THE OHIO STATE UNIVERSITY

TEXAS
The University of Texas at Austin

cisco

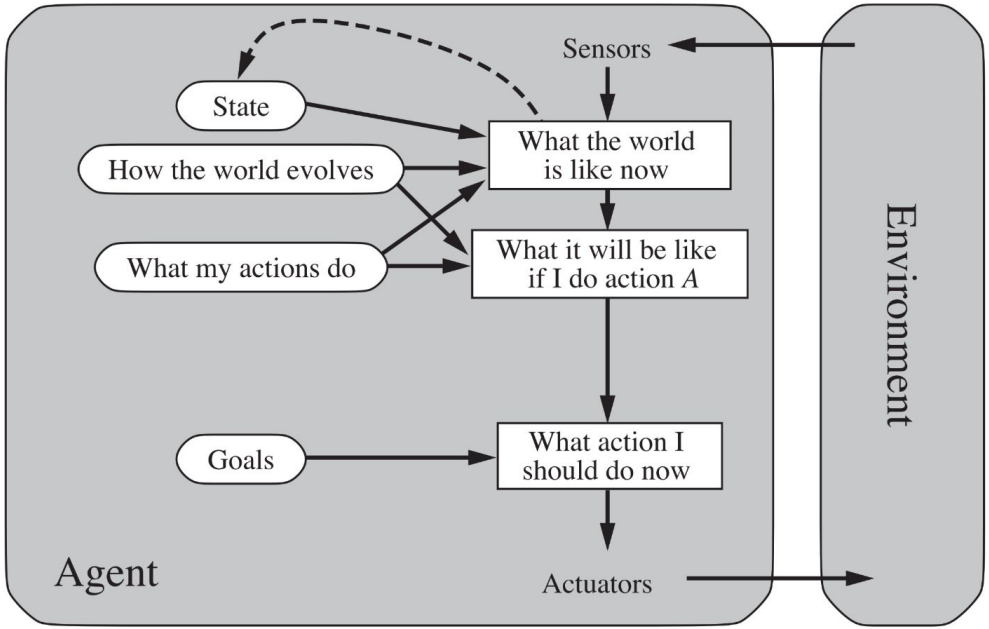
When is Tree Search Useful for LLM Planning? It Depends on the Discriminator

