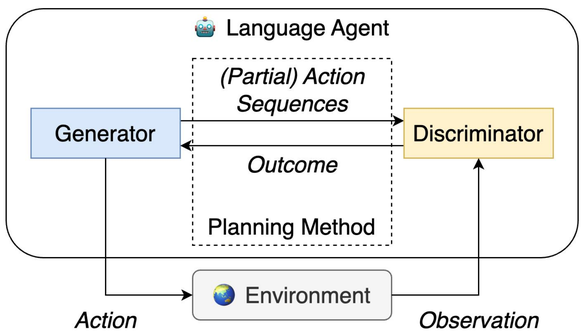
In LLM planning, the discriminator needs to get up to 90% accuracy for tree search to start outperforming simple re-ranking.

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

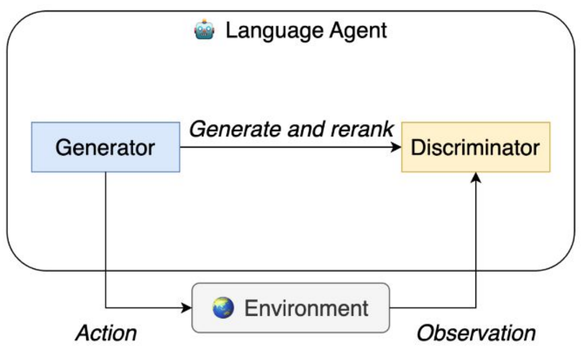
Ziru Chen, Michael White, Ray Mooney, Ali Payani, Yu Su, Huan Sun

The Ohio State University · TEXAS The University of Texas at Austin · CISCO

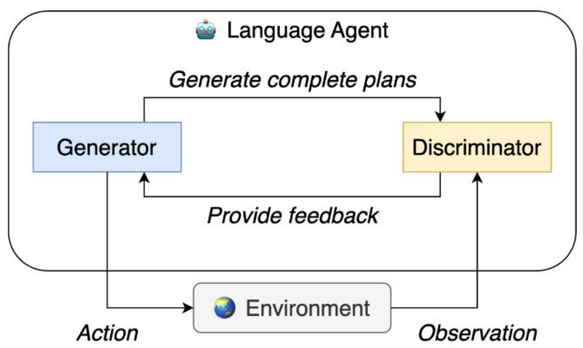
A Generator-Discriminator Framework of Language Agents



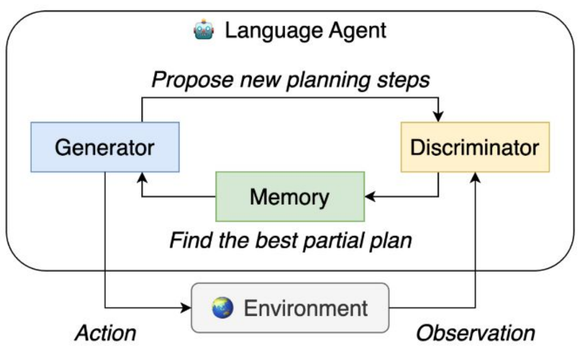
- **Generator**: Propose (partial) action sequences.
- **Discriminator**: Evaluates the outcomes of these actions.
- **Planning method**: Ranks the actions according to their outcomes and manage the interaction between the two models.



(a) Re-ranking.

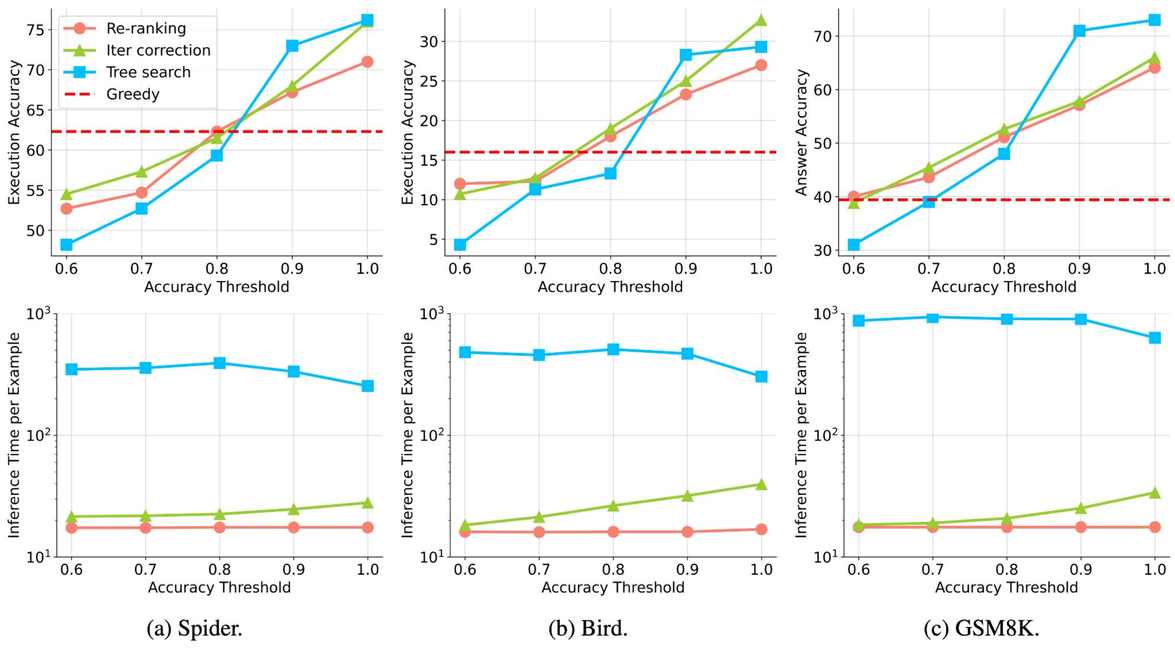


(b) Iterative Correction.



