

Verbal RL via Self-Reflection

Noah Shinn

Write a sentence with 11 words.

GPT-3.5-Turbo: The sky was ablaze with a stunning array of colors.

*Actual count = 10 words

Write a sentence with 11 words.

GPT-3.5-Turbo: The sky was ablaze with a stunning array of colors.

User: Verbosely check if this sentence has exactly 11 words.

GPT-3.5-Turbo: 1. The 2. sky [...] 9. of 10. colors.

Yes, the sentence has exactly 10 words. I apologize for the previous mistake. Here is an alternative sentence with 11 words:

"I embarked on a journey to discover the hidden wonders of nature."

*Actual count = 12 words

Write a sentence with 11 words.

[truncated]

GPT-3.5-Turbo: I embarked on a journey to discover the hidden wonders of nature.

User: Verbosely check if this sentence has exactly 11 words.

GPT-3.5-Turbo: 1. I 2. embarked [...] 11. of 12. nature.

Apologies for the previous oversight. This sentence indeed has 12 words, not 11. Here is an alternative sentence with exactly 11 words:

"She diligently practiced the piano, striving for perfection in every note."

*Actual count = 11 words

Write a sentence with 11 words.

[truncated]

GPT-3.5-Turbo: She diligently practiced the piano, striving for perfection in every note.

User: Verbosely check if this sentence has exactly 11 words.

GPT-3.5-Turbo: 1. She 2. diligently [...] 10. every 11. note.

Yes, the sentence has exactly 11 words.

Task completed

Reasoning in language space

Humans

Q: What is $30,000 + 50,000$

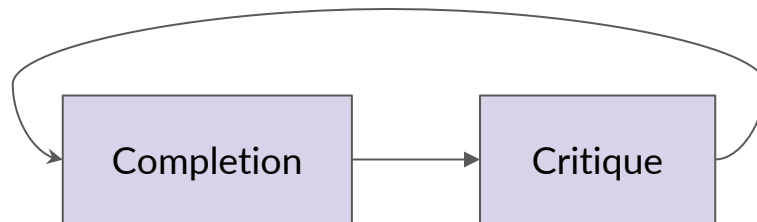
A: 80,000

LLMs

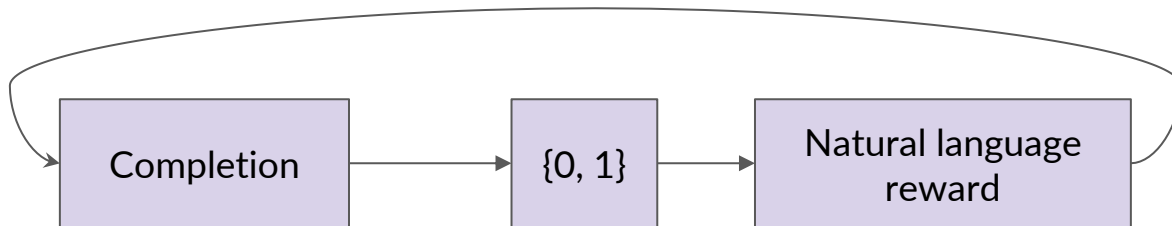
Q: What is $30,000 + 50,000$

A: 80,000

Critique methods



Natural language can amplify rewards



Verbal RL

$$\pi_{\theta}(a_i|s_i), \theta = \{M_a\}$$

↓

$$\pi_{\theta}(a_i|s_i), \theta = \{M_a, \text{mem}\}$$

Algorithm 1 Reinforcement via self-reflection

Initialize Actor, Evaluator, Self-Reflection:

