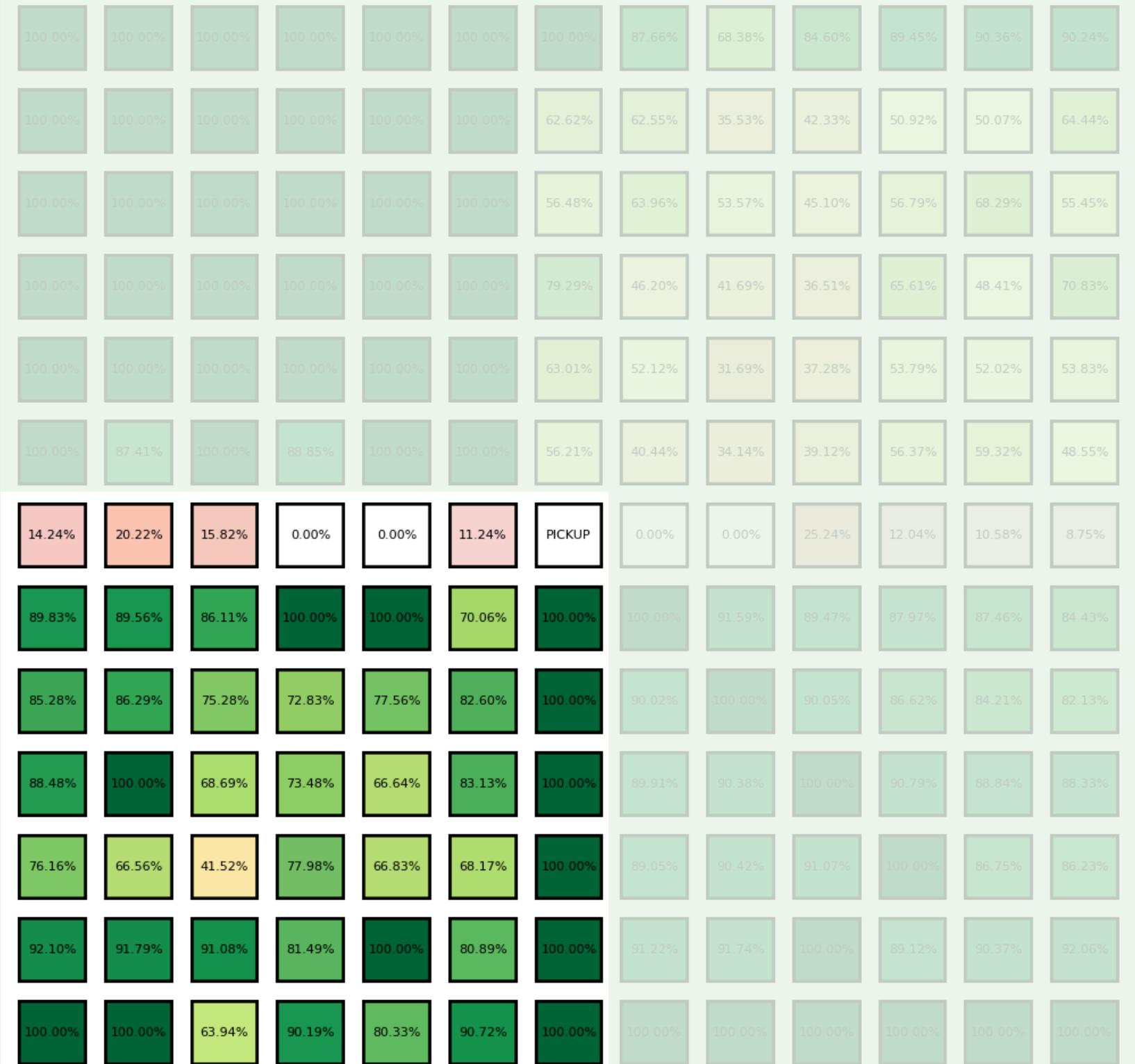
UNIVERSITÀ
DI TRENTO

Common Uncertainty Patterns #1.2



