Massachusetts Institute of Technology

A Collaborative Numeric Task Planning Framework based on Constraint Translations using LLMs

HAXP: Workshop on Human-Aware and Explainable Planning

November 10, 2025

Anthony Favier, **Ngoc La**, Pulkit Verma, Julie A. Shah

Interactive Robotics Group

Massachusetts Institute of Technology

A Collaborative Numeric Task Planning Framework based on Constraint Translations using LLMs

LM4Plan: Workshop on Planning in the Era of LLMs

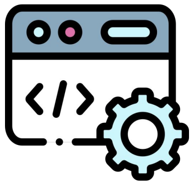
November 11, 2025

Anthony Favier, **Ngoc La**, Pulkit Verma, Julie A. Shah

Interactive Robotics Group

MIT AEROASTRO MIT CSAIL

Introduction



Formal automated planning

Limited accessibility

Requires:

- programming knowledge
- or technical expert interventions

BARRIER

Users,
Domain experts

Credit: Freepik

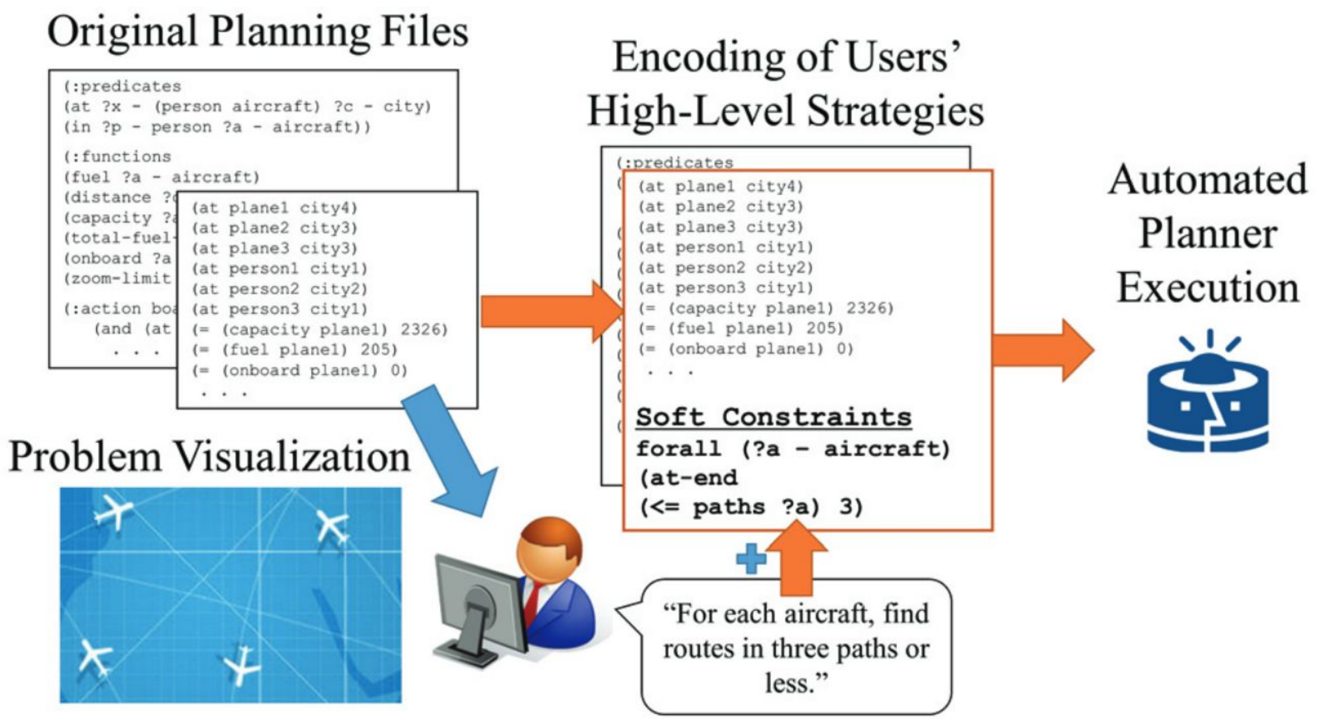
- Problematic for **time-constrained** problem solving, such as **disaster response** scenarios.
- Computing the **optimal** solution is **extremely challenging**.
- Focus on **only** finding best **valid** solution, given a **limited time budget**.

Introduction

Leveraging domain experts intervention
Collaborative Planning

Significant **potential** for higher
quality solutions and **efficiency**.

(Kim, Banks, Shah. AAAI 2017)



User High-Level Input Guidance is translated by **programming experts** in PDDL3 soft constraints (preferences)

Average translation time = 3 min (SD=1.3)

Kim, J.; Banks, C.; Shah, J. Collaborative Planning With Encoding of Users' High-Level Strategies. AAAI 2017

Large Language Models

Promising and improving results in reasoning tasks.
(OpenAI. Gpt-4 technical report. arXiv 2023)

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot

Table 2. Performance of GPT-4 on academic benchmarks.

Large Language Models

Still can't plan reliably on their own.
(Kambhampati et al. ICML 2024)

Domain	Method	Instances correct					
		GPT-4o	GPT-4-Turbo	Claude-3-Opus	LLaMA-3 70B	Gemini Pro	GPT-4
Blocksworld (BW)	One-shot	170/600 (28.33%)	138/600 (23%)	289/600 (48.17%)	76/600 (12.6%)	68/600 (11.3%)	206/600 (34.3%)
	Zero-shot	213/600 (35.5%)	241/600 (40.1%)	356/600 (59.3%)	205/600 (34.16%)	3/600 (0.5%)	210/600 (34.6%)
Mystery BW (Deceptive)	One-shot	5/600 (0.83%)	5/600 (0.83%)	8/600 (1.3%)	15/600 (2.5%)	2/500 (0.4%)	26/600 (4.3%)
	Zero-shot	0/600 (0%)	1/600 (0.16%)	0/600 (0%)	0/600 (0%)	(0/500) (0%)	1/600 (0.16%)

Table 1. Results of LLMs for Plan Generation with prompts in natural language.

Kambhampati et al. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. ICML 2024

Large Language Models

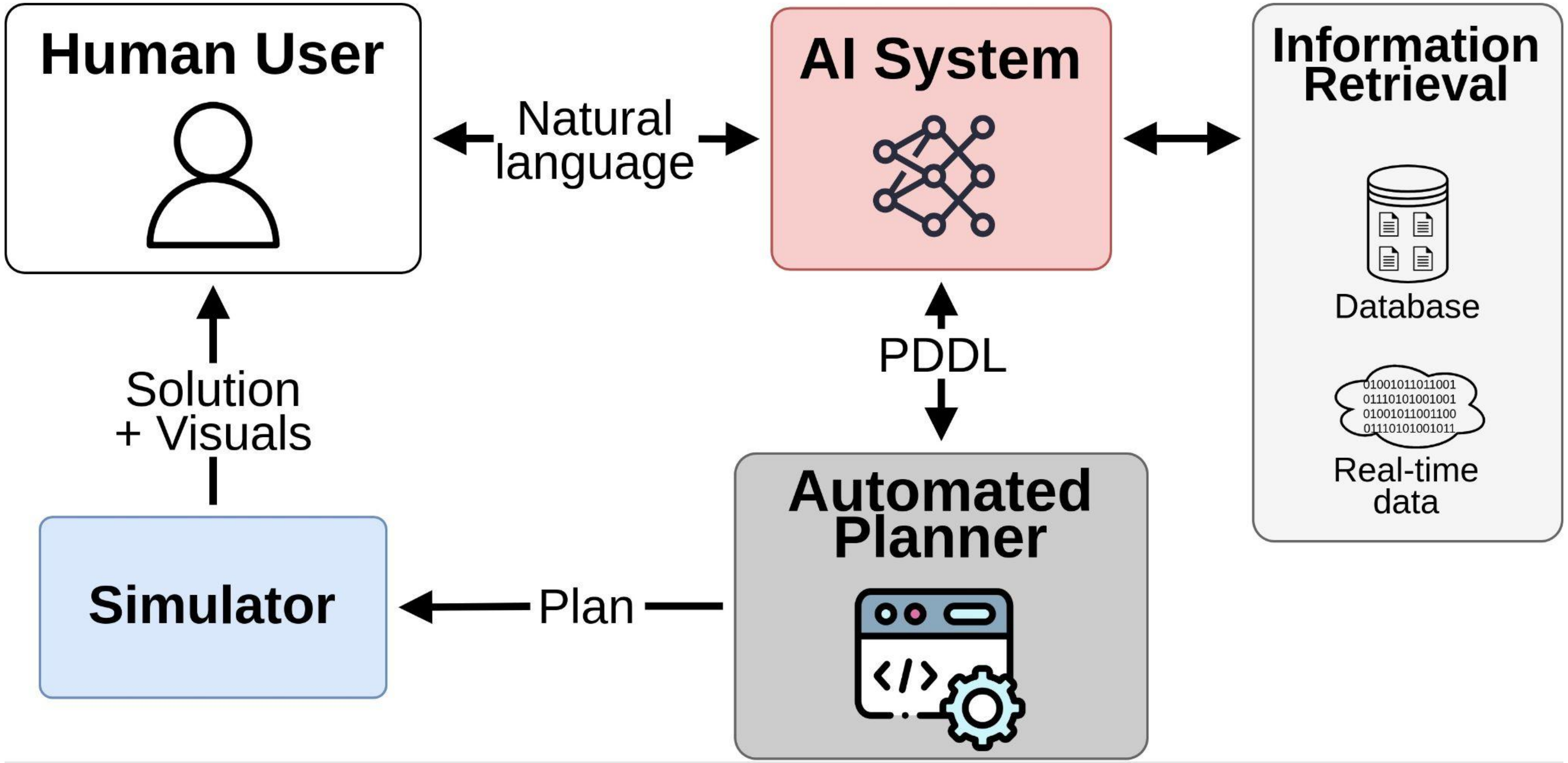
LLM can act as a **bridge** to **improve accessibility**

Our motivation

Improve planning accessibility
to better
leverage human expertise and intuition

- **Avoid** “intuitively bad solutions” and **focus** on “promising directions”.
- **Explore** specific strategies in a “Let’s try this and rollback” approach.
- Dynamically **refine** solutions and **iterate** toward more effective outcomes.

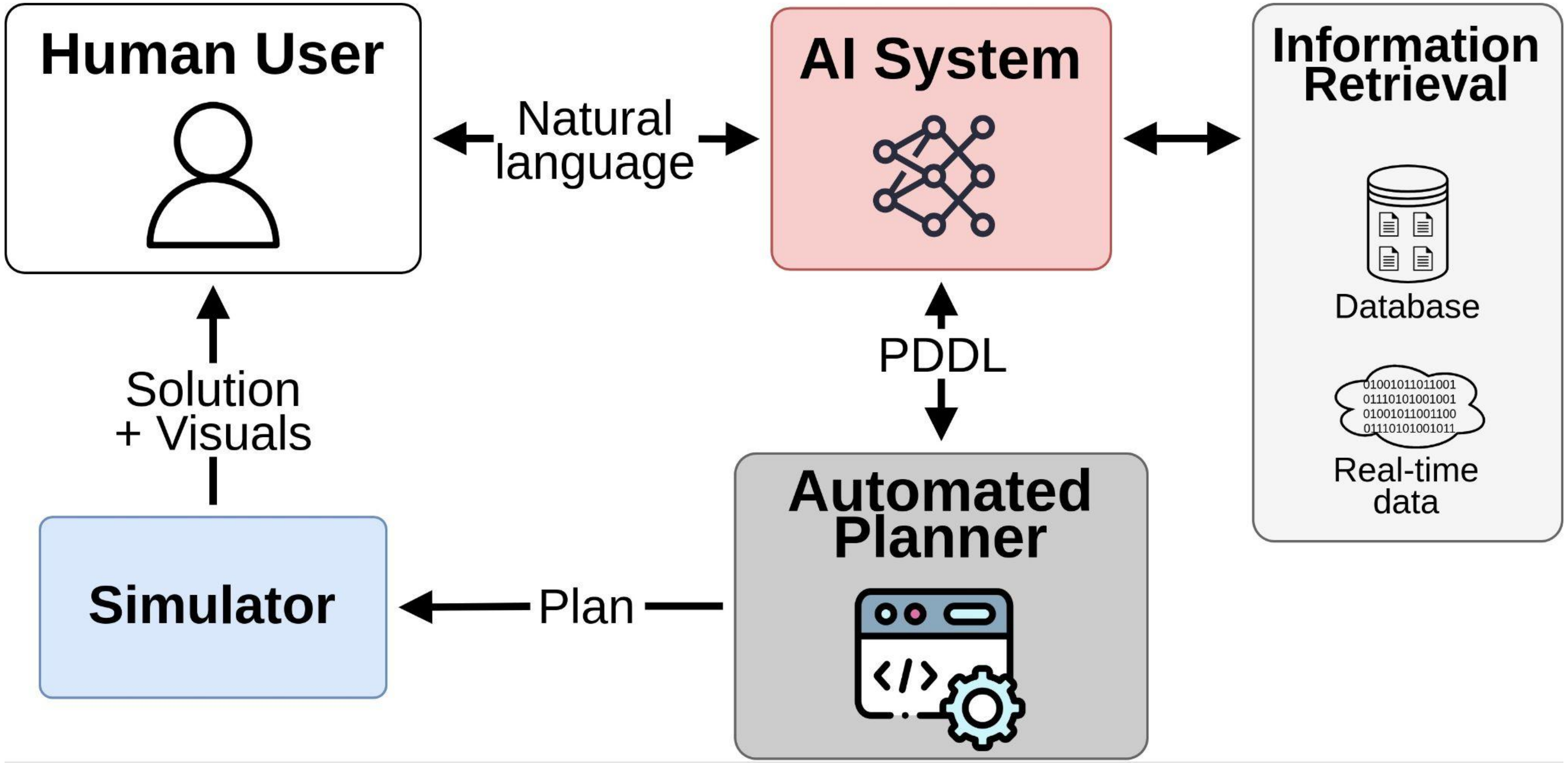
Contribution: Hybrid Collaborative Planning Framework



