Benchmarks and Evaluation

Tutorial on Complex Reasoning in Natural Language

ACL 2023

Evaluation of complex reasoning is hard!

- 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 evaluation settings that measure what we want in a reliable and informative way.

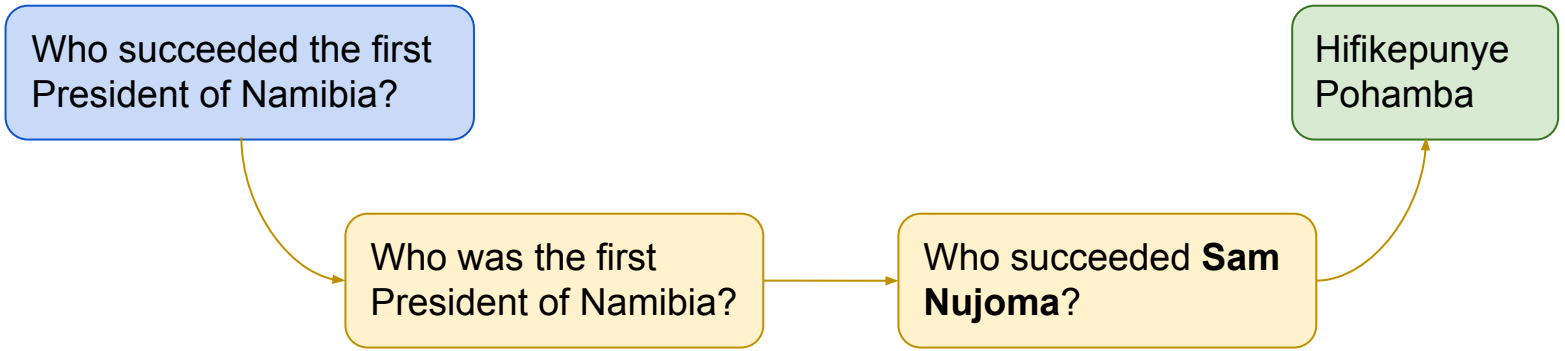
Exciting tasks, that even the best models today often fail on!

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



⚠️ GPT3 davinci-002 with self-ask prompting and an external search tool obtains **15.2 EM** on MuSiQue 2-hop questions.

MuSiQue dataset, Trivedi et al., TACL 2022.

Arithmetic reasoning

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.

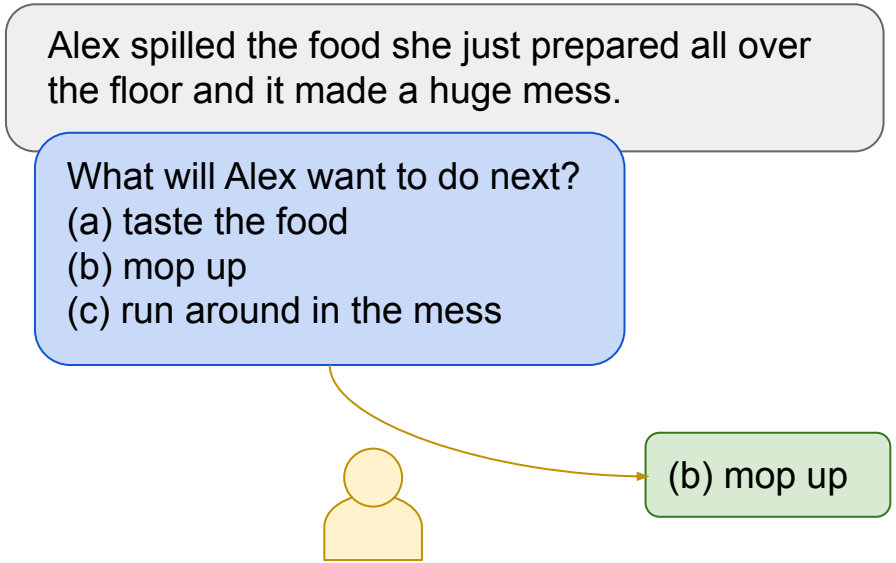
How much did the booth earn for 5 days after paying the rent and the cost of ingredients?