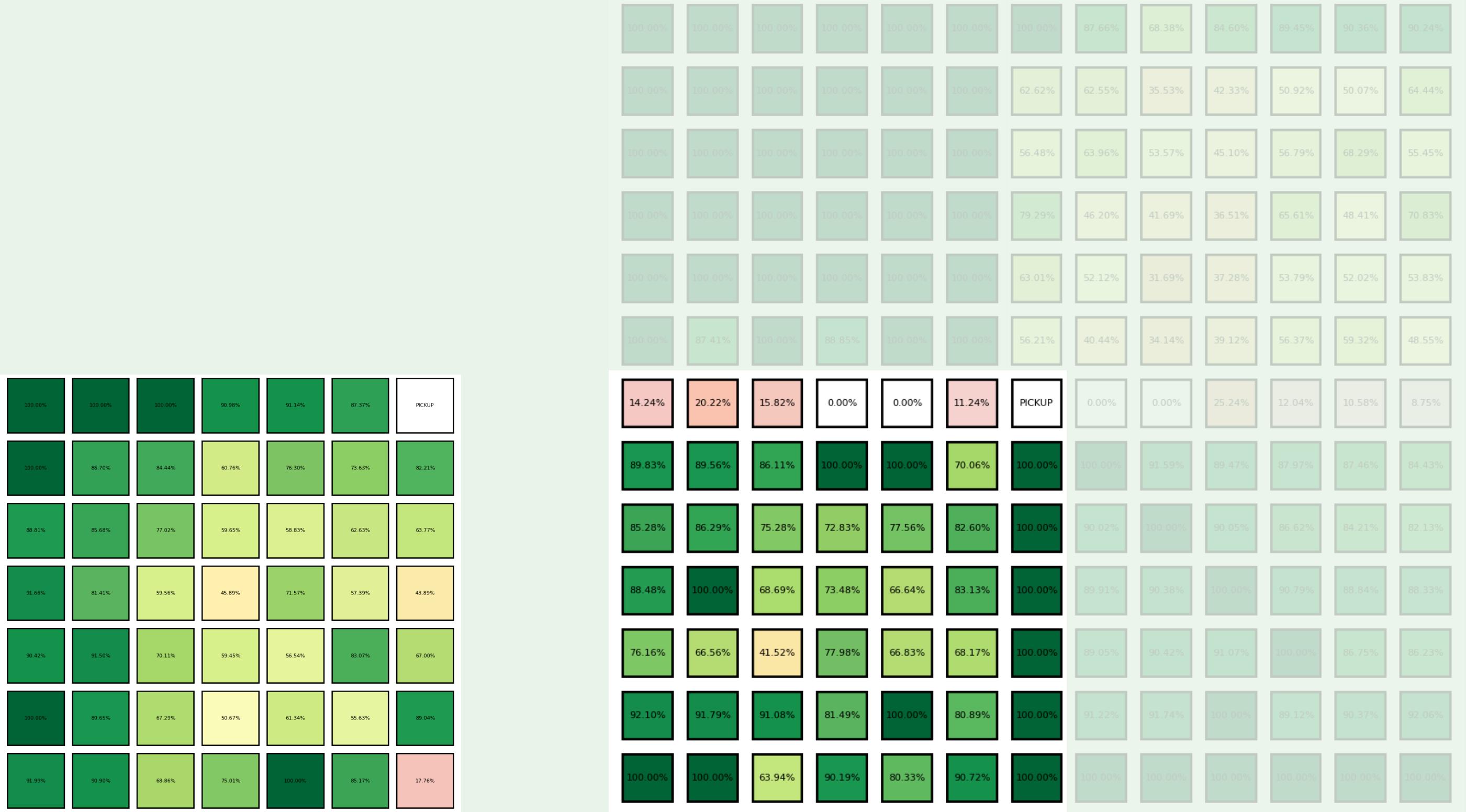
4. Findings

Common Uncertainty Patterns #1.2



