Automatic Reasoning and Tool-use (ART)

Combining CoT prompting and tools in an interleaved manner has shown to be a strong and robust approach to address many tasks with LLMs. These approaches typically require hand-crafting task-specific demonstrations and carefully scripted interleaving of model generations with tool use. Paranjape et al., (2023) propose a new framework that uses a frozen LLM to automatically generate intermediate reasoning steps as a program.

ART works as follows:

- given a new task, it select demonstrations of multi-step reasoning and tool use from a task library
- at test time, it pauses generation whenever external tools are called, and integrate their output before resuming generation

ART encourages the model to generalize from demonstrations to decompose a new task and use tools in appropriate places, in a zero-shot fashion. In addition, ART is extensible as it also enables humans to fix mistakes in the reasoning steps or add new tools by simply updating the task and tool libraries. The process is demonstrated below:

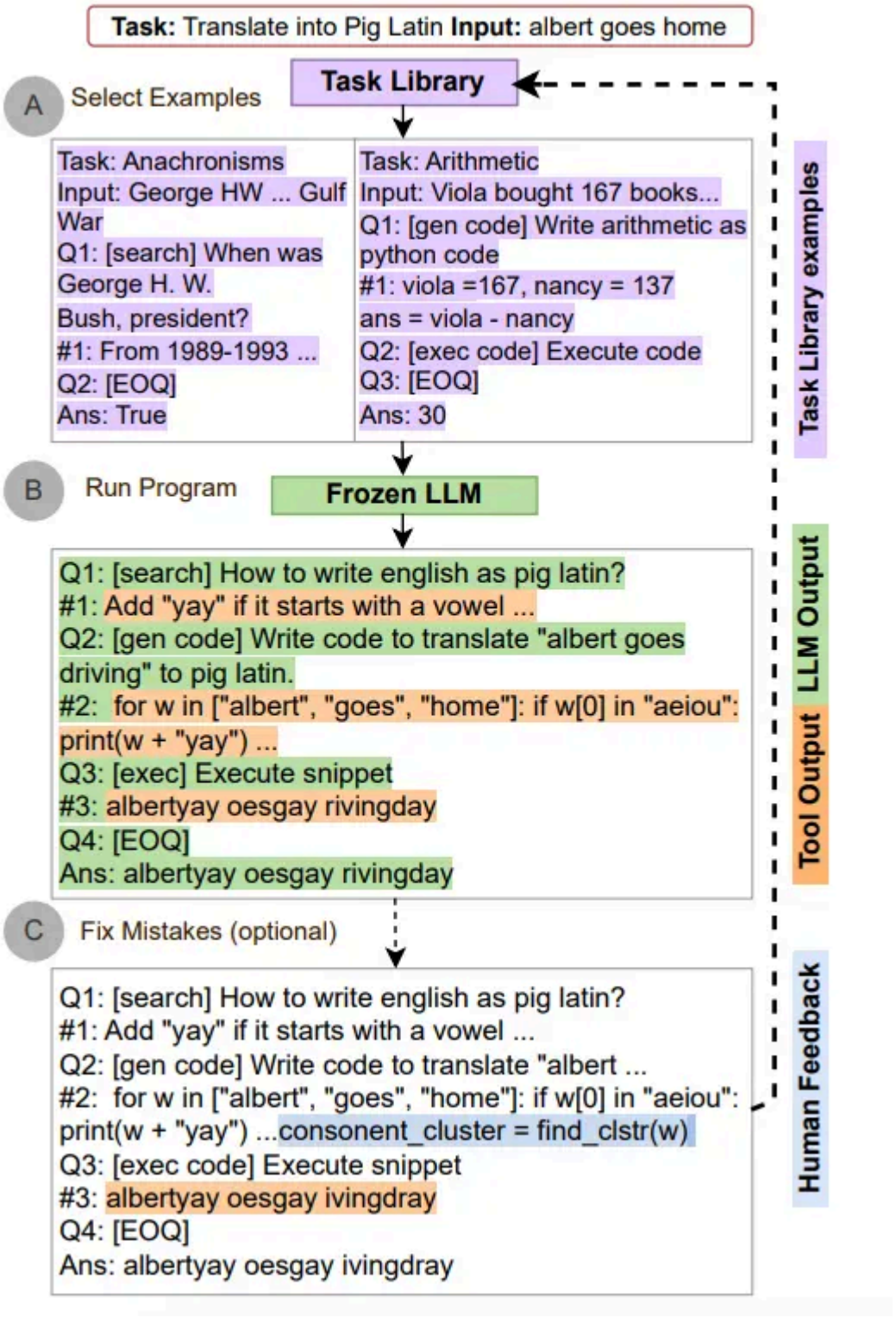

Image Source: [Paranjape et al., (2023)]

