A Data Set for Training QA Systems to Answer Questions about Novels

Yonah Mann

A THESIS

In

Data Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Master of Science in Engineering

2020

Chris Callison-Burch

Chris Callison-Burch

Clayton Greenberg

Acknowledgements

I would like to thank Prof. Chris Callison-Burch for supervising this thesis, Clayton Greenberg for co-advising me and Bhavna Saluja, Gaurav Kumar, and Reno Kriz with whom I collaborated on this project.

Abstract

Answering questions about novel-length texts opens up many new opportunities. A user could go beyond asking their virtual assistant to play the news headlines, and instead do something more sophisticated like asking questions about a novel they are reading in school. But, this task has proven to be extremely challenging since most neural models can only process short texts, and novels are much longer than current models can handle. This thesis presents a new question-answering dataset with data taken from Sparknotes and Project Gutenberg. Following previous work on answering questions on long texts, we provide high quality questions written about full-length novels which can be answered with either using short summaries of the novels or the full text of the novels themselves. We analyze the performance of multiple models on this new dataset and find that the dataset presents a significant challenge. Finally, we demonstrate that models trained on other tasks via transfer learning also perform poorly on our dataset, indicating that our dataset is challenging for current state-of-the-art methodologies in neural question answering.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Natural language processing (NLP) is a subfield of computer science focusing on teaching machines to process and comprehend human language. Within NLP, an important subfield is reading comprehension which focuses on teaching machines to process and comprehend texts and then understand those texts on a deep level. One way to test this comprehension is using the question answering format whereby a system is given a document, or set of documents, and a corresponding question and is then tasked with answering that question.

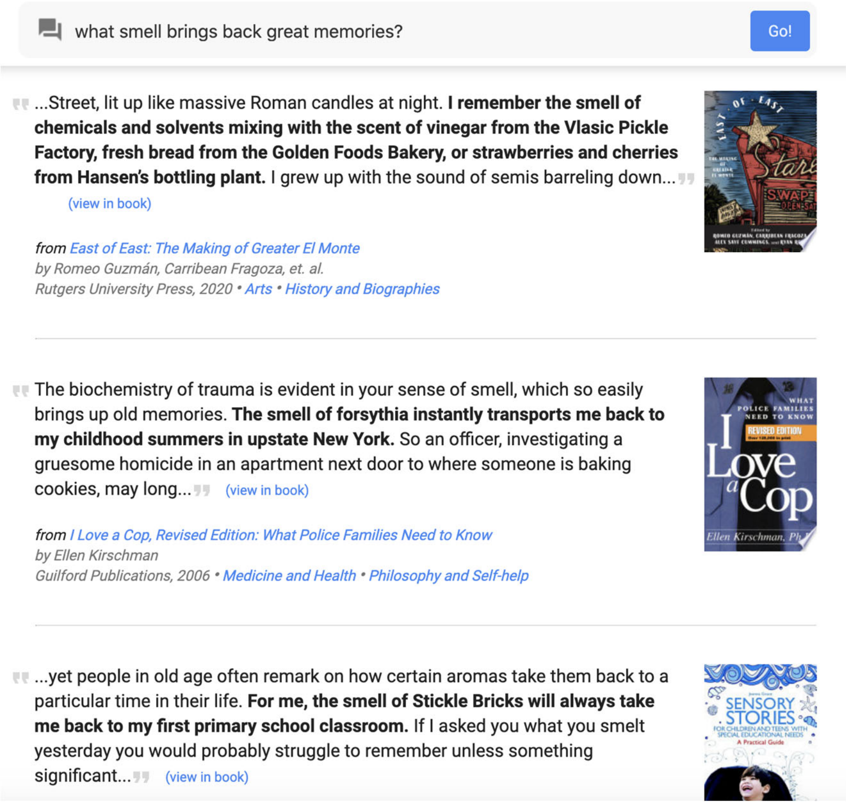
Question answering has also been used in information retrieval. Information retrieval is the activity of identifying information in documents that is relevant to a given query. Google Search is a classic example of an information retrieval task. Here, we can think of asking a question and then tasking a system with identifying a subset of an input document - or a set of input documents - to address this question. For example, if we enter the query, “who is the coach of the New York Knicks?” into Google Search, it will return an answer - Mike Miller - and not just a document containing an answer. Recently, more efforts have been made at combining information retrieval and reading comprehension tasks. The closest research effort to this master's thesis is a project called NarrativeQA. Kocisky et al. (2017) released the NarrativeQA dataset, which contains questions on full-length novels (see Table 1.1 for an example). They argue that answering abstract questions on novels requires synthesizing knowledge over many paragraphs. In contrast many question answering datasets can be solved using pattern matching. Additionally, stories are generally self-contained - they do not require outside knowledge to understand.

In NarrativeQA, systems were given a question and tasked with processing these novels and identifying the passages relevant to the question - an information retrieval task - and then comprehending these passages and answering the question accurately - a reading comprehension task.

Title	Armageddon 2419 A.D.
Question	In what year did Rogers awaken from his deep slumber?
Answer	2419
Summary snippet	. . . Rogers remained in sleep for 492 years. He awakens in 2419 and,. . .
Story snippet	I should state therefore, that I, Anthony Rogers, am, so far as I know, the only man alive whose normal span of eighty-one years of life has been spread over a period of 573 years. To be precise, I lived the first twenty-nine years of my life between 1898 and 1927; the other fifty-two since 2419. The gap between these two, a period of nearly five hundred years, I spent in a state of suspended animation, free from the ravages of katabolic processes, and without any apparent effect on my physical or mental faculties. When I began my long sleep, man had just begun his real conquest of the air…

Table 1.1 An example question from NarrativeQA

With the introduction of more sophisticated deep learning models in recent years, question answering has become a popular task in NLP. Question answering offers significant challenges to researchers, and also benefits for normal technology users as it provides utility to a variety of real-life tasks. Conversations with virtual assistants, for example, often require the assistant to comprehend a question posed by the user and then retrieve and comprehend the relevant information. For example, Google's "Talk to Books" project[1] responds to questions or statements from a user by looking at every sentence in over 100,000 books, before attempting to return an answer to the user's question. The system learns from books how real human conversations flow and uses this knowledge to predict how likely a given statement would follow another. Thus, the system is both a dialog agent and a question answering system. For example, we can ask the system: *"what smell brings back great memories?"* The system will know to answer with a passage from one of the books. We can see the top three options returned by the system in Figure 1.1.



what smell brings back great memories?

"...Street, lit up like massive Roman candles at night. **I remember the smell of chemicals and solvents mixing with the scent of vinegar from the Vlasic Pickle Factory, fresh bread from the Golden Foods Bakery, or strawberries and cherries from Hansen's bottling plant.** I grew up with the sound of semis barreling down..." (view in book)

from East of East: The Making of Greater El Monte
by Romeo Guzmán, Carribean Fragoza, et. al.
Rutgers University Press, 2020 • Arts • History and Biographies

"The biochemistry of trauma is evident in your sense of smell, which so easily brings up old memories. **The smell of forsythia instantly transports me back to my childhood summers in upstate New York.** So an officer, investigating a gruesome homicide in an apartment next door to where someone is baking cookies, may long..." (view in book)

from I Love a Cop, Revised Edition: What Police Families Need to Know
by Ellen Kirschman
Guilford Publications, 2006 • Medicine and Health • Philosophy and Self-help

"...yet people in old age often remark on how certain aromas take them back to a particular time in their life. **For me, the smell of Stickle Bricks will always take me back to my first primary school classroom.** If I asked you what you smelt yesterday you would probably struggle to remember unless something significant..." (view in book)

[1] https://research.google.com/semanticexperiences/about.html

Figure 1.1 An example from the Talk To Books project.

In this thesis, we focus on datasets from two sources: Sparknotes and Project Gutenberg. Sparknotes is a website that provides study guides with a focus on literature. Sparknotes splits novels, plays, and long poems into sections and then provides summaries and quizzes for each section. They also provide in-depth analysis of each work. Project Gutenberg is a website that offers novels that are out of copyright free of charge. Copyright laws vary by country, but in the United States any book published more than 95 years ago is considered to be in the public domain and not copyrighted. They offer novels in ebook formats and also as plain text files for download.

1.2 Main Contributions

The following are the main contributions of this thesis:

- A new question answering dataset from Sparknotes.
- A new question answering dataset linking Sparknotes and Project Gutenberg.
- We show that a BERT model trained on the SuperGLUE corpus performs significantly worse on our questions than it does on the SuperGLUE RTE dataset.
- We provide several models, the best of which achieves approximately 58% accuracy compared to 25% accuracy from random guessing.

1.3 Document Structure

This thesis is structured as follows:

- Chapter 2 is a literature review of the current state of research in question answering and especially the challenge datasets currently used.

