Pedro Rodriguez and **Jordan Boyd-Graber**. **Evaluation Paradigms in Question Answering**. *Empirical Methods in Natural Language Processing*, 2021, 5 pages.


```
    @inproceedings{Rodriguez:Boyd-Graber-2021,
Title = {Evaluation Paradigms in Question Answering},
Author = {Pedro Rodriguez and Jordan Boyd-Graber},
Location = {Punta Cana},
Booktitle = {Empirical Methods in Natural Language Processing},
Year = {2021},
Url = {http://cs.umd.edu/~jbg//docs/2021_emnlp_paradigms.pdf},
}
```


Accessible Abstract: Why do we answer questions? Sometimes it's to provide information, which has been the interpretation of the computer science community. But sometimes it's to probe or test intelligence. This paper argues we should think more about that application of question answering and its connection to the foundations of artificial intelligence: The Turing Test. We thus argue that in addition to the long-standing Cranfield paradigm popularized by information retrieval, this paper proposes an alternative "Manchester paradigm" closer to the Turing test, trivia games, and education.

Links:

- Research Talk [https://youtu.be/NZdKG3loB0M]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Evaluation Paradigms in Question Answering

Pedro Rodriguez
University of Maryland
Facebook Reality Labs
me@pedro.ai

Jordan Boyd-Graber
CS, LSC, iSchool, UMIACS
University of Maryland
jbg@umiacs.umd.edu

Abstract

Question answering (QA) primarily descends from two branches of research: (1) Alan Turing's investigation of machine intelligence at Manchester University and (2) Cyril Cleverdon's comparison of library card catalog indices at Cranfield University. This position paper names and distinguishes these paradigms. Despite substantial overlap, subtle but significant distinctions exert an outsize influence on research. While one evaluation paradigm values creating more intelligent QA systems, the other paradigm values building QA systems that appeal to users. By better understanding the epistemic heritage of QA, researchers, academia, and industry can more effectively accelerate QA research.

1 Introduction

This position paper seeks to answer the question *why do we do question answering* and understand the consequences of different answers to this question. Our primary contribution is to outline two distinct and common reasons that motivate researchers to pursue question answering (QA)—the Cranfield and Manchester paradigms. The Cranfield paradigm is not new (Section 2): it has a long and storied history in information retrieval (Voorhees, 2019). Here, we describe why a large share of QA is implicitly motivated by serving the needs of *users*, which is exactly the Cranfield paradigm (although most do not say so explicitly).

Section 3 christens another paradigm—the Manchester paradigm—at home in the more eclectic corners of QA: to test and inculcate *intelligence*.

These paradigms have much in common (Section 4), which helps explains why this distinction is not immediately apparent. However, the differences (Section 5) are ignored at your own peril. Section 6 articulates how the community can better heed the distinction and how the paradigms can inform each other.

2 Serving Users: The Cranfield Paradigm

Let's start with the Cranfield paradigm, named after Cranfield University in Bedfordshire. The Cranfield "experimental tradition founded by a librarian, working with card indexes, a half-century ago" spurred a revolution in information retrieval evaluation (Robertson, 2008).

In information retrieval, a "system locates information that is relevant to a user's query" (Sanderson and Croft, 2012). The most natural IR evaluation is to *ask* users whether documents satisfy their information need. However, much like annotation in NLP, this is expensive and time-consuming, so Cleverdon (1967) proposes an alternative. Rather than have users interact with every potential system, build *re-usable* test collections and evaluate all systems by re-using the same collection. Although "obvious" to twenty-first century readers, the use of offline test collections for evaluation was controversial (Taube, 1965), and the approach is still debated (Saracevic, 2007).

Rather than putting users in front of every IR system,

> in the Cranfield paradigm, researchers perform experiments on test collections to compare the relative effectiveness of different retrieval approaches (Voorhees, 2002b)

Cranfield paradigm datasets approximate users' searches, and the better your algorithm satisfies those queries, the better (Spärck Jones, 2001) the algorithm.

As IR systems conquered retrieving documents for short queries, researchers turned to finding short answers instead of whole documents (Voorhees, 2000b; Sanderson and Croft, 2012), which naturally lends itself to answering questions and "move[s] retrieval systems closer to *information retrieval* as opposed to *document* retrieval" (Voorhees and Tice, 2000). Under the Cranfield paradigm, a

good QA system should answer the questions users ask. What more could you want?

