Benchmarks and Evaluation

Part 2 of the Tutorial on Complex Reasoning in Natural Language

https://wenting-zhao.github.io/complex-reasoning-tutorial/
Many types of reasoning skills
Many ways in which skills can be combined
Many possible settings and tasks in which skills can be applied
Hard to define settings that evaluate complex reasoning processes in a reliable and informative way.

But there are exciting tasks that even the best models today often fail on!
Plan

1. Core reasoning skills
2. Multimodal reasoning
3. Reasoning with background knowledge
4. Designing evaluations for complex reasoning
Core reasoning skills
Compositional (multi-hop) reasoning

Who succeeded the first President of Namibia?

Who was the first President of Namibia?

Who succeeded Sam Nujoma?

Hifikepunye Pohamba

⚠ GPT3 davinci-002 with self-ask prompting and an external search tool obtains 15.2 EM on MuSiQue 2-hop questions.
A carnival snack booth made $50 selling popcorn each day. It made three times as much selling cotton candy. For a 5-day activity, the booth has to pay $30 rent and $75 for the cost of the ingredients.

How much did the booth earn for 5 days after paying the rent and the cost of ingredients?

How much did the booth make selling cotton candy each day? ($50 \times 3 = $150)
How much did the booth make in a day? ($150 + $50 = $200)
How much did the booth make in 5 days? ($200 \times 5 = $1000)
How much did the booth have to pay? ($30 + $75 = $105)
How much did the booth earn after paying the rent and the cost of ingredients? ($1000 - $105 = $895)
Social reasoning

Alex spilled the food she just prepared all over the floor and it made a huge mess.

What will Alex want to do next?
(a) taste the food
(b) mop up
(c) run around in the mess

(b) mop up
Temporal reasoning

A: Yes, sir. May I help you?
B: Please I'd like a ticket to New York.
A: For today?
B: No, early Saturday morning.
A: We have a flight that we'll put you there at _______. Is that ok?
B: Nothing earlier? I prefer flight at 9 thirty.
A: I'm afraid not, unless you want a night flight.
B: No, exactly not.

(a) ten AM  
(b) 9:30 PM  
(c) eleven AM  
(d) four AM

⚠️ Gopher in a 5-shot setting obtains **50.9 accuracy** on TIMEDIAL compared to 97.8 human performance.
Multimodal reasoning
Who is taller, LeBron James or Stephen Curry?

In his sophomore season in 2007–08, Curry had grown to his adult height of 6 ft 3 in (1.91 m) and again led the Southern Conference in scoring, averaging 25.5 points per game while adding 4.7 rebounds per game and 2.8 assists per game.

Lebron James is considered one of the greatest basketball players of all time, and his height has been a huge part of his success. At 16 years old, he was 6'7″ and weighed an impressive 240 pounds. As he grew older and entered the NBA, his height increased to 6'8″ and his weight increased to 250 pounds.

Currently, at the age of 35, Lebron stands at 6'9″ and weighs 250 pounds. His height, combined with his athleticism and skill, has allowed him to dominate the sport of basketball for many years.
The 2016 Summer Olympics officially known as the Games of the XXXI Olympiad (Portuguese: Jogos da XXXI Olimpíada) and commonly known as Rio 2016, was an international multi-sport event.

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Season</th>
<th>Flag bearer</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXXI</td>
<td>2016</td>
<td>Summer</td>
<td>Yan Naing Soe</td>
</tr>
<tr>
<td>XXX</td>
<td>2012</td>
<td>Summer</td>
<td>Zaw Win Thet</td>
</tr>
<tr>
<td>XXIX</td>
<td>2008</td>
<td>Summer</td>
<td>Phone Myint Tayzar</td>
</tr>
<tr>
<td>XXVII</td>
<td>2004</td>
<td>Summer</td>
<td>Hla Win U</td>
</tr>
<tr>
<td>XXVII</td>
<td>2000</td>
<td>Summer</td>
<td>Maung Maung Nge</td>
</tr>
<tr>
<td>XX</td>
<td>1972</td>
<td>Summer</td>
<td>Win Maung</td>
</tr>
</tbody>
</table>

Best methods reach ~66 EM on HybridQA.