- Chapter 3 is a discussion of the dataset we have created. It delves into how the dataset was created, challenges we encountered, and gives relevant examples and summary statistics to give an overview of the dataset.
- Chapter 4 discusses our baseline model. It describes the architecture, experiments, and results that we ran.
- Chapter 5 is a discussion of transfer learning to improve performance and test how well-known NLP tasks and datasets generalize to other domains.
- Chapter 6 suggests future research directions.

I was assisted throughout this project by Gaurav Kumar and Bhavna Saluja. Specifically:

- The work on web crawling was completed by Gaurav Kumar and Bhavna Saluja and work on the Project Gutenberg dataset was completed together with Bhavna Saluja.
- Bhavna and Gaurav configured the RoBERTa baseline model and did hyperparameter tuning.

Chapter 2

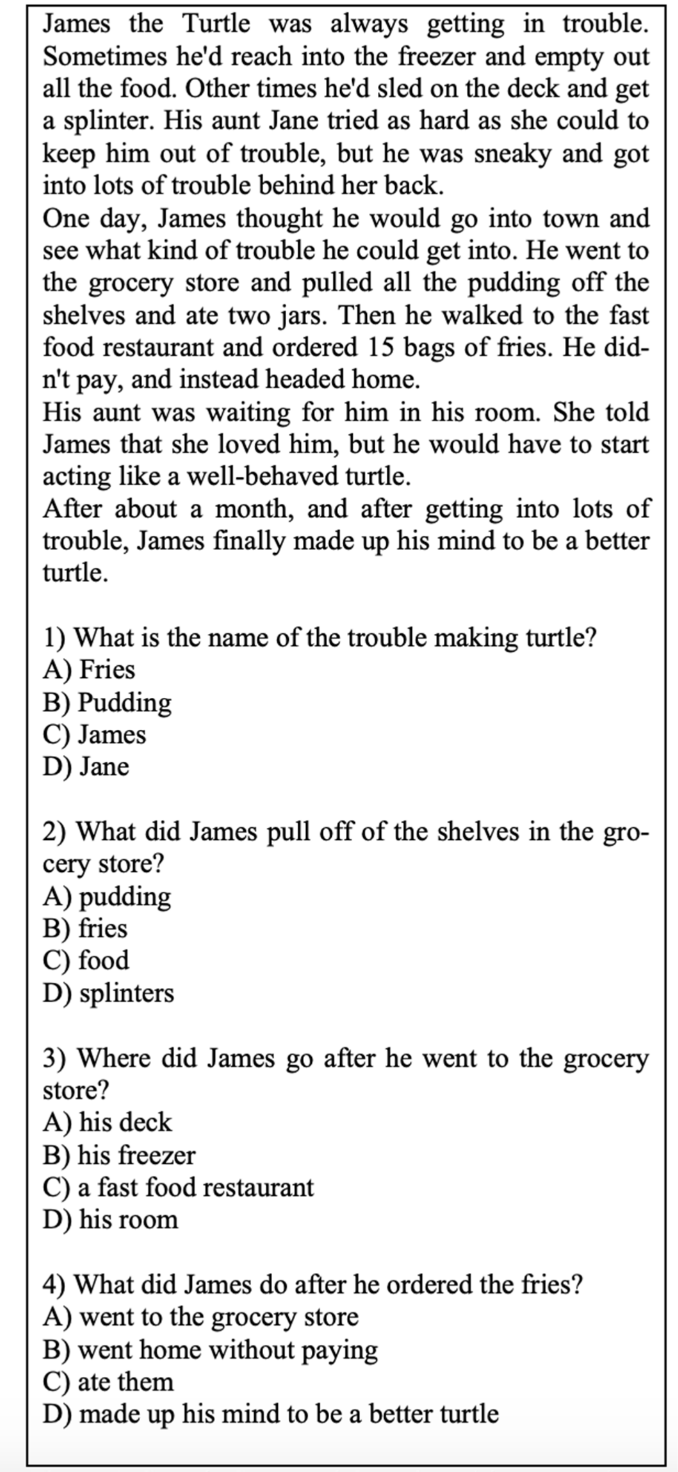
Literature Review

Below we review previous research on question answering datasets. This section is split up by question answering task with an additional sub-section on current using pre-trained models in question answering.

2.1 Multiple Choice

Multiple-choice tests are one common approach to formulating the question answering task. Multiple choice problems are modeled as a tuple (q,D,A) where q is the given questions, D is the document or set of documents containing the answer and A is the collection of incorrect and correct answers.

In the past decade, many question-answering tasks have been proposed to best test reading comprehension. Richardson et al., 2013 created a database of fictional short stories and written on a level for a child in grade school. To do this, workers, using Amazon's Mechanical Turk, were tasked with creating 150-300 word short stories. The workers were then asked to write multiple choice questions to this short story and for each question to provide three incorrect answers and the correct answer. Figure 2.1 shows an example story and questions. Such an approach was limited by several factors. Even with quality controls in place, workers had a hard time coming up with reasonable incorrect answers. Moreover, stories for children are inherently simplistic and so the task was too simple. Thus, even baseline heuristic methods achieved over 50% accuracy.



James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

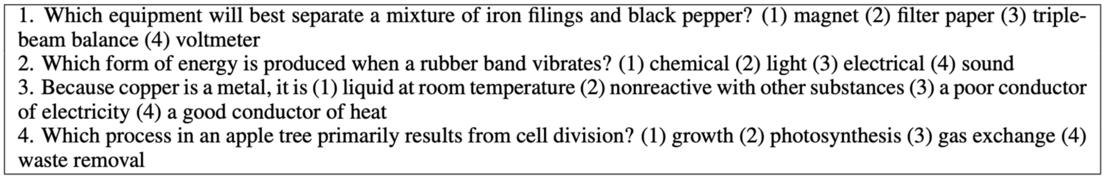
3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

Figure 2.1 An example taken from Richardson et al. (2013)

More recently, Lai et al. (2017) constructed a dataset of English exams for middle and high school Chinese students aged between 12 to 18. They used existing texts and questions and answers which were written by trained English instructors. In contrast to many other datasets, answers to these questions were not directly extracted from the context as text spans contained inside the larger text. Additionally, passages were not limited to any specific domain. This made the task more challenging than simpler tasks such as that introduced by Hill et al. (2015) which focused only on children's books. The authors' state-of-the-art model had accuracy of only 43% compared to human accuracy of 95%. That said, with recent advances in pre-trained models such as BERT (Devlin et al., 2018), current state-of-the-art models seemingly outperform Amazon Mechanical Worker performance (Yang et al., 2019).

In 2019, Clark et al. showed that a system could be trained to achieve 90% on non-graphical multiple choice questions from the Grade 8 New York Regents Science Exam. Non-graphical questions are questions that do not contain diagrams or graphs that students are asked to interpret. Standardized tests are written to challenge students to think critically and show a deep understanding of what they read, making this is a significant milestone.

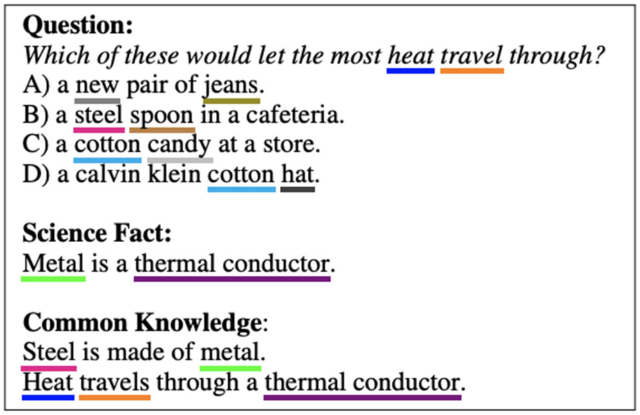


1. Which equipment will best separate a mixture of iron filings and black pepper? (1) magnet (2) filter paper (3) triple-beam balance (4) voltmeter
2. Which form of energy is produced when a rubber band vibrates? (1) chemical (2) light (3) electrical (4) sound
3. Because copper is a metal, it is (1) liquid at room temperature (2) nonreactive with other substances (3) a poor conductor of electricity (4) a good conductor of heat
4. Which process in an apple tree primarily results from cell division? (1) growth (2) photosynthesis (3) gas exchange (4) waste removal

Figure 2.2 Example questions from the Grade 8 New York Regents Science Exam

TriviaQA (Joshi et al., 2017) introduced a new problem formulation by using a set of documents as reference information, rather than a single document. In their formulation, models had to identify the correct document before identifying the correct answer. But, an assumption is made that each answer is a substring of some document in the set of documents. This simplifies the problem greatly and current state-of-the-art already outperforms humans (Back et al., 2018).

Finally, OpenBookQA (Mihaylov et al., 2018) provides a list of 1326 facts and asks questions that require combining multiple facts together. For example, given the question "Which of these would let the most heat travel through?" one needs to put together two facts from the dataset: that metals conduct heat and that a steel spoon is metal. This is one of the more difficult datasets available due to it requiring more complex understanding of facts and not relying on answers as substrings of the given texts. Currently, no one has managed to achieve human-level performance on this dataset.



