ReAct: Synergizing Reasoning and Acting in Language Models

AID 6기 강민석

기계공학부 4학년

Github: @myeolinmalchi

Mail: rkd2274@pusan.ac.kr

부산대학교
PUSAN NATIONAL UNIVERSITY

1. Introduction

인간 지능의 특징:

1. Task-oriented actions with verbal reasoning
2. Cognition for self-regulation, strategization, working memory

Ex: 주방에서 요리를 하는 상황

인간은 구체적 Action 사이에 언어적인 추론을 한다.

1. Tracking progress:
   모든 재료를 썰었으니, 물을 데워야겠다.
2. Exception handling or plan adjustment:
   소금이 없으니 간장과 후추를 대신 사용해야겠다.
3. External information need:
   반죽은 어떻게 만들어야 하지? 인터넷에서 찾아봐야겠다.

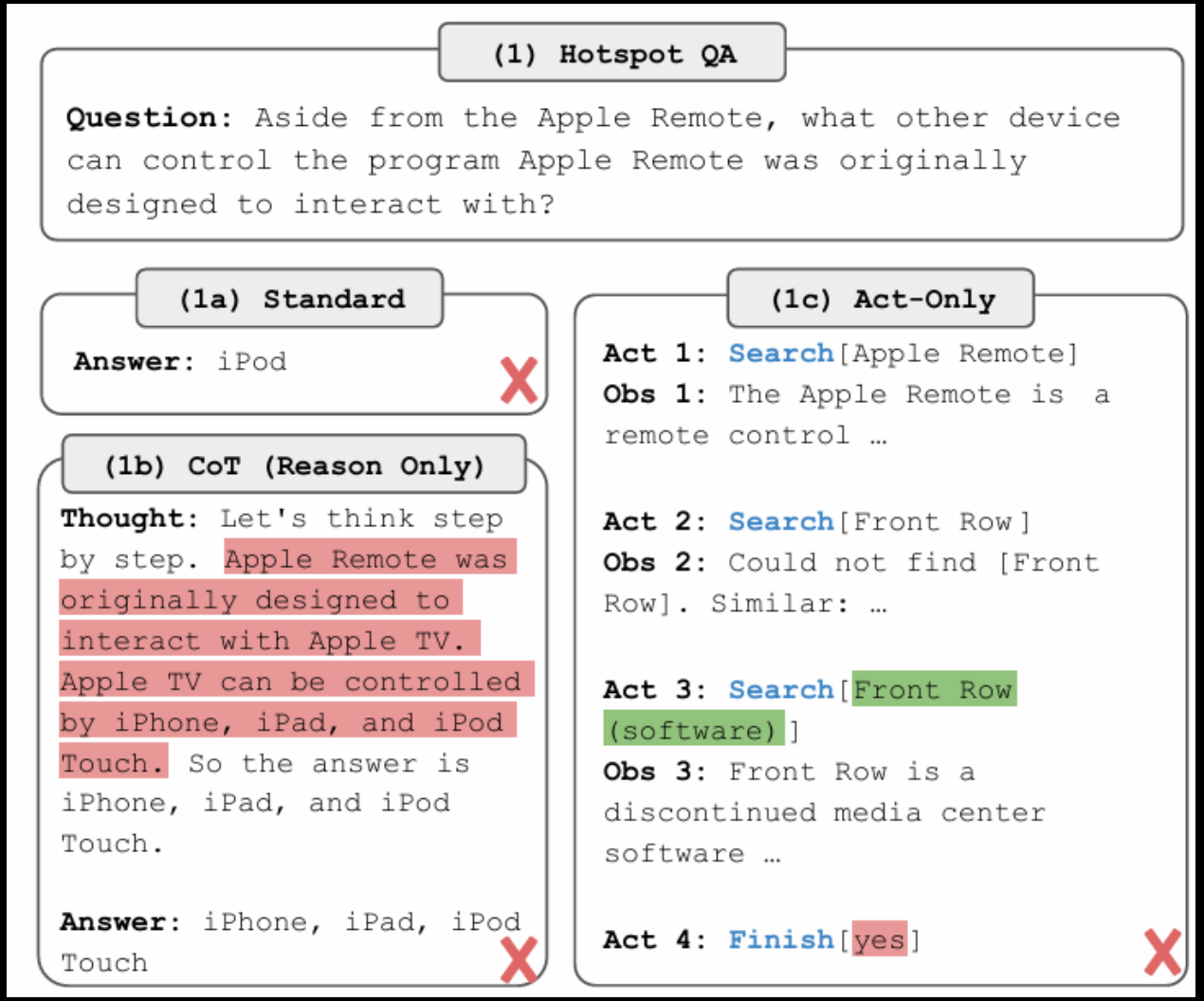
>> "Acting"과 "Reasoning"의 시너지로 낯선 상황에서 빠르게 과업 수행