A **symbolic** automated planner computes plans

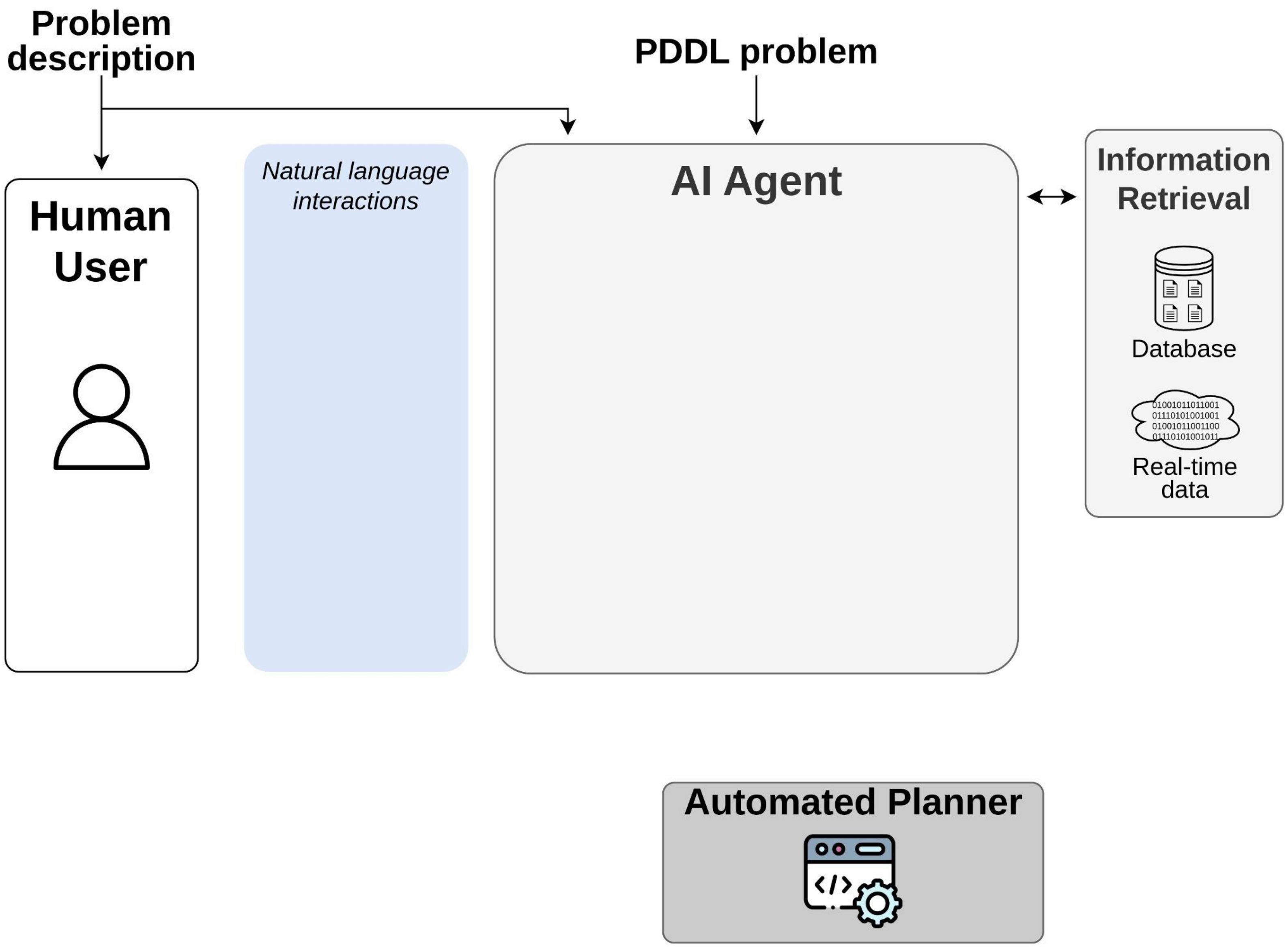
An **LLM-based system** acts as an **interface** and for model elicitation

Contribution: Hybrid Collaborative Planning Framework



Human can **influence** problem solving,
without technical expertise requirements

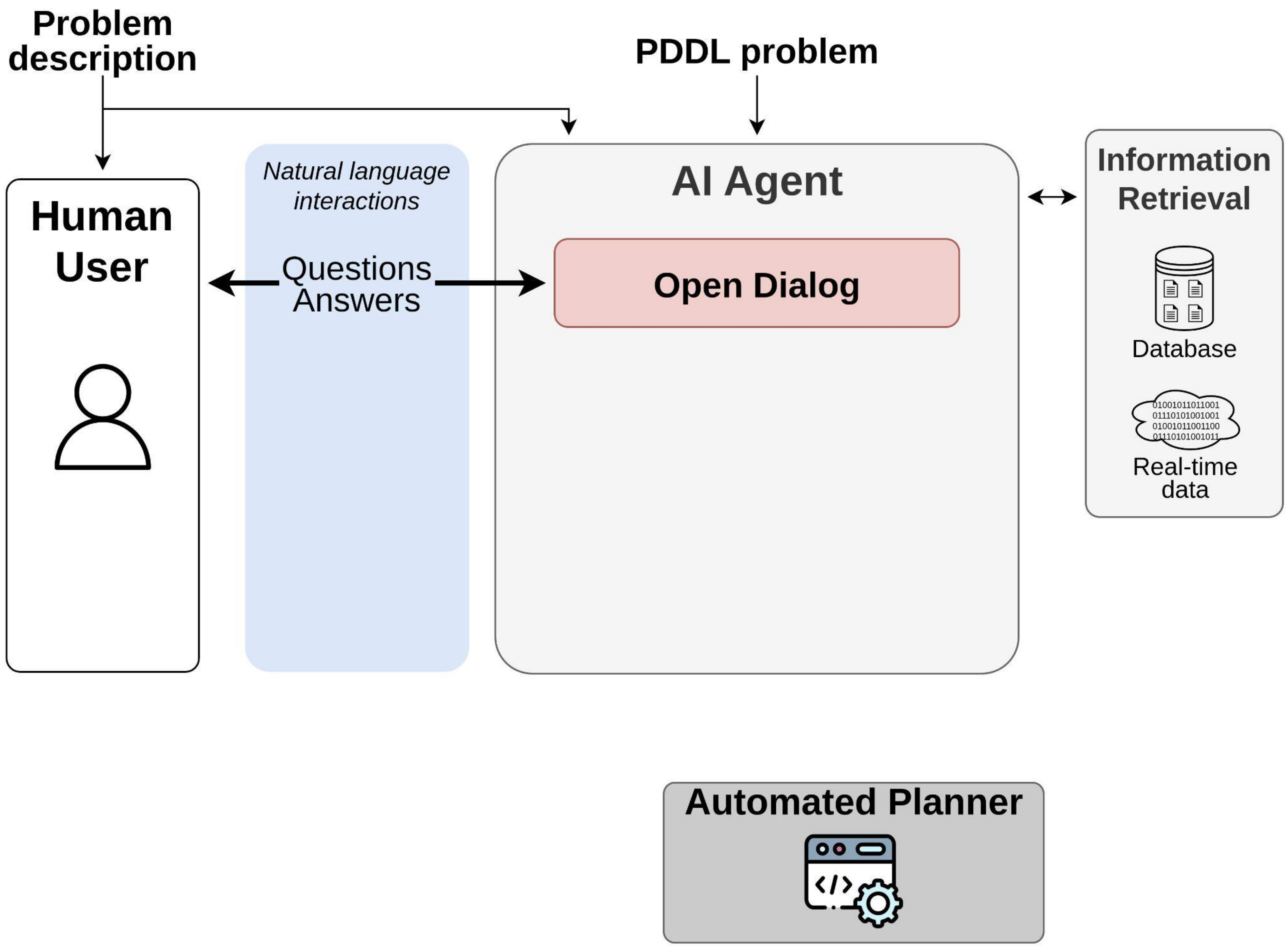
Main capabilities: Chat, Suggestions, Translation



Main capabilities: **Chat,** Suggestions, Translation

Chat

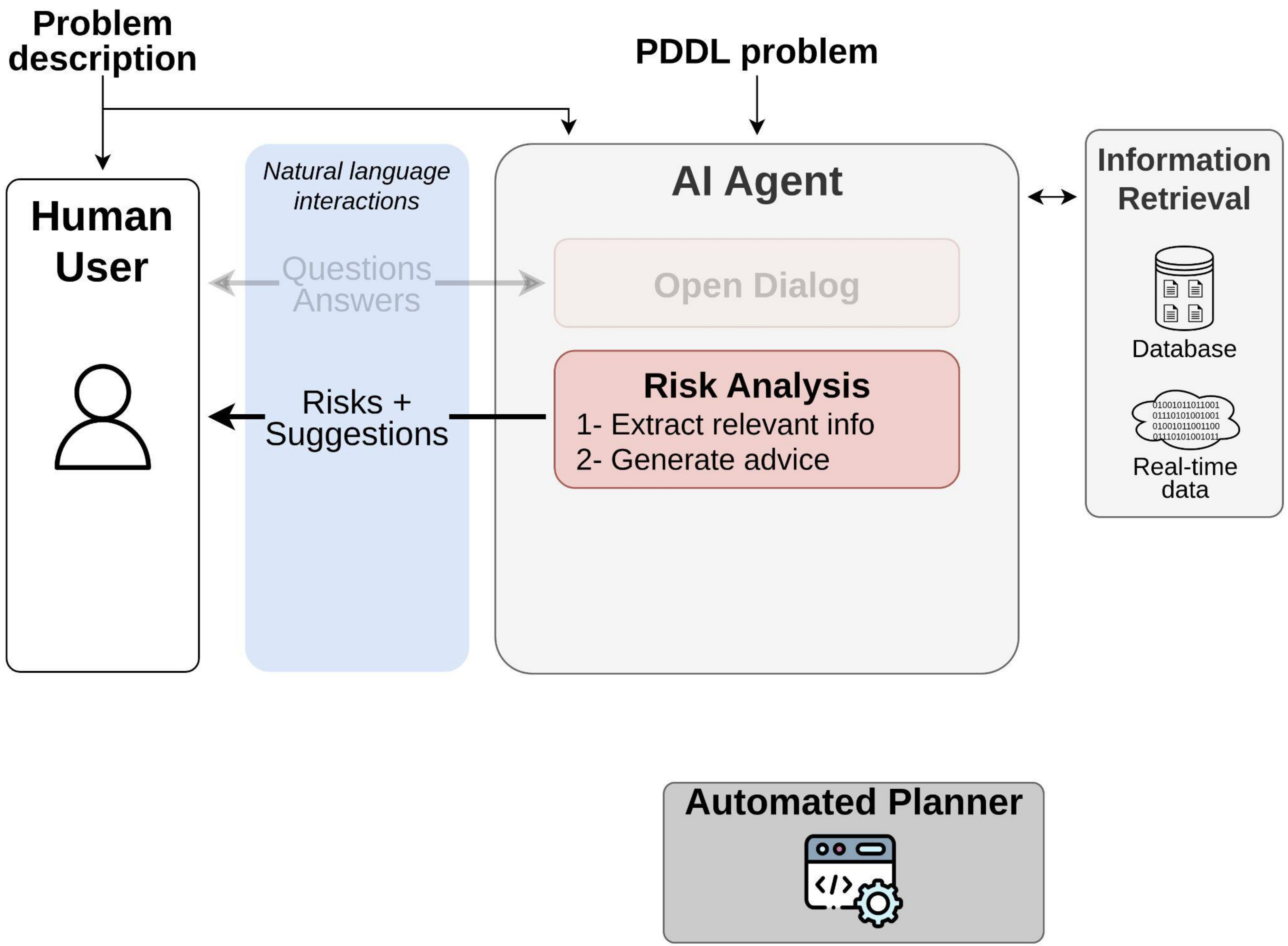
- Get insight on the problem
- Summarize problem
- Modify existing plans
- **No PDDL** for user