M_a, M_e, M_{sr}

Initialize policy $\pi_{\theta}(a_i|s_i)$, $\theta = \{M_a, mem\}$

Generate initial trajectory using π_{θ}

Evaluate τ_0 using M_e

Generate initial self-reflection sr_0 using M_{sr}

Set $mem \leftarrow [sr_0]$

Set $t = 0$

while M_e not pass or $t < \text{max trials}$ **do**

Generate $\tau_t = [a_0, o_0, \dots, a_i, o_i]$ using π_{θ}

Evaluate τ_t using M_e

Generate self-reflection sr_t using M_{sr}

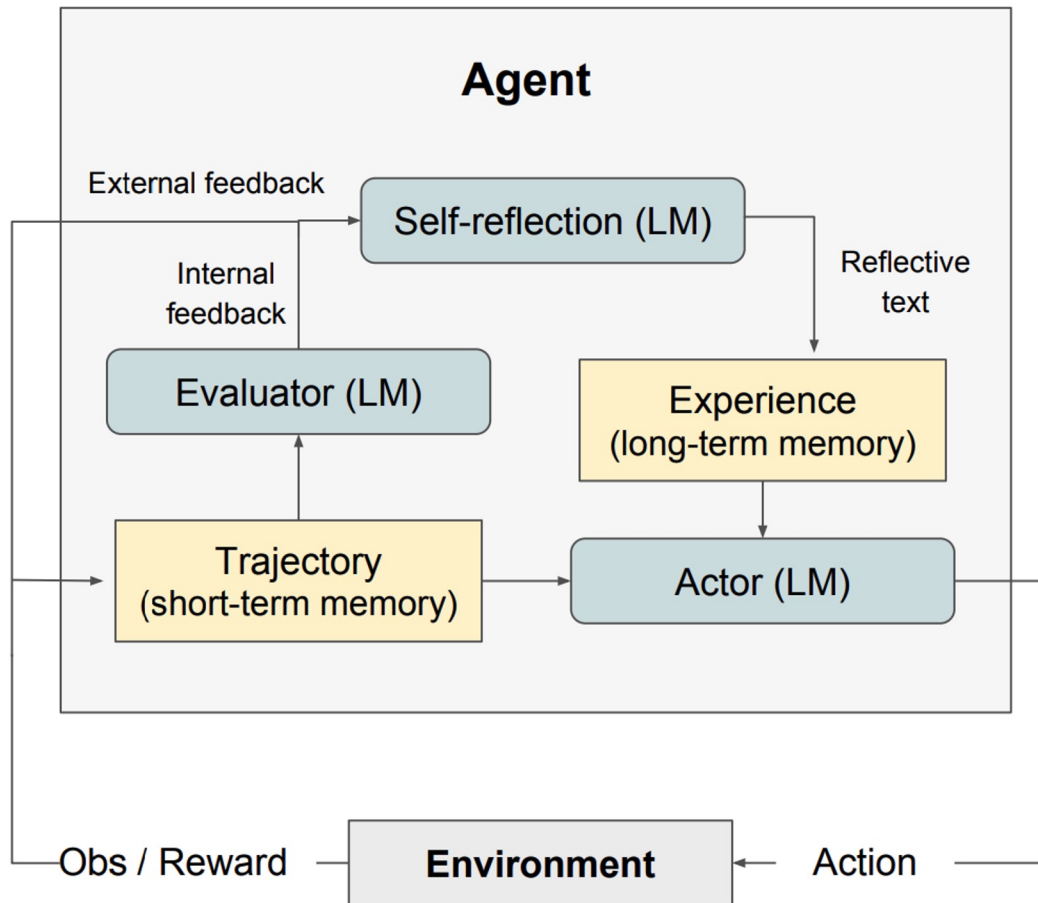
Append sr_t to mem

Increment t

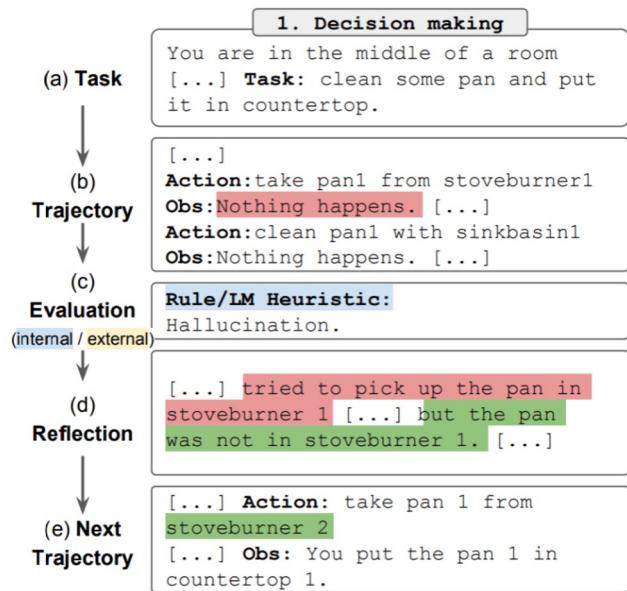
end while