(c) Tree Search.

Simulation Experiments with Oracle



(a) Spider. (b) Bird. (c) GSM8K.

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

Experiments with LLM-Based Discriminators

Discriminators	Spider (Greedy Gen = 62.3)			Bird (Greedy Gen = 16.0)		
	Re-ranking	Iter. Correct.	Tree Search	Re-ranking	Iter. Correct.	Tree Search
CodeLlama-13B	**57.5**	51.7	55.5	**13.3**	**13.3**	**13.3**
GPT-3.5-Turbo	**58.3**	52.7	56.2	**18.0**	17.3	14.0
CodeLlama-13B-FT	**61.5**	51.7	56.0	**14.3**	13.0	13.0
CodeLlama-13B^E	**65.5**	62.0	62.5	21.0	**24.3**	22.7
GPT-3.5-Turbo^E	67.0	**67.5**	66.0	22.3	**25.0**	22.7
CodeLlama-13B-FT^E	**70.3**	68.0	67.5	23.7	**26.3**	21.7
Oracle Simulation ($\tau = 1.0$)	71.0	76.0*	76.2*	27.0	32.7*	29.3

End-to-end execution accuracy on text-to-SQL parsing.

Error Type	Spider		Bird		GSM8K	
	Iter. Correct.	Tree Search	Iter. Correct.	Tree Search	Iter. Correct.	Tree Search
Discrimination	29 (78.4%)	17 (60.7%)	9 (52.9%)	12 (50.0%)	30 (62.5%)	6 (66.7%)
Exploration	8 (21.6%)	11 (39.3%)	8 (47.1%)	12 (50.0%)	18 (37.5%)	3 (33.3%)
Total	37	28	17	24	48	9

Error analysis of examples where re-ranking outperforms advanced planning methods.

(1) **Discrimination error**: The discriminator assigns a higher score for wrong programs than correct ones, which is not recoverable by any planning method.

(2) **Exploration error**: The planning method has not found the correct program before termination.

Intrinsic Evaluation of LLM-Based Discriminators

Models	Spider				Bird				GSM8K‡			
	Acc	F1	H@1	MRR	Acc	F1	H@1	MRR	Acc	F1	H@1	MRR
CodeLlama-7B	54.0	37.1	56.0	62.3	44.6	**46.7**	13.0	18.0	48.6	**38.7**	36.2	46.9
CodeLlama-13B	58.2	37.1	57.0	63.1	49.4	**46.7**	12.7	18.3	**62.2**	**38.7**	**41.8**	**51.0**
CodeLlama-7B-FT	62.4	60.3	59.5	64.6	52.4	**46.7**	14.3	19.1	-	-	-	-
CodeLlama-13B-FT	**69.7**	**67.2**	**61.3**	**65.7**	**62.1**	**46.7**	**16.0**	**20.5**	-	-	-	-
GPT-3.5-Turbo	67.0	47.3	59.0	64.3	64.3	35.7	16.0	20.5	72.1	49.1	46.6	54.0
GPT-4-Turbo	**76.5**	**54.9**	**63.0**	**66.7**	**76.2**	**50.1**	**20.3**	**23.0**	**93.8**	**91.1**	**59.8**	**61.6**

Intrinsic evaluation results of naive LLMs' discrimination abilities.

	CodeLlama-13B			GPT-3.5-Turbo			CodeLlama-13B-FT	
	Spider	Bird	GSM8K	Spider	Bird	GSM8K	Spider	Bird
Naive Discriminator	58.2	49.4	62.2	67.0	64.3	72.1	69.7	62.1
+ Executability Check	78.7	78.8	64.5	84.8	86.3	73.2	83.6	82.2
++ Execution Result	**83.6**	**79.6**	**70.6**	**90.0**	**89.2**	**76.5**	**88.5**	**85.1**
Improvement	25.4	30.2	8.4	23.0	24.9	4.4	18.8	23.0

Discrimination accuracy of observation-enhanced LLMs. The best performance is achieved using both kinds of environmental observations.

Conclusions

- Advanced planning methods, i.e., iterative correction and tree search, demand highly accurate discriminators (up to 90% accuracy) to achieve decent improvements over the simpler method, re-ranking.
- Using environmental feedback, we improve the discrimination accuracy of LLMs. Yet, our end-to-end evaluations suggest they have barely met the need for advanced planning methods to show significant improvements over re-ranking.
- Advanced planning methods may not adequately balance accuracy and efficiency when using LLM-based discriminators. In our experiments, compared to the other two methods, tree search is at least 10–20 times slower but leads to negligible performance gains.