1. Introduction

Chain-of-Thought(CoT)의 한계:

1. 모델의 추론 과정이 "내부 세계"에만 의존
2. "외부 세계"와는 단절되므로
   Knowledge Update가 불가능
3. Hallucination 및 Error propagation 문제

>> Reasoning 및 Acting을 결합하여 General task solving에 적용 시도



2. ReAct: Synergizing Reasoning + Acting

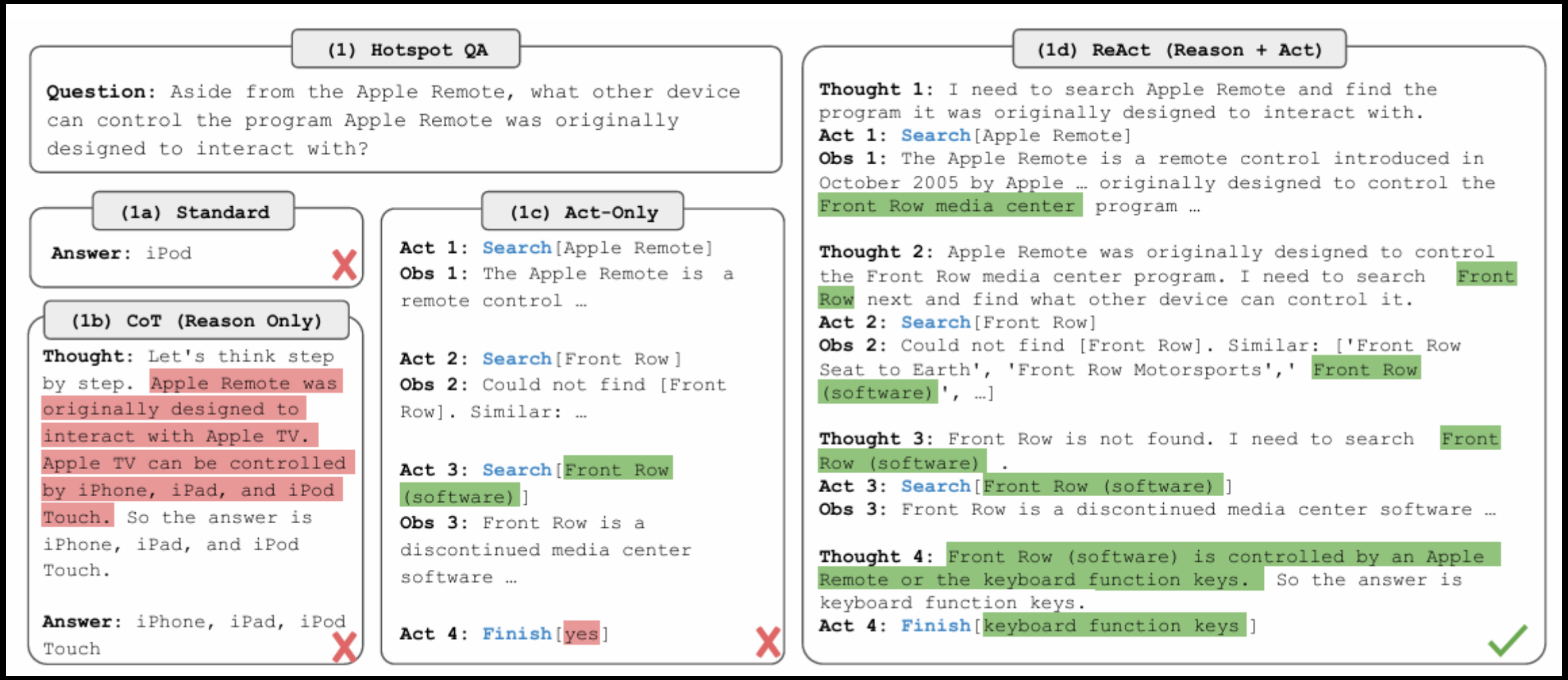
General task solving with environment:

1. Agent는 환경(environment)과 상호작용한다.
2. Observation($o_t \in O$) 및 Policy를 바탕으로 Action($o_t \in O$) 수행
3. Policy는 문맥 기반: $\pi(a_t|c_t),\ c_t = (o_1, a_1, \ldots, o_{t-1}, a_{t-1}, a_t)$,

The idea of ReAct is simple:

1. Agent의 Action 공간(A)을 확장:
   >> $A' = A \cup L$, L은 외부 언어 공간(Language space)
2. 환경에 영향을 주지 않는 Though or reasoning trace: $\hat{a} \in L$
3. context에 대한 reasoning 수행에서 유용한 정보 역할을 하거나,
   context를 update할 때 사용된다: $c_{t+1} = (c_t, \hat{a}_t)$

2. ReAct: Synergizing Reasoning + Acting



3. Knowledge-Intensive Reasoning Tasks

Domains:

1. HotPotQA:
   다수의 문서를 종합하여 질문에 답하고 근거를 제시하도록 설계된,
   설명 가능한 **다중 문서 추론 질문 응답 데이터셋**
2. FEVER:
   위키피디아 문서를 기반으로 주어진 주장(claim)이 사실인지 검증하고,
   그 **근거 문장을 함께 제시해야 하는 사실 검증 데이터셋**

Action Space:

1. search[entity]: 위키피디아에서 entity에 대한 문서 내용 반환
2. lookup[string]: 해당 문자열이 포함된 문서 내에서 다음 문장을 반환
3. finish[answer]: 현재 작업을 중단하고 최종 답변 제출