return

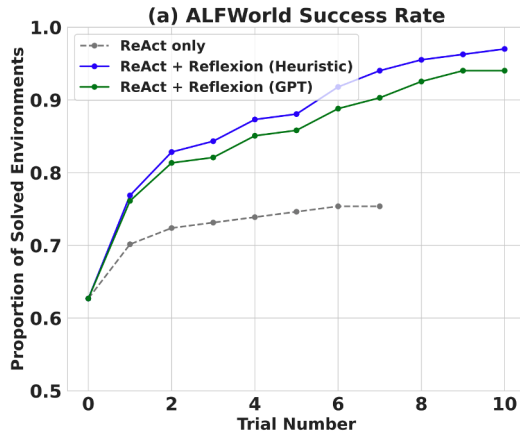
Verbal RL



Decision-making



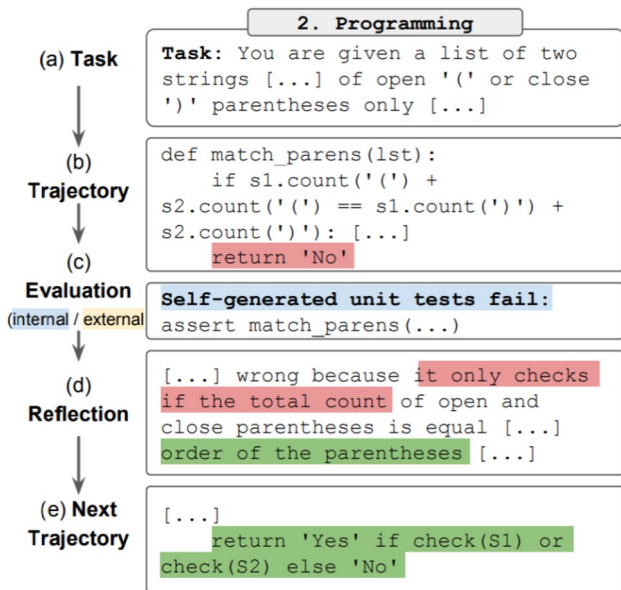
1. (Task)
2. (Action + observation trajectory)
3. (Evaluation)
4. (Self-reflection)
5. (Repeat)



```
Welcome!  
  
You are in the middle of the room.  
Looking around you, you see  
a diningtable, a stove,  
a microwave, and a cabinet.  
  
Your task is to:  
Put a pan on the diningtable.  
  
> goto the cabinet  
  
You arrive at the cabinet.  
The cabinet is closed.  
  
> open the cabinet  
  
The cabinet is empty.
```