$895

How much did the booth make selling cotton candy each day? ($50 x 3 = $150)
How much did the booth make in a day? ($150 + $50 = $200)
How much did the booth make in 5 days? ($200 x 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)

GSM8K dataset, Cobbe et al., 2021.

Social reasoning



Social IQA dataset, Sap et al., EMNLP 2019.

Temporal reasoning

⚠️ Gopher in a 5-shot setting obtains **50.9 accuracy** on TIMEDIAL compared to 97.8 human performance.

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

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

TIMEDIAL dataset, Qin et al., ACL 2021.

Multimodal reasoning

Who is taller, LeBron James or Stephen Curry?

#	Name	length
1	Mo Bamba LAL | C	10.75
2	Talen Horton-Tucker UTA | SG	10.75
3	Jalen Williams OKC | SG	9.75
270	LeBron James LAL | PF	3.75
381	Stephen Curry GSW | PG	1.5

From: craftednba.com

Photo from Sports Illustrated

6 2

Born	December 30, 1984 (age 38) Akron, Ohio, U.S.
Listed height	6 ft 9 in (2.06 m)
Listed weight	250 lb (113 kg)

Personal information	
Born	March 14, 1988 (age 35) Akron, Ohio, U.S.
Listed height	6 ft 2 in (1.88 m)
Listed weight	185 lb (84 kg)

⚠ Best methods reach **~66 EM** on HybridQA.

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 ……

Name	Year	Season	Flag bearer
XXXI	2016	Summer	Yan Naing Soe
XXX	2012	Summer	Zaw Win Thet
XXIX	2008	Summer	Phone Myint Tayzar
XXVIII	2004	Summer	Hla Win U
XXVII	2000	Summer	Maung Maung Nge
XX	1972	Summer	Win Maung

Yan Naing Soe (born **31 January 1** 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 : မြင့်တင်စိုးဖုန်း) 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 …

Hardness

Q: In which year did the judoka bearer participate in the Olympic opening ceremony?	A: 2016
Q: Which event does the 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

HybridQA dataset, Chen et al., Findings of EMNLP, 2020.

United States House of Representatives Elections, 1972

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

Entailed Statement

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 Statement

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.

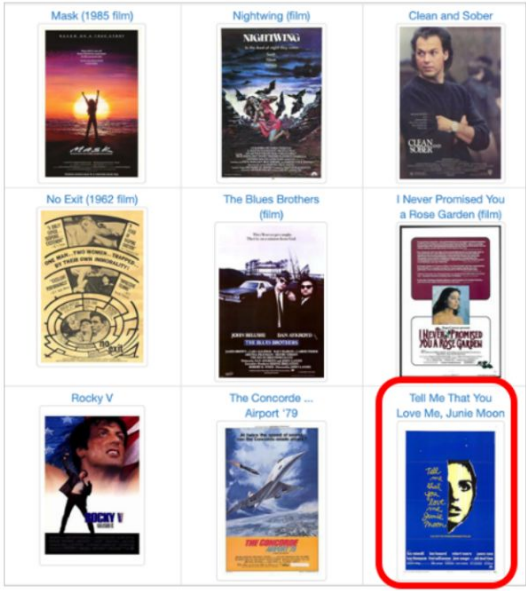
TabFact dataset, Chen et al., ICLR, 2020.

Multimodal Context

Steal 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. ... Guida. 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 ... l 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

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



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

MultiModalQA dataset, Talmor et al., ICLR, 2021.

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?