Question:
Which of these would let the most heat travel through?
A) a new pair of jeans.
B) a steel spoon in a cafeteria.
C) a cotton candy at a store.
D) a calvin klein cotton hat.

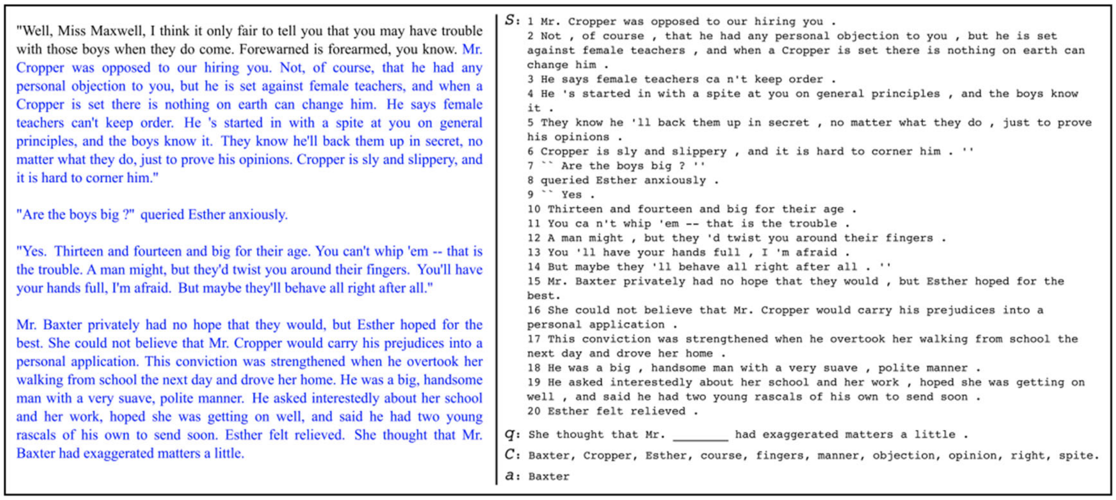
Science Fact:
Metal is a thermal conductor.

Common Knowledge:
Steel is made of metal.
Heat travels through a thermal conductor.

Figure 2.3 An example question from OpenBookQA

2.2 Cloze-Style

Cloze-style problems involve teaching a model to perform a "fill in the blank" task. In other words, we are given a document or set of documents, D, a sentence with a key word missing, s, and we want to predict the correct word to complete the sentence given the document. Hill et al. (2015) first introduced this task using children's books. The task involved filling in the blank from a sentence in the text given the 20 preceding sentences.



"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

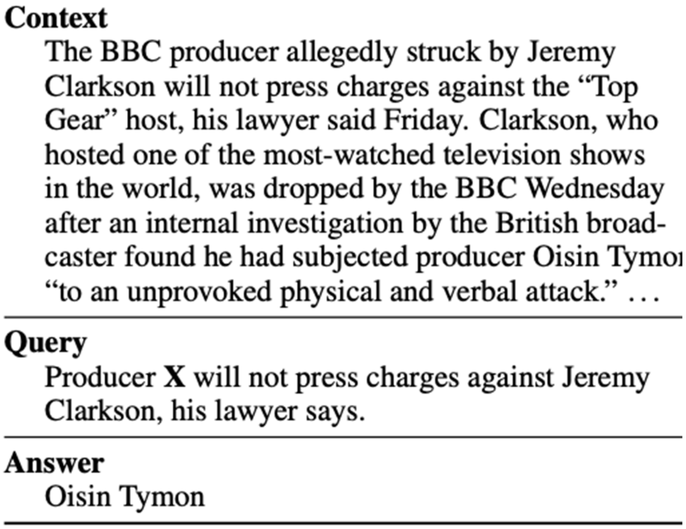
"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

```
S: 1 Mr. Cropper was opposed to our hiring you .
   2 Not , of course , that he had any personal objection to you , but he is set
   against female teachers , and when a Cropper is set there is nothing on earth can
   change him .
   3 He says female teachers ca n't keep order .
   4 He 's started in with a spite at you on general principles , and the boys know
   it .
   5 They know he 'll back them up in secret , no matter what they do , just to prove
   his opinions .
   6 Cropper is sly and slippery , and it is hard to corner him . ''
   7 `` Are the boys big ? ''
   8 queried Esther anxiously .
   9 `` Yes .
   10 Thirteen and fourteen and big for their age .
   11 You ca n't whip 'em -- that is the trouble .
   12 A man might , but they 'd twist you around their fingers .
   13 You 'll have your hands full , I 'm afraid .
   14 But maybe they 'll behave all right after all . ''
   15 Mr. Baxter privately had no hope that they would , but Esther hoped for the
   best.
   16 She could not believe that Mr. Cropper would carry his prejudices into a
   personal application .
   17 This conviction was strengthened when he overtook her walking from school the
   next day and drove her home .
   18 He was a big , handsome man with a very suave , polite manner .
   19 He asked interestedly about her school and her work , hoped she was getting on
   well , and said he had two young rascals of his own to send soon .
   20 Esther felt relieved .
q: She thought that Mr. ________ had exaggerated matters a little .
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.
a: Baxter
```

Figure 2.4 An example cloze question from the children's books data set from Hill et al. (2015)

Other well known cloze-style tasks include the CNN/ Daily Mail dataset (Hermann et al., 2015) which uses CNN and Daily Mail articles as documents and their corresponding bullet sentence summaries as sentences and the “Who-did-What” dataset (Onishi et al., 2016) which uses two independent articles to generate the document and corresponding sentence.



Context

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .

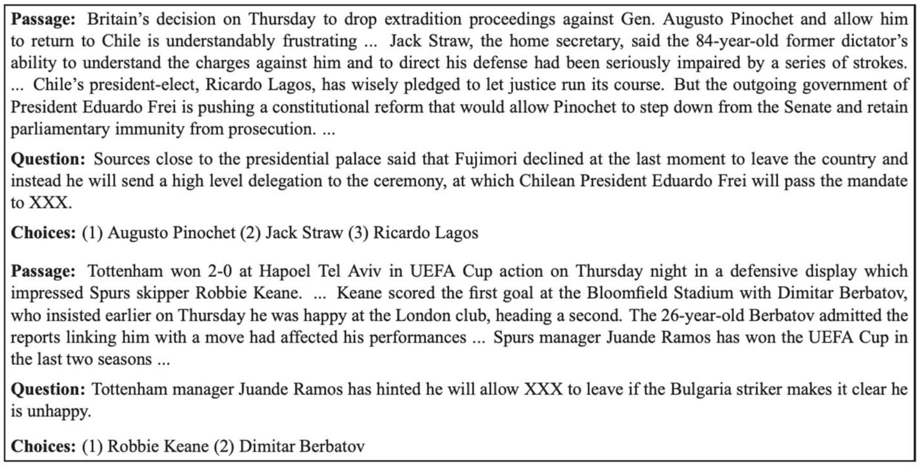
Query

Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer

Oisin Tymon

Figure 2.5 An example from the CNN/ Daily Mail dataset.



Passage: Britain's decision on Thursday to drop extradition proceedings against Gen. Augusto Pinochet and allow him to return to Chile is understandably frustrating ... Jack Straw, the home secretary, said the 84-year-old former dictator's ability to understand the charges against him and to direct his defense had been seriously impaired by a series of strokes. ... Chile's president-elect, Ricardo Lagos, has wisely pledged to let justice run its course. But the outgoing government of President Eduardo Frei is pushing a constitutional reform that would allow Pinochet to step down from the Senate and retain parliamentary immunity from prosecution. ...

Question: Sources close to the presidential palace said that Fujimori declined at the last moment to leave the country and instead he will send a high level delegation to the ceremony, at which Chilean President Eduardo Frei will pass the mandate to XXX.

Choices: (1) Augusto Pinochet (2) Jack Straw (3) Ricardo Lagos

Passage: Tottenham won 2-0 at Hapoel Tel Aviv in UEFA Cup action on Thursday night in a defensive display which impressed Spurs skipper Robbie Keane. ... Keane scored the first goal at the Bloomfield Stadium with Dimitar Berbatov, who insisted earlier on Thursday he was happy at the London club, heading a second. The 26-year-old Berbatov admitted the reports linking him with a move had affected his performances ... Spurs manager Juande Ramos has won the UEFA Cup in the last two seasons ...

Question: Tottenham manager Juande Ramos has hinted he will allow XXX to leave if the Bulgaria striker makes it clear he is unhappy.