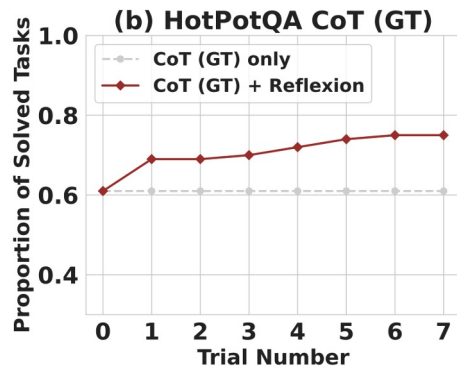
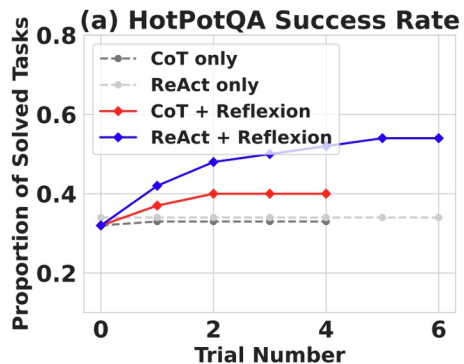
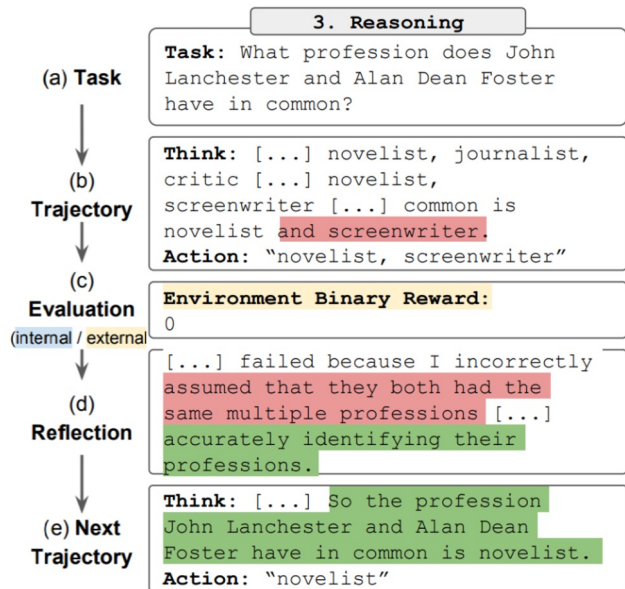
Programming



Benchmark + Language	Prev SOTA Pass@1	SOTA Pass@1	Reflexion Pass@1
HumanEval (PY)	65.8 (CodeT [4] + GPT-3.5)	80.1 (GPT-4)	91.0
HumanEval (RS)	–	60.0 (GPT-4)	68.0
MBPP (PY)	67.7 (CodeT [4] + Codex [5])	80.1 (GPT-4)	77.1
MBPP (RS)	–	70.9 (GPT-4)	75.4
Leetcode Hard (PY)	–	7.5 (GPT-4)	15.0

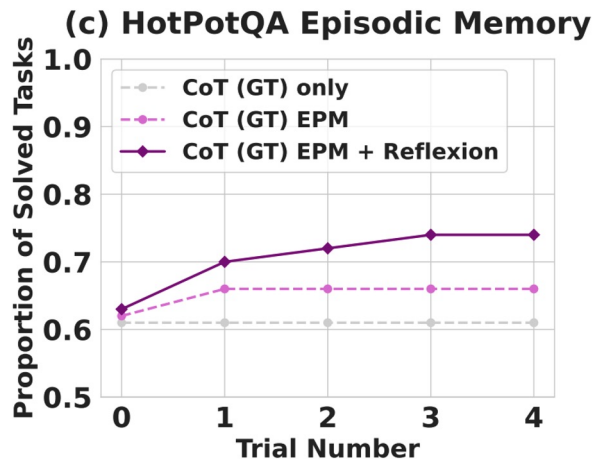
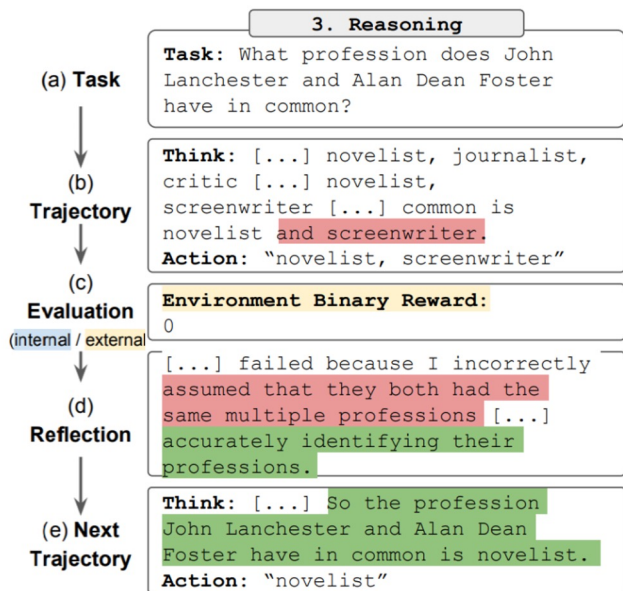
1. (Instruction)
2. (Function implementation)
3. (Unit test feedback) ← execution
4. (Self-reflection)
5. (Repeat)