Yan Naing Soe (born 31 January 1995) competed at the 2016 Summer Olympics in the men’s 100 kg event. He was the flag bearer for Myanmar at the Parade of Nations.

Zaw Win Thet (born 1 March 1991 in Kyonpyaw, Pathein District, Ayeyarwady Division, Myanmar) is a Burmese runner who.

Myint Tayzar Phone (Burmese: မြင်သာဇီ ဖိနှိပ်) born July 2, 1978 is a sprint canoer from Myanmar who competed in the late 2000s.

Win Maung (born 12 May 1949) is a Burmese footballer. He competed in the men’s tournament at the 1972 Summer Olympics.

Q: In which year did the judoka bearer participate in the Olympic opening ceremony? A: 2016
Q: Which event does the XXXI Olympic flag bearer participate in? A: men’s 100 kg event
Q: Where does the Burmesse jodoka participate in the Olympic opening ceremony as a flag bearer? A: Rio
Q: For the Olympic event happening after 2014, what session does the Flag bearer participate? A: Parade of Nations
Q: For the XXXI and XXX Olympic event, which has an older flag bearer? A: XXXI
Q: When does the oldest flag Burmese bearer participate in the Olympic ceremony? A: 1972
United States House of Representatives Elections, 1972

<table>
<thead>
<tr>
<th>District</th>
<th>Incumbent</th>
<th>Party</th>
<th>Result</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>California 3</td>
<td>John E. Moss</td>
<td>democratic</td>
<td>re-elected</td>
<td>John E. Moss (d) 69.9% John Rakus (r) 30.1%</td>
</tr>
<tr>
<td>California 5</td>
<td>Phillip Burton</td>
<td>democratic</td>
<td>re-elected</td>
<td>Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%</td>
</tr>
<tr>
<td>California 8</td>
<td>George Paul Miller</td>
<td>democratic</td>
<td>Lost renomination democratic hold</td>
<td>Pete Stark (d) 52.9% Lew M. Warden, Jr. (r) 47.1%</td>
</tr>
<tr>
<td>California 14</td>
<td>Jerome R. Waldie</td>
<td>republican</td>
<td>re-elected</td>
<td>Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%</td>
</tr>
<tr>
<td>California 15</td>
<td>John J. Mcfall</td>
<td>republican</td>
<td>re-elected</td>
<td>John J. Mcfall (d) unopposed</td>
</tr>
</tbody>
</table>

Entailed Statements:
1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statements:
1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.
**Multimodal Context**

**Stall This Movie!**
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".

**La liceale**
La liceale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini. It was followed by "La liceale nella classe dei ripetenti".

**Sage Stallone**
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balboa Jr., the onscreen son of his father's title character. He did not, however, ...

... After that, he acted in lesser profile films.

**Pierino contro tutti**
Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an...

... I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

---

**Ben Piazza - Filmography**

<table>
<thead>
<tr>
<th>Year</th>
<th>Title</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1957</td>
<td>A Dangerous Age</td>
<td>David</td>
</tr>
<tr>
<td>1959</td>
<td>The Hanging Tree</td>
<td>Rune</td>
</tr>
<tr>
<td>1962</td>
<td>No Exit</td>
<td>Camarero</td>
</tr>
<tr>
<td>1970</td>
<td>Tell Me That You Love Me, Junie Moon</td>
<td>Jesse</td>
</tr>
<tr>
<td>1972</td>
<td>The Outside Man</td>
<td>Desk Clerk</td>
</tr>
<tr>
<td>1985</td>
<td>Mask</td>
<td>Mr. Simms</td>
</tr>
<tr>
<td>1988</td>
<td>Clean and Sober</td>
<td>Kramer</td>
</tr>
<tr>
<td>1990</td>
<td>Rocky V</td>
<td>Doctor</td>
</tr>
<tr>
<td>1991</td>
<td>Guilty by Suspicion</td>
<td>Darryl Zanuck</td>
</tr>
</tbody>
</table>

Q: Which **B. Piazza** title came earlier: the movie **S. Stallone's son** starred in or the movie with **half of a lady's face on the poster**?