Choices: (1) Robbie Keane (2) Dimitar Berbatov

Figure 2.6 Example cloze problem from the Who-did-What dataset

2.3 Free Form

Free form QA tasks involve a tuple (q, D, a) of a question, document or set of documents and answer - similar to other QA tasks. In free form tasks, the question is open-ended and the answer given is a sentence or collection of sentences, rather than as a multiple-choice answer.. Free form answers allow for answering more abstract and complex questions that rely on a deeper understanding of the text. There are two types of free-form tasks: generative, such as NarrativeQA which is described below, and span, such as Natural

Questions which is also described below. Generative tasks require the system to produce an answer for the question and then compares this to the true answer, while span-based tasks assume the answer is a span of text in the document.

One challenge for generative tasks is that they are difficult to evaluate, due to a lack of reliable automatic evaluation metrics. In other words, we have no good way of detecting the quality of our answers, and so their abstractions, in an automated manner. Currently, the BLEU score is used for generative question-answer systems (Papineni et al., 2002), by comparing system output to one or more reference answers created by humans. BLEU is a string matching algorithm that produces a score that can be used to measure the degree of closeness between a machine generated sentence and a true sentence. But, the BLEU score has been found to be unreliable in a variety of NLP tasks including the machine translation task it was originally designed for (Callison-Burch et al., 2006; Sulem et al., 2018). As an example, consider a question taken from the NarrativeQA dataset:

Q: *How is Oscar related to Dana?*

A_1: *Oscar is Dana's son.*

A_2: *Diana is Oscar's mother.*

Both of these answers are correct, though different. Thus, BLEU with a single reference answer would not suffice here.

Given these difficulties with evaluating generative tasks, a more common method of performing free form QA is using spans of the document as the correct answer. Thus, given a question, q and a document D, the task becomes to find the smallest possible span of text in D containing the answer to q. The answer, a, is a tuple of (start index, end index) marking the location of the correct span in the document.

For example, we might have a context like:

There$_0$ *once*$_1$ *was*$_2$ *a*$_3$ *man*$_4$ *named*$_5$ *Bob.*$_6$ *Bob*$_7$ *had*$_8$ *black*$_9$ *hair*$_{10}$ *and*$_{11}$ *blue*$_{12}$ *eyes.*$_{13}$*”*

And a question:

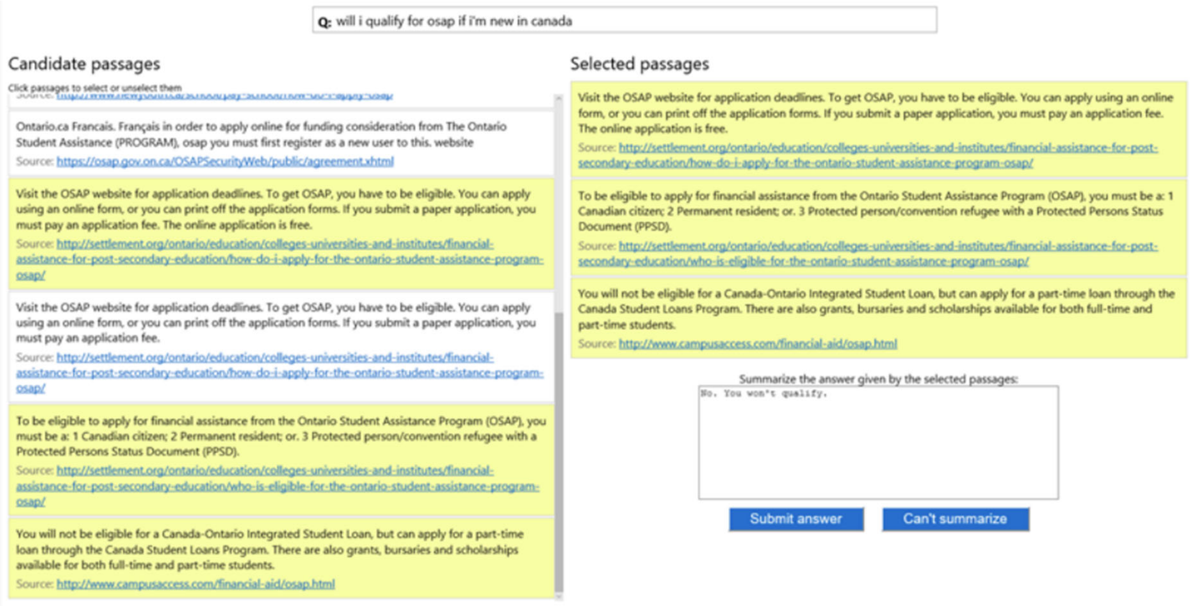
What did Bob look like?

The answer, would be the span containing the correct answer, in this case the span might be the range of words (9, 13) – *black hair and blue eyes*.

Two well known and well-studied datasets that use spans for answers are the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) and the newer 2.0 release of the SQuAD dataset (Rajpurkar et al., 2018). The SQuAD dataset consists of over 100,000 questions written by crowd workers after reading Wikipedia documents. Each document-question pair is matched with an answer which is a text span from the document. Jia and Liang (2017) showed that the question writers for SQuAD often based the wording of their questions very closely on the text of the Wikipedia document that they were given. This makes the task easy for the system and the system simply learns to pattern match, rather than to understand the text. Jia and Liang were able to show that by inserting distractor sentences into the text, they could fool systems trained on the SQuAD dataset into pattern matching on to the wrong sentence - even though the original, correct answer was still left in the text. With this in mind, SQuAD 2.0 was created to address this flaw. Rajpurkar et al. (2018) hypothesize that a root cause for SQuAD’s emphasis on pattern matching and for its failure to be robust in the face of distractor sentences, is that they focus on questions for which a correct answer is guaranteed to exist in the context document. Thus, models trained on the SQuAD dataset look for spans that most closely match the wording of the question rather than for true entailment between the question and the span. To fix this, SQuAD 2.0 includes unanswerable questions that the system must recognize are unanswerable. Then, systems learn to better identify irrelevant information in a text.

More recently, NarrativeQA (Kocisky et al., 2017), and HotpotQA (Yang et al., 2018) were developed as harder challenges than SQuAD and SQuAD 2.0. NarrativeQA is a dataset consisting of over 1,500 stories taken from project Gutenberg and movie scripts scraped from the web. Plot summaries were then taken from Wikipedia and questions were crowdsourced based on the plot summaries. The idea was that questions from the summaries would naturally cover high level abstractions and not just facts. But, current systems tested NarrativeQA perform poorly and it is speculated that, given the abstract nature of the questions and the length of the novels, perhaps NarrativeQA is too difficult for the current state of NLP research (Kwiatkowski et al., 2019).

Another source of quality QA datasets are from companies that develop them in order to improve the quality of their search engines. Microsoft Marco (Nguyen et al., 2016) created a large dataset using Bing queries. Over 100,000 queries were selected and passed to workers together with potentially relevant passages from the documents. The workers were then asked to write a short answer using these passages. Finally, systems are tasked with generating answers similar to those written by the annotators. A similar dataset, DuReader (He et al., 2018) was created on Baidu search results.



Q: will i qualify for osap if i'm new in canada

Candidate passages

Click passages to select or unselect them

Ontario.ca Francais. Français in order to apply online for funding consideration from The Ontario Student Assistance (PROGRAM), osap you must first register as a new user to this. website
Source: https://osap.gov.on.ca/OSAPSecurityWeb/public/agreement.xhtml

Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.
Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee.
Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or. 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD).
Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/

You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans Program. There are also grants, bursaries and scholarships available for both full-time and part-time students.
Source: http://www.campusaccess.com/financial-aid/osap.html

Selected passages

Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.
Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or. 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD).
Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/

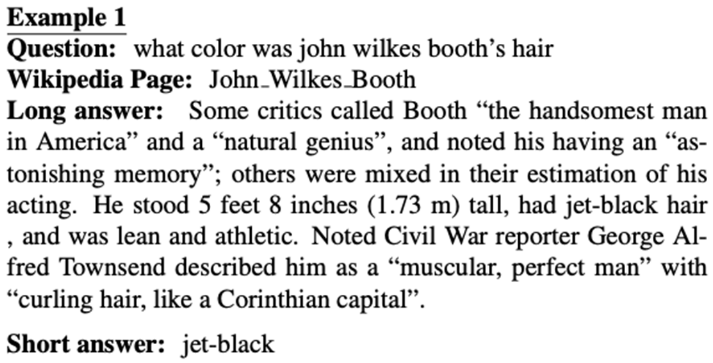
You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans Program. There are also grants, bursaries and scholarships available for both full-time and part-time students.
Source: http://www.campusaccess.com/financial-aid/osap.html

Summarize the answer given by the selected passages:

No. You won't qualify.

Figure 2.7 An example of the passage selection and summarization UI for MS Marco

The Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) was recently introduced to address many of the issues in the QA tasks before it. NQ was created similar to MS Marco - by taking search queries from Google and using them as questions and using the documents produced by Google as the sources. Each query in NQ consists of four parts: a question, wikipedia page, long answer and short answer. The long answer is a subset of a document that contains all of the necessary information to answer the question completely while the short answer is one or more entities or a boolean that provides a short answer to the question. Thus, the task tests both more complex abstract understanding via the long answers as well as simpler generative ability through the short answers. All labels were created using annotators with strict quality control that was more sophisticated than that of any of the previous datasets.