Ziru Chen, Michael White, Raymond Mooney,
Ali Payani, Yu Su, Huan Sun



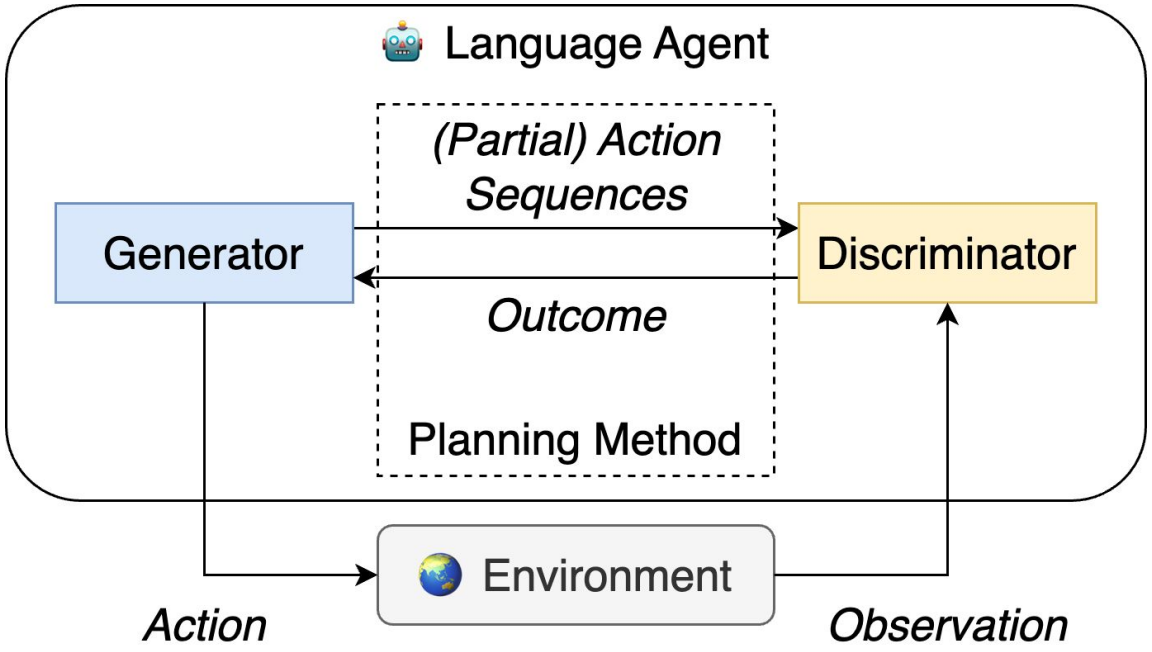



AI Agents for Problem-Solving

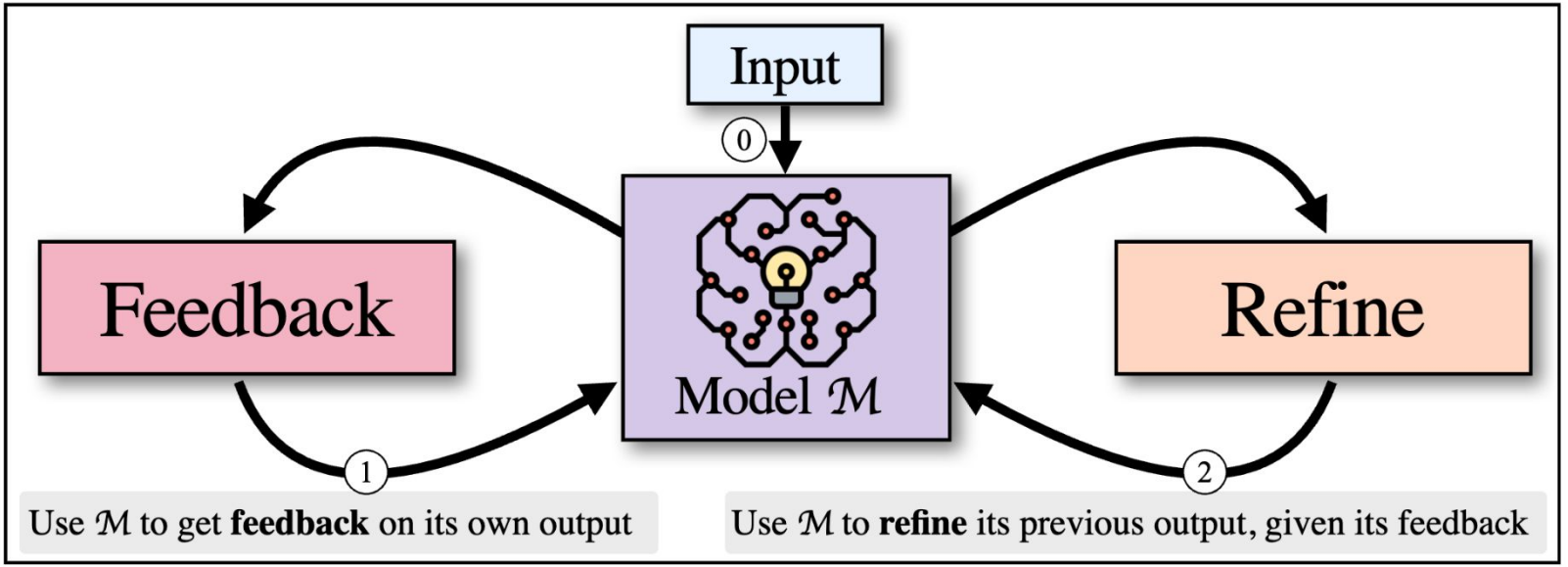


Stuart Russell and Peter Norvig. 2010. Artificial Intelligence: A Modern Approach, 3rd edition. Prentice Hall.

LLM-Based Agents for Problem-Solving

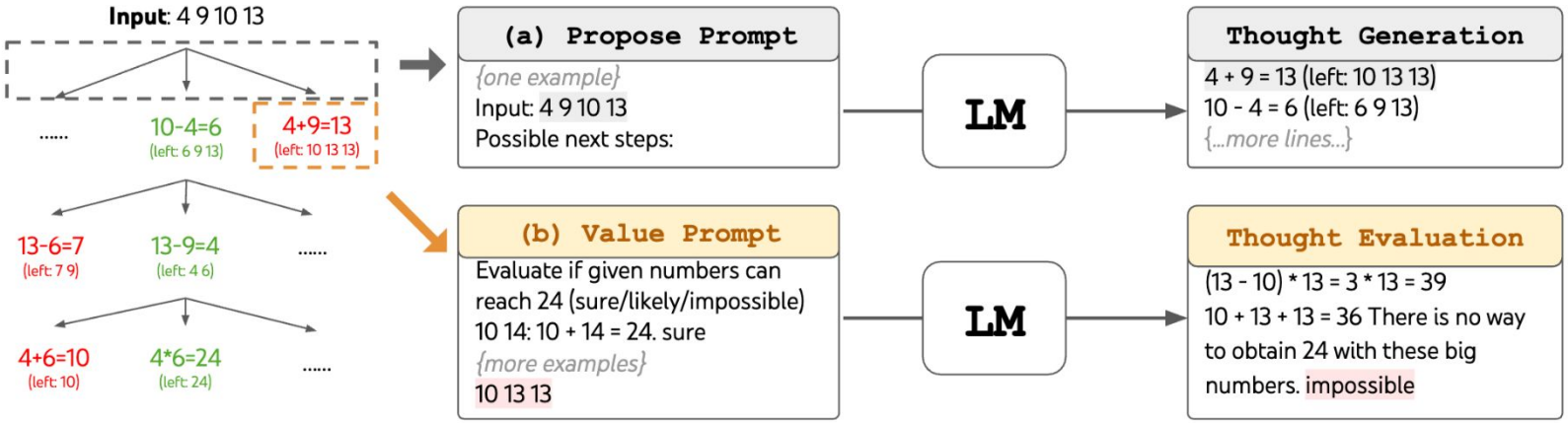


Advanced Planning Methods



Aman Madaan et al. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In NeurIPS 2023.