Main capabilities: Chat, **Suggestions,** Translation

Highlight information, Make suggestions

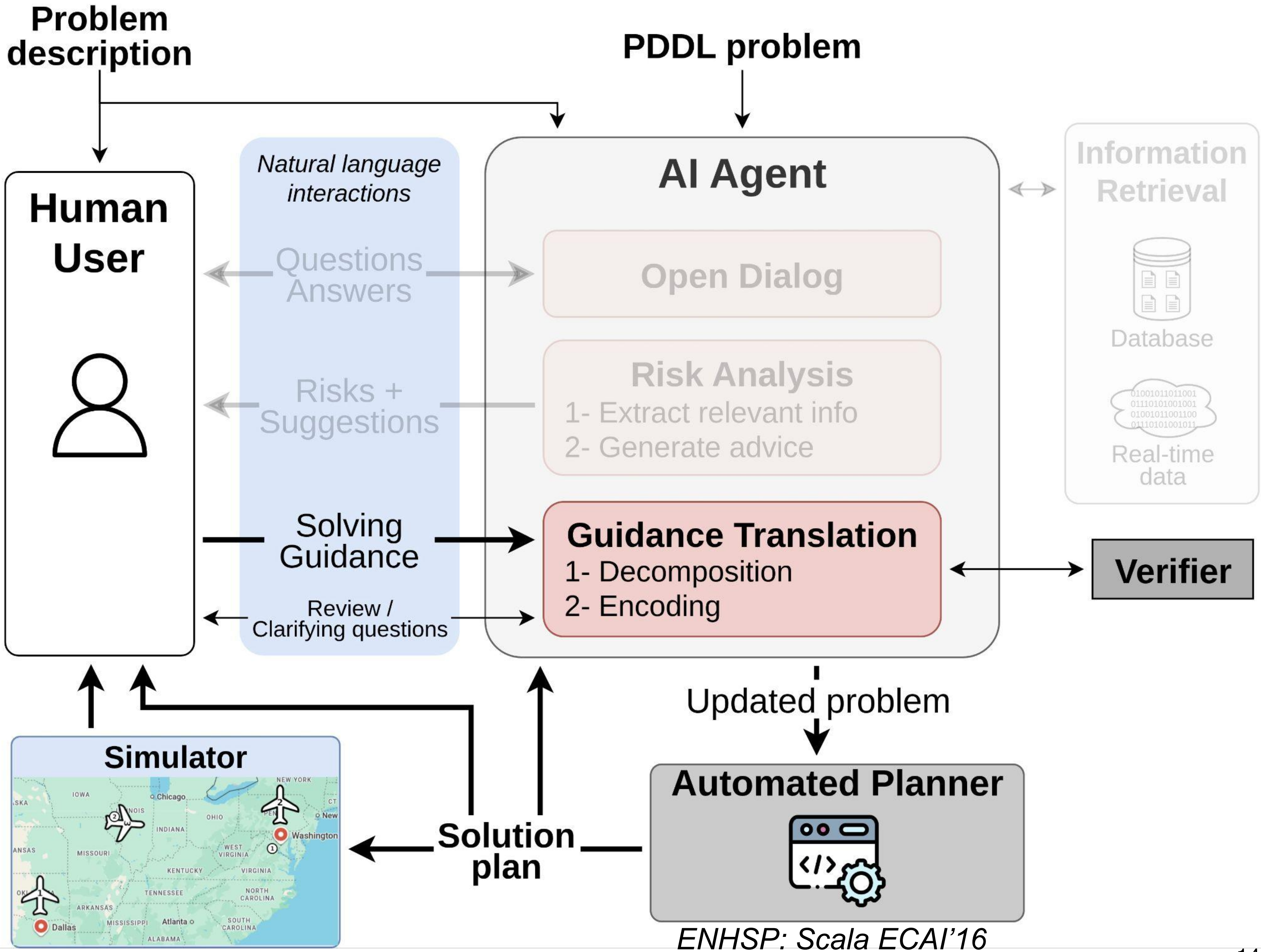
- Retrieve **external** information (RAG) and real-time APIs.
- Can generate real-time **weather** constraints, not modeled in original PDDL problem



Main capabilities: Chat, Suggestions, Translation

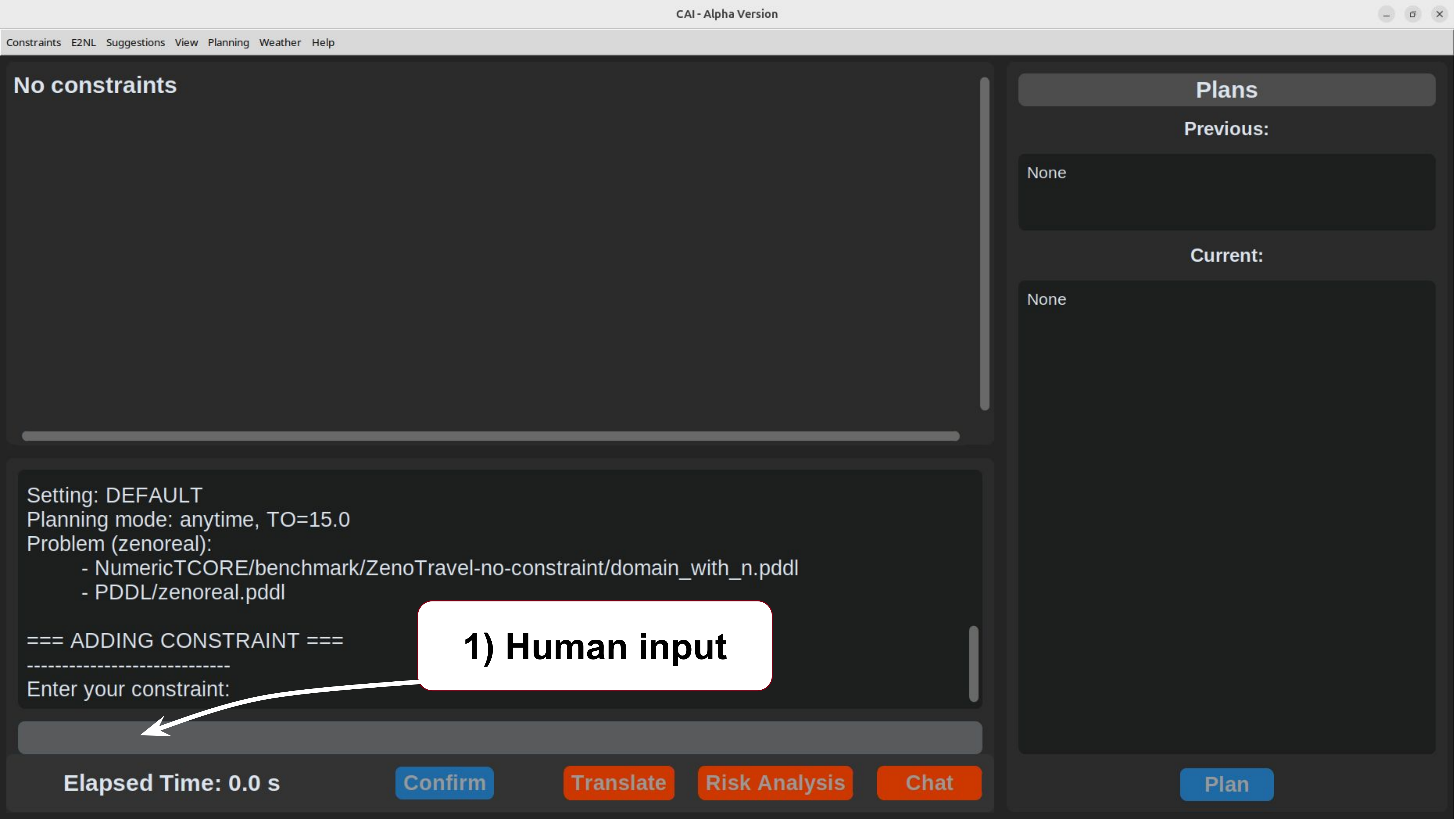
Main contribution:
Planning + Translation

- **Translate** human guidance into planning constraints
- The updated problem is **solved** by **symbolic** planner
- **Simulator** to **visualize** plan