It is then unsurprising that Google and Microsoft adopted this setting for Natural Questions (Kwiatkowski et al., 2019) and MS MARCO (Nguyen et al., 2016). They found questions people asked online and answered them. A good QA system should automate that process (Chen and Yih, 2020).

3 Probing and Pushing Answerers: The Manchester Paradigm

In the Manchester paradigm, we create tasks and datasets whose questions push answerers to better understand the world and create evaluations that probe for human-like capabilities. Since we identify and name the paradigm, we give three justifications that highlight three distinct reasons people ask questions beyond information seeking: to teach, to compare, and to probe.

3.1 Why the Name Manchester?

For symmetry with the Cranfield paradigm, our proposed name is also an English city: Manchester. Because there are multiple aspects to the Manchester paradigm, we provide three connections between the city and question answering: in the nineteenth century, the city's regiment used the mythical Sphinx as a symbol, it is the home to *University Challenge*, and it is where Alan Turing outlined the Turing Test. We discuss each of these reasons for the name Manchester paradigm (saving the best for last).

To Teach: The Sphinx, Manchester's Standard Manchester's Regiment used the Sphinx as its symbol (Farmer, 1901). In Greek myth, the Sphinx asked everyone who entered the city a riddle (Renger, 2013). A Grammy-winning interpretation of the riddle states it as:

> What starts out on four legs then goes round on two Then finishes on three before it's through (Schickele, 1990)

And the answer is a human (crawling baby, walking, and then walking with a cane).

We have neither the space nor the desire to spoil Œdipus's journey, but by answering the question, the hero revealed not just his intelligence to his questioner but learned a tool to uncover his shrouded history. In a Whitman commencement, a classicist contrasted Google's NQ with professors' Sphinx-like riddles (emphasis added):

> Nobody told Œdipus who he was; he figured that out for himself... You can do a Google search to find out the capital of Arkansas (Little Rock?) but you can't... find out who you are, what you're good at, what makes you happy, what matters for your life... [The Sphinx's riddle has a] hybrid nature: the curse of being forced to solve a riddle is also *the gift of the ability to solve riddles*. (Burgess, 2013)

This aspect of question answering also brings us to another Greek connection to asking questions: teaching through the Socratic method (Trepanier, 2017). Through asking the right questions, a teacher guides the answerer to understanding. While Cranfield questioners are seeking information from a more knowledgeable answerer, Manchester questioners often test less knowledgeable answerers. Similarly, perhaps by asking the right questions, the QA community can coax computers to understand more than they do now (Dunietz et al., 2020; Perez et al., 2020).

To Compare: Granada Studios Another inspiration for this question answering approach is Manchester's Granada Studios, creator of *University Challenge* (Taylor et al., 2012; Baber, 2015). This television programme juxtaposes two universities to see who is smarter.

Just like the Sphinx, the dapper host of this game show, Bamber Gasgoine, knows the answer. Thus it is not an information-seeking task *a la* the Cranfield paradigm. It, like the riddle of the Sphinx, is a test of those *answering* the questions.

It's also a tried and true test of question answering researchers' mettle, as when IBM Watson bested Ken Jennings on *Jeopardy!* (Ferrucci et al., 2010). While Cranfield focuses on users' satisfaction, Manchester is at its heart an evaluation of the underlying capabilities of question answerers (either systems or humans): which is smarter, which is worthy? And as we discuss in Section 5, the Manchester paradigm is better suited to discriminating between answerers.

To Probe: The Turing Test The final reason is a paper written by Alan Turing while at the University of Manchester. Rather than create a test for intelligence and be forced to face the substantial challenge of defining intelligence, Turing proposes an indistinguishability test (Turing, 1950).[1] Building

[1] Likewise, we follow in Turing's footsteps and sidestep the definition of intelligence. Like Brooks (1991), we would argue that the best way to test and refine any definition of intelligence would be to have an increasingly difficult sequence of QA challenges in the Manchester paradigm.

off what he imagined would be a fun Victorian-era party game called "the Imitation Game" (Bishop, 2010), a skilled interrogator would ask questions to either a machine or a computer. An intelligent computer should—at minimum—be able to make itself indistinguishable from a human. This competition, the Turing Test, has been called AI-complete (Yampolskiy, 2013) and when taken literally is the implicit basis for claims of "super-human AI" (Cuthbertson, 2018). Its ubiquity extends beyond computer science to popular culture. A variant in *Blade Runner* tests empathy—rather than intelligence—with probing questions (Joerden, 2012).