A: **Tell Me That You Love Me, Junie Moon**
Reasoning with background knowledge
Wilhelm Müller was born on 7 October 1794 in Dessau, the son of a tailor. In 1813-1814 he took part, as a volunteer in the Prussian army, in the national rising against Napoleon. He participated in the battles of Lützen, Bautzen, Hanau and Kulm. In 1814 he returned to his studies at Berlin. Müller's son, Friedrich Max Müller, was an English orientalist who founded the comparative study of religions.

Which battle Wilhelm Müller fought in while in the Prussian army had the most casualties?

- **Battle of Lützen (1813)**
  Napoleon lost 19,655 men, while the Prussians lost 8,500 men and the Russians lost 3,500 men.

- **Battle of Bautzen**
  Losses on both sides totaled around 20,000.

- **Battle of Hanau**
  Overall, 4,500 French soldiers and 9,000 allied soldiers were lost in the battle.

- **Battle of Kulm**
  The French lost more than half of the pursuing force of 34,000; the allies lost approximately 13,000 soldiers.
Implicit reasoning

Was Aristotle alive when the laptop was invented?  
Did Aristotle use a laptop?

1. When did Aristotle live?  
2. When was the laptop invented?  
3. Is #2 before #1?

No

StrategyQA dataset, Geva et al., TACL, 2021.
More implicit reasoning examples

Would a pear sink in water?
- The density of a raw pear is about 0.59 g/cm\(^3\).
- The density of water is about 1 g/cm\(^3\).
- Objects only sink if they are denser than the surrounding fluid.
- **No**

Would a monocle be appropriate for a cyclop?
- Cyclops have one eye.
- A monocle helps one eye at a time.
- **Yes**

⚠️ GPT3 and PaLM 540B obtain ~77 accuracy on the StrategyQA dataset.

StrategyQA dataset, Geva et al., TACL, 2021.
Retrieval + commonsense reasoning over entity knowledge

Claim: Harry Potter can teach classes on how to fly on a broomstick.  
Harry Potter is a wizard ... He plays Quidditch while riding on a broomstick.  
Someone who’s good at something can teach it.  
TRUE

Claim: One can drive La Jolla to New York City in less than two hours.  
La Jolla is in California. NYC is in New York.  
It takes 5h with airplane to fly from California to New York.  
FALSE

CREAK dataset, Onoe et al., NeurIPS, 2021.
Best models on the CSQA2 leaderboard obtain ~78 accuracy.
Designing evaluations for complex reasoning
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

The Oberoi Group is a hotel company with its head office in Delhi. Founded in 1934, the company owns and/or operates 30+ luxury hotels and two river ...

The Oberoi family is part of a hotel company that has a head office in what city?

Delhi

Chen and Durrett, NAACL, 2019.
Various forms of bias in our crowdsourcing practices...
Instructions given to crowdworkers often include biases

Generate 5 data instances for “Temporal reasoning”.

The question has to require “temporal” understanding regarding some events in the sentence. In this task, we ask you to create five different questions for finding out Event duration. Event Duration for events is the understanding of duration of event such as “brushing teeth”, usually takes few minutes.

Examples:
- **How long** did Jack play basketball?
- **How long** has the software been having this issue?
- **How long** is a drive from Toronto to New York City?
Instruction bias limits the coverage of the “real” task

Dominant pattern in the Quoref dataset: What is the first name ...

- Instruction Examples: 36.4% without Pattern, 63.6% with Pattern
- Training Set: 57.0% without Pattern, 43.0% with Pattern
- Evaluation Set: 60.0% without Pattern, 40.0% with Pattern

What can we do about it?

- **Improve our data collection protocols** – instructions should be diverse, measure correlation between model performance and input patterns, model-in-the-loop, etc.

- **Build and evaluate on contrast sets / counterfactuals**

- **Evaluating the reasoning process itself**

- **More “realistic” setups** – combining reasoning skills, testing reasoning as part of “real” tasks.
Next: Knowledge-Augmented Approaches after Pretraining
References
Core reasoning skills


Core reasoning skills


Multimodal reasoning


Reasoning with background knowledge


Designing evaluations for complex reasoning


Designing evaluations for complex reasoning


