Appendix A

Annotation Schema

Here, we describe our annotation schema in greater detail. We present the respective phenomenon, give a short description and present an example that illustrates the feature. Examples for categories that occur in the analysed samples are taken directly from observed data and therefore do not represent the views, beliefs or opinions of the authors. For those categories that were not annotated in the data we construct a minimal example.

Supporting Fact

We define and annotate "Supporting fact(s)" in line with contemporary literature as the (minimal set of) sentence(s) that is required in order to provide an answer to a given question. Other sources also call supporting facts "evidence".

Answer Type

Span We mark an answer as span if the answer is a text span from the paragraph.
Question: Who was freed from collapsed roadway tunnel?
Passage: [...] The quake collapsed a roadway tunnel, temporarily trapping about 50 construction workers. [...]
Expected Answer: 50 construction workers.

Paraphrasing We annotate an answer as paraphrasing if the expected correct answer is a paraphrase of a textual span. This can include the usage of synonyms, altering the constituency structure or changing the voice or mode.
Question: What is the CIA known for?
Passage: [...] The CIA has a reputation for agility [...]
Expected Answer: CIA is known for agility.

Unanswerable We annotate an answer as unanswerable if the answer is not provided in the accompanying paragraph.
Question: average daily temperature in Beaufort, SC
Passage: The highest average temperature in Beaufort is June at 80.8 degrees. The coldest average temperature in Beaufort is February at 50 degrees [...].

Generated We annotate an answer as generated, if and only if it does not fall into the three previous categories. Note that neither answers that are conjunctions of previous categories (e.g. two passage spans concatenated with "and") nor results of concatenating passage spans or restating the question in order to formulate a full sentence (i.e. enriching it with pronomina) are counted as generated answers.
Question: How many total points were scored in the game?
Passage: [...] as time expired to shock the Colts 27-24.
Expected Answer: 51.

Quality

Debatable We annotate an answer as debatable, either if it cannot be deduced from the paragraph, if there are multiple plausible alternatives or if the answer is not specific enough. We add a note with the alternatives or a better suiting answer.
Question: what does carter say?
Passage: [...] "From the time he began, [...]" the former president [...] said in a statement. "Jody was beside me in every decision I made [...]"
Expected Answer: "Jody was beside me in every decision I made [...]" (*This is an arbitrary selection as more direct speech is attributed to Carter in the passage.*)

Wrong We annotate an answer as wrong, if the answer is factually wrong. Further, we denote why the answer is wrong and what the correct answer should be.
Question: What is the cost of the project?
Passage: [...] At issue is the [...] platform, [...] that has cost taxpayers $1.2 billion in earmarks since 2004. It is estimated to cost at least $2.9 billion more [...].
Expected Answer: $2.9 Billion. (*The overall cost is at least $ 4.1 Billion*)

Linguistic Features

We annotate occurrences of the following linguistic features in the supporting facts. On a high-level, we differentiate between syntax and lexical semantics, as well as variety and ambiguity. Naturally, features that concern question and corresponding passage context tend to fall under the variety category, while features that relate to the passage only are typically associated with the ambiguity category.

Lexical Variety

Redundancy We annotate a span as redundant, if it does not alter the factuality of the sentence. In other words the answer to the question remains the same if the span is removed (and the sentence is still grammatically correct).

Question: When was the last time the author went to the cellars?
Passage: I had not, [if I remember rightly]$_{Redundancy}$, been into [the cellars] since [my hasty search on]$_{Redundancy}$ the evening of the attack.

Lexical Entailment We annotate occurrences, where it is required to navigate the semantic fields of words in order to derive the answer as lexical entailment. In other words we annotate cases, where the understanding of words' hypernymy and hyponomy relationships is necessary to arrive at the expected answer.

Question: What [food items]$_{LexEntailment}$ are mentioned?
Passage: He couldn't find anything to eat except for [pie]$_{LexEntailment}$! Usually, Joey would eat [cereal]$_{LexEntailment}$, [fruit]$_{LexEntailment}$ (a [pear]$_{LexEntailment}$), or [oatmeal]$_{LexEntailment}$ for breakfast.

Dative We annotate occurrences of variance in case of the object (i.e. from dative to using preposition) in the question and supporting facts.

Question: Who did Mary buy a gift for?
Passage: Mary bought Jane a gift.

Synonym and Paraphrase We annotate cases, where the question wording uses synonyms or paraphrases of expressions that occur in the supporting facts.

Question: How many years longer is the life expectancy of [women]$_{Synonym}$ than [men]$_{Synonym}$?
Passage: Life expectancy is [female]$_{Synonym}$ 75, [male]$_{Synonym}$ 72.

