1 Token Efficiency for Game of 24

RAFA is superior in terms of token complexity. Methods that lack planning like Reflexion has a low token demand, however, it is not enough to compensate for the drop in performance. Methods that lack in-context learning like ToT would generate unnecessarily repeated trials due to lack of reflection and improvement, which makes the method token inefficient.

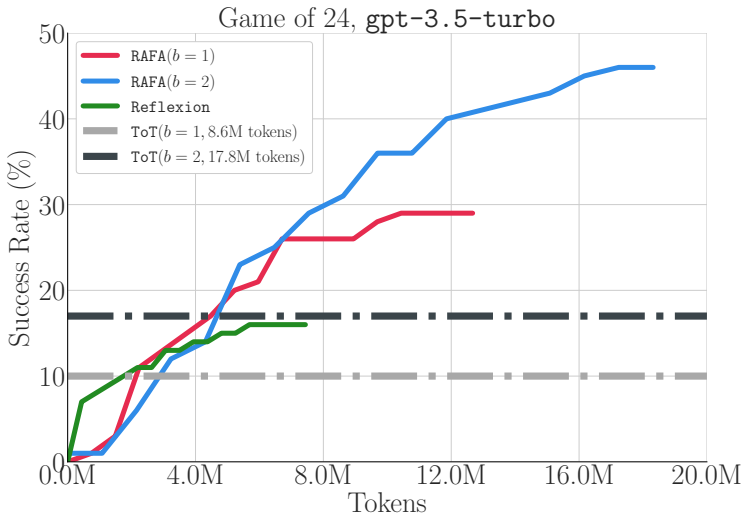

Figure 1: Token efficiency on Game of 24 using GPT-3.5-turbo.

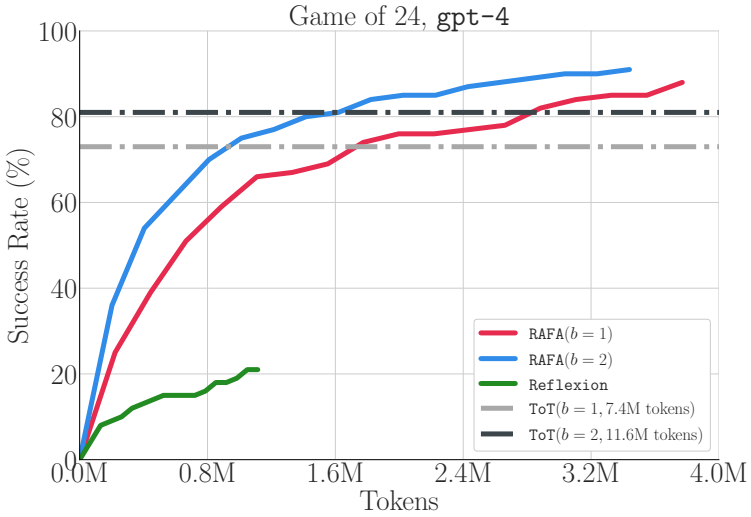

Figure 2: Token efficiency on Game of 24 using GPT-4

2 Token Efficiency for BlocksWorld

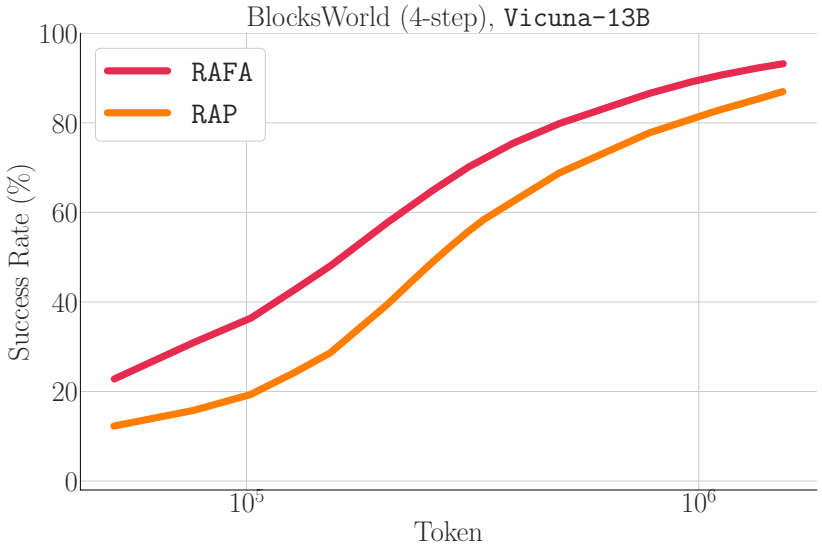

Figure 3: Token efficiency for RAFA and RAP on 4-step task of BlocksWorld, where the used LLM is Vicuna-13B (v1.3).