Example 1

Question: what color was john wilkes booth's hair

Wikipedia Page: John_Wilkes_Booth

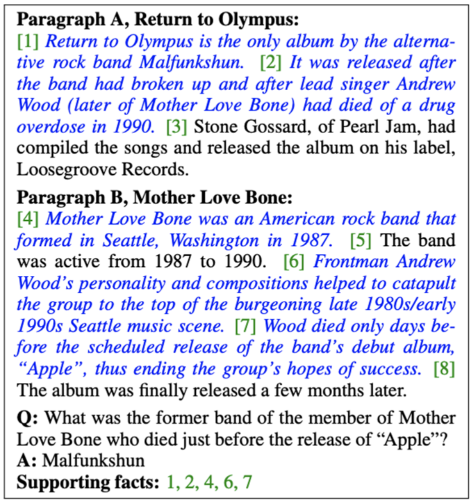
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Figure 2.8 A training example from the Natural Questions dataset

2.4 Multi-Hop QA

One major goal of Natural Language Understanding is to teach systems to comprehend and synthesize information from multiple sources. Multi-hop QA tasks are one way of testing such reasoning. In such tasks, systems are given a question and a set of documents and are required to synthesize information from multiple documents in order to find the correct answer. OpenBookQA (Mihaylov et al., 2018), described in the section on multiple choice tasks, is one example of such a task - another is HotpotQA. HotpotQA aims to challenge systems by providing questions that require "hopping" through multiple Wikipedia articles to find the correct answer. Answers are generally one or two words so that evaluation is simple and doesn't require using BLEU scores.



Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

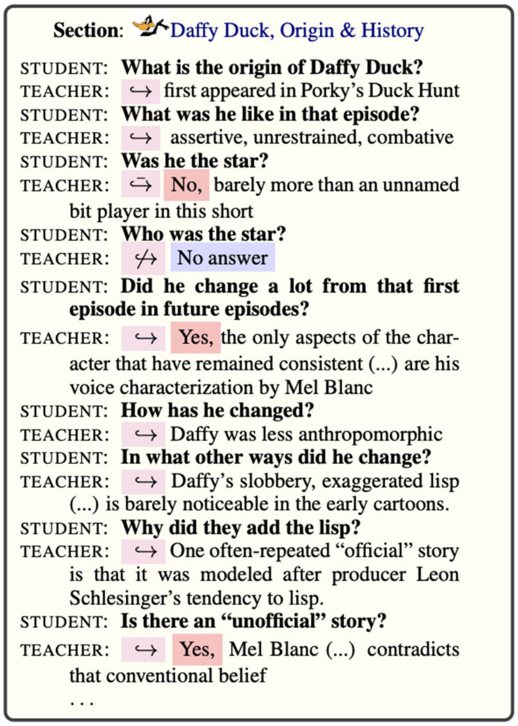
Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 2.9 An example of the multi-hop questions in HotpotQA. The supporting facts are italicized in blue.

2.5 Question Answering and Dialogues

One of the main reasons that conversations with virtual assistants feel so artificial is their inability to hold a conversation over multiple conversational turns. One of the newest areas of research in QA is conversational question answering which aims to tackle this problem. The idea with conversational QA is to mimic conversations between two people by teaching a system to answer questions that incrementally build on one another. This differs significantly from the single turn question answering task. Single turn question answering often requires understanding a large body of text while conversational question answering generally moves incrementally through a passage (Kwiatkowski et al., 2019).



Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪̄ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↪̸ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

Figure 2.10 An example dialog from QuAC (Choi et al., 2018)

Dialogue datasets generally present their data as a series of questions and answers where each question builds on information gleaned in the previous questions. These sometimes require linguistic processing like coreference resolution of pronouns. Each conversation is modeled as a tuple of the form (D, QA) where D is a passage from a text and QA is a list of question-answer pairs. Two well-known datasets for this task are CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018). QuAC shows a questioner a title from a Wikipedia text and shows the answerer the entire passage. The questioner then is free to ask questions about the passage which the answerer can then answer. CoQA collects data from a variety of domains - Wikipedia, children's books, middle and high school exams, etc. There are some differences between the two tasks - With CoQA the questioner can also see the entire task, thus ensuring the questions are from the passage and probably making the task easier. But, with CoQA answers are shorter in length and so easier to evaluate.

2.6 Using BERT with QA Systems

Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), or BERT, is a pre-trained model that has been used to great success in many NLP tasks. BERT uses the Transformer, a popular attention model, to learn contextual relations between words. The Transformer includes an encoder which BERT uses to generate a language model. BERT has been trained on a 3.3-billion-word corpus, including BooksCorpus (800 million words) and English Wikipedia (2.5 billion words), and encodes huge amounts of information on language and grammar. For question answering, we can use the pre-trained BERT as a base for our model. The pre-trained model is then fine-tuned on the question answering task. This has been shown to significantly improve the accuracy of NLP models.

Fine-tuning is a process whereby we take a pre-trained model and then add on extra layers to teach it to solve a different task. Figure 2.11 illustrates an example of this.

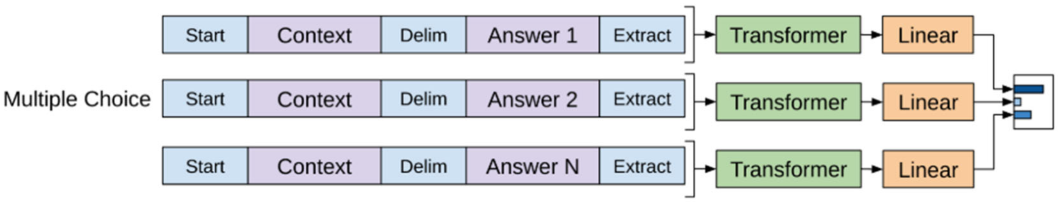


Figure 2.11 Architecture for fine-tuning a pre-trained model for solving the multiple choice task (Radford et al., 2018)

The transformer layer has been pre-trained on a variety of NLP tasks. For fine-tuning, we add on an untrained linear layer and train this new model for our new classification task. BERT already encodes most the information about language that we need, this training is just to tune the model to our new task. This is why it is generally recommended that one fine-tune for a short number of epochs, since most of the knowledge is already contained in the pre-trained layers. In the case of the example in Figure 2.11, our linear layer outputs values for each of the possible multiple choice answers. These values are then run through a softmax layer to obtain the final probabilities.

Chapter 3

SparknotesQA

In this chapter, we describe a novel question answering dataset generated using data from Sparknotes. We chose to use data from Sparknotes because all of the content on the website is created by trained professionals. We believe that the quality of the summaries and quizzes is higher and more uniform than previous datasets.

3.1 NarrativeQA

Prior to our work, Kocisky et al (2017) introduced NarrativeQA which uses full-length texts as input documents. With our dataset, we aimed to provide a new dataset that complements NarrativeQA. NarrativeQA aims to encourage deeper comprehension of language by providing systems with questions that cannot be answered from any individual part of a text together with a text that's too long for a system to memorize. Thus, the system is forced to comprehend a long document and synthesize information from different pieces. Specifically, it is emphasized that most questions require reading segments that are multiple paragraphs long, if not longer.

The dataset was generated by creating a set of movie scripts and novels. Then, annotators on Mechanical Turk were given summaries of each text taken from Wikipedia and instructed to write question-answer pairs based off of the summaries. The idea of writing the questions based on the summaries rather than the full novels was two-fold:

1) Writing questions and answers on a full-length novel is difficult for crowd workers and most of the Mechanical Turk annotators used in the study were not familiar with the texts they were working on beforehand.

2) Questions generated from summaries will not feature verbatim passages from the novel unlike the SQUAD datasets. As a result, a system will require a deeper comprehension of the novel's text to answer that goes beyond the simple string-matching heuristics that are often successful on SQUAD.

Finally, the question-answer pairs were combined with the full-length texts (dropping the summaries) to construct the dataset.

Abstract questions on full-length texts are difficult even for humans to tackle and for machines all the more so and Table 3.1 shows the relatively poor performance of the NarrativeQA baseline models.