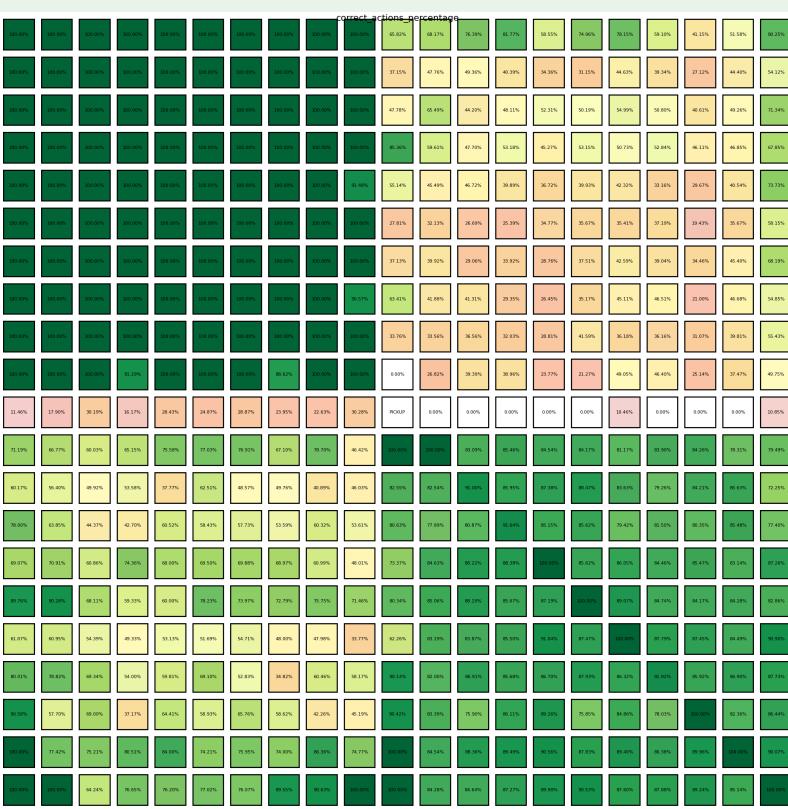
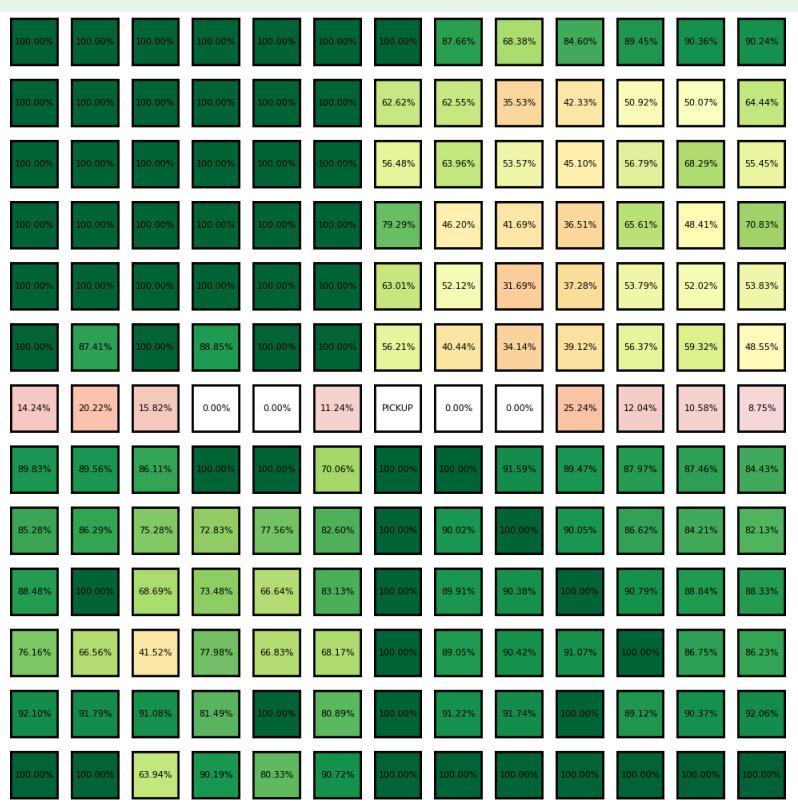