Advanced Planning Methods



Shunyu Yao et al. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In NeurIPS 2023.

Are Advanced Planning Methods the Solution?

"*Models outperform humans in generation but underperform humans in discrimination*."

THE GENERATIVE AI PARADOX:
"What It Can Create, It May Not Understand"

Peter West[1*] **Ximing Lu**[1,2*] **Nouha Dziri**[2*] **Faeze Brahman**[1,2*] **Linjie Li**[1*]
Jena D. Hwang[2] **Liwei Jiang**[1,2] **Jillian Fisher**[1] **Abhilasha Ravichander**[2]
Khyathi Raghavi Chandu[2] **Benjamin Newman**[1]
Pang Wei Koh[1] **Allyson Ettinger**[2] **Yejin Choi**[1,2]

[1]University of Washington [2]Allen Institute for Artificial Intelligence

{pawest,linjli}cs.washington.edu
{ximinglu,nouhad,faezeb}allenai.org

Are Advanced Planning Methods the Solution?

"*LLMs struggle to self-correct their responses without external feedback*."

THE GENERATIVE AI PARADOX:
"What It Can Create, It May Not Understand"

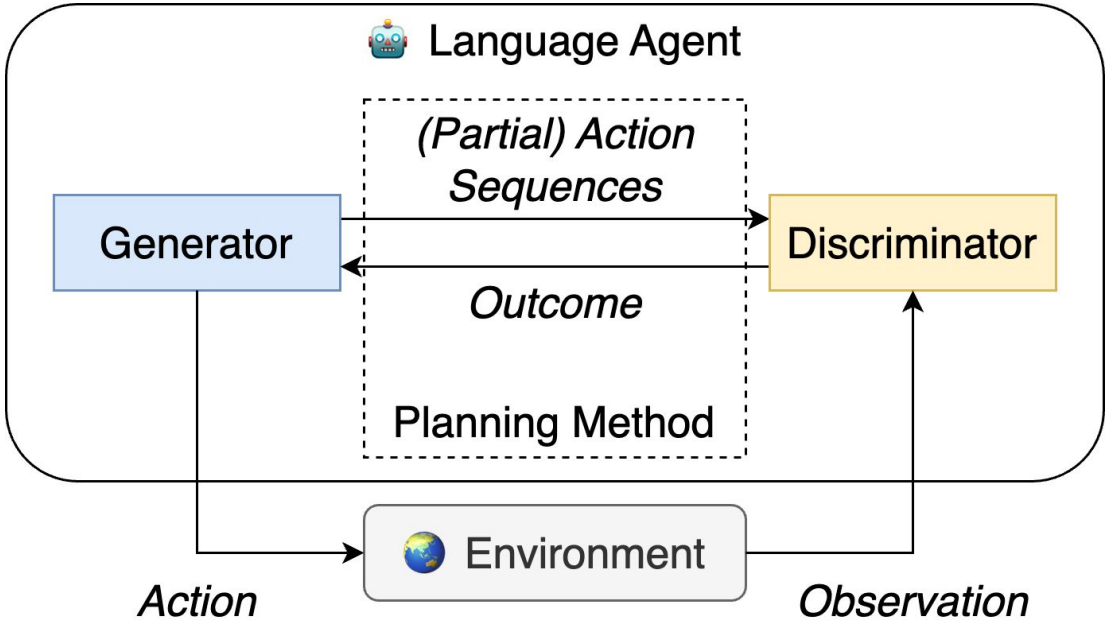
Peter West[1*] Ximing Lu[1,2*] Nouha Dziri[2*] Faeze Brahman[1,2*] Linjie Li[1*]
Jena D. Hwang[2] Liwei Jiang[1,2] J
Khyathi Raghavi Chandu[2] Benja
Pang Wei Koh[1] Allyson Ettinger[2]
[1]University of Washington [2]Allen In
{pawest,linjli}cs.washing
{ximinglu,nouhad,faezeb}a

LARGE LANGUAGE MODELS CANNOT SELF-CORRECT REASONING YET

Jie Huang[1,2*] Xinyun Chen[1*] Swaroop Mishra[1] Huaixiu Steven Zheng[1] Adams Wei Yu[1]
Xinying Song[1] Denny Zhou[1]
[1]Google DeepMind [2]University of Illinois at Urbana-Champaign
jeffhj@illinois.edu, {xinyunchen, dennyzhou}@google.com

Are Advanced Planning Methods the Solution?

We hypothesize that the ***discriminator*** may be more important in LLM planning.



