NeurIPS 2025

Feedback-Aware MCTS for Goal-Oriented Information Seeking

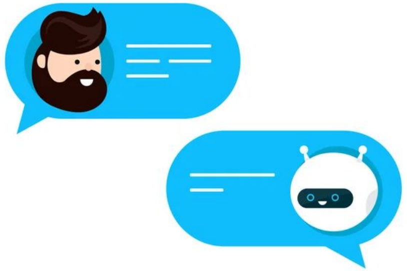
Harshita Chopra & Chirag Shah

{hchopra3, chirags}@cs.washington.edu

UNIVERSITY of WASHINGTON

Background

- User often start the conversation with partial or vague information.
  > *"My laptop won't start. What's wrong?"*
- Effective problem-solving requires **identifying and acquiring missing information**.
- Goal-oriented dialogue systems must ask the **right questions** to reach the answer efficiently.
- Poor questioning leads to **long, unhelpful interactions.**



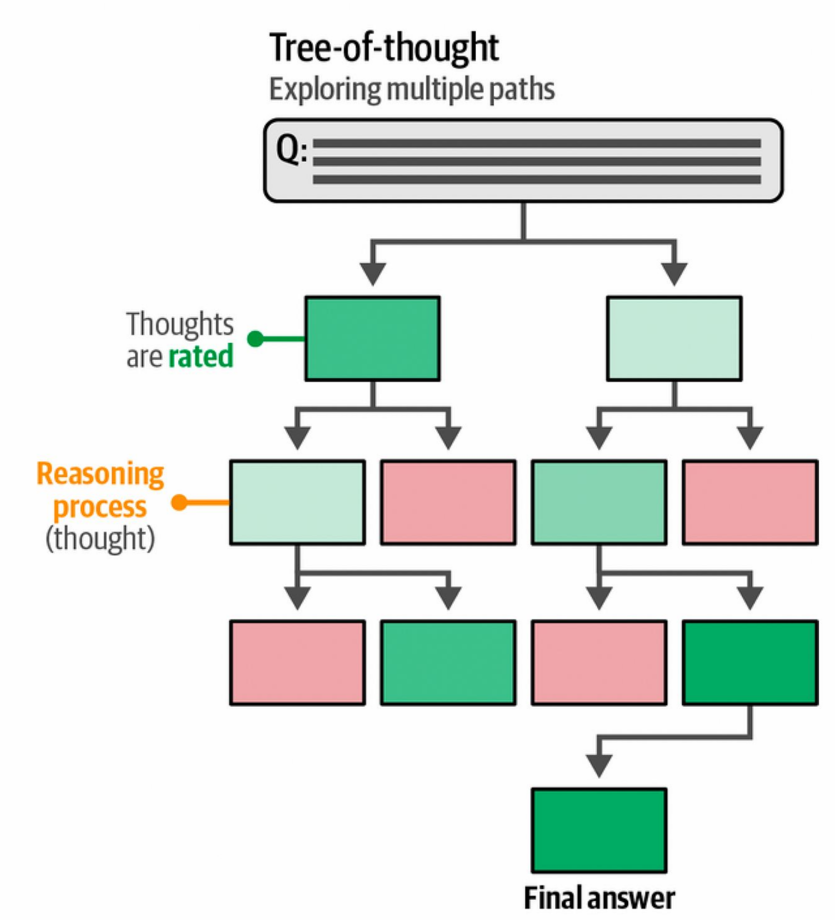
Motivation

Existing planning methods struggle with:

- > **Uncertainty** – large space of possibilities
- > **Lack of adaptation** to historical interaction patterns

Tree-based planning is **powerful** but **expensive**.

- > **Can we balance exploration-exploitation?**
- > **Can we learn from prior interactions?**