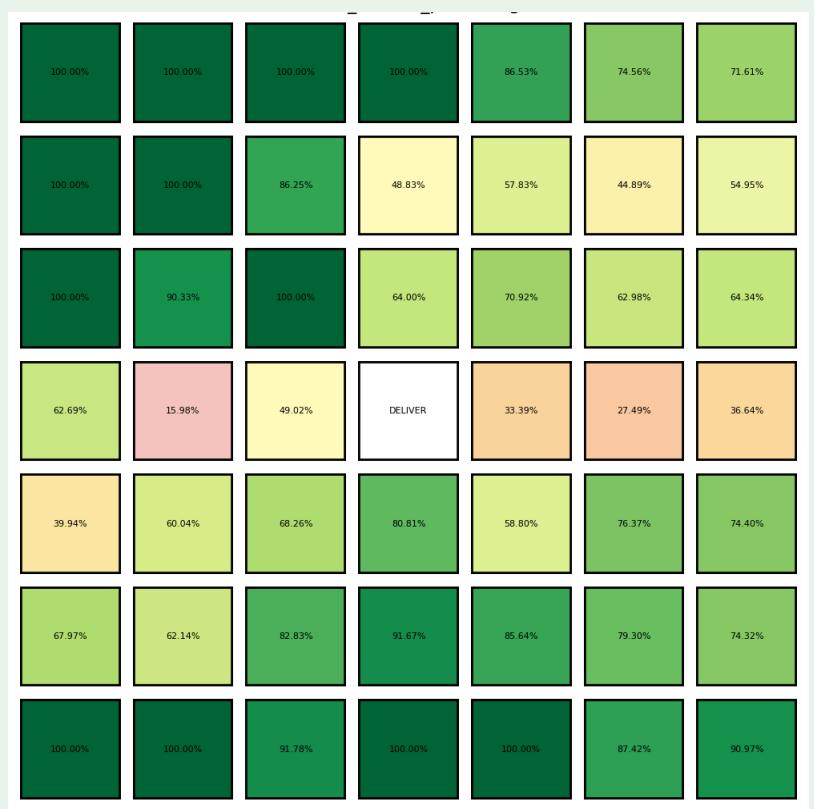
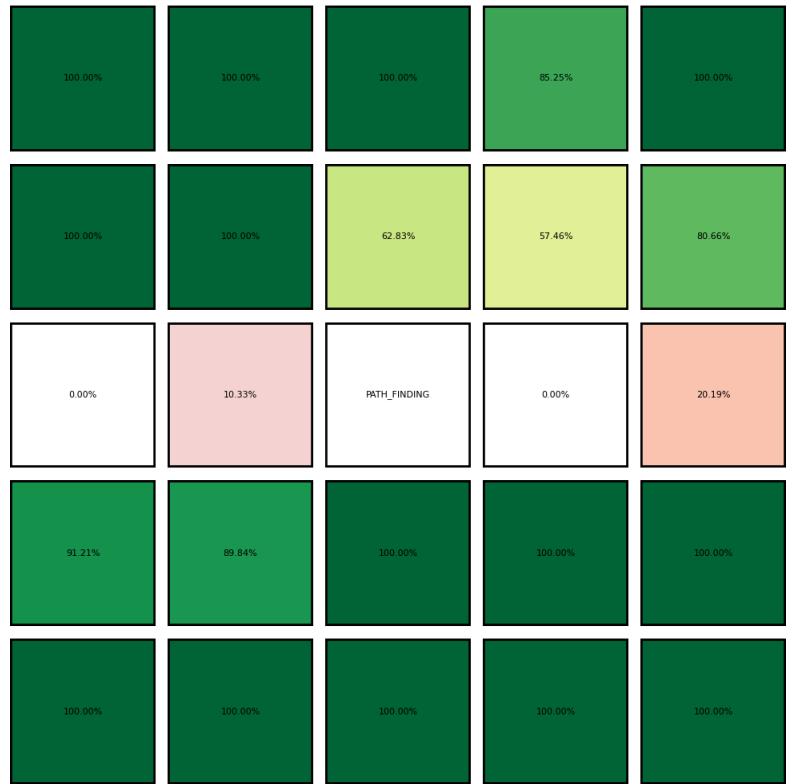
4. Findings