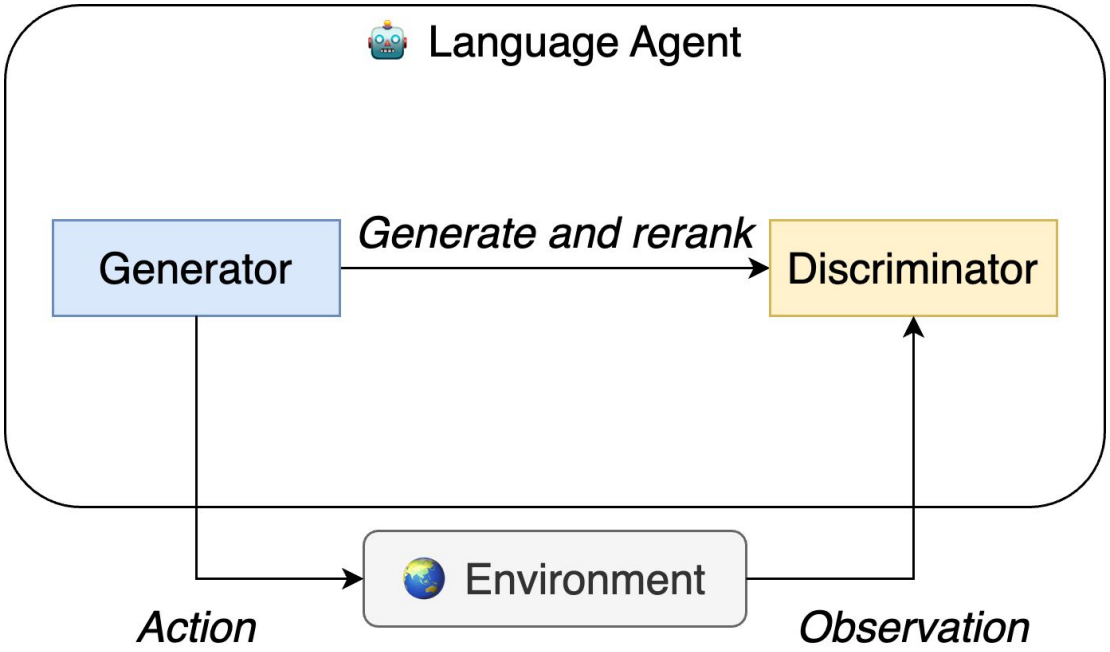
Our Contributions

- Investigation of three planning methods under a unified language agent framework

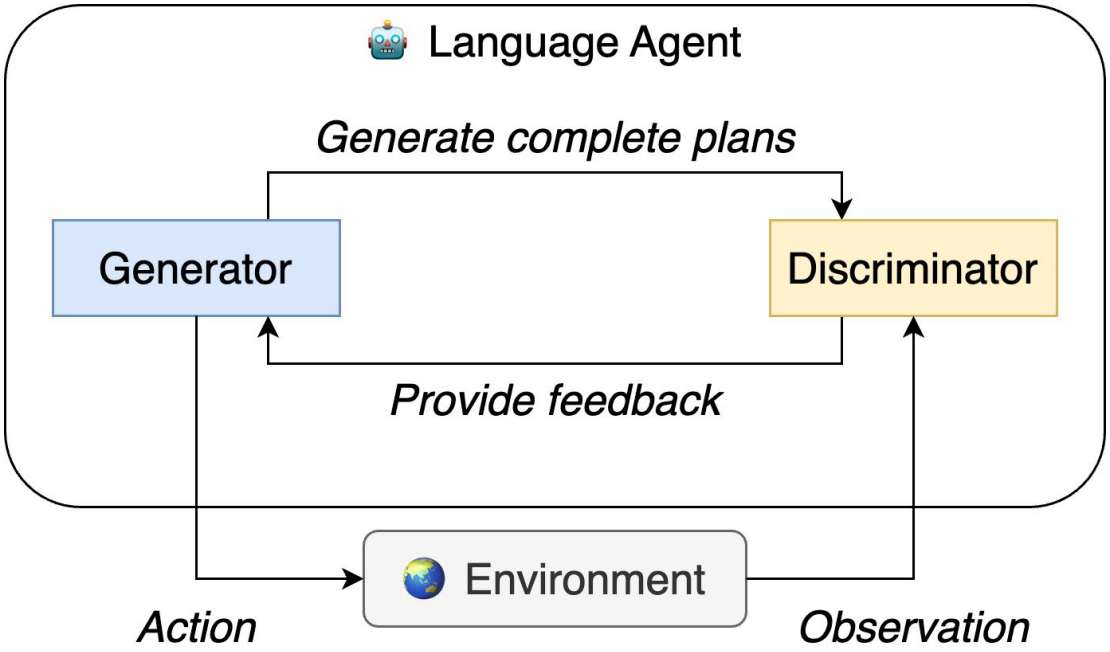
- Comprehensive experiments on two real-world tasks, text-to-SQL parsing and math reasoning

- Empirical analysis of LLMs' discrimination abilities and their impact on LLM planning

