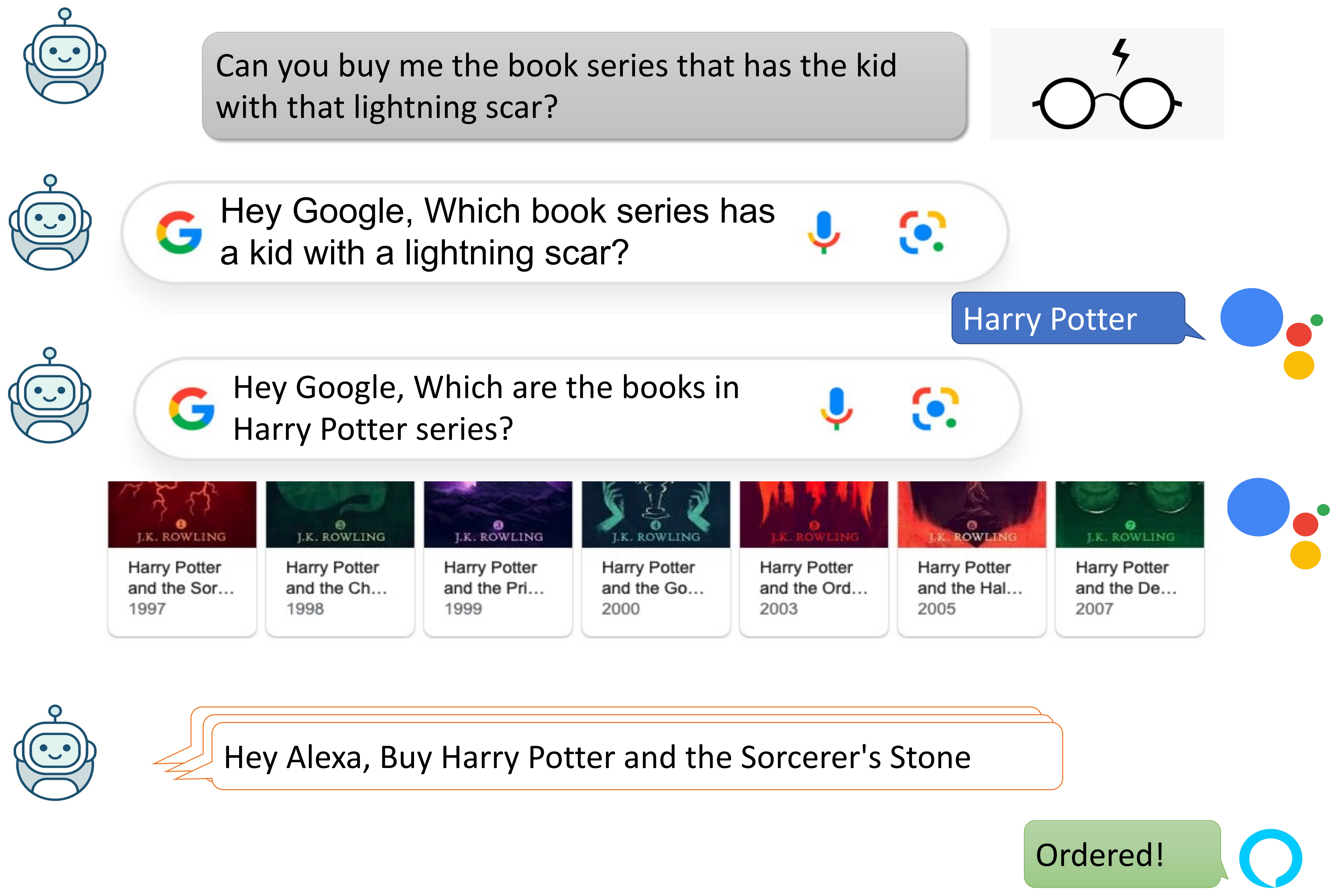
ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Hey AI, Can You Solve Complex Tasks by Talking to Agents?

Tushar Khot, Kyle Richardson, Daniel Khashabi, Ashish Sabharwal

AI2

Motivation



Can you buy me the book series that has the kid with that lightning scar?

Hey Google, Which book series has a kid with a lightning scar?

Harry Potter

Hey Google, Which are the books in Harry Potter series?

Harry Potter and the Sor… 1997
Harry Potter and the Ch… 1998
Harry Potter and the Pri… 1999
Harry Potter and the Go… 2000
Harry Potter and the Ord… 2003
Harry Potter and the Hal… 2005
Harry Potter and the De… 2007

Hey Alexa, Buy Harry Potter and the Sorcerer's Stone

Ordered!