Likewise, for tests of intelligence in the Manchester paradigm, the Turing Test "represents what it is that AI must endeavo[u]r eventually to accomplish scientifically" (Harnad, 1992). Methodologically, the Manchester paradigm iteratively imagines tasks where machines should rival humans (Levesque, 2014), develops systems, and then determines if systems pass the test.

3.2 Examples

Questions derived from education (Clark, 2015), puzzles (Littman et al., 2002), and trivia competitions (Joshi et al., 2017) are in the Manchester camp (full categorisation in Appendix A). However, prominent Manchester paradigm questions were first composed for computers: the Winograd schema challenge (Levesque et al., 2011) and its successor the Winogrand challenge (Sakaguchi et al., 2020). In this task, changing one word between two nearly identical binary questions also changes the answer.[2] Should a machine fail such questions, it does not evince intelligence—at least not like humans.

While we set these paradigms in opposition to each other, we next discuss the swath of research that advances the goals of both.

4 What Cranfield and Manchester Share

Although these paradigms have different core goals, research advancing the goals of one often advances the goals of the other.

While there are differences between QA datasets *across* paradigms (Cambazoglu et al., 2020; Zeng et al., 2020; Dzendzik et al., 2021), these differences are overshadowed *within a paradigm* by dataset-specific quirks. Thus, a paradigm-agnostic blueprint for QA (Chen and Yih, 2020) is to combine sparse (Chen et al., 2017) or dense retrieval (Guu et al., 2020; Karpukhin et al., 2020) followed by span selection (Seo et al., 2017) or generation (Lewis et al., 2020). As a consequence, researchers indifferent to which questions are answered can improve representations and algorithms for both paradigms (although as interactions become richer, this may not be the case, as we discuss at the end of Section 6). The paradigms' evaluations also overlap; they benefit from expert annotators (Gardner et al., 2020; Feng and Boyd-Graber, 2019), crowd annotators, and alternative evaluations like behavioural testing (Ribeiro et al., 2020).

Similarly, both paradigms value robustness (Dalvi et al., 2004; Jia, 2020). Additionally, answering infrequently asked questions is important for search engines (Baeza-Yates et al., 2007), and building models that learn more from less qualifies as intelligent behaviour (Linzen, 2020). Creating systems robust to spelling mistakes (Wang and Pedersen, 2011) is a worthy goal. From the Cranfield perspective, systems hobbled by spelling mistakes lead to a poor user experience. On the other side, humans are impressively robust to poor spelling (Rayner et al., 2006), so from the Manchester perspective this form of robustness is also valuable. But this has its limits; in the next section, we argue why adversarial examples are more consistent with the Manchester paradigm.

5 Ignore the Distinction at your Peril

How Adversarial is too Much? Common ground has its limits. That there is not a dichotomy between these two approaches can sometimes mask the importance of distinguishing motivations. Other proposals for robustness postulate that models should be robust to input modifications users would not make (Feng et al., 2018), challenging yet unnatural adversarial questions that users are unlikely to ask (Jia and Liang, 2017; Wallace et al., 2019; Bartolo et al., 2020; Kiela et al., 2021), and testing a concept in multiple ways (Gardner et al., 2020; Kaushik et al., 2020). While solving these challenges may eventually improve Cranfield-motivated systems, in the short term solving these challenges does not directly contribute to improving the user experience: researchers who build overly artificial datasets are likely going to be ignored by the Cranfield-focused community.

[2] In "the trophy would not fit in the brown case because **it** was too big. What was too big?" with possible answers "trophy" and "suitcase." Changing the underlined word to small would change the answer from "trophy" to "suitcase."

Users are the Customers Many future business dissertations will survey IBM Watson's circuitous route from TREC system (Chu-Carroll et al., 2002) to *Jeopardy!* spectacle to embattled spin-off (Deutscher, 2021) despite "IBM [having] bragged to the media that Watson's question-answering skills are good for more than annoying Alex Trebek." (Jennings, 2011). One challenge may have been transitioning between paradigms. One aspect that made the transition more difficult was that the *tour de force* victory on *Jeopardy!* was firmly on the side of the Manchester paradigm, but to be a successful commercial application, it needed to make the shift to the Cranfield paradigm.