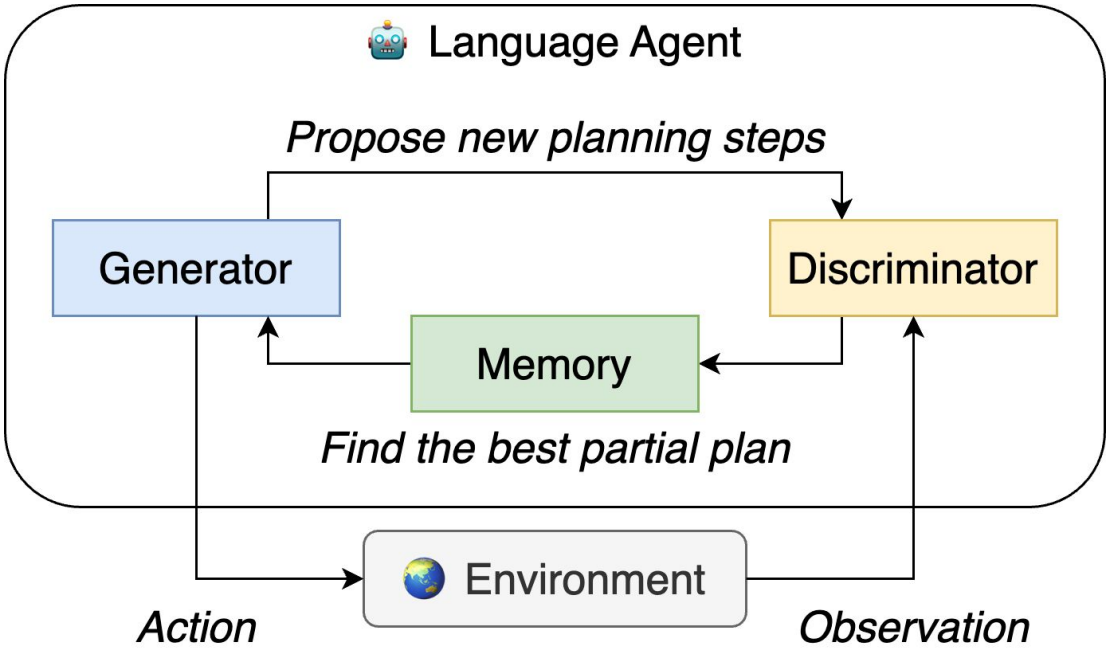
Unified View of Planning - Re-ranking



Unified View of Planning - Iterative Correction



Unified View of Planning - Tree Search



Research Questions

- **RQ1**: How does discrimination accuracy affect the performance of language agents using different planning methods?

- **RQ2**: Can LLM-based discriminators correctly assess language agents' actions in practical settings?

Research Questions

- **RQ1**: How does discrimination accuracy affect the performance of language agents using different planning methods?

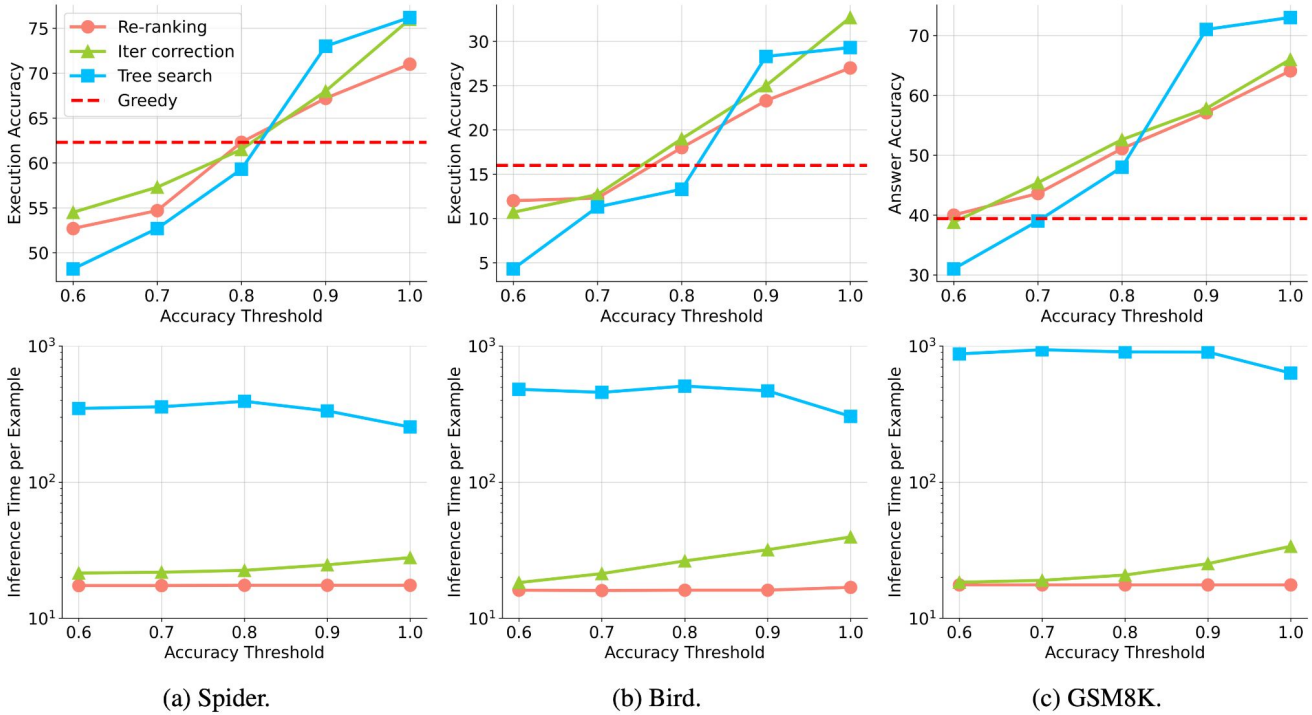
- **RQ2**: Can LLM-based discriminators correctly assess language agents' actions in practical settings?