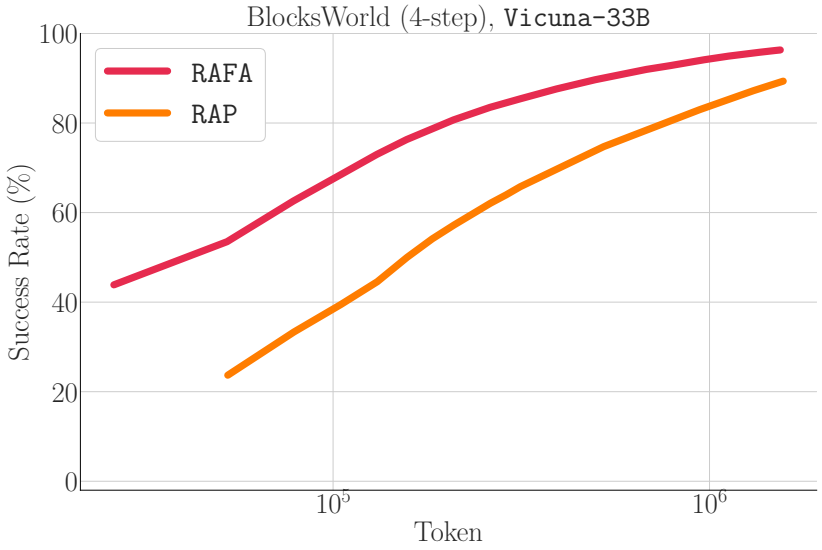

Figure 4: Token efficiency for RAFA and RAP on 4-step task of BlocksWorld, where the used LLM is Vicuna-33B (v1.3).

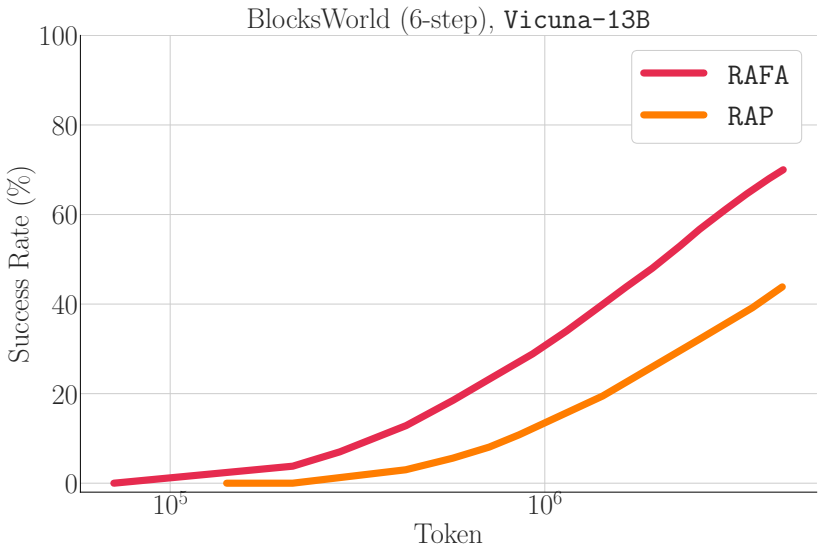

Figure 5: Token efficiency for RAFA and RAP on 6-step task of BlocksWorld, where the used LLM is Vicuna-13B (v1.3).

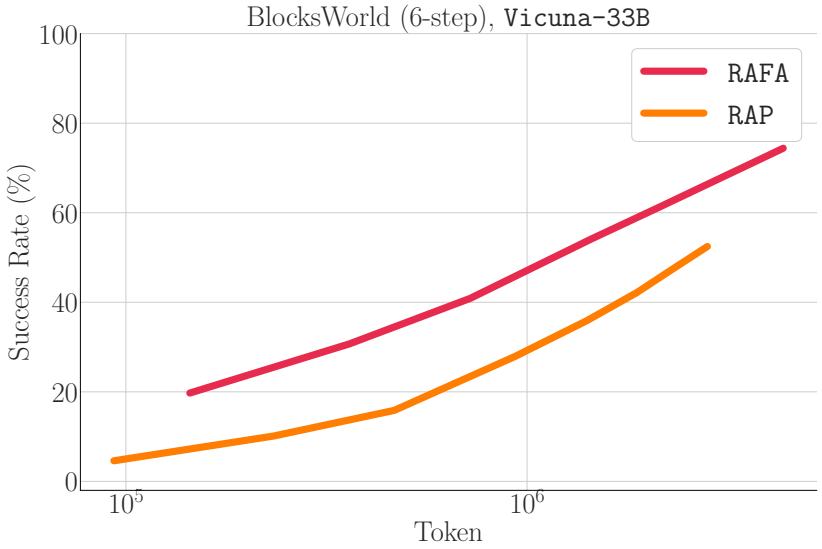

Figure 6: Token efficiency for RAFA and RAP on 6-step task of BlocksWorld, where the used LLM is Vicuna-33B (v1.3).

3 Token Efficiency for ALFWorld