Abbreviation We annotate cases where the correct resolution of an abbreviation is required, in order to arrive at the answer.

Question: How many [touchdowns]$_{Abbreviation}$ did the Giants score in the first half?
Paragraph: [...] with RB Brandon Jacobs getting a 6-yard and a 43-yard [TD]$_{Abbreviation}$ run [...]

Symmetry, Collectivity and Core arguments We annotate the argument variance for the same predicate in question and passage such as argument collection for symmetric verbs or the exploitation of ergative verbs.

Question: Who married John?
Passage: John and Mary married.

Syntactic Variety

Nominalisation We annotate occurrences of the change in style from nominal to verbal (and vice versa) of verbs (nouns) occurring both in question and supporting facts.

Question: What show does [the host of]$_{Nominalisation}$ The 2011 Teen Choice Awards ceremony currently star on?
Passage: The 2011 Teen Choice Awards ceremony, [hosted by]$_{Nominalisation}$ Kaley Cuoco, aired live on August 7, 2011 at 8/7c on Fox.

Genitives We annotate cases where possession of an object is expressed by using the genitive form ("'s") in question and differently (e.g. using the preposition "of") in the supporting facts (and vice versa).

Question: Who used Mary's computer?
Passage: John's computer was broken, so he went to Mary's office where he used the computer of Mary.

Voice We annotate occurrences of the change in voice from active to passive (and vice versa) of verbs shared by question and supporting facts.

Question: Where does Mike Leach currently [coach at]$_{Voice}$?
Passage: [The 2012 Washington State Cougars football team] was [coached]$_{Voice}$ by by first-year head coach Mike Leach [...].

Lexical Ambiguity

Restrictivity We annotate cases where restrictive modifiers need to be resolved in order to arrive at the expected answers. Restrictive modifiers – opposed to redundancy – are modifiers that change the meaning of a sentence by providing additional details.

Question: How many dogs are in the room?
Passage: There are 5 dogs in the room. Three of them are brown. All the [brown]$_{Restrictivity}$ dogs leave the room.

Factivity We annotate cases where modifiers – such as verbs – change the factivity of a statement.

Question: When did it rain the last time?
Passage: Upon reading the news, I realise that it rained two days ago. I believe it rained yesterday.
Expected Answer: two days ago

Coreference We annotate cases where intra- or inter-sentence coreference and anaphora need to be resolved in order to retrieve the expected answer.
Question: What is the name of the psychologist who is known as the originator of social learning theory?
Passage: Albert Bandura OC (born December 4, 1925) is a psychologist who is the David Starr Jordan Professor Emeritus of Social Science in Psychology at Stanford University. [...] He is known as the originator of social learning theory and the theoretical construct of self-efficacy, and is also responsible for the influential 1961 Bobo doll experiment.

Ellipsis/Implicit We annotate cases where required information is not explicitly expressed in the passage.
Question: How many years after producing Happy Days did Beckett produce Rockaby?
Passage: [Beckett] produced works [...], including [...], Happy Days $[(1961)]_{Implicit}$, and Rockaby $[(1981)]_{Implicit}$. *(The date in brackets indicates the publication date implicitly.)*

Syntactic Ambiguity
Preposition We annotate occurrences of ambiguous prepositions that might obscure the reasoning process if resolved incorrectly.
Question: What tool do you eat spaghetti with?
Passage: Let's talk about forks. You use them to eat spaghetti with meatballs.

Listing We define listing as the case where multiple arguments belonging to the same predicate are collected with conjunctions or disjunctions (i.e. "and" or "or"). We annotate occurrences of listings where the resolution of such collections and mapping to the correct predicate is required in order to obtain the information required to answer the question.
Passage: $[\text{She is also known for her roles}]_{Predicate}$ [as White House aide Amanda Tanner in the first season of ABC's "Scandal"]$_{Argument}$ $[\text{and}]_{Listing}$ [as attorney Bonnie Winterbottom in ABC's "How to Get Away with Murder"]$_{Argument}$.

Coordination Scope We annotate cases where the scope of a coordination may be interpreted differently and thus lead to a different answer than the expected one. *Question:* Where did I put the marbles?
Passage: I put the marbles in the box and the bowl on the table. *Depending on the interpretation, the marbles were either put both in the box and in the bowl that was on the table, or the marbles were put in the box and the bowl was put on the table.*

Relative clause, adverbial phrase and apposition We annotate cases that require the correct resolution of relative pronomina, adverbial phrases or appositions in order to answer a question correctly.
Question: José Saramago and Ivo Andrić were recipients of what award in Literature?
Passage: Ivo Andrić [...] was a Yugoslav novelist, poet and short story writer $[\text{who}]_{Relative}$ won the Nobel Prize in Literature in 1961.