Kulm

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

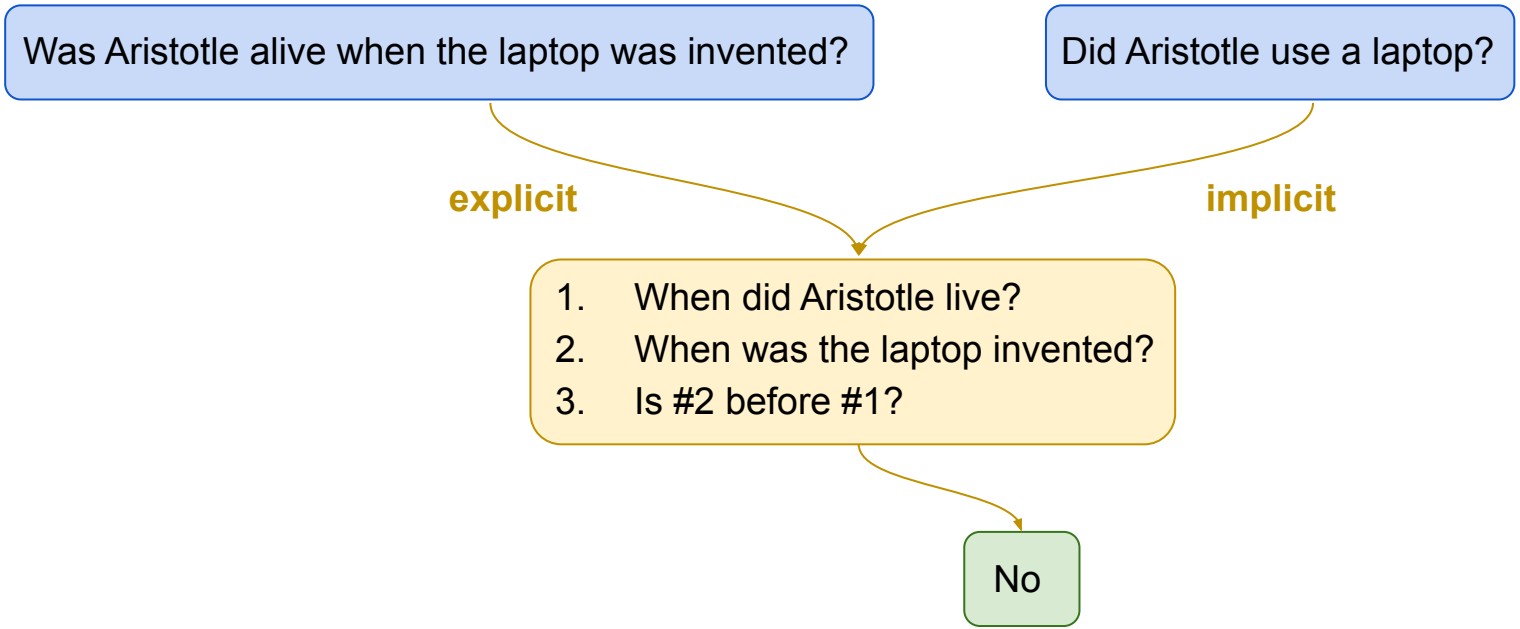
Battle of Lützen (1813)
Napoleon lost 19,655 men, while the Prussians lost 8,500 men and the Russians lost 3,500 men

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.

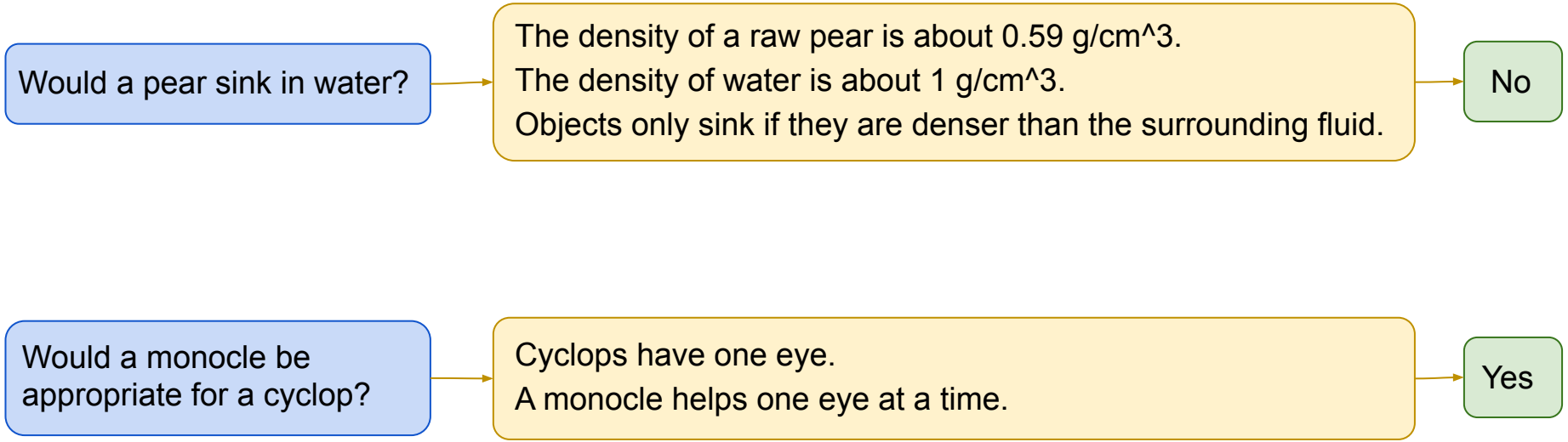
IIRC. Ferguson et al., EMNLP, 2020.

