Chapter 13: Ask Me Anything

1. TRUE/FALSE - The computer on the starship *Enterprise* – with its vast store of knowledge and seamless understanding of the questions put to it – has long been a lodestar for human-computer interaction, envied by *Star Trek* fans and AI researchers alike (and the intersection of these groups is, shall we say, not insignificant).
2. While computers can accurately transcribe our requests, they remain far from the "final frontier" of question answering systems. What, according to Melanie Mitchell, is that final frontier?

Melanie Mitchell states that the final frontier of question answering systems is to get the systems to understand the semantics of the questions they are asked (215).

3. What is the name of the natural-language researcher who led the Watson development team at IBM?

The natural-language researcher who led the Watson development is David Ferucci (215).

4. In just two or three sentences, describe Watson.

Watson is a question-answering system that uses natural-language processing for question parsing and classification combined with a large knowledge base of a broad range of subjects to play *Jeopardy!* in its original design. Using the aforementioned features, Watson generates several answers with a confidence percentage and buzzes in if any of those percentages exceed a specified threshold.

5. In conceding Watson's inevitable victory, he added the following pop-culture reference to his answer card for the final Jeopardy clue: "I for one welcome our new computer overlords." Later, he quipped, "To my surprise, losing to an evil quiz-show-playing computer turned out to be a canny career move. Everyone wanted to know What It All Meant, and Watson was a terrible interview, so suddenly I was the one writing think pieces and giving TED Talks … Like Kasparaov before me, I now make a reasonable living as a professional human loser." Who is he?

Who is the "Greatest Of All Time" Jeopardy! player, Ken Jennings? (217).

6. TRUE/FALSE - Following Watson's win, the AI community was divided as to whether Watson was a true advance in AI or a "publicity stunt" or "parlor trick," as some called.

While most people agreed that Watson's performance on *Jeopardy!* was extraordinary, the question remained: Was Watson actually solving a genuinely hard problem – responding to sophisticated questions posed in colloquial language? Or is the task of responding to *Jeopardy!* clues, with their very particular linguistic format and fact-driven answers, actually not so hard for a computer with built-in access to *Wikipedia*, among other huge data repositories? Not to mention that the computer has been trained on a hundred thousand *Jeopardy!* clues with formats very similar to the ones it was faced with.

7. TRUE/FALSE - Beyond sharing the same name, the same planet-with-swirling lights logo, and the well-known pleasant robotic voice, the "Watson" that IBM's marketing department is pitching today has very little in common with the "Watson" that beat Ken Jennings and Brad Rutter at "Jeopardy!" in 2011. Moreover, today the name Watson refers not to one coherent AI system but rather to a suite of services that IBM offers to its customers – mainly businesses – under the Watson brand. In short, Watson essentially refers to whatever IBM does in the space of AI while bestowing on these services the valuable halo of the *Jeopardy!* winner.
8. TRUE/FALSE - For a variety of reasons, IBM Watson Group, as advanced and useful as its products might be, has seemed to struggle more than other tech companies. Some of the company's high-profile contracts with customers (for example, Houston's MD Anderson Cancer Center) have been canceled. A raft of negative articles about Watson have been published, often quoting disgruntled former employees arguing that some executives and marketers at IBM have far overpromised what the technology can deliver. Overpromising and under-delivering are, of course, an all-too-common story in AI; IBM is far from being the only guilty part. Only the future can tell what IBM's contribution will be in AI's spread to health care, law, and other areas in which automated question-answering systems could have a huge impact. But for now, in addition to its *Jeopardy!* win, Watson may be a contender for the "most notorious hype" award, a dubious achievement in the history of AI.
9. What is SQuAD?

SQuAD stands for the "Stanford Question Answering Dataset." It is an assessment of a machine's level of reading comprehension. Given a question about a paragraph from Wikipedia, machines must extract the answer from the paragraph for its response (222-223).

10. TRUE/FALSE - In 2018, two groups – one from Microsoft's research lab and the other from the Chinese company Alibaba – produced programs that exceeded Stanford's measure of human accuracy on the task of reading comprehension. Soon after, both Microsoft and Alibaba issued statements suggesting that their programs perform reading comprehension tasks at least as well as humans do.
11. Please recount MM's sarcastic description of a recurring recipe for AI research.

Mitchell writes, "Define a relatively narrow, though useful, task and collect a large data set for testing machine performance on this task. Perform a limited measure of human ability on this data set. Set up a competition in which AI systems can vie to outperform one another on this data set, until the human performance measure is met or exceeded. Report only on the genuinely impressive and useful achievement, but also claim, falsely, that the winning AI systems have human-level performance on a more general task (for example, 'reading comprehension')," (223-224).

12. What is a "Winograd Schema?"

The Winograd Schema is a method of measuring reading comprehension in machines by presenting two sentences and changing one word such that the word replacement changes which pronoun is being referenced (226-227).

13. TRUE/FALSE - In 2011, three AI researchers – Hector Levesque, Ernest Davis, and Leora Morgenstern – proposed using a large set of Winograd schemas as an alternative to the Turing test. The authors argued that, unlike the Turing test, a test that consists of Winograd schemas forestalls the possibility of a machine giving the correct answer without actually understanding anything about the sentence. The three researchers hypothesized (in notably cautious language) that "with a very big probability, anything that answers correctly is engaging in behavior that we would say show thinking in people." The researchers continued, "Our [Winograd schema] challenge does not allow a subject to hide behind a smokescreen of verbal tricks, playfulness, or canned responses … What we have proposed here is certainly less demanding than an intelligent conversation about sonnets (say), as imagined by Turing; it does, however, offer a test challenge that is less subject to abuse."
14. How do computer programs compare with humans at answering Winograd schema questions?

The best machines have 61% accuracy when answering 250 Winograd Schemas. Humans have 100% accuracy – if they are paying attention (227).

15. TRUE/FALSE - Similar to computer-vision programs, NLP systems can be vulnerable to "adversarial examples."
16. TRUE/FALSE - MM believes it to be extremely unlikely that machines will ever reach the level of humans on translation, reading comprehension, and the like by learning exclusively from online data, with essentially no real understanding of the language they process, which relies on commonsense knowledge and understanding the world.