Common Uncertainty Patterns #1.2

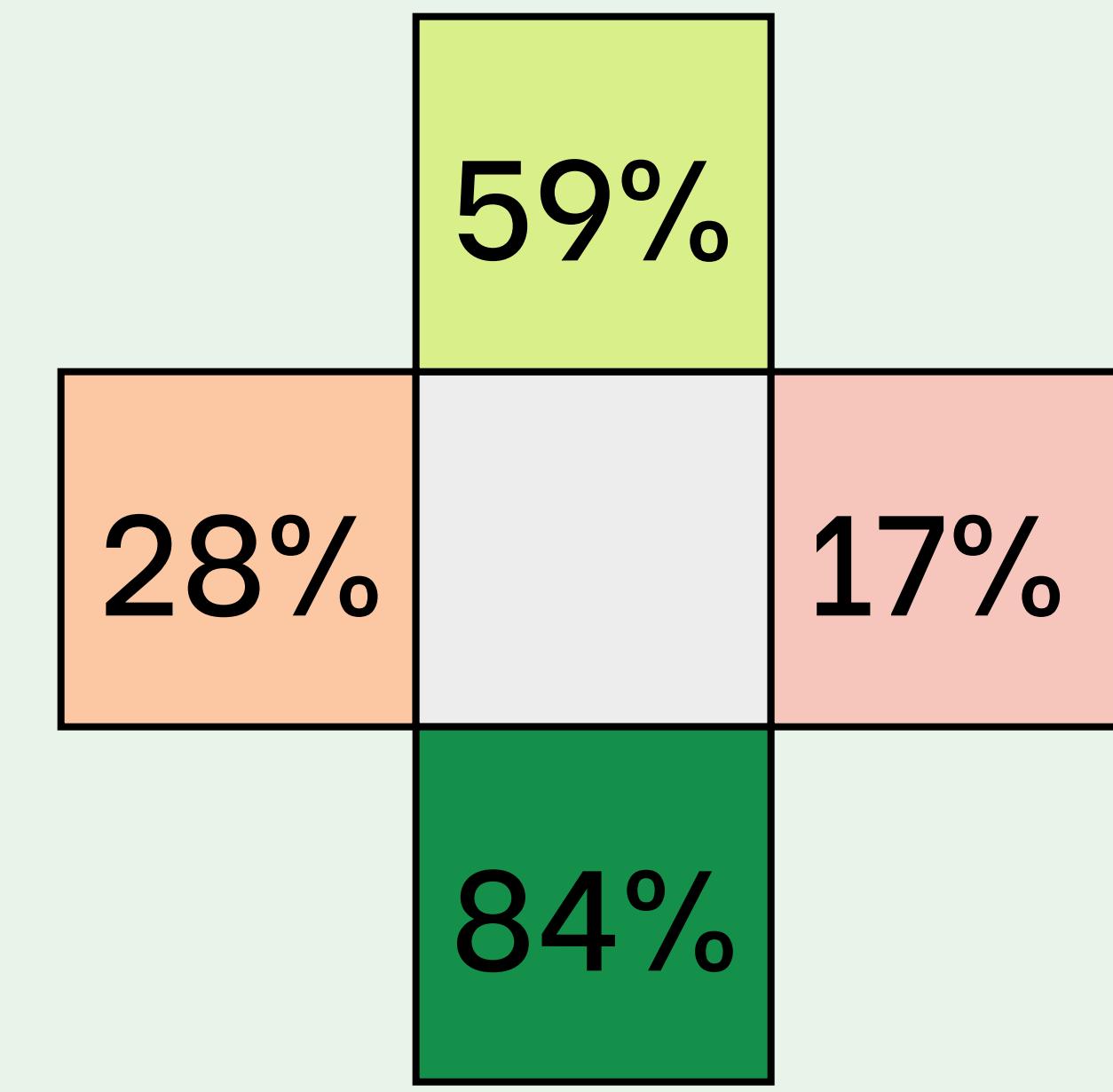


4. Findings

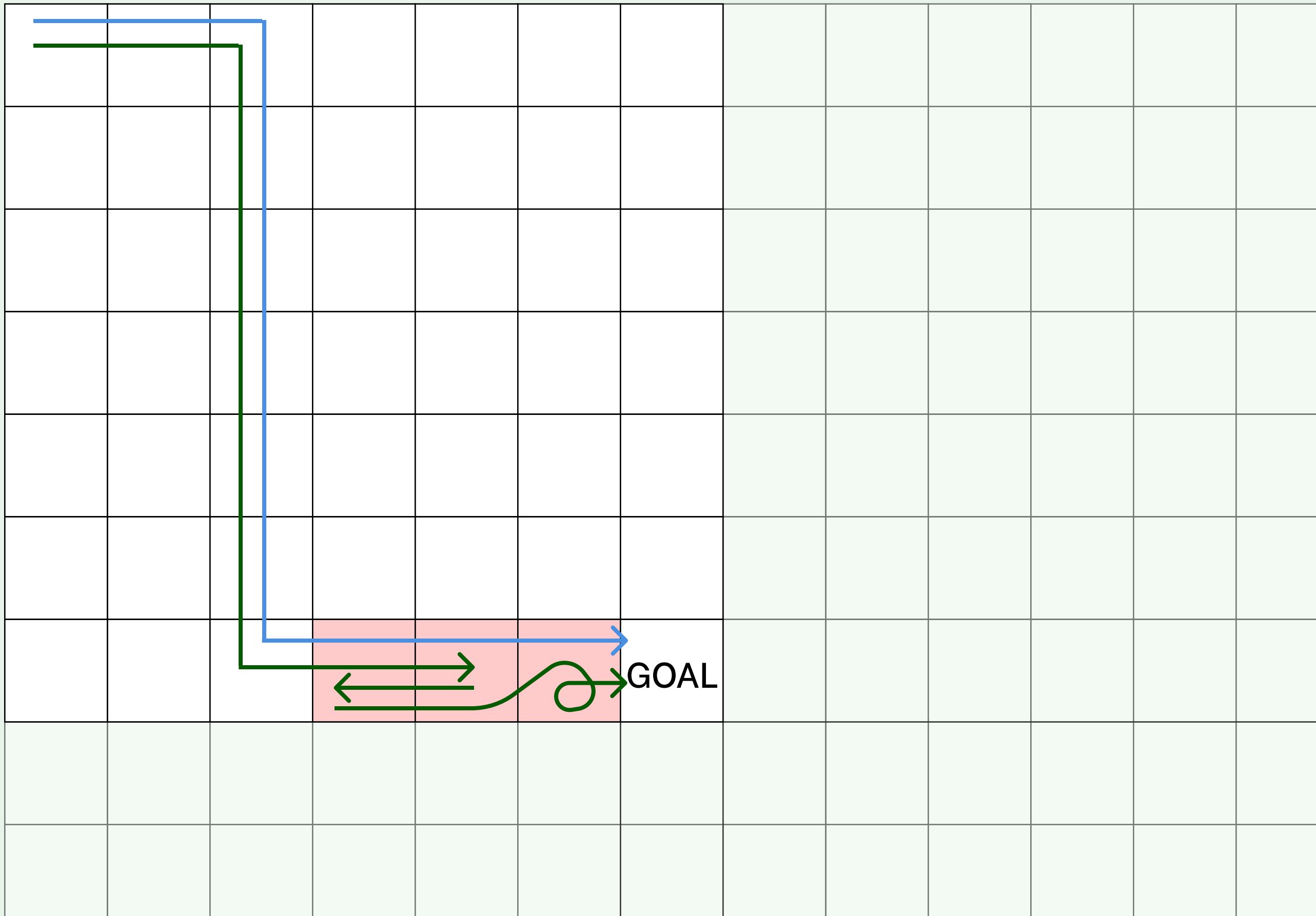
Common Uncertainty Patterns #2



Average



Stateful Example



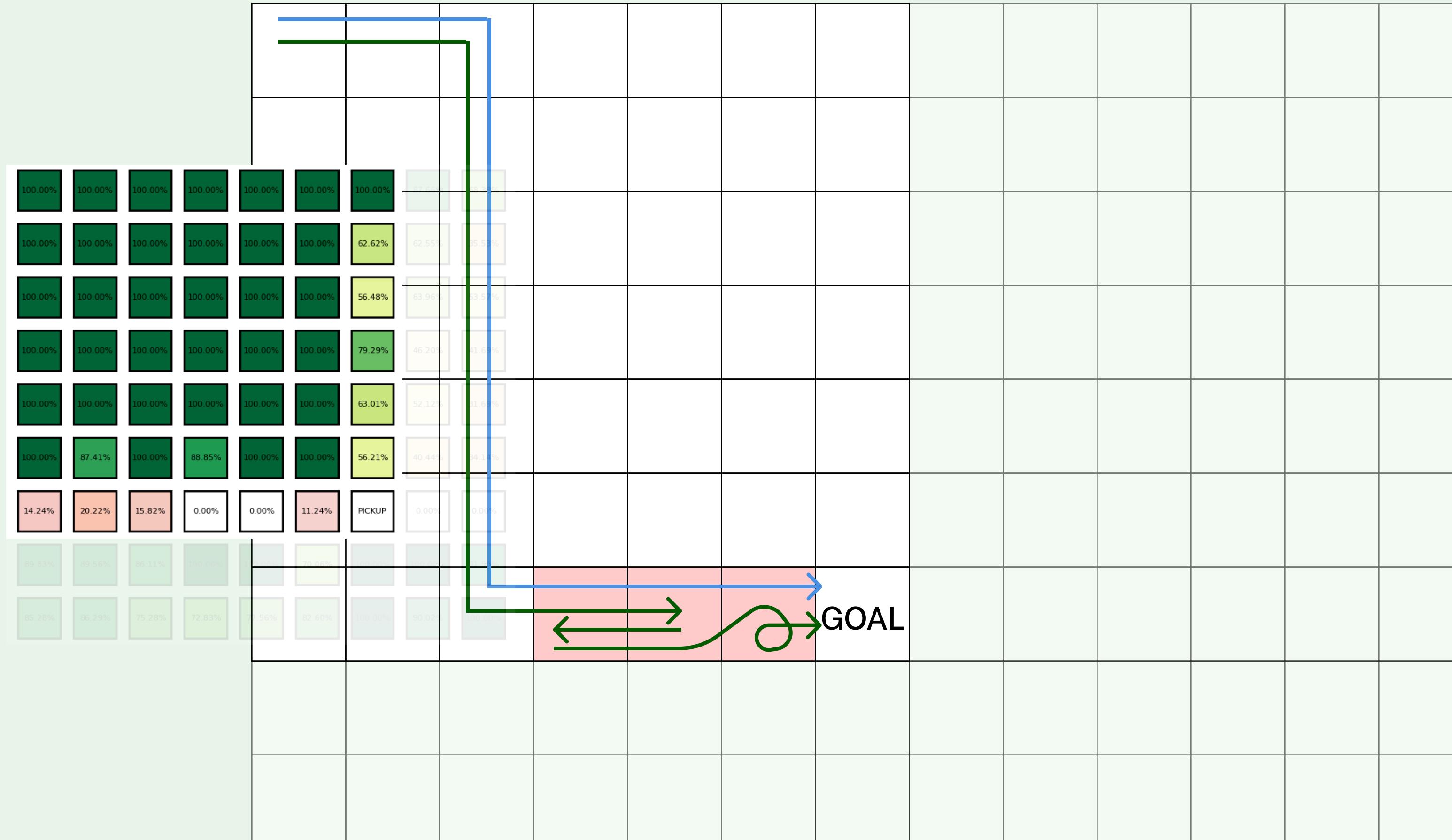
- Optimal path
- Our path
- Repeated cells

Shared Nodes: 100%

Ratio: 81%

4. Findings

Stateful Example



Optimal path
Our path
Repeated cells
Shared Nodes: 100%
Ratio: 81%

4. Findings

Stateful

Map Size	Stateless	Stateful	Difference
13x13	41	39	-5%
7x7	27	21	-29%
5x5	14	11	-27%
3x3	11	9	-22%

4. Findings



Conclusions

Strengths

Effective

~ 20% Less action required
with history

Retrieving goals in big maps

Better models, better results

Similar % error as the size
increase

Weaknesses

Limited Explainability

Prone to error near the goal

Consistent problematic zones

Context size is a limitation

Similar % error as the size
increase, but real number is a
problem

4. Findings



UNIVERSITÀ
DI TRENTO

Thank You

Exploring the Use of LLMs for Agent Planning: Strengths and Weaknesses