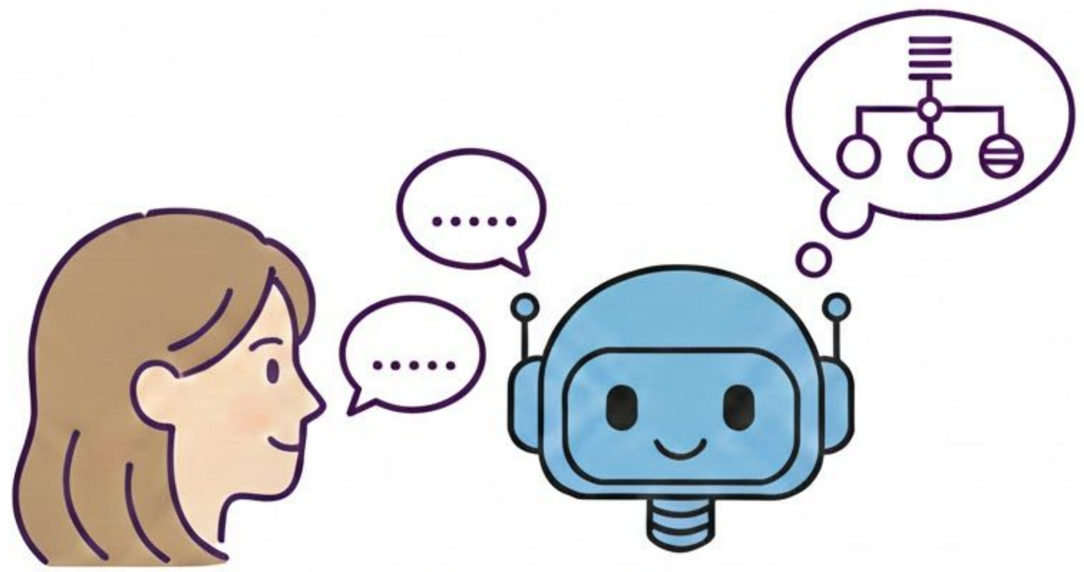


Proposed Solution

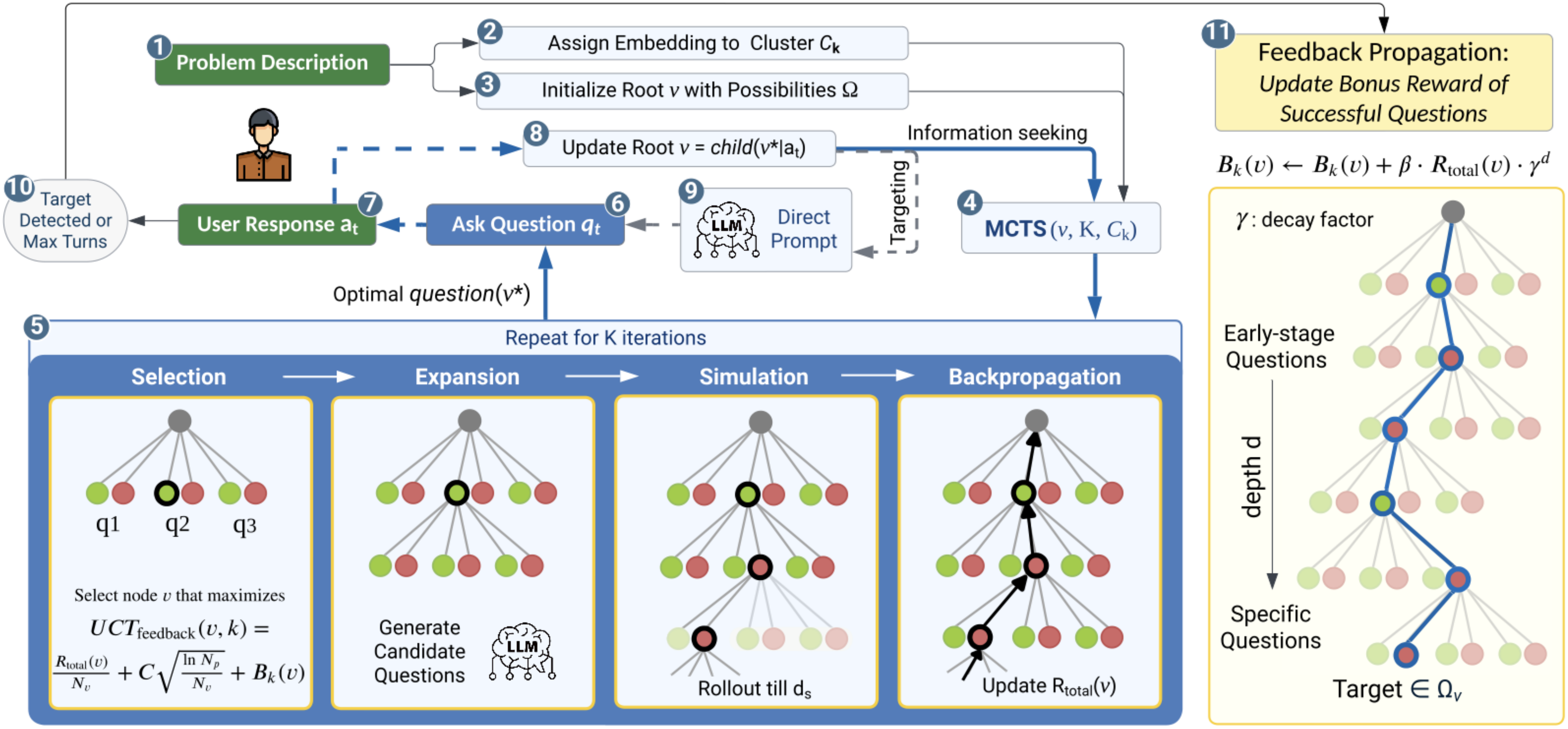
MISQ-HF: **M**onte Carlo Tree Search for **I**nformation **S**eeking **Q**uestions with **H**ierarchical **F**eedback