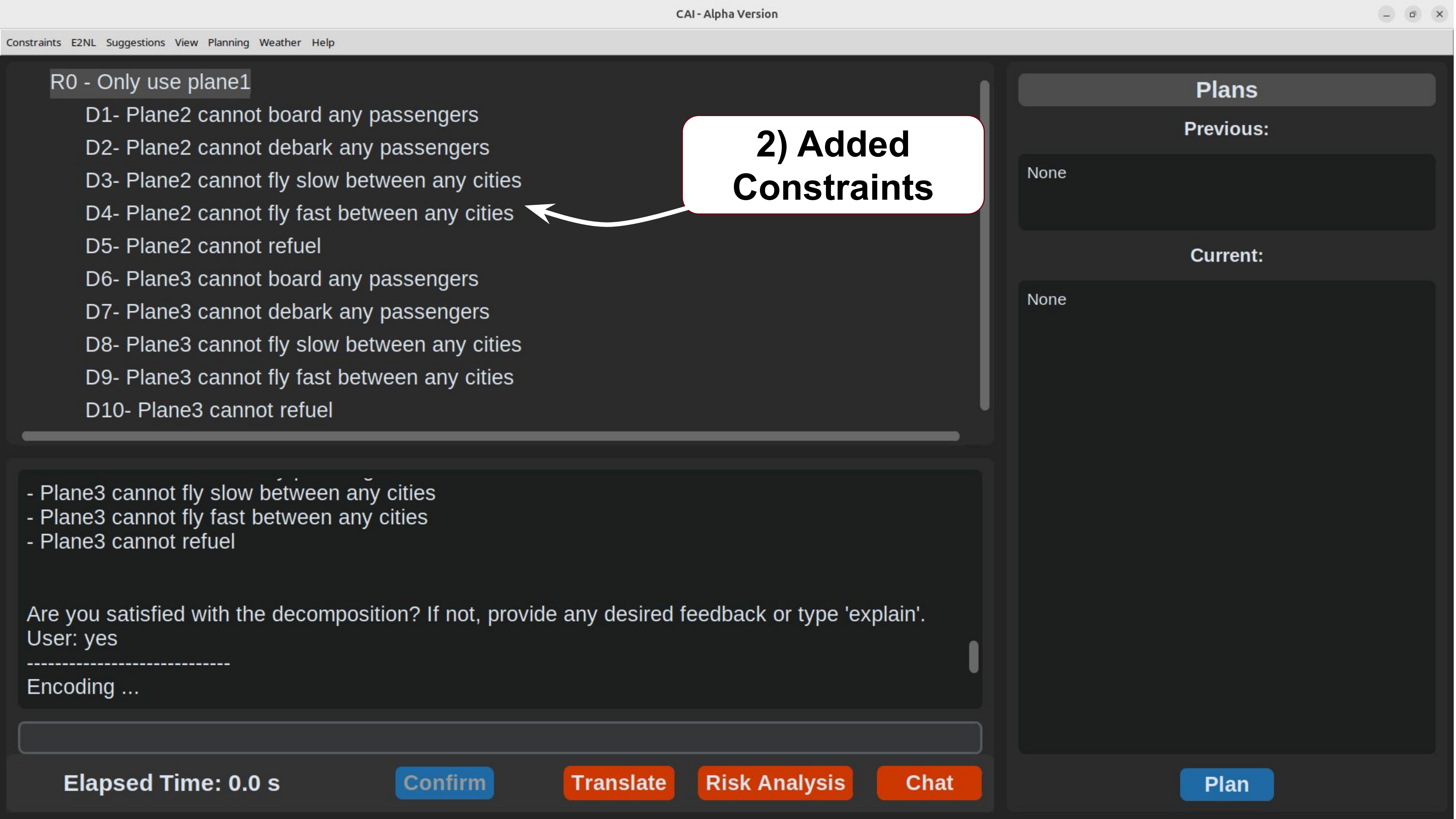


PDSim: Pellegrin ICAPS'24

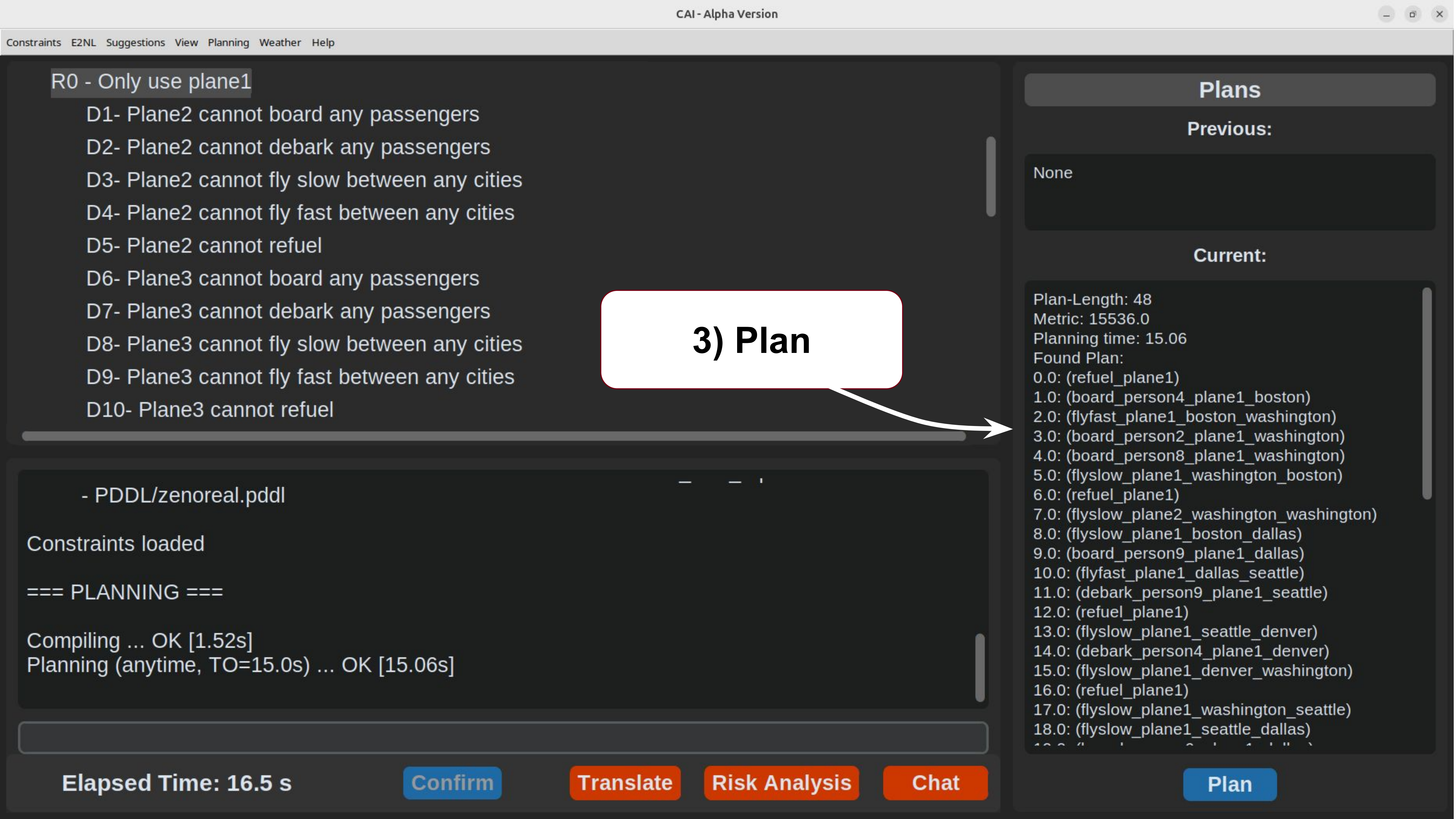
ENHSP: Scala ECAI'16
NTCORE+: Bonassi AAAI'24

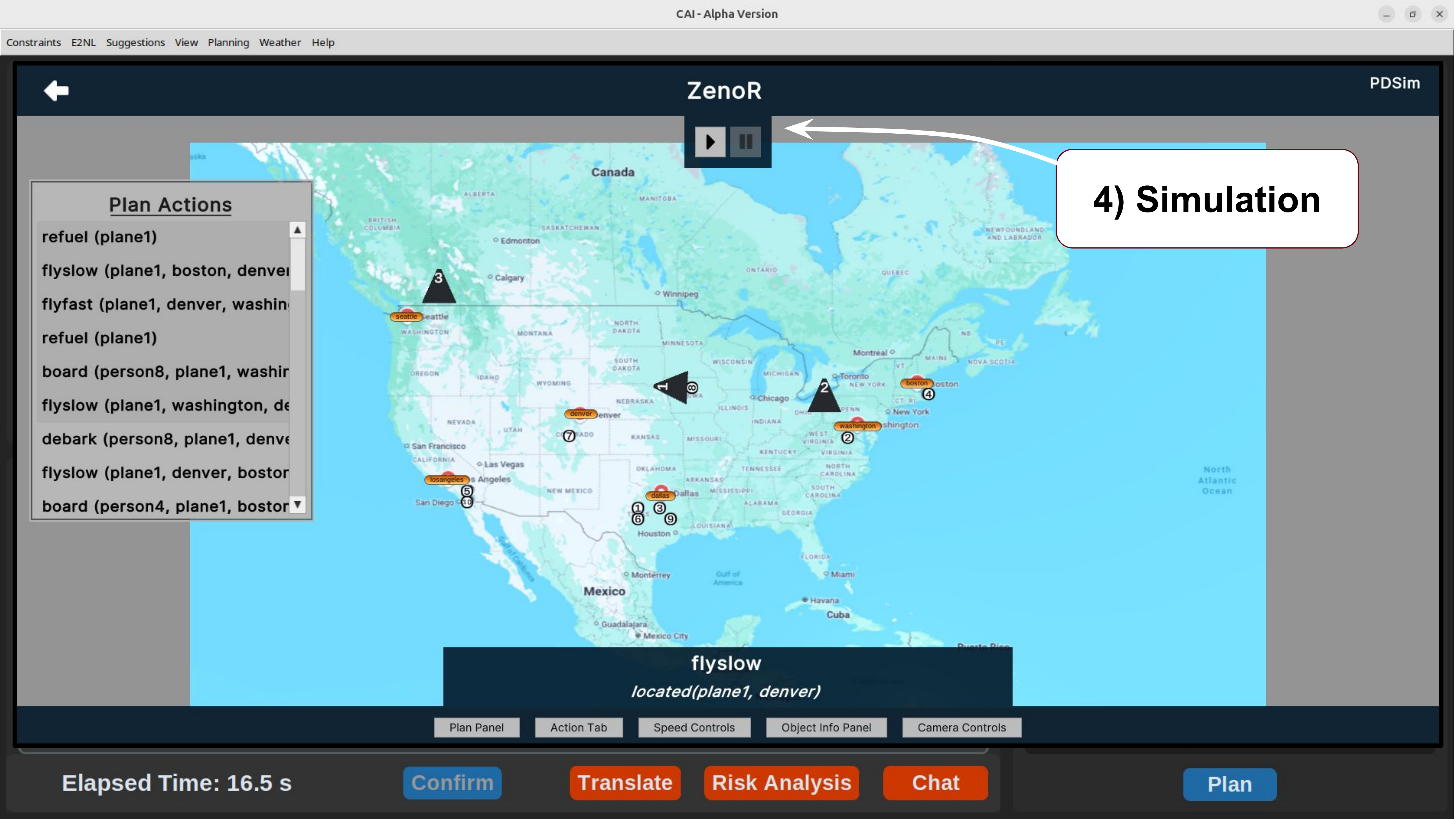

CAI - Alpha Version
Constraints E2NL Suggestions View Planning Weather Help
No constraints
Plans
Previous:
None
Current:
None
Setting: DEFAULT
Planning mode: anytime, TO=15.0
Problem (zenoreal):
- NumericTCORE/benchmark/ZenoTravel-no-constraint/domain_with_n.pddl
- PDDL/zenoreal.pddl
=== ADDING CONSTRAINT ===

Enter your constraint:
1) Human input
Elapsed Time: 0.0 s
Confirm
Translate
Risk Analysis
Chat
Plan


CAI - Alpha Version
Constraints E2NL Suggestions View Planning Weather Help
R0 - Only use plane1
D1- Plane2 cannot board any passengers
D2- Plane2 cannot debark any passengers
D3- Plane2 cannot fly slow between any cities
D4- Plane2 cannot fly fast between any cities
D5- Plane2 cannot refuel
D6- Plane3 cannot board any passengers
D7- Plane3 cannot debark any passengers
D8- Plane3 cannot fly slow between any cities
D9- Plane3 cannot fly fast between any cities
D10- Plane3 cannot refuel
2) Added Constraints
- Plane3 cannot fly slow between any cities
- Plane3 cannot fly fast between any cities
- Plane3 cannot refuel
Are you satisfied with the decomposition? If not, provide any desired feedback or type 'explain'.
User: yes

Encoding ...
Elapsed Time: 0.0 s
Confirm
Translate
Risk Analysis
Chat
Plans
Previous:
None
Current:
None
Plan

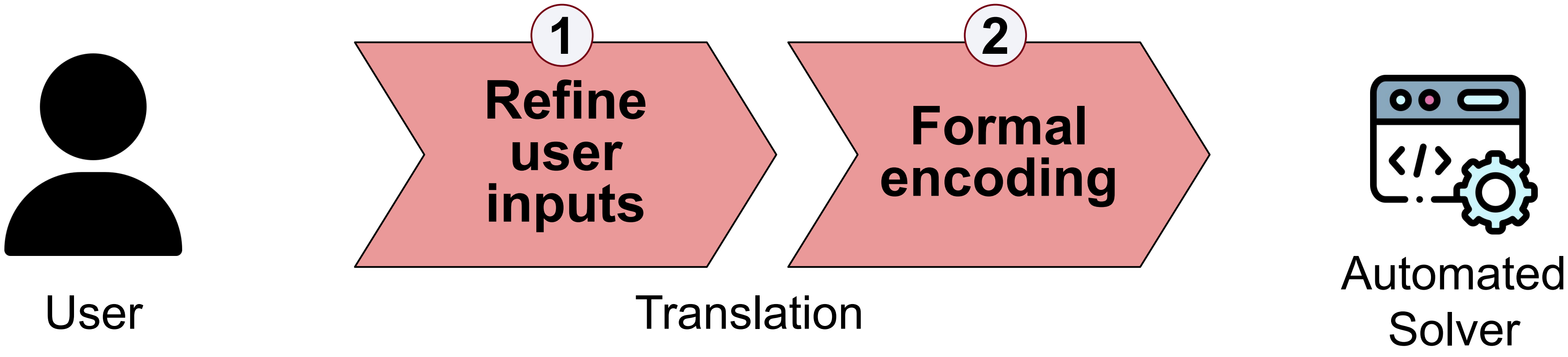

CAI - Alpha Version
Constraints E2NL Suggestions View Planning Weather Help
R0 - Only use plane1
D1- Plane2 cannot board any passengers
D2- Plane2 cannot debark any passengers
D3- Plane2 cannot fly slow between any cities
D4- Plane2 cannot fly fast between any cities
D5- Plane2 cannot refuel
D6- Plane3 cannot board any passengers
D7- Plane3 cannot debark any passengers
D8- Plane3 cannot fly slow between any cities
D9- Plane3 cannot fly fast between any cities
D10- Plane3 cannot refuel
3) Plan
- PDDL/zenoreal.pddl
Constraints loaded
=== PLANNING ===
Compiling ... OK [1.52s]
Planning (anytime, TO=15.0s) ... OK [15.06s]
Elapsed Time: 16.5 s
Confirm
Translate
Risk Analysis
Chat
Plans
Previous:
None
Current:
Plan-Length: 48
Metric: 15536.0
Planning time: 15.06
Found Plan:
0.0: (refuel_plane1)
1.0: (board_person4_plane1_boston)
2.0: (flyfast_plane1_boston_washington)
3.0: (board_person2_plane1_washington)
4.0: (board_person8_plane1_washington)
5.0: (flyslow_plane1_washington_boston)
6.0: (refuel_plane1)
7.0: (flyslow_plane2_washington_washington)
8.0: (flyslow_plane1_boston_dallas)
9.0: (board_person9_plane1_dallas)
10.0: (flyfast_plane1_dallas_seattle)
11.0: (debark_person9_plane1_seattle)
12.0: (refuel_plane1)
13.0: (flyslow_plane1_seattle_denver)
14.0: (debark_person4_plane1_denver)
15.0: (flyslow_plane1_denver_washington)
16.0: (refuel_plane1)
17.0: (flyslow_plane1_washington_seattle)
18.0: (flyslow_plane1_seattle_dallas)
Plan