Simulation Experiments with Oracle

- Simulate a perfect discriminator with gold answers

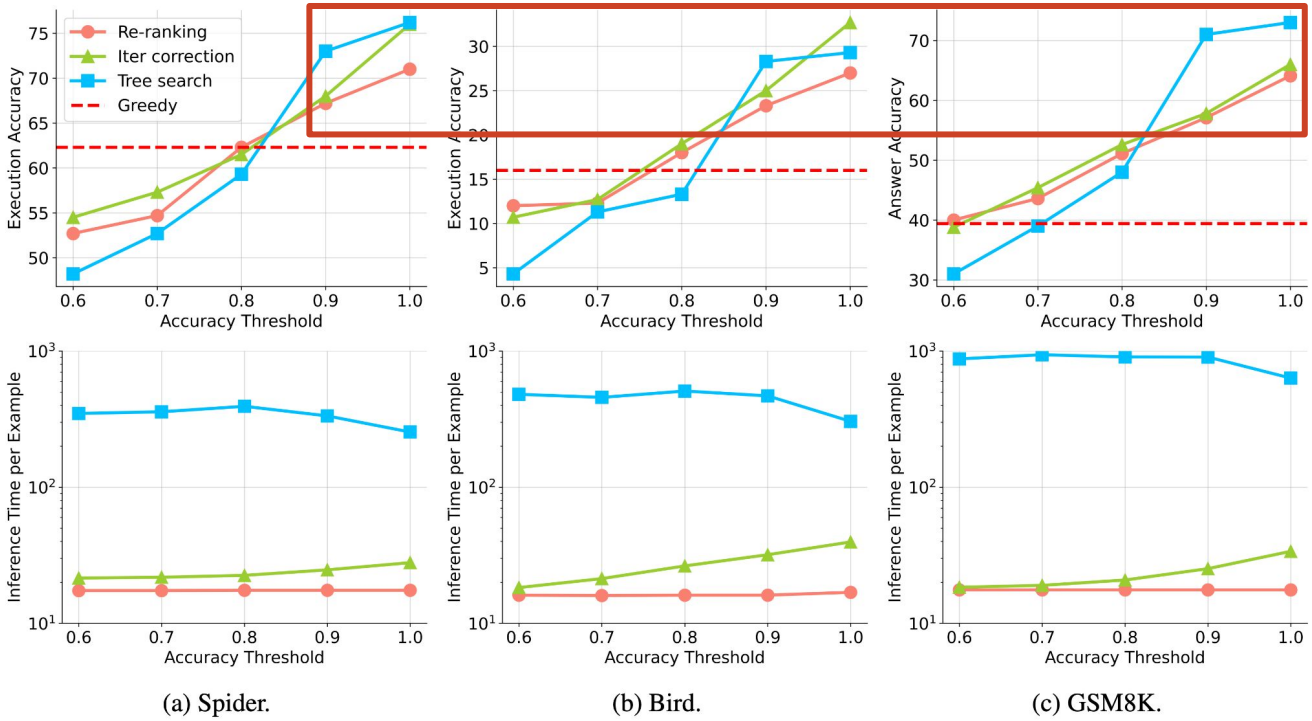
- Control the accuracy with a probability-based threshold
  - Sample a random number between 0 and 1
  - If the number is smaller than our threshold, we follow the discriminator’s score
  - Otherwise, we inverse the score