3.2 Methods

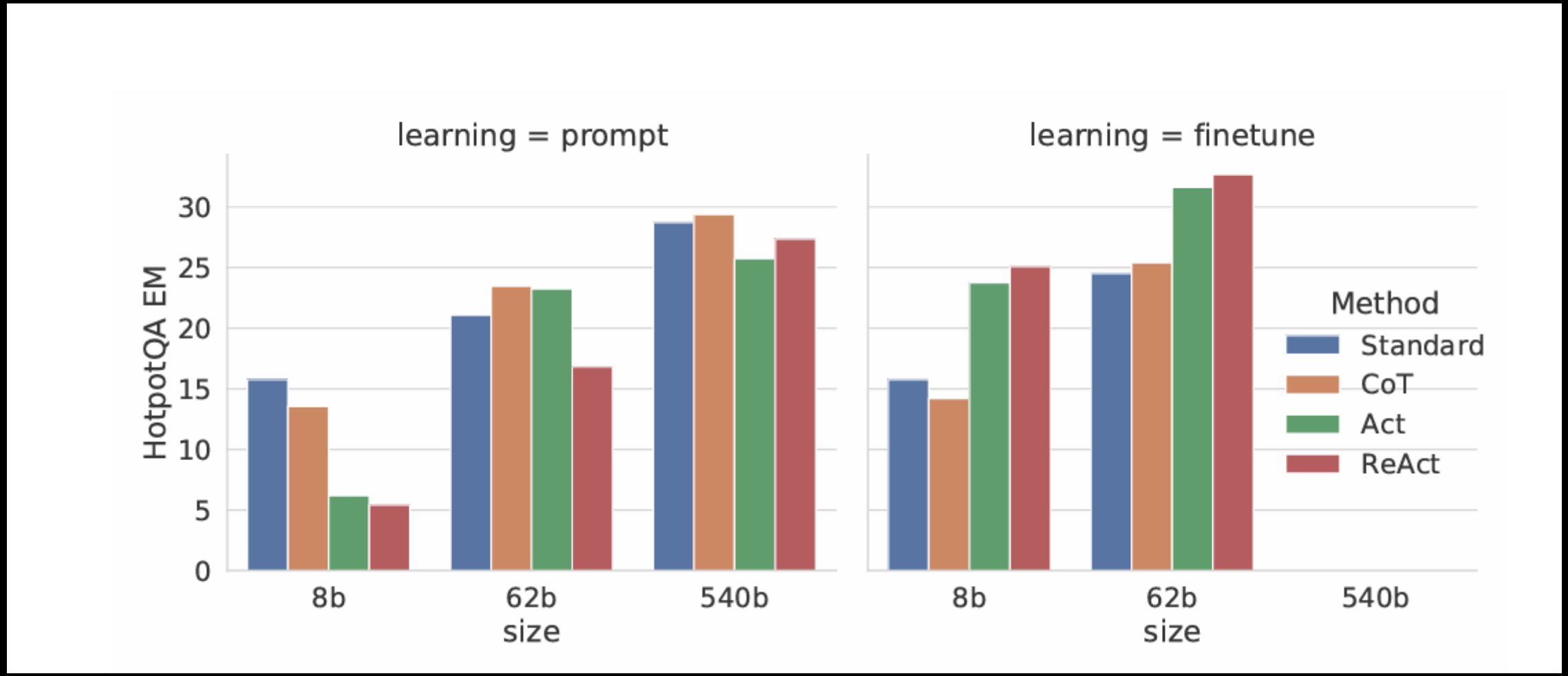
ReAct Prompting For HotpotQA and Fever, we randomly select 6 and 3 cases[2] from the training set and manually compose ReAct-format trajectories to use as few-shot exemplars in the prompts. Similar to Figure 1(d), each trajectory consists of multiple thought-action-observation steps (i.e. dense thought), where free-form thoughts are used for various purposes. Specifically, we use a combination of thoughts that decompose questions ("I need to search x, find y, then find z"), extract information from Wikipedia observations ("x was started in 1844", "The paragraph does not tell x"), perform commonsense ("x is not y, so z must instead be...") or arithmetic reasoning ("1844 < 1989"), guide

Baselines We systematically ablate ReAct trajectories to build prompts for multiple baselines (with formats as Figure 1(1a-1c)): (a) **Standard prompting** (Standard), which removes all thoughts, actions, observations in ReAct trajectories. (b) **Chain-of-thought prompting** (CoT) (Wei et al., 2022), which removes actions and observations and serve as a reasoning-only baseline. We also build a self-consistency baseline (CoT-SC) (Wang et al., 2022a;b) by sampling 21 CoT trajectories with decoding temperature 0.7 during inference and adopting the majority answer, which is found to consistently boost performance over CoT. (c) **Acting-only prompt** (Act), which removes thoughts in ReAct trajectories, loosely resembling how WebGPT (Nakano et al., 2021) interacts with the Internet to answer questions, though it operates on a different task and action space, and uses imitation and reinforcement learning instead of prompting.