Similarly, SQuAD was written by people (Mechanical Turkers) who knew the answers. . . just like most of the questions in the Manchester paradigm. However, it did not follow the same principles of the Manchester paradigm, which led to the "shortcuts" that other investigators have discovered in the years since (Weissenborn et al., 2017). For example, priming made exploitable clues more frequent and Mechanical Turkers write each question as quickly as possible. Levesque (2014) anticipates this behaviour, specifically avoiding "cheap tricks" in their Manchester-paradigm Winograd challenge. In other Manchester paradigm questions, trivia question writers frequently take pride in well-crafted questions (Boyd-Graber and Börschinger, 2020).

Comparisons One of the primary inspirations for the Manchester paradigm is competitions (e.g., *University Challenge*). Because these competitions are meant to determine who the smartest answerer is, they are remarkably efficient. The world accepted the judgement that Watson was smarter than Ken Jennings and Brad Rutter after 122 answers. Why not? These competitions are designed to discriminate between player abilities. In contrast, the dev and test sets of Cranfield-inspired datasets have thousands of questions, and even that may not be enough (Card et al., 2020; Rodriguez et al., 2021).

6 Call to Action

Our central plea is that researchers in QA and NLP more broadly should have a clear answer to the question: "Why are you working on this?". This is of particular importance as QA datasets proliferate (Rogers et al., 2021), and NLP practitioners "lost in dozens of recent datasets" want to know what datasets measure (Rogers and Rumshisky, 2020). While Gardner et al. (2019) offer a trenchant enumeration of QA uses,[3] we think the onus of definition should fall on dataset creators, not on post-hoc analyses. Other than more explicitly naming two of the uses of QA after English University towns, our goal is to encourage researchers to recognize the tensions between these two uses and the opportunities created from recognizing the distinction.

Make what you Value Explicit Each of these paradigms value different skills and embed these values in datasets and tasks. To make machine learning useful to society and adopt value-sensitive design (Dotan and Milli, 2020), developers of datasets should make their goals clear from the outset (Bender and Friedman, 2018; Gebru et al., 2018). In the Cranfield paradigm, aligning these evaluations with user satisfaction is essential (Spärck Jones, 2001). Industry is naturally financially motivated towards this goal, and they have the user data (Zhang et al., 2019)—only a fraction of which is published to protect privacy (Barbaro and Zeller, 2006). Still, strategic and thoughtful partnerships like the Cranfield-inspired TREC workshops are valuable; without TREC, it is estimated that "US Internet users would have spent up to 3.15 billion additional hours using web search engines between 1999 and 2009" (Tassey et al., 2010). One of the goals of the Manchester paradigm should be to identify the linguistic phenomena or ethnic and linguistic groups (Peskov et al., 2021) that are not well-served by Cranfield-focused data.

Thus, before you begin your question answering research, make it clear what your goal is: are you trying to build AGI[4] or to serve users? That answer will then inform your evaluation methology.

Academia's Special Role It is no coincidence that our paradigms are named after the homes of universities, and universities are where the Manchester paradigm will thrive. Thus, there is a strategic opportunity for academia and funding agencies to support Manchester-aligned work abjured by

[3] These are: (1) fill human information needs, (2) probe a system's understanding of some context (Weston et al., 2016), and (3) to transfer learned parameters. While the first are analogous to Cranfield and Manchester paradigms, we do not discuss the third use—transferring parameters—as that is model/architecture specific.

[4] Or more generally, are you trying to build systems with better language understanding and intelligence, even if it is not necessarily AGI?

deep-pocketed industry. Lovingly crafted questions by trivia experts (Rodriguez et al., 2019; Boyd-Graber and Börschinger, 2020) and adversarial questions (Wallace et al., 2019; Bartolo et al., 2020; Kiela et al., 2021) are unlikely to change the way a smart assistant answers a question, but they might expose blind spots of QA systems or improve evaluation. Moreover, asking questions in public is not just entertaining; it can generate data (von Ahn and Dabbish, 2008) and help the public better understand the possibilities and limitations of AI (hsiung Hsu et al., 1995; Silver et al., 2016). Thus, those in the Manchester paradigm can game show-ify evaluations to make question answering more fun and illuminating.

Build for the Future We do not advocate for firewalling these interests: they are ideally synergistic. Cranfield-inspired tasks can identify the most helpful capabilities that Manchester-inspired tasks can work towards. However, evaluating systems on users' *current* information needs may leave much on the table. Users' habits and low expectations encourage users to avoid difficult questions (Ng, 2015; Moorhead, 2015): e.g., avoiding complex syntax or hard to recognize named entities (Peskov et al., 2019) with voice recognition.