CAI - Alpha Version
Constraints E2NL Suggestions View Planning Weather Help
ZenoR
PDSim
4) Simulation
Plan Actions
refuel (plane1)
flyslow (plane1, boston, denve
flyfast (plane1, denver, washin
refuel (plane1)
board (person8, plane1, washir
flyslow (plane1, washington, de
debark (person8, plane1, denve
flyslow (plane1, denver, bostor
board (person4, plane1, bostor
Canada
Mexico
Cuba
flyslow
located(plane1, denver)
Plan Panel
Action Tab
Speed Controls
Object Info Panel
Camera Controls
Elapsed Time: 16.5 s
Confirm
Translate
Risk Analysis
Chat
Plan

Guidance Translation

Translate user inputs as **guidance** for the **solver**

Two-step process:



Guidance Translation

1

Refinements of user inputs

Example: “*Only use robot2*”

– Simple but **not straightforward**: must be rephrased to “*Never use robot1*”
– Must **clarify** what “*using a robot*” means
– Planner only supports **state-based** constraints: can't directly constrain actions
– Refined inputs:
  ○ *“robot1 must always be located at initial location”*
  ○ *“robot1 tools must always be turned off”*

- Ask **clarifying questions**
- **Decompose** user input into independent, simpler **sub-constraints**
- **Rephrase** to match problem **characteristics**
- User **reviews** decomposition to **identify** any **misinterpretation**

Guidance Translation

2

Formal encoding

- **Parallel translation** of each sub-constraint into PDDL3
- Leverage **automated symbolic verifier** checking **syntax**
- **Back translation**: each encoded constraints are translated back into natural language for human review

- Translated constraints are **added** to the system and can be **activated** and **combined** at user's discretion
- All **activated constraints** are considered when **planning**

Evaluation of translation quality: Ablation Study

4 Settings to evaluate our translation pipeline:

ECODING: LLM alone

+ VERIFIER: Symbolic syntax checker

+ DECOMP: Constraint decomposition

+ HUMAN: Human interventions on decomposition

Evaluation of translation quality: Ablation Study

Model:
Claude Sonnet 4
(thinking enabled)

Setting	Translation			Human interventions
	Parsable	Correct	Time (s)	
Encoding	26	19	29.3 ± 12.3	0
+ Verifier	30	20	35.8 ± 13.5	0
+ Decomposition				
+ Human				

Table 1: Ablation study reporting syntax and semantic accuracy ($N = 30$)

Correct Syntax

- LLM alone makes syntax mistakes
- Symbolic verifier feedback fixes syntax mistakes

Evaluation of translation quality: Ablation Study

Model:
Claude Sonnet 4
(thinking enabled)

Setting	Translation			Human interventions
	Parsable	Correct	Time (s)	
Encoding	26	19	29.3 ± 12.3	0
+ Verifier	30	20	35.8 ± 13.5	0
+ Decomposition	30	20	55.0 ± 26.2	0
+ Human	30	27	81.9 ± 53.7	12

Table 1: Ablation study reporting syntax and semantic accuracy ($N = 30$)

Satisfying semantic accuracy

- Decomposition no direct effect
- But allows for human review
- Human intervention significantly improves correctness

Evaluation of translation quality: Ablation Study

Model:
Claude Sonnet 4
(thinking enabled)

Setting	Translation			Human interventions
	Parsable	Correct	Time (s)	
Encoding	26	19	29.3 ± 12.3	0
+ Verifier	30	20	35.8 ± 13.5	0
+ Decomposition	30	20	55.0 ± 26.2	0
+ Human	30	27	81.9 ± 53.7	12

Table 1: Ablation study reporting syntax and semantic accuracy ($N = 30$)

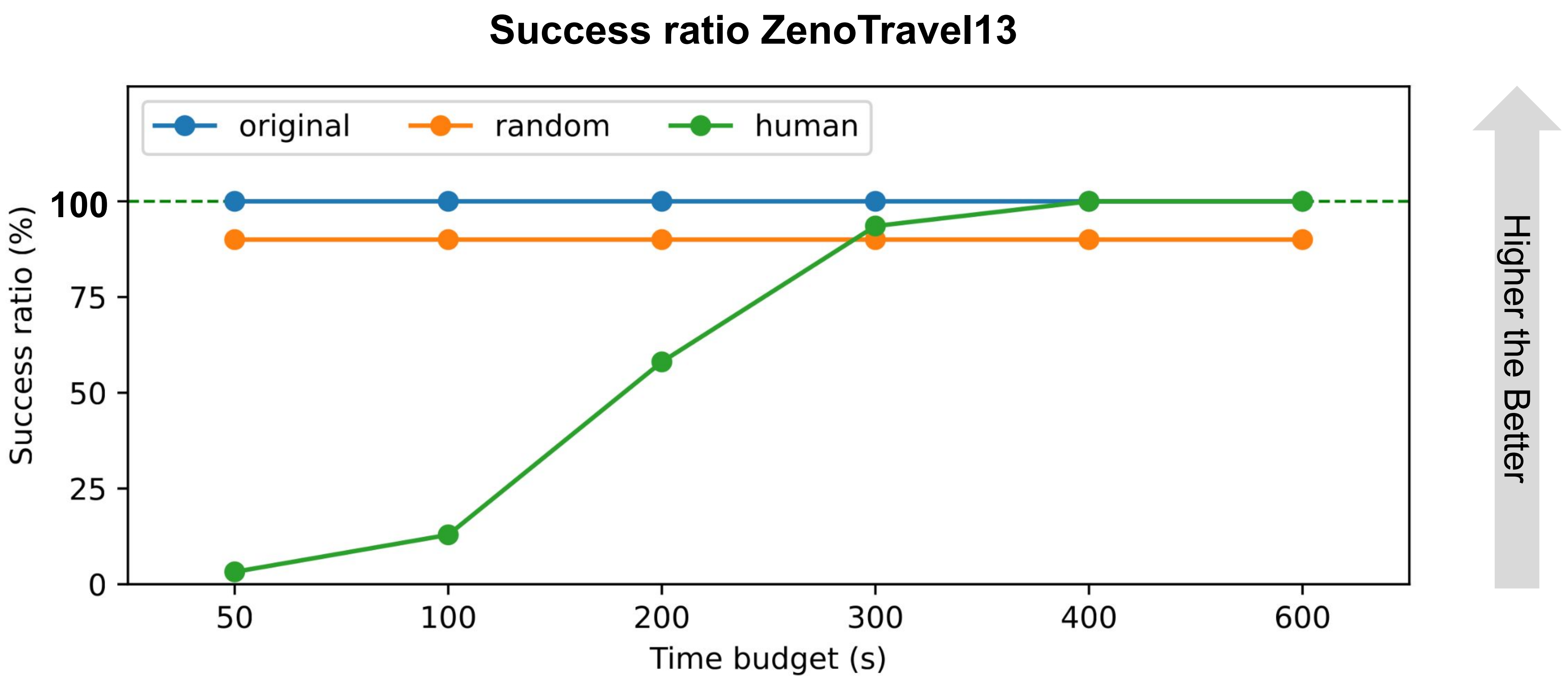
<u>Seems</u> faster than human experts

- Ours ~82s (SD=53.7) vs. Prior work 180s (SD=78)
- But comparison maybe unfair
  - similar but not identical constraints

Effects on plan cost - Experiment Setup

- Baselines:
  - Solving original problem (***original***) (N=10)
  - Using random valid constraints (***random***) + AND/OR combinations (N=30)
- Our approach (***human***):
  - Using relevant complementary constraints + all AND combinations (N=31)
- Use limited time budget from 50s to 600s
  - **Constraints** induce **delays** (translation / compilation), reducing effective planning time
- Measure:
  - Planning success ratio
  - Plan quality / cost (e.g. fuel consumption)