Implicit reasoning



StrategyQA dataset, Geva et al., TACL, 2021.

More implicit reasoning examples



⚠️ 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. **TRUE**

Harry Potter is a wizard ...
He plays Quidditch while riding on a broomstick.

Someone who's good at something can teach it.

Claim: One can drive La Jolla to New York City in less than two hours. **FALSE**

La Jolla is in California.
NYC is in New York.

It takes 5h with airplane to fly from California to New York.

CREAK dataset, Onoe et al., NeurIPS, 2021.

⚠️ Best models on the CSQA2 leaderboard obtain **~78 accuracy**.

CommonsenseQA2 dataset, Talmor et al., NeurIPS, 2021.

Designing evaluations for complex reasoning

Reasoning “shortcuts”

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?

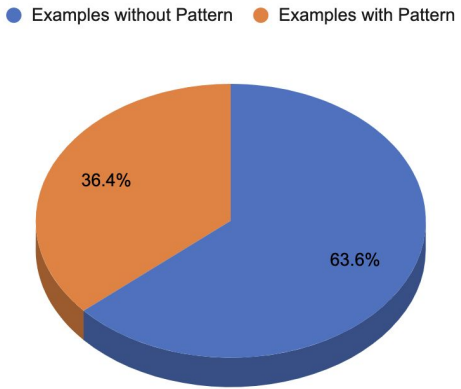
Delhi

Chen and Durrett, NAACL, 2019.

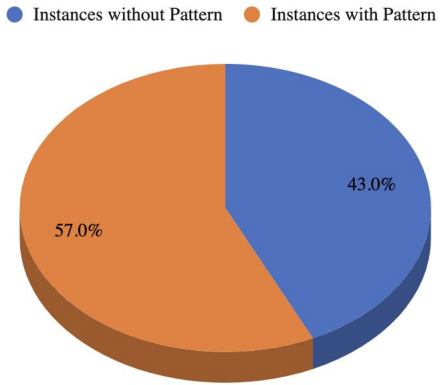
Various forms of bias in our crowdsourcing practices…

Instruction bias limits the coverage of the “real” task

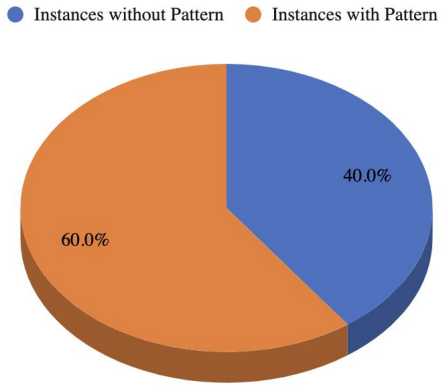
Dominant pattern in the Quoref dataset: `What is/was the first/last/full name …`



Instruction Examples



Training Set



Evaluation Set

Parmar et al. EACL, 2023.

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**
- **Automatic evaluation methods for reasoning**
- **More “realistic” setups** – combining reasoning skills, testing skills as part of existing tasks.

References

Core reasoning skills

[1] HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, Christopher D. Manning. EMNLP, 2018.

[2] MuSiQue: Multihop Questions via Single-hop Question Composition. Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, Ashish Sabharwal. TACL, 2022.