Begin a Dialog with Users Regardless of which paradigm you favor, QA is at its heart an interaction with users. In the Cranfield paradigm, the user knows less than the system. In the Manchester paradigm, the user knows more and takes the role of a teacher or an evaluator. In both cases, Shneiderman (2021) argues that responsible AI should enable an interactive, responsive conversation between the system and the user.

In the Cranfield paradigm, this is an opportunity to correct false presuppositions: "when did Raphael paint the *Mona Lisa*" could flag that Da Vinci painted it in 1503 and to explain multiple interpretations of a question (Min et al., 2020). In the Manchester paradigm, this can use dialog to train systems (Choi et al., 2018), guide the system to semantically equivalent answers (Si et al., 2021), or to learn from how humans answer the same questions (He et al., 2016). For example, if a computer answers Bush to the question "Who appointed Scalia to the supreme court", a Manchester inquisitor would rightly follow up with "can you be more specific", to which the system would hopefully respond George W. Bush.

7 Conclusion

We identify two core motivations for QA research over the past twenty years. We link one to the user-centered goals of the Cranfield paradigm and propose the Manchester paradigm to describe research working towards building human-like, intelligent QA systems. In at least the short-term, this distinction is important as it illuminates the goals of industry and academic stakeholders; ultimately, this makes it easier to ensure that both research agendas are valued. In the long term, we suspect that the best QA agents will benefit from the insights of user-oriented tasks and the longer-range efforts towards natural language understanding (Bender and Koller, 2020; Linzen, 2020).

Acknowledgements

For helpful comments and whose work was an inspiration for this paper, we thank Ellen Voorhees. For insightful discussions and ideas we thank Doug Oard, Shi Feng, and Alexander Hoyle. For feedback on prior versions of this paper, we thank members of the UMD CLIP lab, Robin Jia, Patrick Lewis, Joe Barrow, Eleftaria Briakou, Adina Williams, Douwe Kiela, and John P. Lalor. We thank our anonymous EMNLP reviewers and meta-reviewer for suggestions and comments. Boyd-Graber and Rodriguez's work at UMD were supported by NSF Grant IIS-1822494. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

David Baber. 2015. *Television Game Show Hosts: Biographies of 32 Stars*. McFarland.

Ricardo Baeza-Yates, Aristides Gionis, Flavio Junqueira, Vanessa Murdock, Vassilis Plachouras, and Fabrizio Silvestri. 2007. The impact of caching on search engines. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Michael Barbaro and Tom Zeller. 2006. A face is exposed for aol searcher no. 4417749.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy S. Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mark Bishop. 2010. The imitation game. *Kybernetes*, 39:88.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rodney A. Brooks. 1991. Intelligence without representation. *Artificial Intelligence*, 47(1–3):139–159.

Dana L Burgess. 2013. The gift of the sphinx. Whitman College Convocation Address.

B. Barla Cambazoglu, Mark Sanderson, Falk Scholer, and Bruce Croft. 2020. A review of public datasets in question answering research. *ACM SIGIR Forum*, 54(2).

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Association for Computational Linguistics*.

Danqi Chen and Wen-Tau Yih. 2020. Open-Domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jennifer Chu-Carroll, John Prager, Cristopher Welty, Krzysztof Czuba, and David Ferrucci. 2002. A Multi-Strategy and Multi-Source approach to question answering. In *Proceedings of the Text REtrieval Conference*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural Yes/No questions. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Peter Clark. 2015. Elementary school science and math tests as a driver for AI: take the aristo challenge! In *Association for the Advancement of Artificial Intelligence*, pages 4019–4021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cyril Cleverdon. 1967. The cranfield tests on index language devices. In *Aslib proceedings*. MCB UP Ltd.

Anthony Cuthbertson. 2018. Robots can now read better than humans, putting millions of jobs at risk. *Newsweek*.

Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Knowledge Discovery and Data Mining*. Association for Computing Machinery.

Hoa Trang Dang, Diane Kelly, and Jimmy Lin. 2007. Overview of the trec 2007 question answering track. In *Proceedings of the Text REtrieval Conference*.

Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2006. Overview of the trec 2006 question answering track. In *Proceedings of the Text REtrieval Conference*.

Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. 2019. QUOREF: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maria Deutscher. 2021. IBM reportedly mulling spinoff of $1B watson health unit. *Silicon Angle*.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*.

Ravit Dotan and Smitha Milli. 2020. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jesse Dunietz, Gregory Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and David Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the Association for Computational Linguistics*.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Daria Dzendzik, Carl Vogel, and Jennifer Foster. 2021. English machine reading comprehension datasets: A survey. *arXiv preprint arXiv:2101.10421*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

John S. Farmer. 1901. *The Regimental Records of the British Army: A Historical Résumé Chronologically Arranged of Titles, Campaigns, Honours, Uniforms, Facings, Badges, Nicknames, Etc.* Grant Richards.

Shi Feng and Jordan Boyd-Graber. 2019. What AI can do for me: Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building watson: An overview of the deepqa project. *AI Magazine*, 31:59–79.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna 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, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; When is it useful? *arXiv preprint arXiv:1909.11291*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. Datasheets for datasets.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented language model Pre-Training. In *Proceedings of the International Conference of Machine Learning*.

Stevan Harnad. 1992. The turing test is not a trick: Turing indistinguishability is a scientific criterion. *SIGART Bull.*, 3(4):9–10.

He He, Jordan L. Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *Proceedings of the International Conference of Machine Learning*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of Advances in Neural Information Processing Systems*.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over wikipedia. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the International Conference on Learning Representations*.

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: a reading comprehension system. In *Proceedings of the Association for Computational Linguistics*.

Feng hsiung Hsu, Murray Campbell, and A. Joseph Hoane. 1995. Deep blue system overview. In *Proceedings of the 9th International Conference on Supercomputing*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of Empirical Methods in Natural Language Processing*.

Ken Jennings. 2011. My puny human brain. *Slate Magazine*.

Robin Jia. 2020. *Building Robust Natural Language Processing Systems*. Ph.D. thesis, Stanford University.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jan C. Joerden. 2012. Maschinen mit würde? thesen zu einem turing-test für würde. *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics*, 20:311–318.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with Counterfactually-Augmented data. In *International Conference on Learning Representations*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *Association for the Advancement of Artificial Intelligence*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hector J Levesque. 2014. On our best behaviour. *Artificial intelligence*, 212(1):27–35.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of Advances in Neural Information Processing Systems*.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Michael L. Littman, Greg A. Keim, and Noam Shazeer. 2002. A probabilistic approach to solving crossword puzzles. *Artif. Intell.*, 134(1-2):23–55.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Patrick Moorhead. 2015. NVIDIA GTC: The race to perfect voice recognition using GPUs. *Forbes*.

Andrew Ng. 2015. Deep learning: What's next.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MAchine Reading COmprehension dataset. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, and Roser Morante. 2013. QA4MRE 2011-2013: Overview of question answering for machine reading evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Keith Rayner, Sarah J White, Rebecca L Johnson, and Simon P Liversedge. 2006. Raeding wrods with jubmled lettres: there is a cost. *Psychological science*, 17(3):192–193.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Almut-Barbara Renger. 2013. *Oedipus and the Sphinx: The Threshold Myth from Sophocles through Freud to Cocteau*. University of Chicago Press.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matthew Richardson, Christopher J C Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson. 2008. On the history of evaluation in IR. *Journal of Information Science and Engineering*, 34(4):439–456.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. 2020. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. *Association for the Advancement of Artificial Intelligence*, 34(05):8722–8731.

Anna Rogers and Anna Rumshisky. 2020. A guide to the dataset explosion in QA, NLI, and commonsense reasoning. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 27–32, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *Association for the Advancement of Artificial Intelligence*.

M Sanderson and W B Croft. 2012. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451.

Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915–1933.

Peter Schickele. 1990. Oedipus Tex and other choral calamities.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. *Proceedings of the International Conference on Learning Representations*.

Ben Shneiderman. 2021. Responsible AI: Bridging from ethics to practice. *Communications of the ACM*, 64(8):32–35.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? Answer equivalence for open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Karen Spärck Jones. 2001. Automatic language and information processing: rethinking evaluation. *Natural Language Engineering*, 7(1):29–46.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-Tau Yih, and Ashish Sabharwal. 2019. QUAREL: A dataset and models for answering questions about qualitative relationships. *Association for the Advancement of Artificial Intelligence*, 33:7063–7071.