ART substantially improves over few-shot prompting and automatic CoT on unseen tasks in the BigBench and MMLU benchmarks, and exceeds performance of hand-crafted CoT prompts when human feedback is incorporated.

Below is a table demonstrating ART's performance on BigBench and MMLU tasks:

Task Name (Cluster)	Few Shot	AutoCot	ART w/o Tool Use	ART	GPT-3 Best
Test Tasks					
Sentence Ambiguity (Search)	70.67[5]	51.47	71.00	73.33	-
Strategy QA (Search)	55.49[5]	27.22	59.37	66.44	-
Physics (Search)	70.09[5]	61.83	59.13	67.55	-
Δ with ART (Search)	**+3.7**	**+22.27**	**+ 5.9**		
Physics Questions (Arithmetic)	7.02[5]	5.56	6.30	20.37	-
Operators (Arithmetic)	71.23[7]	75.52	71.80	92.00	-
Unit interpretation (Arithmetic)	58.2[7]	41.20	51.4	53.99	-
Repeat copy logic (Arithmetic)	50.01[7]	15.63	31.25	44.38	-
Object Counting (Arithmetic)	39.2[7]	26.80	42.2	87.00	81.20[1]
Penguins in a table (Arithmetic)	58.23[7]	40.40	68.86	77.85	72.34[1]
Reasoning about objects (Arithmetic)	71.00[7]	33.33	45.35	64.34	52.69[1]
Tracking shuffled objects (Arithmetic)	22.39[7]	19.44	18.14	37.67	36.32[1]
Δ with ART (Arithmetic)	**+19.0**	**+36.7**	**+ 23.1**		**+6.1**
Word Unscramble (String)	40.72[7]	32.44	23.03	42.7	-
Simple Text Editing (Code)	35.31[5]	30.21	20.74	27.65	-
CS Algorithms (Code)	73.48[7]	0.0	41.59	88.11	-
Sports Understanding (CoT)	69.74[5]	51.47	92.89	-	86.59[1]
Snarks (CoT)	54.58[5]	57.24	57.13	-	65.2[1]
Disambiguation QA (Free-form)	55.03[5]	48.45	55.89	-	60.62[1]
Temporal sequences (CoT)	55.80[7]	19.70	49.5	-	81.8[1]
Ruin names (CoT)	71.01[5]	55.28	60.22	-	-
Δ with ART (Misc)	**2.4**	**22.5**	**24.37**		-9.4
Δ with ART (Overall)	**+6.9**	**+24.6**	**+16.7**		-1.7
	MMLU				
College Computer Science (Search)	41.00	43.99	63.40	67.80	63.6[6]
Astronomy (Search)	62.10	41.48	76.71	79.1	62.5[6]
Business Ethics (Search)	61.60	48.8	77.17	81.16	72.7[6]
Virology (Search)	50.03	49.52	71.60	71.49	50.72[6]
Geography (Search)	77.67	57.07	70.30	71.71	81.8[6]
Mathematics (Arithmetic)	36.67	33.77	39.50	45.66	34.5[6]
Δ with ART (MMLU)	**+14.6**	**+23.7**	**+3.0**		**+8.5**

Table 3: ART performance on BigBench tasks and MMLU tasks. ([1] Human-crafted CoT (Wei et al., 2022; Suzgun et al., 2022), [5] InstructGPT (Ouyang et al., 2022), [6] Scaled instruction finetuning (Chung et al., 2022), [7] Code-davinci-002 (Chen et al., 2021)).

Image Source: Paranjape et al., (2023)

Last updated on September 19, 2024