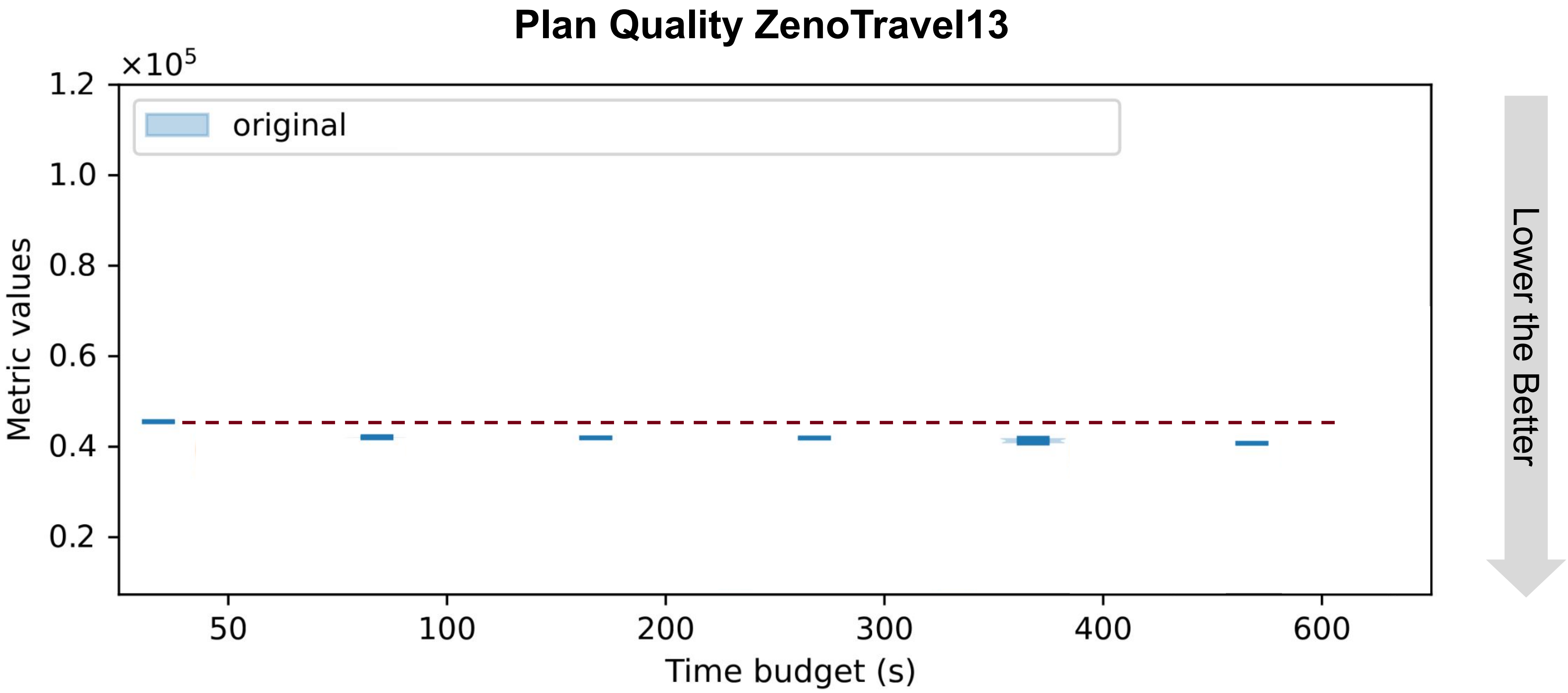
Effects on plan cost - Positive Results



Translation takes time
⇒ reduced effective planning time

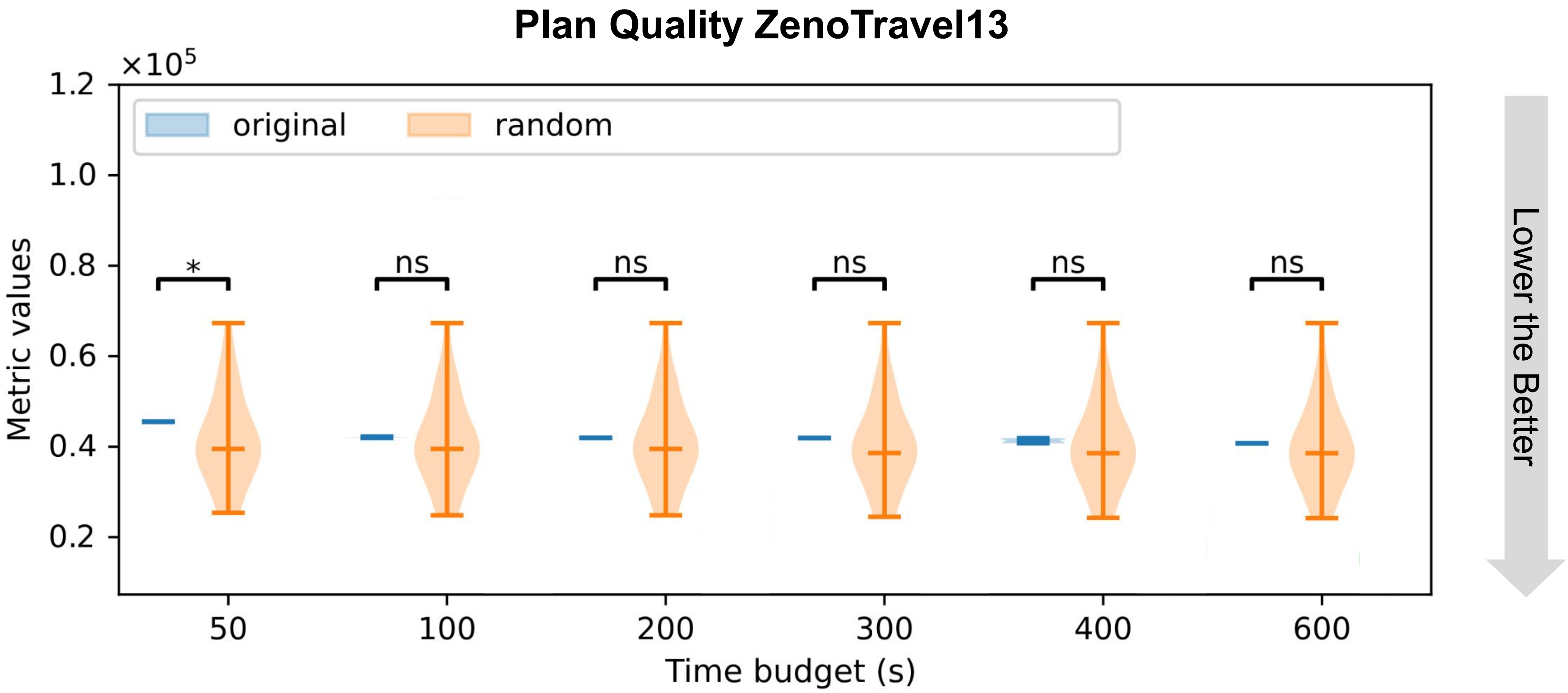
Eventually solves all problems

Effects on plan cost - Positive Results



Original barely improves with time

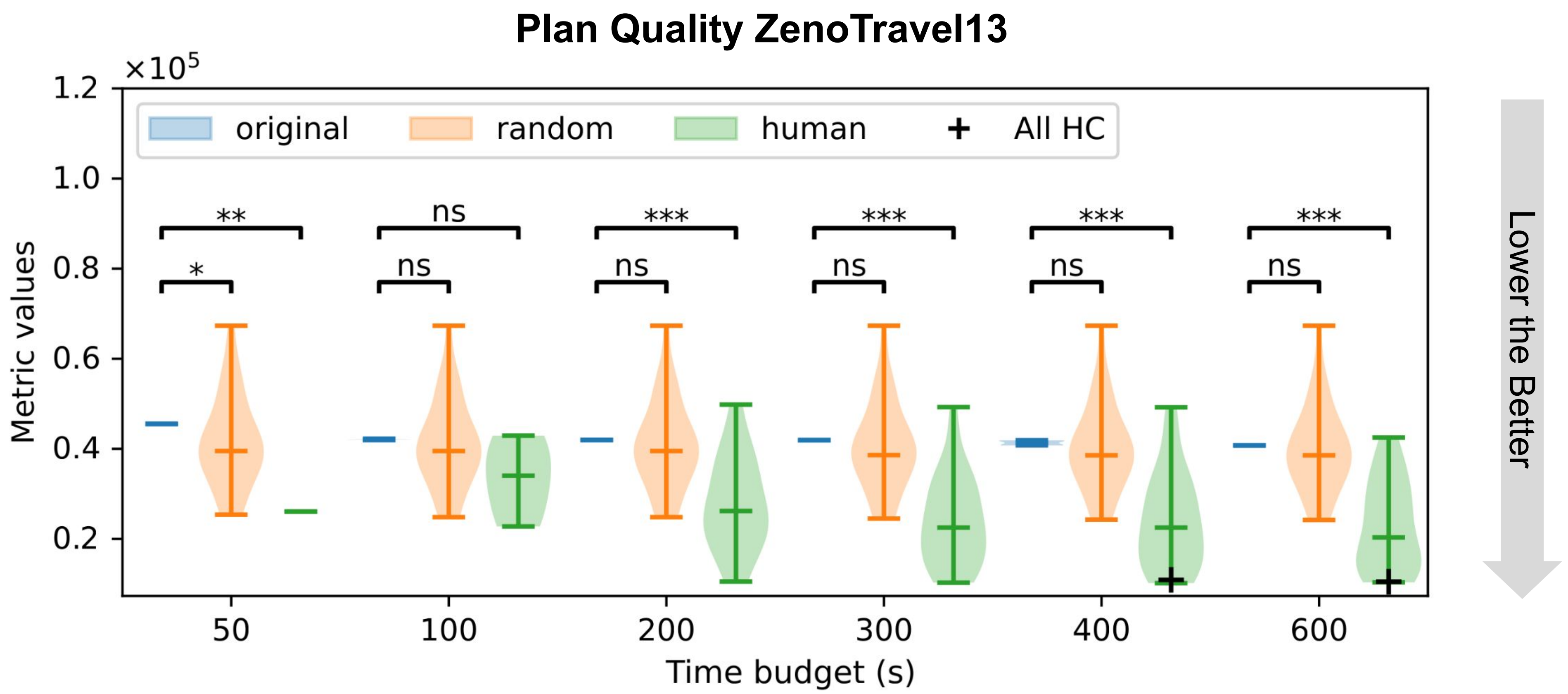
Effects on plan cost - Positive Results



Original barely improves with time

Random constraints have random effects

Effects on plan cost - Positive Results



Original barely improves with time

Random constraints have random effects

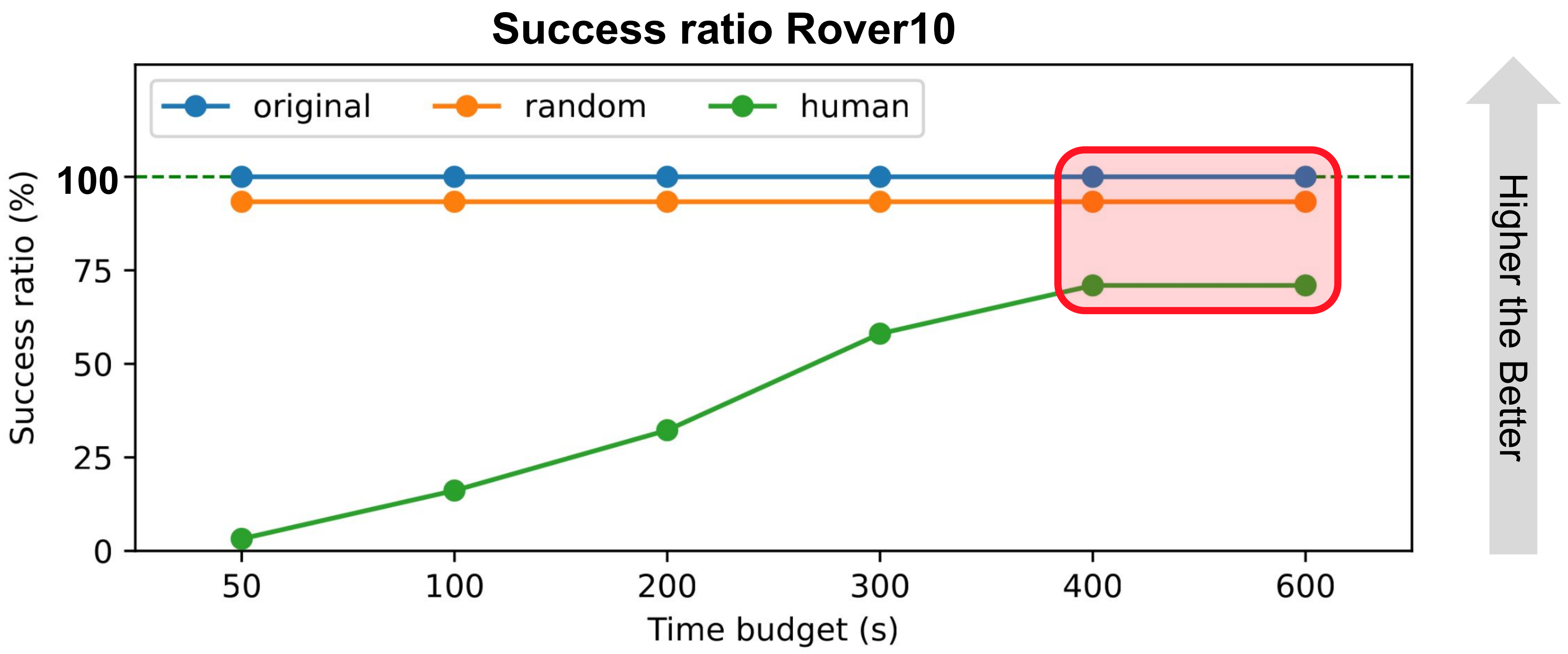
Our approach leads to significant improvements