Reasoning



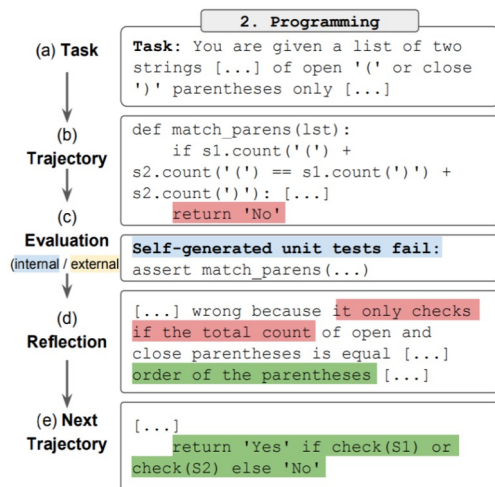
1. (Question)
2. (Knowledge retrieval)
3. (Answer)
4. (Exact match evaluation)
5. (Self-reflection)
6. (Repeat)

Is episodic memory enough for self-improvement?



1. (Question)
2. (Knowledge retrieval)
3. (Answer)
4. (Exact match evaluation)
- ~~5. (Self reflection)~~
6. (Repeat)

Is self-debugging enough for self-improvement?



Test generation omission

1. (Instruction)
2. (Function implementation)
- ~~3. (Unit test feedback) ← execution~~
4. (Self-reflection)
5. (Repeat)

Self-reflection omission

1. (Instruction)
2. (Function implementation)
3. (Unit test feedback) ← execution
- ~~4. (Self reflection)~~
5. (Repeat)

Approach	Test Generation	Self-reflection	Pass@1 (Acc)
Base model	False	False	0.60
Test generation omission	False	True	0.52
Self-reflection omission	True	False	0.60
Reflexion	True	True	0.68

More directions

- Various exploration techniques exist
- Persisting memory can store new ideas
- Few-shot demonstrations are powerful
- Transfer learning in semantic space

Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." arXiv preprint arXiv:2305.10601 (2023).

Wang, Guanzhi, et al. "Voyager: An open-ended embodied agent with large language models." arXiv preprint arXiv:2305.16291 (2023).

Du, Yilun, et al. "Improving Factuality and Reasoning in Language Models through Multiagent Debate." *arXiv preprint arXiv:2305.14325* (2023).

References

1. Chen, Bei, et al. "Codet: Code generation with generated tests." arXiv preprint arXiv:2207.10397 (2022).
2. Du, Yilun, et al. "Improving Factuality and Reasoning in Language Models through Multiagent Debate." *arXiv preprint arXiv:2305.14325* (2023).
3. Goodman, Noah. "Meta-Prompt: A Simple Self-Improving Language Agent." <https://noahgoodman.substack.com/p/meta-prompt-a-simple-self-improving>
4. Saunders, William, et al. "Self-critiquing models for assisting human evaluators." *arXiv preprint arXiv:2206.05802* (2022).
5. Shinn, Noah, et al. "Reflexion: Language Agents with Verbal Reinforcement Learning." arXiv preprint arXiv:2303.11366 (2023).
6. Shridhar, Mohit, et al. "Alfworld: Aligning text and embodied environments for interactive learning." arXiv preprint arXiv:2010.03768 (2020).
7. Wang, Guanzhi, et al. "Voyager: An open-ended embodied agent with large language models." arXiv preprint arXiv:2305.16291 (2023).
8. Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." arXiv preprint arXiv:2201.11903 (2022).
9. Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).
10. Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." arXiv preprint arXiv:2305.10601 (2023).