Alon Talmor and Jonathan Berant. 2018. The web as a Knowledge-Base for answering complex questions. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *Computer Vision and Pattern Recognition*.

Gregory Tassey, Brent R Rowe, Dallas W Wood, Albert N Link, and Diglio A Simoni. 2010. Economic impact assessment of NIST's text retrieval conference (TREC) program. Technical report, National Institute of Standards and Technology.

Mortimer Taube. 1965. A note on the pseudo-mathematics of relevance. *American documentation*, 16(2):69–72.

David Taylor, Colin McNulty, and Jo Meek. 2012. Your starter for ten: 50 years of University Challenge.

Lee Trepanier. 2017. *The Socratic Method Today: Student-Centered and Transformative Teaching in Political Science*. Taylor & Francis.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics.

Alan M. Turing. 1950. Computers & thought. chapter Computing Machinery and Intelligence, pages 11–35. MIT Press, Cambridge, MA, USA.

Luis von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

Ellen M. Voorhees. 2000a. Overview of the trec-9 question answering track. In *Proceedings of the Text REtrieval Conference*.

Ellen M Voorhees. 2000b. The TREC-8 question answering track report.

Ellen M. Voorhees. 2001. Overview of the trec 2001 question answering track. In *Proceedings of the Text REtrieval Conference*.

Ellen M. Voorhees. 2002a. Overview of the trec 2002 question answering track. In *Proceedings of the Text REtrieval Conference*.

Ellen M Voorhees. 2002b. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*.

Ellen M. Voorhees. 2003. Overview of the trec 2003 question answering track. In *Proceedings of the Text REtrieval Conference*.

Ellen M. Voorhees. 2004. Overview of the trec 2004 question answering track. In *Proceedings of the Text REtrieval Conference*.

Ellen M. Voorhees. 2019. *The Evolution of Cranfield*, pages 45–69. Springer International Publishing, Cham.

Ellen M. Voorhees and Hoa Trang Dang. 2005. Overview of the trec 2005 question answering track. In *Proceedings of the Text REtrieval Conference*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Eric Wallace, Pedro Rodriguez, Shi Feng, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. In *Transactions of the Association for Computational Linguistics*, pages 387–401.

Kuansan Wang and Jan Pedersen. 2011. Review of MSR-Bing web scale speller challenge. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1339–1340, New York, NY, USA. Association for Computing Machinery.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Conference on Computational Natural Language Learning*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *Proceedings of the International Conference on Learning Representations*.

Roman V. Yampolskiy. 2013. *Turing Test as a Defining Feature of AI-Completeness*, pages 3–17. Springer Berlin Heidelberg, Berlin, Heidelberg.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension: Tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21).

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic intent representation in web search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

A Categorizing QA Datasets by Paradigm

To make our QA evaluation paradigms idea more concrete, we categorize fifty-six QA datasets as either primarily motivated by the Cranfield paradigm or the Manchester paradigm (Table 1). As is expected, the TREC QA tasks fall under the Cranfield paradigm while trivia-based datasets like *Jeopardy!* (SearchQA), Quizbowl, and TriviaQA fall under the Manchester paradigm. Many of the datasets that fall under the Manchester paradigm attempt to probe for "understanding" of some context; SQuAD for example probes for "understanding" of a context paragraph. Other datasets like ELI-5 are also clearly Cranfield since they are sourced specifically from questions that real users have asked. Although Table 1 likely does not enumerate all QA datasets, it nonetheless represents a extensive survey of the most prominent QA datasets. For more extensive QA surveys, see Cambazoglu et al. (2020) and Rogers et al. (2021) or a tutorial by Chen and Yih (2020).