Simulation Experiments with Oracle



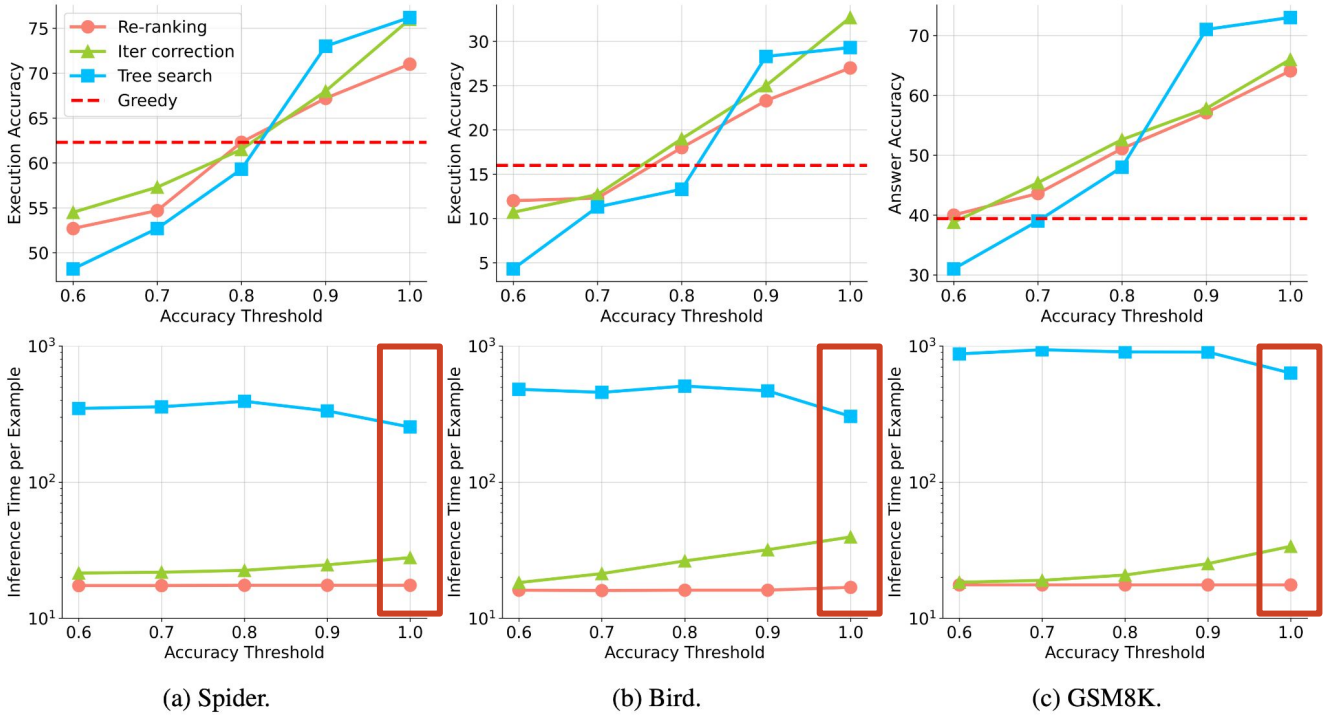
End-to-end evaluation results (the first row) and average inference time in log scale (the second row) of our simulation experiments with oracle-based discriminator.

Simulation Experiments with Oracle



End-to-end evaluation results (the first row) and average inference time in log scale (the second row) of our simulation experiments with oracle-based discriminator.

Simulation Experiments with Oracle



End-to-end evaluation results (the first row) and average inference time in log scale (the second row) of our simulation experiments with oracle-based discriminator.

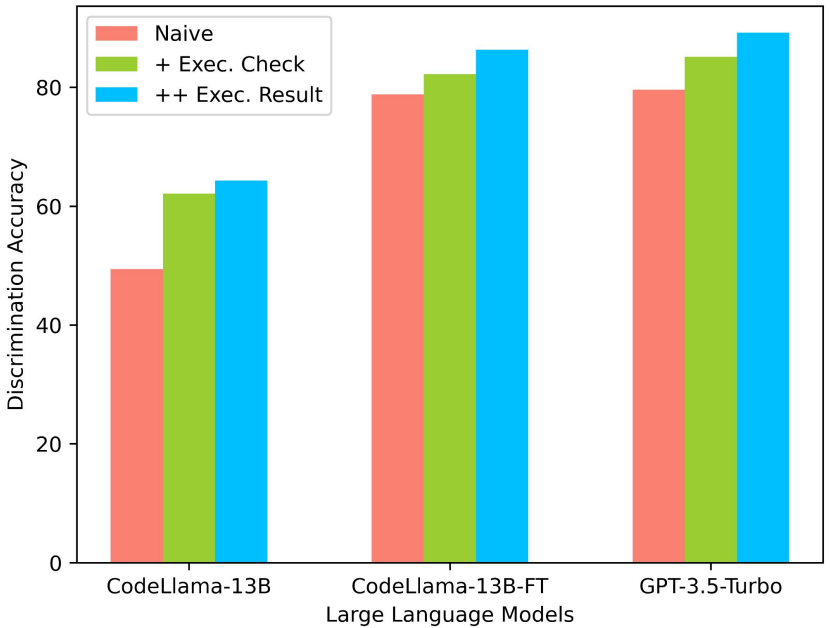
Research Questions

- **RQ1**: How does discrimination accuracy affect the performance of language agents using different planning methods?

- **RQ2**: Can LLM-based discriminators correctly assess language agents' actions in practical settings?