Effects on plan cost - Negative Results

Some human constraints are worse than original

Possible explanation:
Small effective planning time

Effects on plan cost - Negative Results

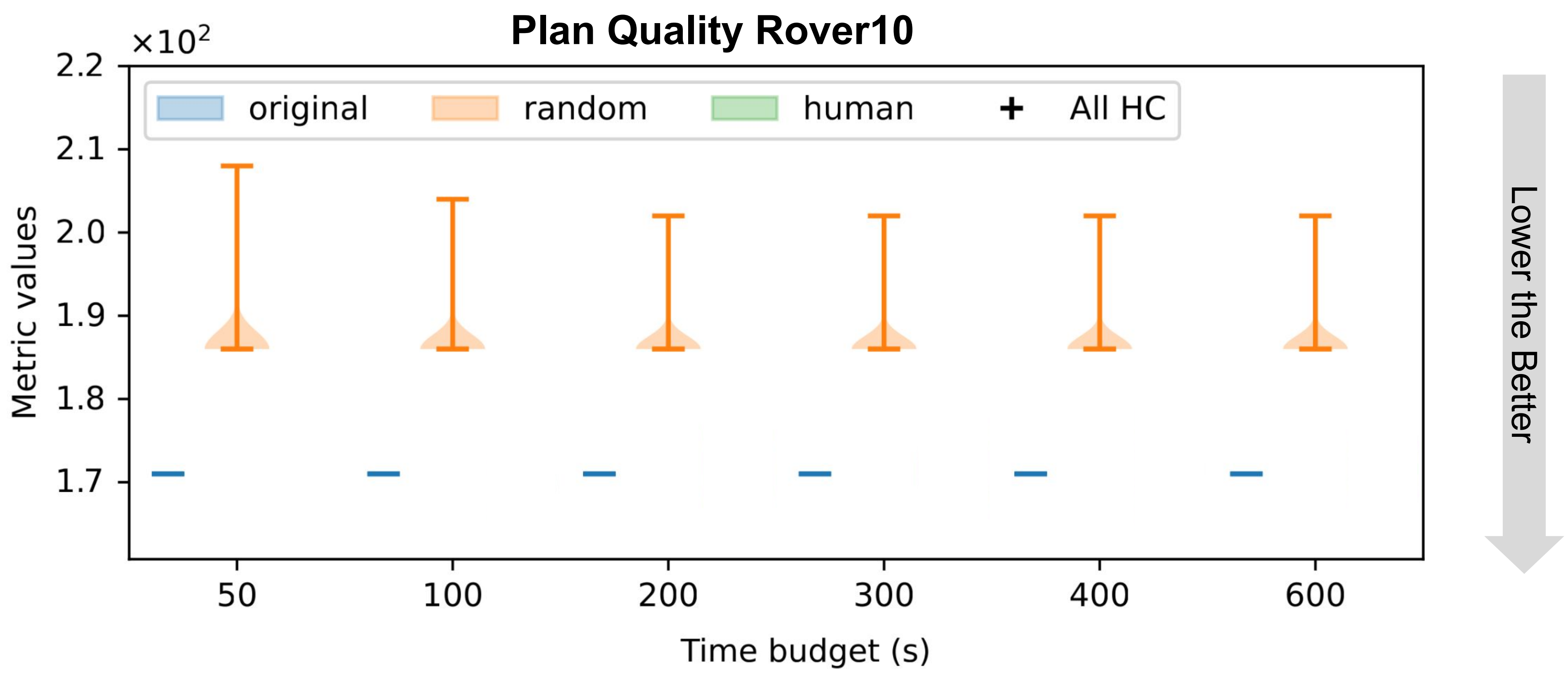


Unexpected lower success results
Planner timeout?

Human constraints are all feasible

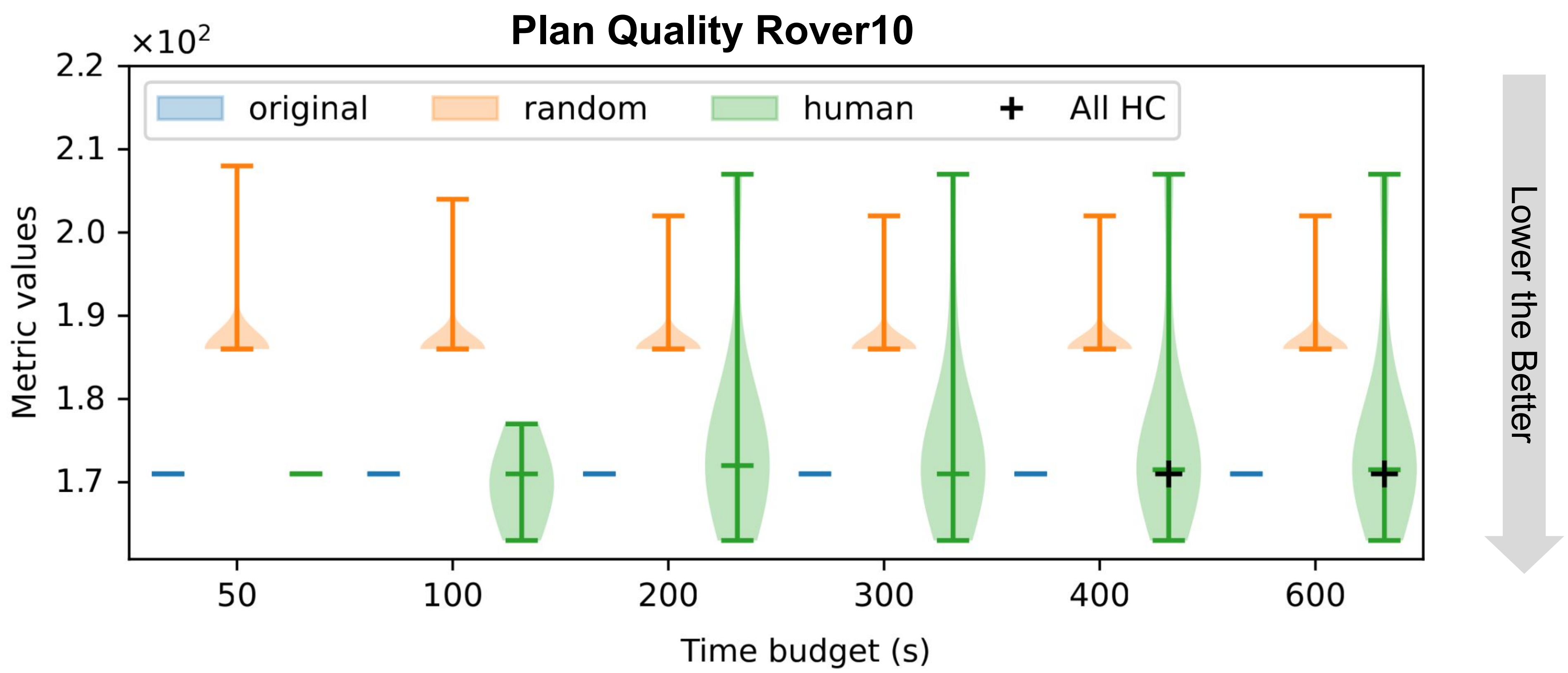
Not effective for all problem

Effects on plan cost - Negative Results



Random constraint are always worse than original

Effects on plan cost - Negative Results



Random constraint are always worse than original

Our approach is often worse for this problem

Discussion

Main assumption is flawed

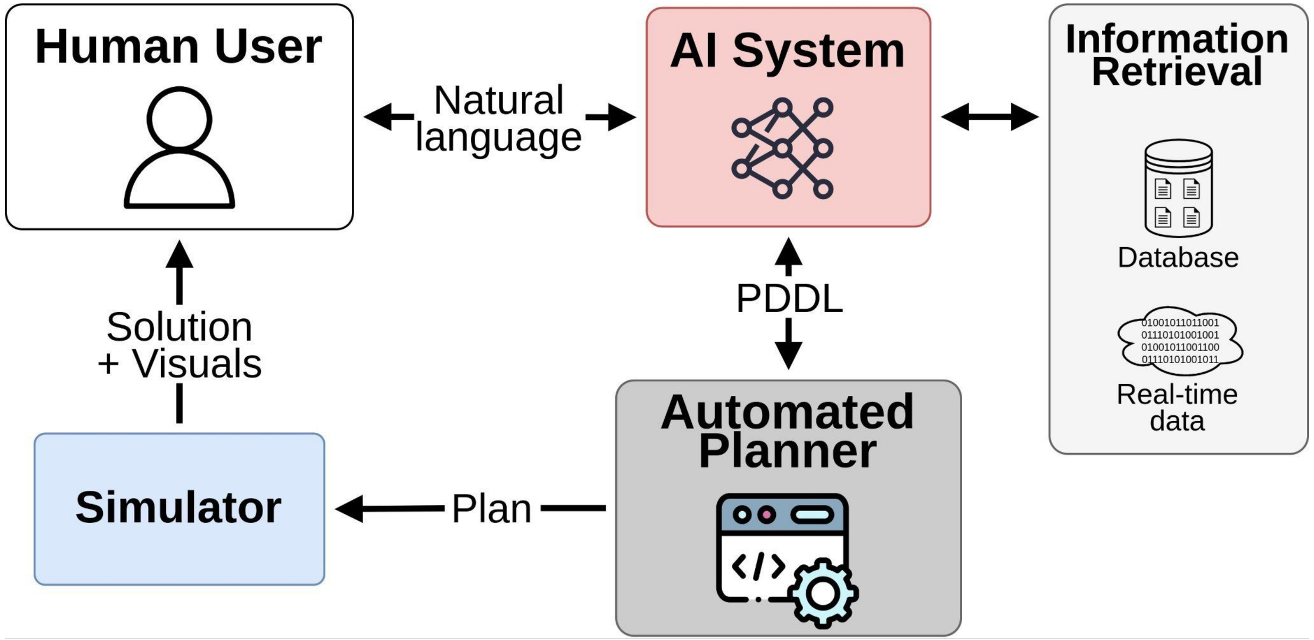
- Trying to reduce the search space with hard constraints, even with “relevant” constraints, does not seem to systematically improve performances in all cases.

Other approaches

- Now looking into other formalisms to better leverage human inputs:
  - soft constraints
  - linear programming

Conclusion

Hybrid collaborative planning framework (neuro-symbolic)



Creates a **collaborative**, **mixed-initiative planning** scheme where the human can influence problem solving, **without** technical expertise requirements

Conclusion

Accessibility

- Never interact with PDDL
- Able to chat and get insights on problem
- Relevant information highlighted

Translation

- Consistent correct syntax
- Improved semantic accuracy
- “Translate faster than technical experts”

Performances

- Translation delays can be worth to do
- But currently not reliable for all problems

Massachusetts Institute of Technology

A Collaborative Numeric Task Planning Framework based on Constraint Translations using LLMs

Q&A

Feel free to reach out!

Anthony Favier: antfav24@mit.edu
Ngoc La: ntmla@mit.edu
Pulkit Verma: pulkitv@mit.edu
Julie A Shah: julie_a_shah@csail.mit.edu

Interactive Robotics Group

Paper PDF

Come see our **Demo!**
"An LLM-powered Collaborative Numeric Task Planning Framework"

Backup Slides

Solution quality: Used constraints and objective

(human) ZenoTravel13 constraints (before AND combinations)

Objective: (:metric minimize (total_fuel_used))

- Only use plane1
- Person7 should never move
- Planes should only fly slowly
- Plane1 should never fly to the same city more than 3 times
- Person1 and person3 should travel together

- human compilation: 7.8s (SD=2.9)
- random compilation: 3.2s (SD=0.5)

(human) Rover10 constraints (before AND combinations)

Objective: (:metric minimize (total-energy-used))

- Rover2 should never be used
- Rover0 should handle soil and rock data from waypoint4
- No rover should ever be in waypoint2 or waypoint5
- Rover1 should take all images
- Waypoint6 should always have the same rock sample

- human compilation: 30.8s (SD=3.2)
- random compilation: 31.3s (SD=3.9)