Combines

- **LLM**s for generating **candidate questions**
- **MCTS** to efficiently plan under **uncertainty**
- **Similarity-based feedback** to reuse questioning strategies from past conversations .



Workflow



1 Problem Description
2 Assign Embedding to Cluster C_k
3 Initialize Root v with Possibilities Ω
8 Update Root v = $child(v^*|a_t)$
Information seeking
Targeting
9 LLM Direct Prompt
4 MCTS (v, K, C_k)
6 Ask Question q_t
7 User Response a_t
10 Target Detected or Max Turns
Optimal $question(v^*)$
5 Repeat for K iterations
Selection
Expansion
Simulation
Backpropagation
q1
q2
q3
Select node v that maximizes
$UCT_{feedback}(v,k) = \frac{R_{total}(v)}{N_v} + C\sqrt{\frac{\ln N_p}{N_v}} + B_k(v)$
Generate Candidate Questions
LLM
Rollout till d_s
Update $R_{total}(v)$
11 Feedback Propagation: Update Bonus Reward of Successful Questions
$B_k(v) \leftarrow B_k(v) + \beta \cdot R_{total}(v) \cdot \gamma^d$
γ : decay factor
Early-stage Questions
depth d
Specific Questions
Target $\in \Omega_v$

Method

Decision Tree Construction:

- LLM generates **candidate questions** to split the current possibilities.
- Node attributes: Question, Answer, Remaining possibilities,
  Total reward, Cluster-based Bonus reward

MCTS Phases:

- **Selection**: Based on a **modified UCT** (Upper Confidence Bound for Trees).

$$\boxed{UCT_{\text{feedback}}(v,k) = \frac{R_{\text{total}}(v)}{N_v} + C\sqrt{\frac{\ln N_p}{N_v}} + B_k(v)}$$

  - *Reward: Expected information gain (uncertainty reduction)*
- **Expansion**: Generate new questions if needed.
- **Simulation**: Simulate random rollouts to estimate long-term reward.
- **Backpropagation**: Update rewards and visit counts of ancestor nodes.

Method

Hierarchical Feedback Mechanism

- After a successful conversation, update bonus rewards for nodes along the successful questioning path.

$$B_k(v) \leftarrow B_k(v) + \beta \cdot R_{\text{total}}(v) \cdot \gamma^{d_v}$$

- Bonus rewards are specific to clusters.
- Early-stage questions (high-level, broadly applicable) get higher bonus (reward decays with depth).

Adaptation: System learns which questions work best for similar problems.