IR baselines	**Bleu-1**	**Bleu-4**	**Meteor**	**Rouge-L**	**MRR**
Bleu-1 given question (1 sentence)	10.48	3.02	11.93	14.34	0.176
Rouge-L given question (8-gram)	11.74	2.18	7.05	12.58	0.168
Cosine given question (1 sentence)	7.49	1.88	10.18	12.01	0
Random rank					0.133
Neural Benchmarks					
Seq2Seq (no context)	16.1	1.40	4.22	13.29	0.211
Attention Sum Reader	23.54	5.90	8.02	23.28	**0.269**
Span Prediction	**33.45**	**15.69**	**15.68**	**36.74**	
Oracle IR Models					
Bleu-1 given answer (ans. length)	54.60	26.71	31.32	58.90	1
Rouge-L given answer (ans. length)	52.94	27.18	30.81	59.09	1
Cosine given answer (ans. length)	46.69	24.25	27.02	44.64	0.836
Human (given summaries)	44.24	18.17	23.87	57.17	

Table 3.1 Baseline results from the original NarrativeQA paper

Since the release of NarrativeQA, the task has been tackled by Tay et al. (2019) who saw only modest improvement over the random baseline models. For this reason, most research on NarrativeQA has focused on using the summaries to answer the questions, rather than the full-length texts. We note that the quality of Wikipedia summaries varies.

Depending on the quality of a book and the author of the Wikipedia text, summaries may be short and uninformative or longer and in-depth.

So, we introduce Sparknotes summaries, which are long - oftentimes multiple pages - and meet a minimum quality level. This can be seen in the difference in detail between Sparknotes summaries and Wikipedia summaries. Wikipedia summaries are on average 570 words for a complete book or movie, whereas the Sparknotes chapter and section summaries are on average 708 words and each book is split up into 10.4 sections on average. Therefore, the combined summary of a book on Sparknotes consists of 7,363 words on average compared to the 570 words in Wikipedia. We contend that Sparknotes summaries are a good challenge for the NLP community: they offer multi-paragraph texts that are too long for processing using a deep learning model, but also shorter than full-length novels so as to be manageable. They are professionally written. They may be linked to the full text of the novels that they summarize. Sparknotes questions are similarly high quality. In the past, questions were written by Mechanical Turk workers.

3.2 Sparknotes Dataset Creation

3.2.1 Crawling Sparknotes[2]

This dataset is composed of three parts - Sparknotes summaries, Sparknotes multiple choice questions and Project Gutenberg full-length stories and novels. To gather the

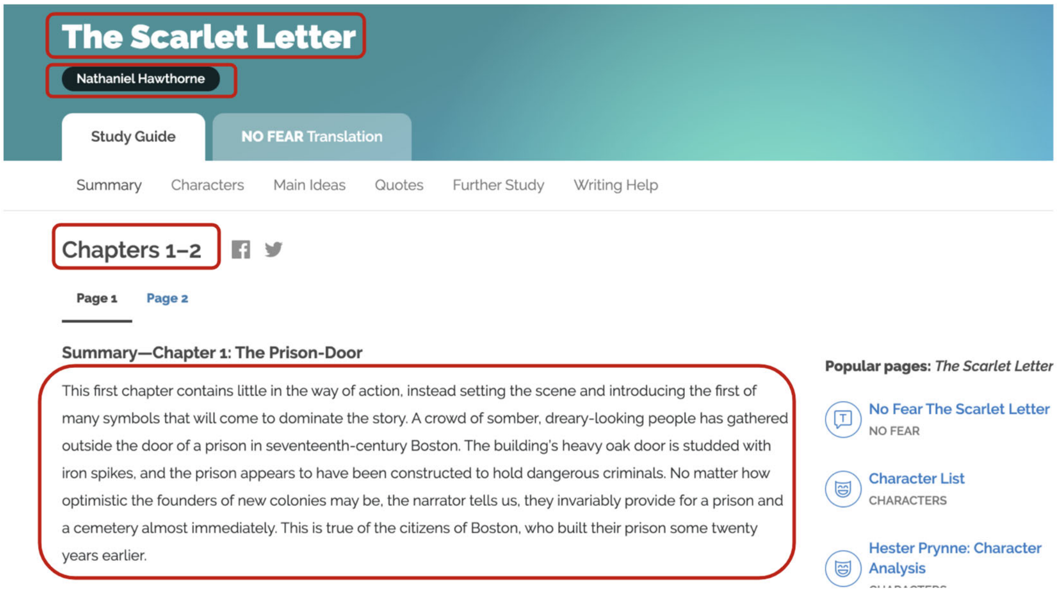
[2] This work was done by Gaurav Kumar and Bhavna Saluja

Sparknotes data, we first crawled sparknotes.com, collecting the HTML pages for all summaries and associated quizzes.

The crawler takes in a starting URL, the directory of the database where the data is stored, and the maximum size of a document and then traverses the links in the HTML pages it encounters. As the crawler encounters new HTML pages, it downloads these pages and traverses their links, continuing until no new links remain.

3.2.2 Data Pre-processing

Once all the HTML pages are collected, they must be filtered out and then processed. We parsed the HTML page URLs and each HTML page was categorized as being a summary page, a quiz page or not relevant. From here, relevant pages were identified to be scraped and processed.



The Scarlet Letter

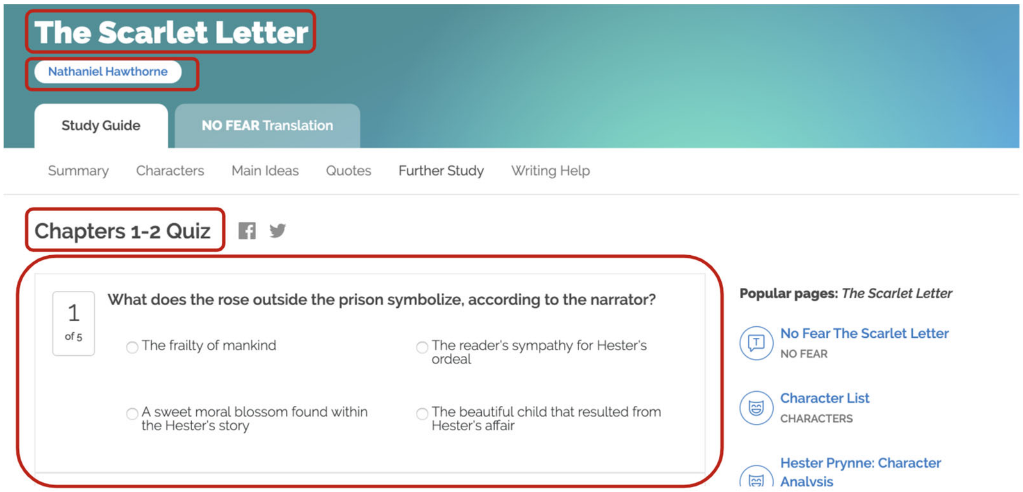
Nathaniel Hawthorne

Chapters 1–2

Summary—Chapter 1: The Prison-Door

This first chapter contains little in the way of action, instead setting the scene and introducing the first of many symbols that will come to dominate the story. A crowd of somber, dreary-looking people has gathered outside the door of a prison in seventeenth-century Boston. The building's heavy oak door is studded with iron spikes, and the prison appears to have been constructed to hold dangerous criminals. No matter how optimistic the founders of new colonies may be, the narrator tells us, they invariably provide for a prison and a cemetery almost immediately. This is true of the citizens of Boston, who built their prison some twenty years earlier.

Figure 3.1 An example Sparknotes summary page. We extract all of the circled data.



The Scarlet Letter

Nathaniel Hawthorne

Chapters 1-2 Quiz

1 of 5

What does the rose outside the prison symbolize, according to the narrator?

○ The frailty of mankind

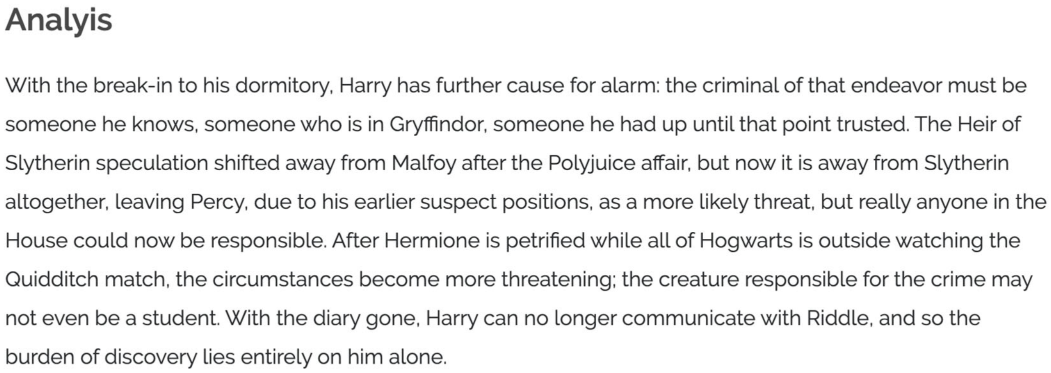
○ The reader's sympathy for Hester's ordeal

○ A sweet moral blossom found within the Hester's story

○ The beautiful child that resulted from Hester's affair

Figure 3.2 An example Sparknotes quiz page. We extract all of the circled data.