3.2 Methods

Combining Internal and External Knowledge As will be detail in Section 3.3, we observe that the problem solving process demonstrated by ReAct is more factual and grounded, whereas CoT is more accurate in formulating reasoning structure but can easily suffer from hallucinated facts or thoughts. We therefore propose to incorporate ReAct and CoT-SC, and let the model decide when to switch to the other method based on the following heuristics: A) **ReAct → CoT-SC**: when ReAct fails to return an answer within given steps, back off to CoT-SC. We set 7 and 5 steps for HotpotQA and FEVER respectively as we find more steps will not improve ReAct performance[3]. B) **CoT-SC → ReAct**: when the majority answer among n CoT-SC samples occurs less than $n/2$ times (i.e. internal knowledge might not support the task confidently), back off to ReAct.

3.3 Result and observations



3.3 Result and observations

	Type	Definition	ReAct	CoT
Success	True positive	Correct reasoning trace and facts	94%	86%
	False positive	Hallucinated reasoning trace or facts	6%	14%
Failure	Reasoning error	Wrong reasoning trace (including failing to recover from repetitive steps)	47%	16%
	Search result error	Search return empty or does not contain useful information	23%	-
	Hallucination	Hallucinated reasoning trace or facts	0%	56%
	Label ambiguity	Right prediction but did not match the label precisely	29%	28%

A) **Hallucination is a serious problem for CoT,**

B) **While interleaving reasoning, action and observation steps improves ReAct's groundedness and trustworthiness, such a structural constraint also reduces its flexibility in formulating reasoning steps,**

C) **For ReAct, successfully retrieving informative knowledge via search is critical.**

3.3 Result and observations

ReAct + CoT-SC perform best for prompting LLMs Also shown in Table 1, the best prompting method on HotpotQA and Fever are ReAct → CoT-SC and CoT-SC → ReAct respectively. Furthermore, Figure 2 shows how different methods perform with respect to the number of CoT-SC samples used. While two ReAct + CoT-SC methods are advantageous at one task each, they both significantly and consistently outperform CoT-SC across different number of samples, reaching CoT-SC performance with 21 samples using merely 3-5 samples. These results indicate the value of properly combining model internal knowledge and external knowledge for reasoning tasks.

4. Decision Making Tasks

Domains:

1. ALF World:
   text-based game으로 6가지 타입의 tasks가 존재한다.
   외부 환경과 자연어로 navigating과 interacting을 한다.
   (e.g. go to coffetable 1, take paper 2, use desklamp 1)

2. WebShop:
   WebShop은 ALFWorld와 달리 엄청 다양한 structured & unstructured texts (상품명, 상품 설명, 옵션 등을 포함한 Amazon에서 크롤링 된 데이터)를 가지고 있으며 1.18M real-world products와 12k human instructions를 가지고 있다. 이 task는 user instruction (e.g. "I am looking for a nightstand with drawers. It should have a nickel finish, and priced lower than $140.")에 따라 적합한 상품을 구매하는 것이다.

4. Decision Making Tasks

Method	Pick	Clean	Heat	Cool	Look	Pick 2	All
Act (best of 6)	88	42	74	67	72	**41**	45
ReAct (avg)	65	39	83	76	55	24	57
ReAct (best of 6)	**92**	58	**96**	86	**78**	**41**	**71**
ReAct-IM (avg)	55	59	60	55	23	24	48
ReAct-IM (best of 6)	62	**68**	87	57	39	33	53
BUTLER$_g$ (best of 8)	33	26	70	76	17	12	22
BUTLER (best of 8)	46	39	74	**100**	22	24	37

Table 3: AlfWorld task-specific success rates (%). BUTLER and BUTLER$_g$ results are from Table 4 of Shridhar et al. (2020b). All methods use greedy decoding, except that BUTLER uses beam search.

Method	Score	SR
Act	62.3	30.1
ReAct	**66.6**	**40.0**
IL	59.9	29.1
IL+RL	62.4	28.7
Human Expert	82.1	59.6

Table 4: Score and success rate (SR) on Webshop. IL/IL+RL taken from Yao et al. (2022).