Discrimination Accuracy of LLMs

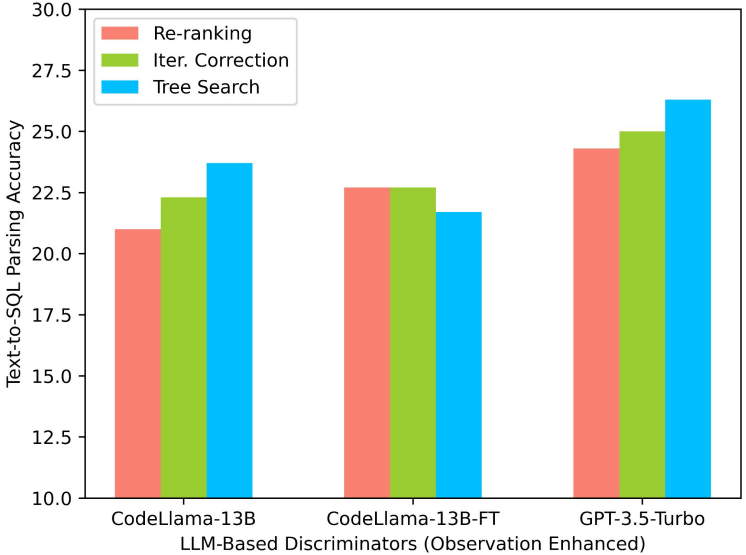
We improve LLMs' discrimination accuracy with environmental observations.

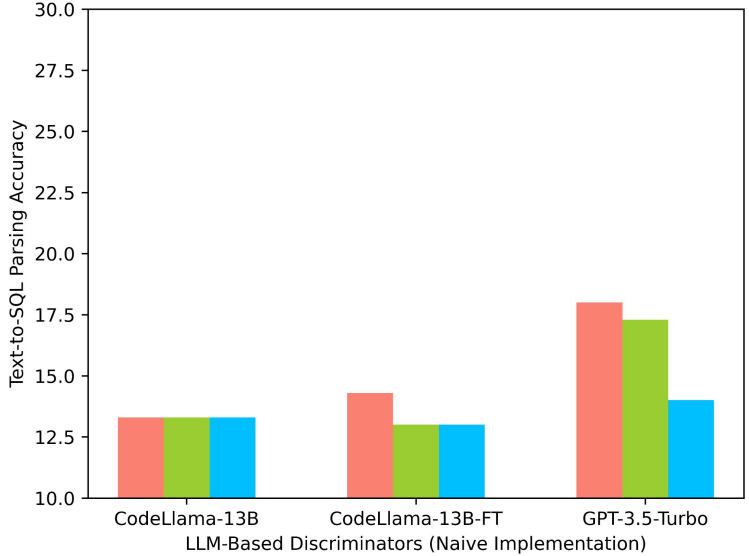


Discrimination accuracy of naive and observation-enhanced LLMs on BIRD-SQL.

End-to-End Evaluation of LLM Planning

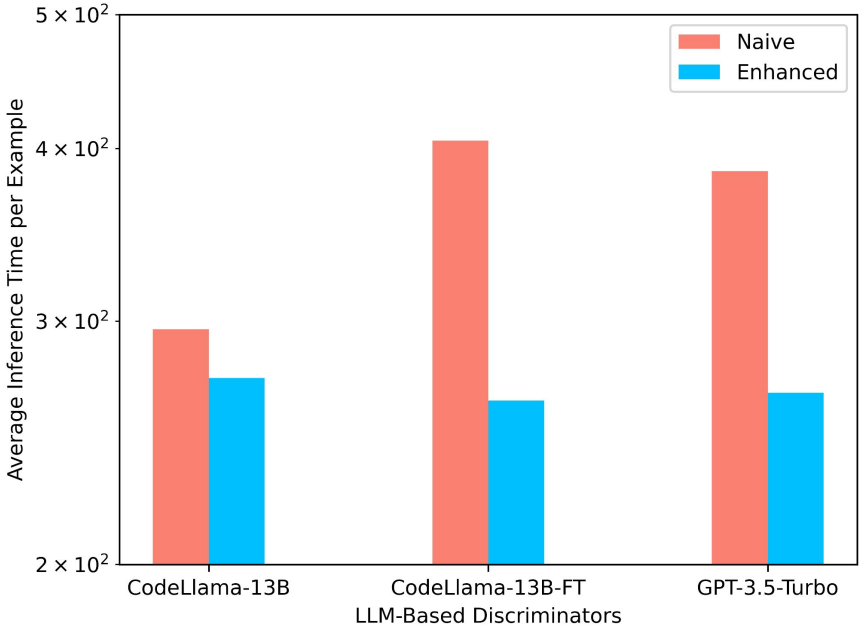
LLM-based discriminators cannot help advance planning methods to achieve significant accuracy improvement yet.





End-to-End Evaluation of LLM Planning

Observation-enhanced discriminators can largely reduce tree search latency.



Average end-to-end inference time (seconds; log scale) per example using tree search on BIRD-SQL.

Conclusions

- Advanced planning methods, i.e., iterative correction and tree search, demand highly accurate discriminators to achieve decent improvements over the simpler method, re-ranking.

- The discrimination accuracy of LLMs may not yet be sufficient for advanced planning methods.

- The accuracy-efficiency trade-off can impede the deployment of advanced planning methods in real-world applications.

Thank you!

Code and Data: https://github.com/OSU-NLP-Group/Auto-SQL-Correction

Email: chen.8336@osu.edu