Dataset	Paradigm	Domain	Author	Citation
Deep Read	Manchester	Stories	🧑	Hirschman et al. (1999)
TREC-8 QA	Cranfield	News	🧑	Voorhees (2000b)
TREC-9 QA	Cranfield	Search	🧑	Voorhees (2000a)
TREC QA 2001	Cranfield	Search	🧑	Voorhees (2001)
TREC QA 2002	Cranfield	Search	🧑	Voorhees (2002a)
TREC QA 2003	Cranfield	Search	🧑	Voorhees (2003)
TREC QA 2004	Cranfield	Search	🧑	Voorhees (2004)
TREC QA 2005	Cranfield	Search	🧑	Voorhees and Dang (2005)
TREC QA 2006	Cranfield	Search	🧑	Dang et al. (2006)
TREC QA 2007	Cranfield	Search	🧑	Dang et al. (2007)
QA4MRE 2011-2013	Manchester	Multiple	🧑	Peñas et al. (2013)
MCTest	Manchester	Stories	🧑	Richardson et al. (2013)
WebQuestions	Cranfield	Search	🧑+👥	Berant et al. (2013)
CNN/Daily mail	Manchester	News	🧑	Hermann et al. (2015)
Simple Questions	Manchester	Freebase	🤖→👥	Bordes et al. (2015)
Children's Book Test	Manchester	Stories	🤖	Hill et al. (2016)
bAbI	Manchester	Stories	🤖	Weston et al. (2016)
SQuAD 1.0	Manchester	Wikipedia	👥	Rajpurkar et al. (2016)
WikiReading	Manchester	Wikipedia	🤖	Hewlett et al. (2016)
MS-MARCO	Cranfield	Search	🧑	Nguyen et al. (2016)
MovieQA	Manchester	Movies	🤖	Tapaswi et al. (2016)
RACE	Manchester	Exams	🧑	Lai et al. (2017)
TriviaQA	Manchester	Trivia	🧑	Joshi et al. (2017)
SearchQA	Manchester	Trivia	🧑	Dunn et al. (2017)
Quasar-T	Manchester	Trivia	🧑	Dhingra et al. (2017)
SciQ	Manchester	Science	👥↔🤖	Welbl et al. (2017)
NewsQA	Cranfield	News	👥	Trischler et al. (2017)
CWQ	Manchester	Wikipedia	🤖→👥	Talmor and Berant (2018)
NarrativeQA	Manchester	Stories	🧑	Kočiský et al. (2018)
DuoRC	Manchester	Movies	👥	Saha et al. (2018)
MultiRC	Manchester	Multiple	👥	Khashabi et al. (2018)
HotpotQA	Manchester	Wikipedia	🧑	Yang et al. (2018)
SQuAD 2.0	Manchester	Wikipedia	👥	Rajpurkar et al. (2018)
QBLink	Manchester	Trivia	🧑	Elgohary et al. (2018)
WikiHop[5]	Manchester	Wikipedia	🤖	Welbl et al. (2018)
OpenBookQA	Manchester	Science	👥↔🤖	Mihaylov et al. (2018)
QASC	Manchester	Science	👥→🤖	Khot et al. (2020)
DROP	Manchester	Wikipedia	👥↔🤖	Dua et al. (2019)
QUOREF	Manchester	Wikipedia	👥↔🤖	Dasigi et al. (2019)
QuAC	Cranfield	Wikipedia	👥	Choi et al. (2018)
BoolQ	Cranfield	Search	🧑	Clark et al. (2019)
ELI-5	Cranfield	Reddit	🧑	Fan et al. (2019)
ARC	Manchester	Science	🧑	Clark et al. (2018)
Record	Manchester	News	🤖→👥	Zhang et al. (2018)
ROPES	Manchester	Wikipedia/Science	👥	Lin et al. (2019)
CoQA	Cranfield	Multiple	👥	Reddy et al. (2019)
CosmosQA	Manchester	Stories	🧑	Huang et al. (2019)
Natural Questions	Cranfield	Search/Wikipedia	🧑	Kwiatkowski et al. (2019)
Quizbowl	Manchester	Trivia	🧑	Rodriguez et al. (2019)
Trickme	Manchester	Trivia	🧑↔🤖	Wallace et al. (2019)
MCScript	Manchester	Commonsense	👥	Ostermann et al. (2018)
QuaRel	Manchester	Stories	👥↔🤖	Tafjord et al. (2019)
CommonSenseQA	Manchester	Commonsense	👥↔🤖	Talmor et al. (2019)
AmbigQA	Cranfield	Search	👥	Min et al. (2020)
Curiosity	Cranfield	Geopolitical	👥↔🤖	Rodriguez et al. (2020)
QuAIL	Manchester	Multiple	👥	Rogers et al. (2020)

Table 1: We categorizeQA datasets by paradigm, domain area, and who authored them. For authorship, we show whether they are authored by non-crowdsourced humans (🧑), crowdsourcing (👥), or are automatically generated (🤖). When humans and machines collaborate on questions, we indicate this and the directionality (e.g., SimpleQuestions generates knowledge base triples that crowdsource workers phrase as questions).