Sparknotes summaries contain both summary paragraphs as well as analysis paragraphs. We chose to limit the supporting document to only the summary portion because we found that the analysis sometimes answers the question explicitly. Thus, the model is forced to infer from the actual summary, rather than recognize patterns in the analysis. Also, Sparknotes web pages are sometimes inconsistently formatted and a small percentage of pages are mislabeled or contain typos (see Figure 3.3). We identified outlier pages that were mislabeled and added manual checks to ensure that they were corrected before being processed. For example, to identify mislabeled paragraphs, we checked for summaries that were empty or contained paragraph headers not recognized by the scraper. We then either added additional logic to the scraper or in a select few cases, created the relevant json entry manually.



Analyis

With the break-in to his dormitory, Harry has further cause for alarm: the criminal of that endeavor must be someone he knows, someone who is in Gryffindor, someone he had up until that point trusted. The Heir of Slytherin speculation shifted away from Malfoy after the Polyjuice affair, but now it is away from Slytherin altogether, leaving Percy, due to his earlier suspect positions, as a more likely threat, but really anyone in the House could now be responsible. After Hermione is petrified while all of Hogwarts is outside watching the Quidditch match, the circumstances become more threatening; the creature responsible for the crime may not even be a student. With the diary gone, Harry can no longer communicate with Riddle, and so the burden of discovery lies entirely on him alone.

Figure 3.3 An example of an analysis paragraph mislabeled as "Analyis" instead of "Analysis"

Once the pages were processed, the first dataset - using the Sparknotes summaries - could be constructed. The dataset is a json file containing all of the information and metadata necessary for the QA task.

3.3 Data Description

The data is summarized in Table 3.2. Summaries vary in length - most are between 3 and 9 paragraphs. Figure 3.4 shows the number of summaries of different paragraph lengths. The main drawback of this dataset is its relatively small size, similar to NarrativeQA. As will be seen, baseline models trained solely on Sparknotes data perform poorly. As such, we believe that this dataset is best suited as an evaluation dataset for measuring the robustness of models trained on other, larger datasets.

# books in Sparknotes with section quizzes	396
# section summaries	4,156
# questions	20,419

avg. # questions per summary	4.9
Avg. summary length (in words)	708

Table 3.2 Summary statistics for Sparknotes dataset

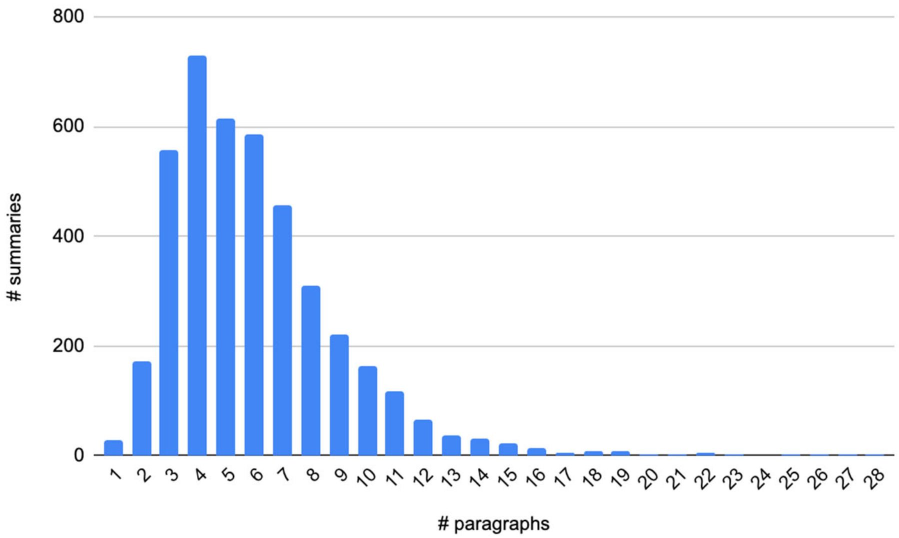
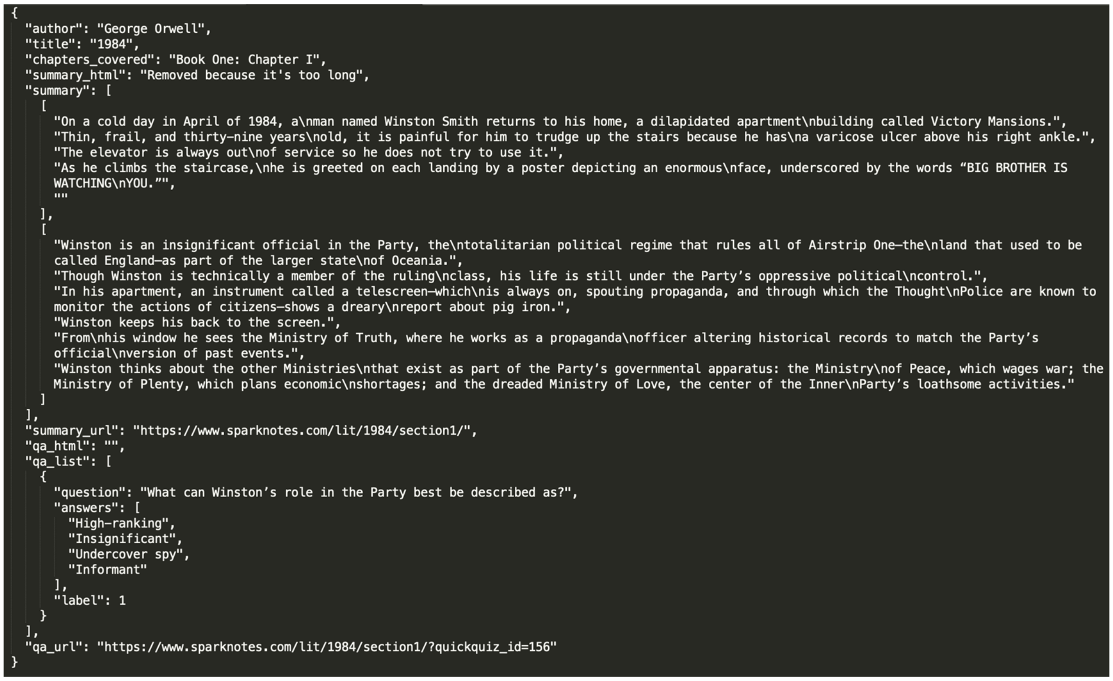


Figure 3.4 A histogram with the length of a summary on the x-axis and the number of summaries of that length on the y-axis

The data is stored as a list of summaries with their associated questions. Each JSON object contains a summary, a list of questions and answers associated with that summary and relevant metadata (see Figure 3.5 for an example).

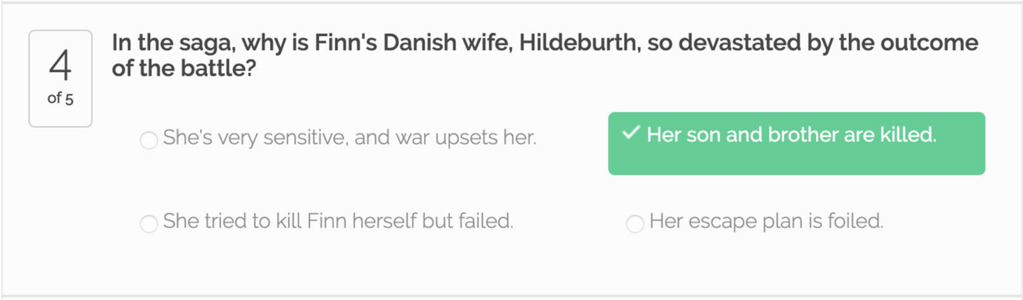


```
{
  "author": "George Orwell",
  "title": "1984",
  "chapters_covered": "Book One: Chapter I",
  "summary_html": "Removed because it's too long",
  "summary": [
    [
      "On a cold day in April of 1984, a\nman named Winston Smith returns to his home, a dilapidated apartment\nbuilding called Victory Mansions.",
      "Thin, frail, and thirty-nine years\nold, it is painful for him to trudge up the stairs because he has\na varicose ulcer above his right ankle.",
      "The elevator is always out\nof service so he does not try to use it.",
      "As he climbs the staircase,\nhe is greeted on each landing by a poster depicting an enormous\nface, underscored by the words “BIG BROTHER IS
      WATCHING\nYOU.”",
      ""
    ],
    [
      "Winston is an insignificant official in the Party, the\ntotalitarian political regime that rules all of Airstrip One—the\nland that used to be
      called England—as part of the larger state\nof Oceania.",
      "Though Winston is technically a member of the ruling\nclass, his life is still under the Party’s oppressive political\ncontrol.",
      "In his apartment, an instrument called a telescreen—which\nis always on, spouting propaganda, and through which the Thought\nPolice are known to
      monitor the actions of citizens—shows a dreary\nreport about pig iron.",
      "Winston keeps his back to the screen.",
      "From\nhis window he sees the Ministry of Truth, where he works as a propaganda\nofficer altering historical records to match the Party’s
      official\nversion of past events.",
      "Winston thinks about the other Ministries\nthat exist as part of the Party’s governmental apparatus: the Ministry\nof Peace, which wages war; the
      Ministry of Plenty, which plans economic\nshortages; and the dreaded Ministry of Love, the center of the Inner\nParty’s loathsome activities."
    ]
  ],
  "summary_url": "https://www.sparknotes.com/lit/1984/section1/",
  "qa_html": "",
  "qa_list": [
    {
      "question": "What can Winston’s role in the Party best be described as?",
      "answers": [
        "High-ranking",
        "Insignificant",
        "Undercover spy",
        "Informant"
      ],
      "label": 1
    }
  ],
  "qa_url": "https://www.sparknotes.com/lit/1984/section1/?quickquiz_id=156"
}
```