Experimental Setup

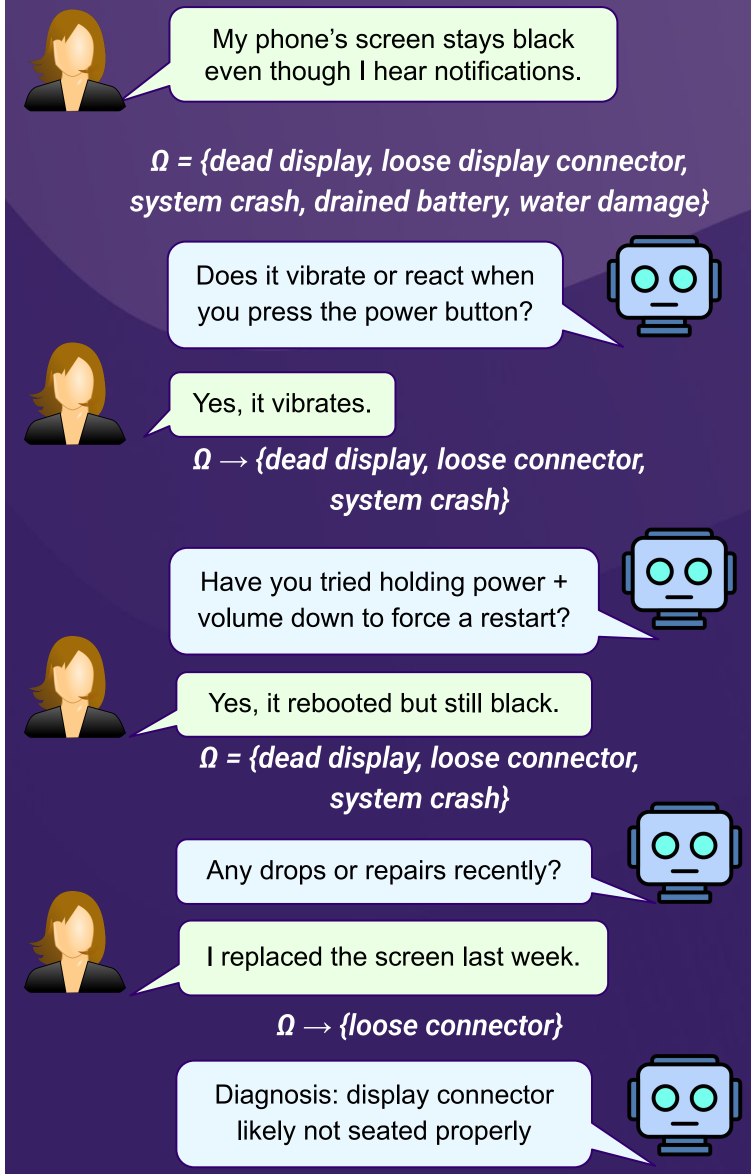
Domains: Medical Diagnosis, Troubleshooting, 20 Questions

Baselines

- Direct Prompting (DP): No planning
- Uncertainty of Thoughts (UoT): Exhaustive tree expansion. [1]
- MISQ: Our method without feedback

Metrics

- Success Rate (SR)
- Mean Conversation Length in Successful Cases (MSC)
- Question Generation Calls (QGC): Number of LLM calls for planning



[1] Hu, et al. "Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models." *NeurIPS 2024*

Results

12% better Success Rate on average

~10x lesser LLM calls for planning

Domain: General — 20-Questions

Method	Ω-aware	Common SR↑	Common MSC↓	Common QGC↓	Thing SR↑	Thing MSC↓	Thing QGC↓
Llama 3.3 70B Instruct							
UoT	×	39.63	8.27	4.08	19.00	9.78	4.48
MISQ	×	41.44	8.43	5.05	23.5	9.57	1.57
DP	✓	45.94	13.70	-	32.50	13.27	-
UoT	✓	61.26	9.94	7.92	35.50	11.43	3.40
MISQ	✓	**74.77**	**9.90**	**4.74**	**59.50**	**10.68**	**3.31**
Mixtral 8*7B Instruct							
DP	✓	8.10	14.33	-	7.50	13.46	-
UoT	✓	28.82	11.56	4.34	12.50	13.52	5.91
MISQ	✓	**37.83**	**11.38**	**2.39**	**20.00**	**11.50**	**0.06**
GPT-4o							
DP	✓	63.06	14.72	-	40.50	14.16	-
UoT	✓	74.77	8.59	5.88	47.00	**9.13**	2.75
MISQ	✓	**85.58**	**8.51**	**4.86**	**55.50**	9.54	**2.19**

Domain: Medical Diagnosis (MD), Troubleshooting (TS)

Model	Method	Ω-aware	MD: DX SR↑	MD: DX MSC↓	MD: DX QGC↓	MD: MedDG SR↑	MD: MedDG MSC↓	MD: MedDG QGC↓	TS: FloDial SR↑	TS: FloDial MSC↓	TS: FloDial QGC↓
Llama 3.3 70B Instruct	UoT	×	72.11	1.54	0.36	79.51	2.09	4.95	34.64	6.84	43.76
	MISQ	×	75.00	2.17	0.05	86.56	3.39	0.40	35.29	9.09	3.99
	MISQ-HF	×	80.76	1.94	0.21	86.78	3.29	0.78	39.86	9.09	4.07
	DP	✓	88.46	3.15	-	84.14	3.93	-	21.56	13.72	-
	UoT	✓	79.80	1.65	0.77	89.86	**2.16**	4.84	60.78	**8.47**	44.61
	MISQ	✓	92.30	**1.28**	0.48	92.29	3.44	3.59	62.74	9.73	5.16
	MISQ-HF	✓	**98.07**	1.84	**0.04**	**93.39**	3.35	**0.54**	**67.97**	9.81	**3.97**
Mixtral 8*7B Instruct	DP	✓	50.00	3.50	-	76.43	3.91	-	16.99	14.23	-
	UoT	✓	76.92	**1.43**	0.45	83.70	**2.19**	5.70	39.21	**7.01**	45.11
	MISQ	✓	63.46	2.63	0.08	76.55	3.33	**0.17**	47.71	10.45	1.66
	MISQ-HF	✓	**76.92**	2.40	**0.06**	**84.58**	3.08	0.33	**49.01**	9.62	**1.46**
GPT-4o	DP	✓	73.07	3.48	-	81.27	3.98	-	43.79	14.86	-
	UoT	✓	82.69	**1.18**	0.17	88.79	**2.03**	1.81	59.47	**8.14**	41.86
	MISQ	✓	87.50	1.97	0.05	89.20	3.46	0.60	**74.50**	10.15	4.10
	MISQ-HF	✓	**99.03**	2.19	**0.03**	**90.30**	3.42	**0.41**	72.54	10.36	**2.94**