Summary

New Task: Learning to Talk to Agents to Solve Complex Tasks

1. Solve a complex task by breaking it down into agent's capabilities
2. Interact with agents in their expected and natural language

Why?

- Green AI: Reuse existing expensive and even proprietary models
- Better Long-Term Bet: No need to learn every task from scratch
- Interpretability: Naturally modular and interpretable systems
- Technical Challenge: Search for solutions by interacting with NL agents

New Dataset: CommaQA: Communicating with Agents for QA

- Synthetic Multi-hop QA Dataset solvable using agents
- Challenging for current black-box models and task baselines

https://github.com/allenai/CommaQA

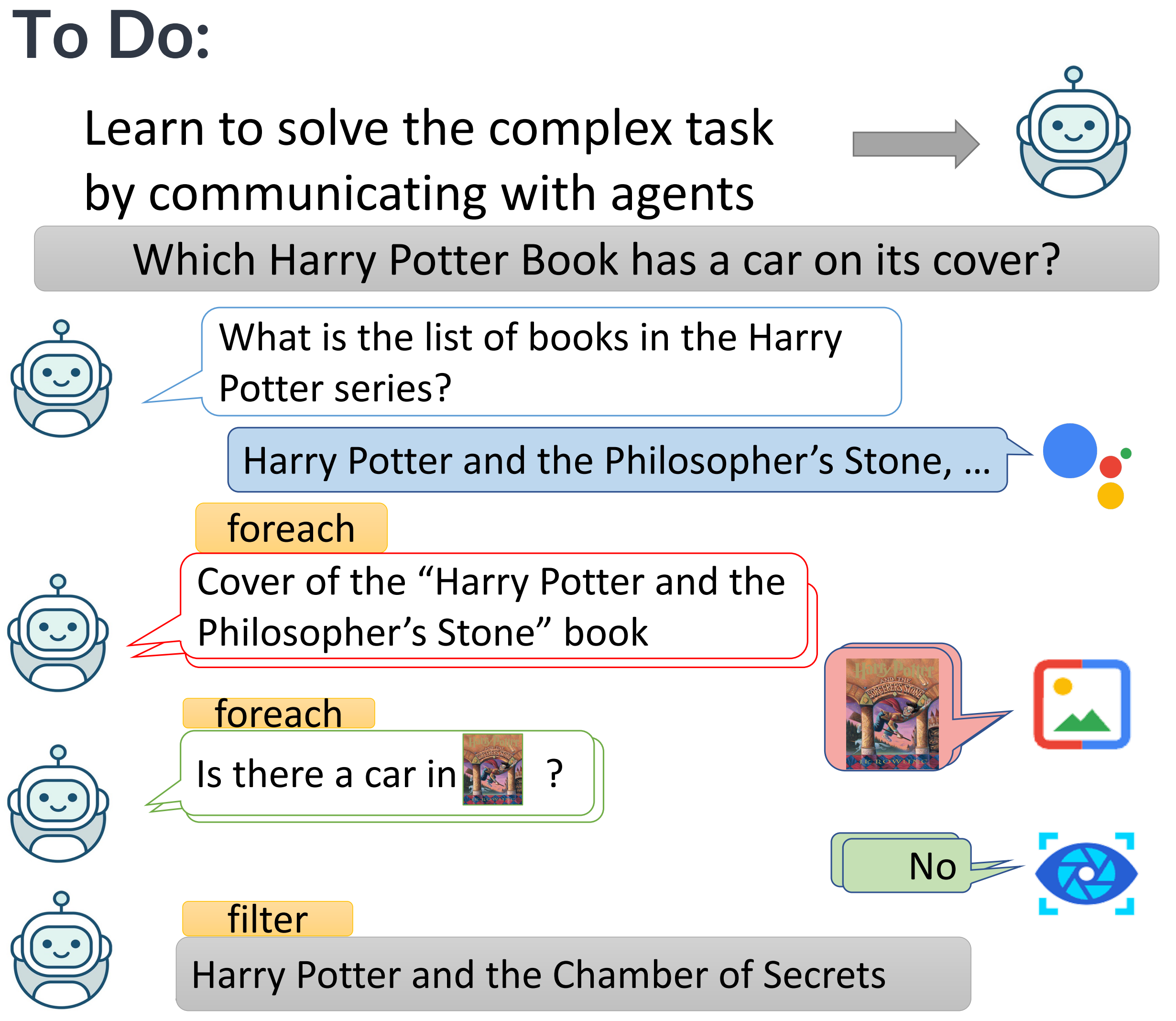
Task: Learning to Talk to Agents to Solve Complex Tasks

Given:

Agents	Examples of Valid Inputs
	What is the list of books in LOTR series Which movies are in the Fast & Furious series …
	Cover of "The Hobbit" book Poster of the "2 Fast 2 Furious" movie …
	Is there a truck in ? …
Complex Task Supervision	Which Harry Potter Book has a car on its cover? => Harry Potter and the Chamber of Secrets

To Do:

Learn to solve the complex task by communicating with agents



Which Harry Potter Book has a car on its cover?