Required Reasoning

Operational Reasoning
We annotate occurrences of the arithmetic operations described below. Operational reasoning is a type of abstract reasoning, which means that we do not annotate passages that explicitly state the information required to answer the question, even if the question's wording might indicate it. For example, we don't regard the reasoning in the question "How many touchdowns did the Giants score in the first half?" as operational (counting) if the passage states "The Giants scored 2 touchdowns in the first half."

Bridge We annotate cases where information to answer the question needs to be gathered from multiple supporting facts, "bridged" by commonly mentioned entities, concepts or events. This phenomenon is also known as "Multi-hop reasoning" in literature.
Question: What show does the host of The 2011 Teen Choice Awards ceremony currently star on?
Passage: [...] The 2011 Teen Choice Awards ceremony, hosted by $[\text{Kaley Cuoco}]_{Entity}$, aired live on August 7, 2011 at 8/7c on Fox. [...] $[\text{Kaley Christine Cuoco}]_{Entity}$ is an American actress. Since 2007, she has starred as Penny on the CBS sitcom "The Big Bang Theory", for which she has received Satellite, Critics' Choice, and People's Choice Awards.

Comparison We annotate questions where entities, concepts or events needs to be compared with regard to their properties in order to answer a question.
Question: What year was the alphabetically first writer of Fairytale of New York born?
Passage: "Fairytale of New York" is a song written by Jem Finer and Shane MacGowan [...].

Constraint Satisfaction Similar to the Join category, we annotate instances that require the retrieval of entities, concepts or events which additionally satisfy a specified constraint.
Question: Which Australian singer-songwriter wrote Cold Hard Bitch?
Passage: ["Cold Hard Bitch"] was released in March 2004 and was written by band-members Chris Cester, Nic Cester, and Cameron Muncey. [...] Nicholas John "Nic" Cester is an Australian singer-songwriter and guitarist [...].

Intersection Similar to the Comparison category, we annotate cases where properties of entities, concepts or events need need to be reduced to a minimal common set.
Question: José Saramago and Ivo Andrić were recipients of what award in Literature?

Arithmetic Reasoning
We annotate occurrences of the arithmetic operations described below. Similarly to operational reasoning, arithmetic reasoning is a type of abstract reasoning, so we annotate it analogously. An example for *non-arithmetic* reasoning is, if the question states "How many total points were scored in the game?" and the passage expresses the required information similarly to "There were a total of 51 points scored in the game."

Substraction *Question:* How many points were the Giants behind the Dolphins at the start of the 4th quarter?
Passage: New York was down 17-10 behind two rushing touchdowns.

Addition *Question:* How many total points were scored in the game?
Passage: [...] Kris Brown kicked the winning 48-yard field goal as time expired to shock the Colts 27-24.

Ordering We annotate questions with this category, if it requires the comparison of (at least) two numerical values (and potentially a selection based on this comparison) to produce the expected answer.
Question: What happened second: Peace of Paris or appointed governor of Artois?
Passage: He [...] retired from active military service when the war ended in 1763 with the Peace of Paris. He was appointed governor of Artois in 1765.

Count We annotate questions that require the explicit enumeration of events, concepts, facts or entities.
Question: How many touchdowns did the Giants score in the first half?
Passage: In the second quarter, the Giants took the lead with RB Brandon Jacobs getting a 6-yard and a 43-yard TD run [...].

Other We annotate any other arithmetic operation that does not fall into any of the above categories with this label.
Question: How many points did the Ravens score on average?
Passage: Baltimore managed to beat the Jets 10-9 on the 2010 opener [...]. The Ravens rebounded [...], beating Cleveland 24-17 in Week 3 and then Pittsburgh 17-14 in Week 4. [...] Next, the Ravens hosted Miami and won 26-10, breaking that teams 4-0 road streak.

Linguistic Reasoning
Negations We annotate cases where the information in the passage needs to be negated in order to conclude the correct answer.
Question: How many percent are not Marriage couples living together?
Passage: [...] 46.28% were Marriage living together. [...]

Conjunctions and Disjunctions We annotate occurrences, where in order to conclude the answer logical conjunction or disjunction needs to be resolved.
Question: Is dad in the living room?
Passage: Dad is either in the kitchen or in the living room.