Evaluation of translation quality: Ablation Study

4 Settings to evaluate our translation pipeline:

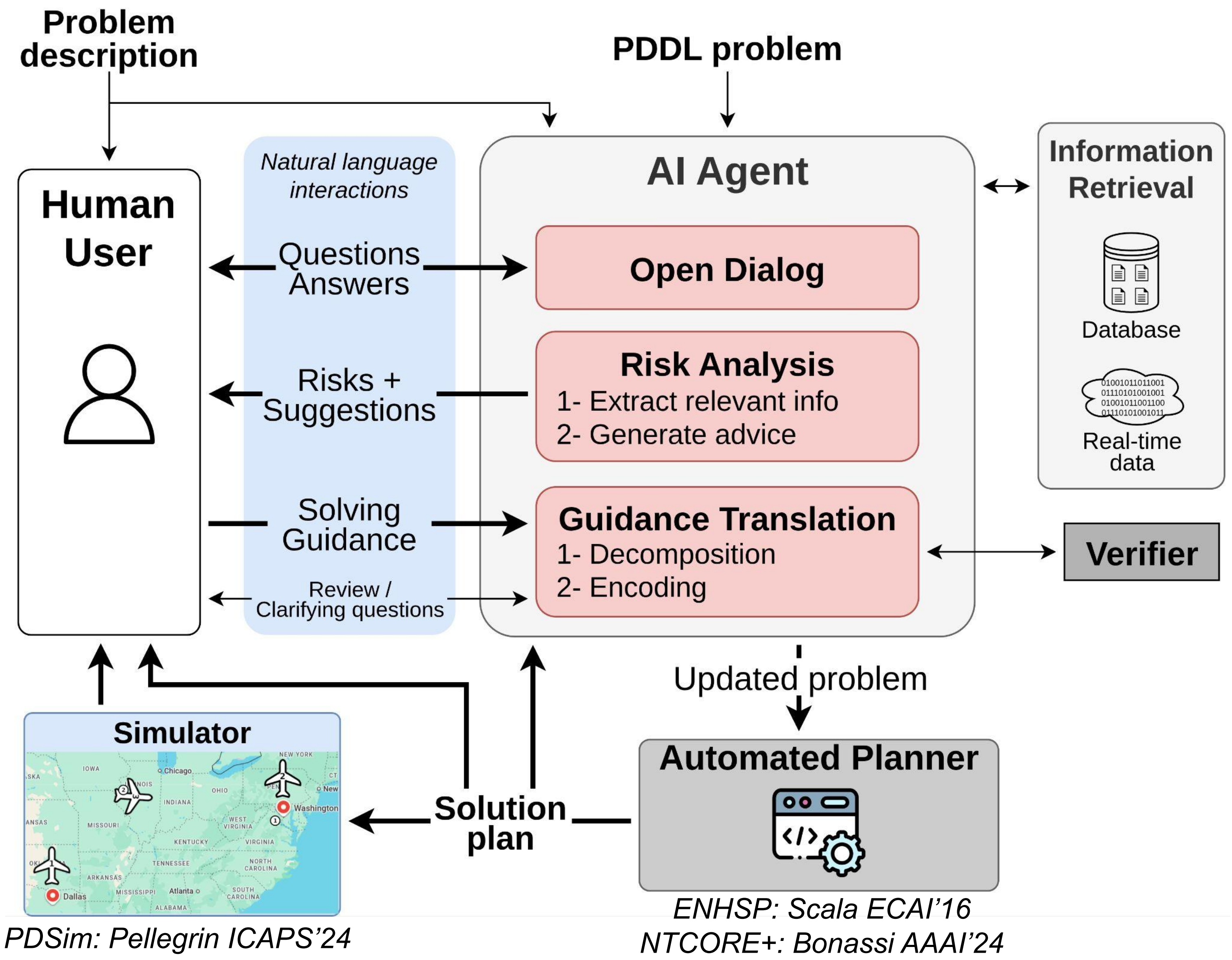
ECODING: LLM alone
+ **VERIFIER**: symbolic syntax checker
+ **DECOMP**: constraint decomposition
+ **HUMAN**: human interventions on decomposition

- 15 predefined constraints for two IPC numerical problems: 8 (ZenoTravel13) + 7 (Rover10)
- Constraints are arbitrarily more or less ambiguous
  - E.g., "*X should always be located at L*" vs. "*Never use X*"
- Run twice for each constraint
- Human interventions were as simple and short as possible.

An LLM-powered Collaborative Numeric Task Planning Framework

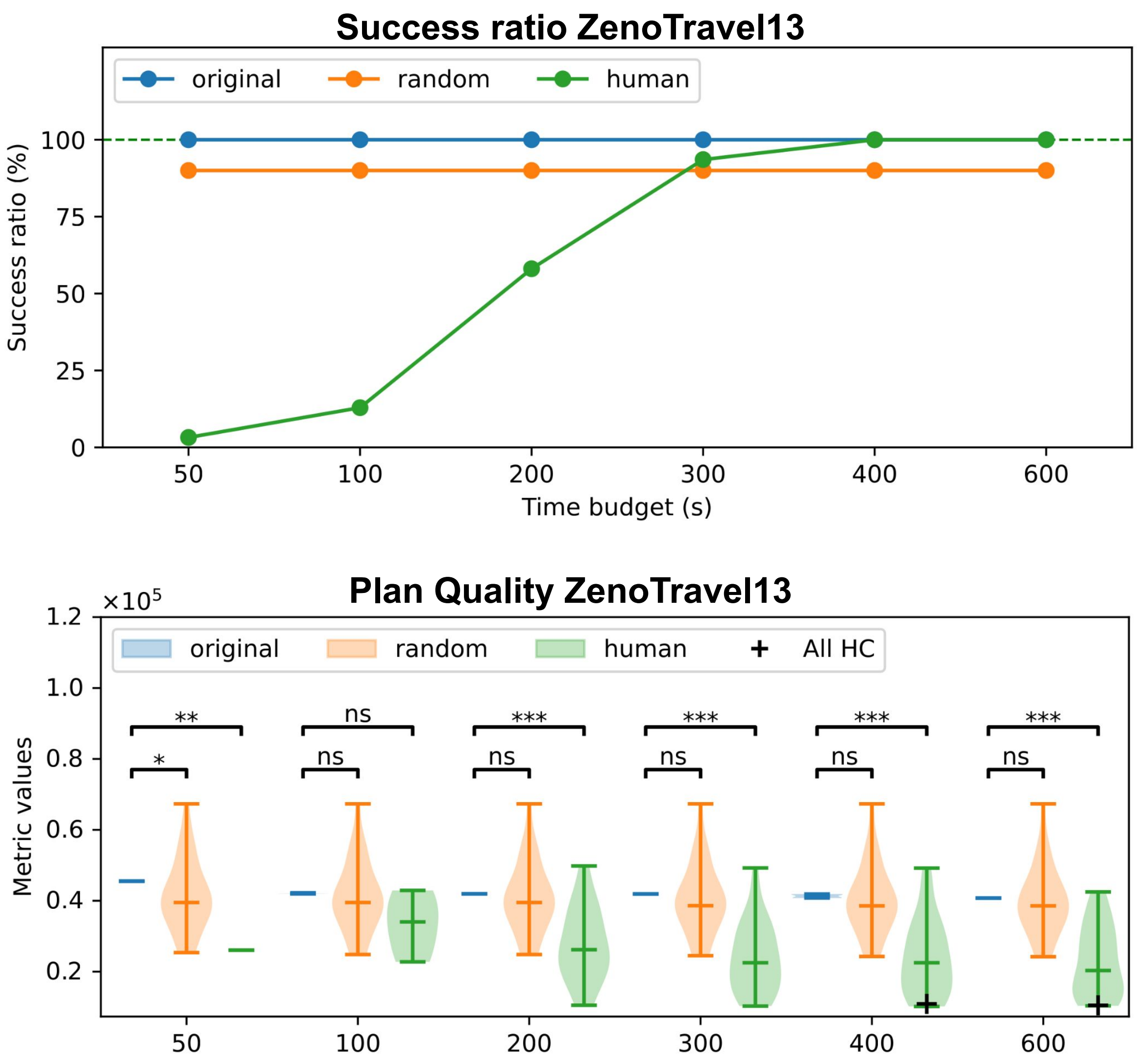
Anthony Favier, Pulkit Verma, Ngoc La, Julie A Shah

Human **guide** can influence planning, **without** technical expertise

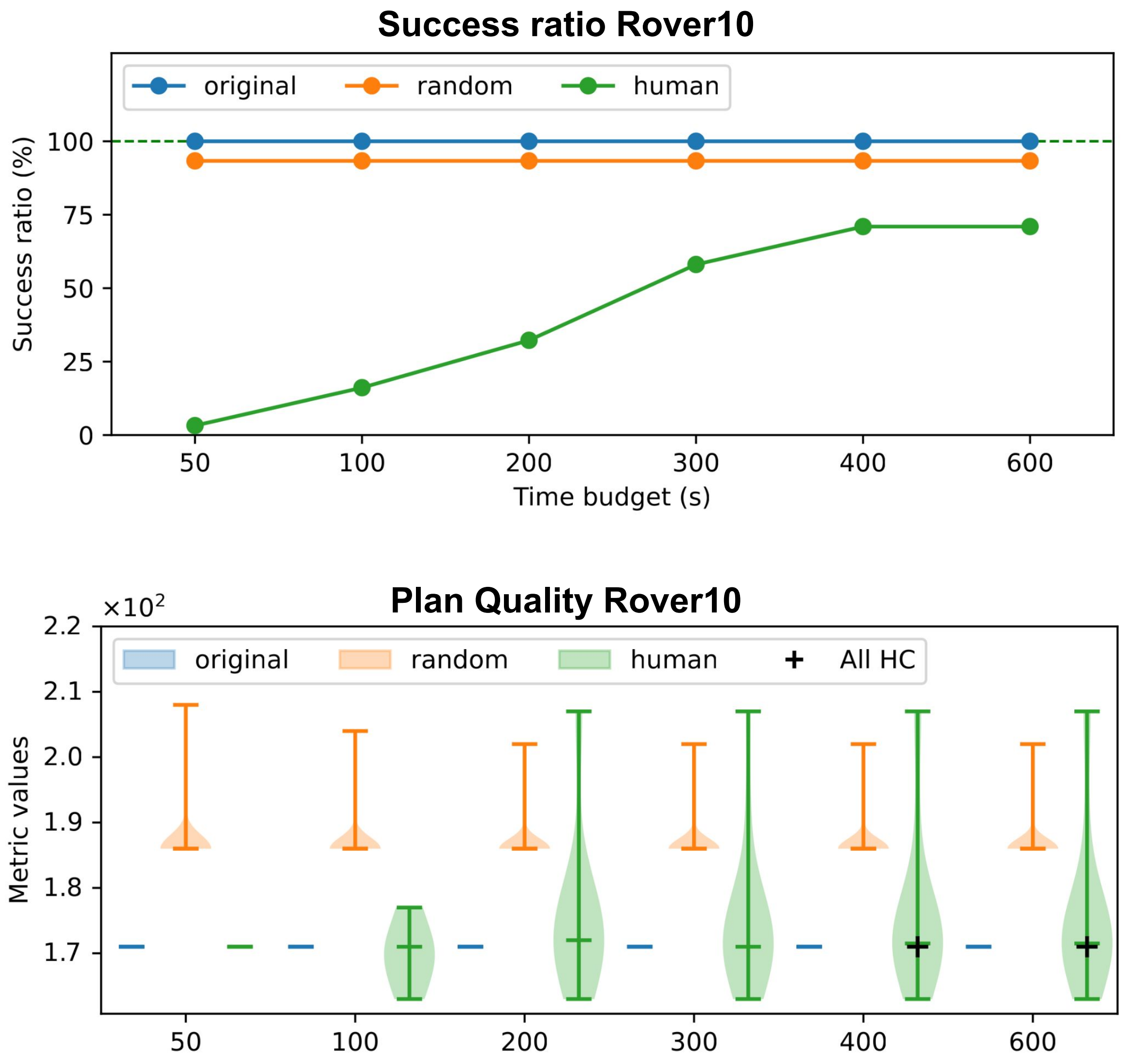


Paper PDF

IPC Problem: ZenoTravel13



IPC Problem: Rover10



Decompositions
LLM
3
CAI - Alpha Version
R0 - Plane1 should never be used.
D1- There must always be no person in plane1.
D2- Plane1 should remain at its initial location (city1) throughout the plan.
D3- The fuel level of plane1 should remain unchanged from its initial value.
Internal PDDL translation
LLM
4
Planning
Solver
5
Add
Delete
Activate / Deactivate
Activate E2NL
Plan
Change Planning Mode
Change Timeout (TO)
Show Encodings
Hide Decomps
Plans
Previous:
None
Current:
Plan-Length:13
Metric (Search):4786.0
Planning Time (msec): 217
Found Plan:
0.0: (board_person3_plane2_city2)
1.0: (refuel_plane2)
2.0: (flyslow_plane2_city2_city0)
3.0: (board_person2_plane2_city0)
4.0: (flyslow_plane2_city0_city1)
5.0: (board_person4_plane2_city1)
6.0: (flyslow_plane2_city1_city3)
7.0: (debark_person2_plane2_city3)
8.0: (debark_person4_plane2_city3)
9.0: (debark_person3_plane2_city3)
10.0: (board_person1_plane2_city3)
11.0: (flyslow_plane2_city3_city2)
(debark_person1_plane2_city2)
=== PLANNING ===
Compiling ... OK
Planning (opt-hmax, TO=15.0s) ... OK
Get advices and insights
LLM+Database
1
User Inputs: NL constraints
Human
2
Elapsed Time: 0.9 s
Risk Analysis
Dialog
Update Sim
Copy
Loop
Visual simulation
6
PDSIM
Person
Plane
city0
city1
city2
city3
2.0: flyslow(plane2, city2, city0)