What is the list of books in the Harry Potter series?

Harry Potter and the Philosopher's Stone, …

foreach

Cover of the "Harry Potter and the Philosopher's Stone" book

foreach

Is there a car in ?

No

filter

Harry Potter and the Chamber of Secrets

Should **NOT** rely on:

Training Data of the Agents

Agents Internal Knowledge

Conversation Supervision

Dataset: CommaQA -- Communicating with Agents for QA

CommaQA-E

TextQA	TableQA
What awards has Sian Heder won? …	What awards has Beyoncé won? …

What awards have the actors of the Oscar-winning movies received? => ["Oscar", …"]

Inspired by HotpotQA (Yang et al. '18), MusiQue (Trivedi et al. '21)

CommaQA-I

TextQA	KBQA
What objects does Apple make? …	What is IPhone a type of? …

What objects has Steve Jobs likely used? => ["Cell Phone", …"]

Inspired by OpenBookQA (Mihaylov et al. '18), StrategyQA (Geva et al. '21)

CommaQA-N

TextQA	MathQA
What was the length of javelin throws by Shey ? …	What is the maximum value in [35.0, 15.0, 25.0]? …

What was the gap between the longest javelin throws from USA and Ireland?=> 2.0

Inspired by DROP (Dua et al. '19)

Handling Complex Answers

project	return [(x, f_i(q(x))) for x in a]	#1=[Jordan, Johnson] q="What were the lengths of throw by #1?" f_i= textqa → [(Jordan, [23, 34]), (Johnson, [45, 56])]
projectValues	return [(k, f_i(q(v))) for (k, v) in a]	#1=[(Jordan, [23, 34]), (Johnson, [45, 56])] q="Which is largest value in #1?" f_i= mathqa → [(Jordan, 34), (Johnson, 56)]
filter	return [x for x in a if f_i(q(x))]	#1=[23, 34, 56] q="Is #1 greater than 50?" f_i= mathqa → [56]

```
What awards have movies written by people born in 1905 won?                    CommaQA-E
  (select) [text] Who were born in the year 1905?                          A: ["Gigafuna"]
  (project_values_flat_unique) [table] What movies has #1 written?         A: ["Pneumodendron", "Pipesia", "Riften"]
  (project_values_flat_unique) [table] Which awards were given to #2?      A: ["Pludgel", "Dessication", "Pianogram"]
```

NOT a standard RC/open-domain QA task – system needs to communicate with the agents

Results

Unsolved using the current baselines that talk to agents

Black-box models struggle even when given access to the agent's private knowledge

But solvable by training on conversation supervision (oracle upper bound)

TMN: Khot et al, '21; T5: Raffel et al. '20; UQA: Khashabi et al. '20

Model	Aux. Info	E	I	N	Avg.
TMN-S_5		0.0	0.0*	0.0	0.0
TMN-S_{10}		17.0	0.0*	0.0	5.7
Auxiliary Supervision Models					
T5-L	KB	0.9	10.2	35.4	15.5
UQA-L		1.0	10.2	39.0	16.7
TMN-G		75.4	36.0	100.0	70.5
TMN-S		100.0	100.0	100.0	100.0

Related Work

Multi-hop QA Datasets: These datasets (Khashabi et al., 2018; Mihaylov et al., 2018; Yang et al., 2018; Dua et al., 2019; Khot et al., 2020; Geva et al., 2021) can be potentially solved by composition of single-hop models but,

- Single-hop shortcuts incentivize non-compositional models (Min et al., 2019a; Trivedi et al., 2020)
- Lack reliable agents to solve single-hop sub-tasks (e.g. list answer QA) (Khot et al., 2021)

Question Decomposition: These approaches solve QA by decomposing complex question but,

- Current approaches generally limited to one (Talmor and Berant, 2018; Min et al., 2019b; Perez et al., 2020) QA model
- Many questions are out-of-scope due to lack of agents(Khot et al., 2021)
- Rely on human annotation of decomposition (Talmor and Berant, 2018; Min et al., 2019b)

Text-Based Games: Also require solving tasks by interacting with agents (often the game environment) but focus on different class of problems with different assumptions on agent's language. (Yuan et al., 2019, 2020; Hausknecht et al., 2020; Ammanabrolu et al., 2021; Jansen, 2021)