Conditionals We annotate cases where the the expected answer is guarded by a condition. In order to arrive at the answer, the inspection whether the condition holds is required.
Question: How many eggs did I buy?
Passage: I am going to buy eggs. If you want some, too, I will buy 6, if not I will buy 3. You didn't want any.

Quantification We annotate occurrences, where it is required to understand the concept of quantification (existential and universal) in order to determine the correct answer.
Question: How many presents did Susan receive?
Passage: On the day of the party, all five friends showed up. [Each friend]$_{Quantification}$ had a present for Susan.

Other types of reasoning
Temporal We annotate cases, where understanding about the succession is required in order to derive an answer. Similar to arithmetic and operational reasoning, we do not annotate questions where the required information is expressed explicitly in the passage. *Question:* Where is the ball? *Passage:* I take the ball. I go to the kitchen after going to the living room. I drop the ball. I go to the garden.

Spatial Similarly to temporal, we annotate cases where understanding about directions, environment and spatiality is required in order to arrive at the correct conclusion. *Question:* What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station? *Passage:* [Marietta Air Force Station] is located 2.1 mi northeast of Smyrna, Georgia.

Causal We annotate occurrences where causal (i.e. cause-effect) reasoning between events, entities or concepts is required to correctly answer a question. We do not annotate questions as causal, if passages explicitly reveal the relationship in a "effect because cause" manner. For example we don't annotate "Why do men have a hands off policy when it comes to black women's hair?" as causal, even if the wording indicates it, because the corresponding passage immideately reveals the relationship by stating "Because women spend so much time and money on their hair, Rock says men are forced to adopt a hands-off policy.".
Question: Why did Sam stop Mom from making four sandwich?
Passage: [...] There are three of us, so we need three sandwiches. [...]

By Exclusion We annotate occurrences (in the multiple-choice setting) where there is not enough information present to directly determine the expected answer, and the expected answer can only be assumed by excluding alternatives.
Question: Calls for a withdrawal of investment in Israel have also intensified because of its continuing occupation of @placeholder territories – something which is illegal under international law.
Answer Choices Benjamin Netanyahu, Paris, [Palestinian]$_{Answer}$, French, Israeli, Partner's, West Bank, Telecoms, Orange

Information Retrieval We collect cases that don't fall under any of the described categories and where the answer can be directly retrieved from the passage under this category.
Question: Officers were fatally shot where?
Passage: The Lakewood police officers [...] were fatally shot November 29 [in a coffee shop near Lakewood]$_{Answer}$.

Knowledge

We recognise passages that do not contain the required information in order to answer a question as expected. These non self sufficient passages require models to incorporate some form of *external knowledge*. We distinguish between factual and common sense knowledge.

Factual

We annotate the dependence on factual knowledge – knowledge that can clearly be stated as a set facts – from the domains listed below.

Cultural/Historic *Question:* What are the details of the second plot on Alexander's life in the Central Asian campaign?
Passage: Later, in the Central Asian campaign, a second plot against his life was revealed, this one instigated by his own royal pages. His official historian, Callisthenes of Olynthus, was implicated in the plot; however, historians have yet to reach a consensus regarding this involvement.
Expected Answer: Unsuccessful

Geographical/Political *Question:* Calls for a withdrawal of investment in Israel have also intensified because of its continuing occupation of @placeholder territories – something which is illegal under international law.
Passage: [...] But Israel lashed out at the decision, which appeared to be related to Partner's operations in the occupied West Bank. [...]
Expected Answer: Palestinian

Legal *Question:* [...] in part due to @placeholder – the 1972 law that increased opportunities for women in high school and college athletics – and a series of court decisions.
Passage: [...] Title IX helped open opportunity to women too; Olympic hopeful Marlen Exparza one example. [...]
Expected Answer: Title IX

Technical/Scientific *Question:* What are some renewable resources?
Passage: [...] *plants are not mentioned in the passage* [...]
Expected Answer: Fish, plants

Other Domain Specific *Question:* Which position scored the shortest touchdown of the game?
Passage: [...] However, Denver continued to pound away as RB Cecil Sapp got a 4-yard TD run, while kicker Jason Elam got a 23-yard field goal. [...]
Expected Answer: RB

Intuitive

We annotate the requirement of intuitive knowledge in order to answer a question common sense knowledge. Opposed to factual knowledge, it is hard to express as a set of facts.
Question: Why would Alexander have to declare an heir on his deathbed?
Passage: According to Diodorus, Alexander's companions asked him on his deathbed to whom he bequeathed his kingdom; his laconic reply was "toi kratistoi"–"to the strongest".
Expected Answer: So that people know who to follow.