Results

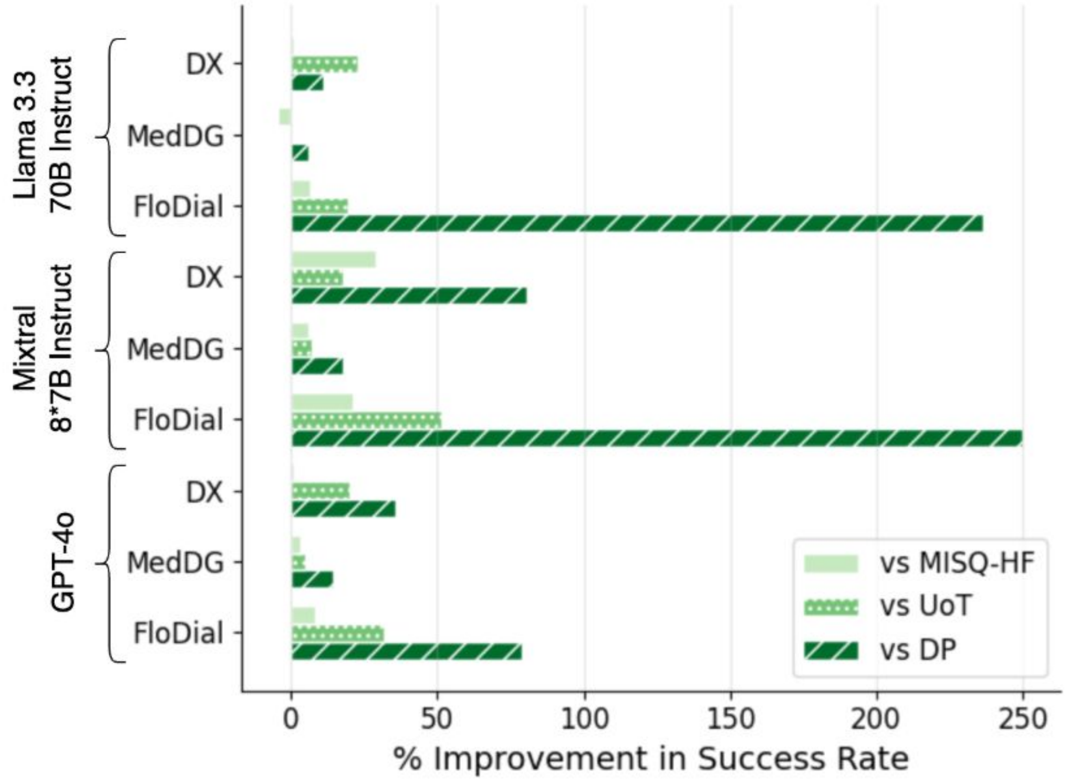


Figure 3: Improvement in Success Rate on MD and TS Domain in a Closed Set scenario, when initializing the root node with the constrained set of possibilities $\Omega_c \subseteq \Omega$.

+8% improvement in Success Rate when starting with a **constrained set of possibilities** at the root node as compared to using the full set.

Conclusion

MISQ-HF .

- Enables **adaptive, efficient information-seeking.**
- **Learns from** historical successes using cluster-based **feedback**.
- Reduces computational cost without sacrificing accuracy.

Future Directions:

- Multi-dimensional Reward for better questions.
- Collect data for RLHF: train policies with positive and negative interactions

Thank you.

Harshita Chopra
PhD Student
hchopra3@cs.washington.edu
UNIVERSITY of WASHINGTON

Scan to view our paper!