[3] Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, Akiko Aizawa. COLIN, 2020.

[4] Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, Dragomir Radev. EMNLP, 2018.

[5] DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, Matt Gardner. NAACL, 2019.

[6] Training Verifiers to Solve Math Word Problems. Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, John Schulman. 2021.

[7] Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, Matt Gardner. EMNLP, 2019.

Core reasoning skills

[8] How much coffee was consumed during EMNLP 2019? Fermi Problems: A New Reasoning Challenge for AI. Ashwin Kalyan, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, Peter Clark. EMNLP, 2019.

[9] TIMEDIAL: Temporal Commonsense Reasoning in Dialog. Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, Manaal Faruqui. ACL, 2021.

[10] Social IQa: Commonsense Reasoning about Social Interactions. Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, Yejin Choi. EMNLP, 2019.

[11] QASC: A Dataset for Question Answering via Sentence Composition. Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, Ashish Sabharwal. AAAI, 2020.

[12] TempQuestions: A Benchmark for Temporal Question Answering. Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, Gerhard Weikum. The Web Conference, 2018.

Multimodal reasoning

[13] HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, William Yang Wang. Findings of EMNLP, 2020.

[14] MultiModalQA: complex question answering over text, tables and images. Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, Jonathan Berant. ICLR, 2021.

[15] TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, Tat-Seng Chua. ACL, 2021.

[16] ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. Darryl Hannan, Akshay Jain, Mohit Bansal. AAAI, 2020.

[17] TabFact: A Large-scale Dataset for Table-based Fact Verification. Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, William Yang Wang. ICLR, 2020.

[18] Compositional Semantic Parsing on Semi-Structured Tables. Panupong Pasupat, Percy Liang. ACL, 2015.

Reasoning with background knowledge

[19] IIRC: A Dataset of Incomplete Information Reading Comprehension Questions. James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, Pradeep Dasigi. EMNLP, 2020.

[20] Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, Jonathan Berant. TACL, 2021.

[21] CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, Greg Durrett. NeurIPS, 2021.

[22] CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, Jonathan Berant. NeurIPS, 2021.

[23] Thinking Like a Skeptic: Defeasible Inference in Natural Language. Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, Yejin Choi. Findings of EMNLP, 2021.

[24] The Web as a Knowledge-Base for Answering Complex Questions. Alon Talmor, Jonathan Berant. NAACL, 2018.

[25] TheoremQA: A Theorem-driven Question Answering dataset. Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, Tony Xia. 2023.

Designing evaluations for complex reasoning

[26] Is Multihop QA in DiRe Condition? Measuring and Reducing Disconnected Reasoning. Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, Ashish Sabharwal. EMNLP, 2020.

[27] Understanding Dataset Design Choices for Multi-hop Reasoning. Jifan Chen, Greg Durrett. NAACL, 2019.

[28] Compositional Questions Do Not Necessitate Multi-hop Reasoning. Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, Luke Zettlemoyer. ACL, 2019.

[29] Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. Mor Geva, Yoav Goldberg, Jonathan Berant. EMNLP, 2019.

[30] Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, Pontus Stenetorp. TACL, 2020.

[31] Don’t Blame the Annotator: Bias Already Starts in the Annotation Instructions. Mihir Parmar, Swaroop Mishra, Mor Geva, Chitta Baral. EACL, 2023.

[32] On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study. Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, Wen-tau Yih. ACL, 2021.

Designing evaluations for complex reasoning

[33] Evaluating Models' Local Decision Boundaries via Contrast Sets. Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, Ben Zhou. Findings of EMNLP, 2020.

[34] Break, Perturb, Build: Automatic Perturbation of Reasoning Paths Through Question Decomposition. Mor Geva, Tomer Wolfson, Jonathan Berant. TACL, 2022.

[35] What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, Samuel R. Bowman. ACL, 2021.

[36] What Will it Take to Fix Benchmarking in Natural Language Understanding? Samuel R. Bowman, George Dahl. NAACL, 2021.

[37] REV: Information-Theoretic Evaluation of Free-Text Rationales. Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, Swabha Swayamdipta. ACL, 2023.

[38] ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, Asli Celikyilmaz. ICLR, 2023.