Figure 3.5 An example summary with its corresponding questions and metadata

3.4 Sparknotes Difficulty Evaluation

To evaluate the Sparknotes dataset, we wanted to see what score a human would get trying to answer the quizzes using the Sparknotes summaries without reading the corresponding novel. So, we reviewed a sample of 100 questions to approximate human performance. From this sample, we correctly answered 95 questions indicating that this is very much a feasible task. The questions we got wrong had been based on the Sparknotes analysis of the chapter which covers themes and ideas not necessarily found in the novel or its summaries. Thus, we can see that there are a few questions taken from the analysis rather than the summaries, but for the most part all of the questions come from the summary text. That said, many questions did not come verbatim out of the text and required inference and a deeper understanding of the summary.



4 of 5

In the saga, why is Finn's Danish wife, Hildeburth, so devastated by the outcome of the battle?

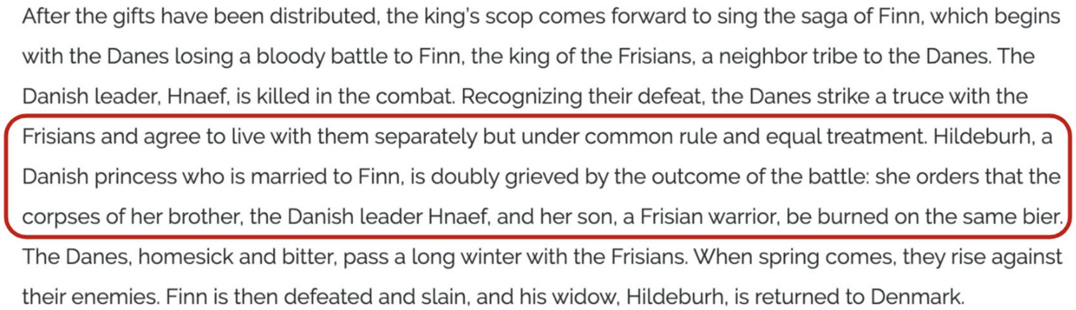
○ She's very sensitive, and war upsets her.

✓ Her son and brother are killed.

○ She tried to kill Finn herself but failed.

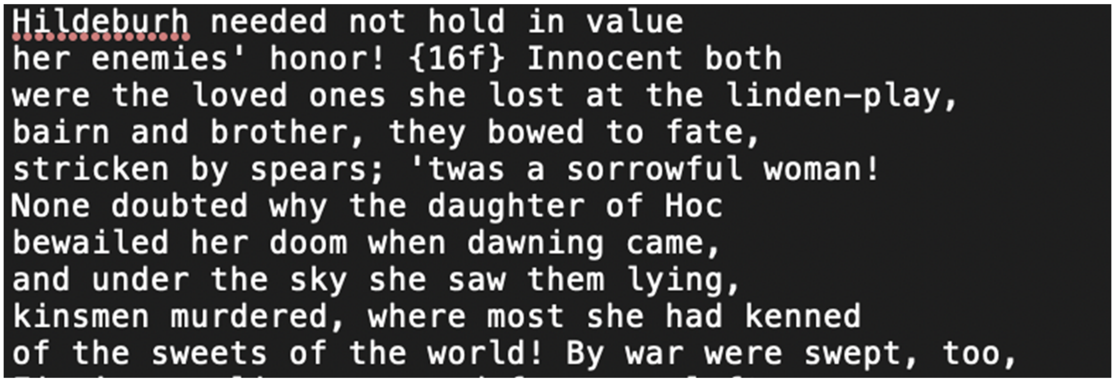
○ Her escape plan is foiled.

Figure 3.6 A question taken from the summaries of Beowulf



After the gifts have been distributed, the king's scop comes forward to sing the saga of Finn, which begins with the Danes losing a bloody battle to Finn, the king of the Frisians, a neighbor tribe to the Danes. The Danish leader, Hnaef, is killed in the combat. Recognizing their defeat, the Danes strike a truce with the Frisians and agree to live with them separately but under common rule and equal treatment. Hildeburh, a Danish princess who is married to Finn, is doubly grieved by the outcome of the battle: she orders that the corpses of her brother, the Danish leader Hnaef, and her son, a Frisian warrior, be burned on the same bier. The Danes, homesick and bitter, pass a long winter with the Frisians. When spring comes, they rise against their enemies. Finn is then defeated and slain, and his widow, Hildeburh, is returned to Denmark.

Figure 3.7 The passage from the Sparknotes summary containing the answer to the question posed in Figure 3.6. The answer is not stated explicitly, but rather must be inferred from the text.



```
Hildeburh needed not hold in value
her enemies' honor! {16f} Innocent both
were the loved ones she lost at the linden-play,
bairn and brother, they bowed to fate,
stricken by spears; 'twas a sorrowful woman!
None doubted why the daughter of Hoc
bewailed her doom when dawning came,
and under the sky she saw them lying,
kinsmen murdered, where most she had kenned
of the sweets of the world! By war were swept, too,
```

Figure 3.8 The passage from Beowulf in Project Gutenberg containing the answer to the question posed in 3.6.

3.4 Project Gutenberg

In addition to the Sparknotes dataset, we also created a dataset of full-length Project Gutenberg texts to be paired with Sparknotes questions. Appendix B contains a list of all of the works we have included. Of the 396 books and 20,419 questions in Sparknotes, there are 135 public domain works in Project Gutenberg with over 5,000 matching

Sparknotes questions. This is a relatively small dataset and, like the Sparknotes dataset, its main drawback is that it serves more as an evaluation dataset due to its small size.

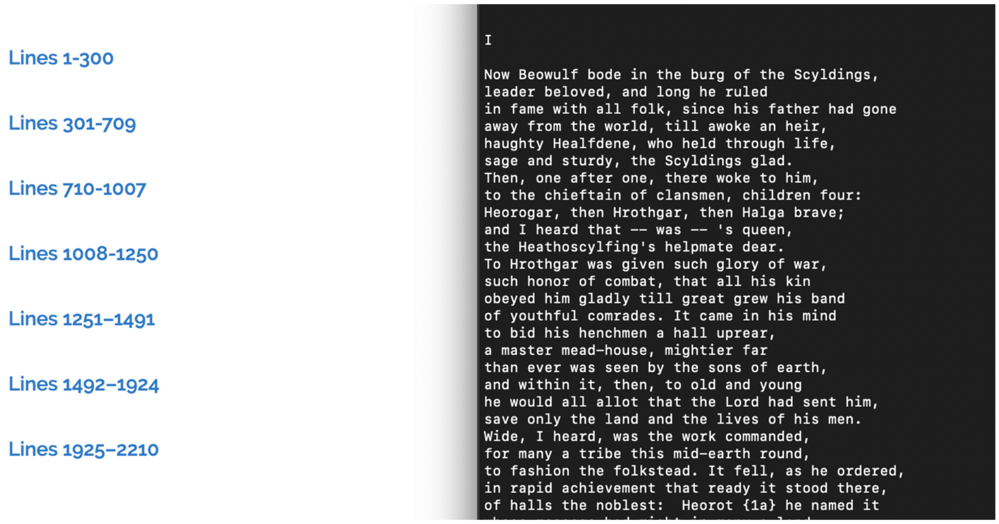
3.4.1 Dataset Overview

The Project Gutenberg texts were downloaded[3] using Project Gutenberg APIs[4]. From here, we extracted the titles of each book and identified titles contained in both Project Gutenberg and our Sparknotes dataset.

The challenging task with creating the Project Gutenberg dataset is splitting up the works in the same way that Sparknotes splits them up. Some novels are split up by book and chapter while others are split up only by chapter. Some plays are split up by act or scene, while others are split up by line number. Additionally, different versions of novels and plays are split up differently. For example, some longer Russian novels can have as many as 10 more chapters in one edition than another.

[3] Thanks to Daphne Ippolito for providing the downloaded texts

[4] https://www.gutenberg.org/wiki/Gutenberg:Information_About_Robot_Access_to_our_Pages



Lines 1-300

Lines 301-709

Lines 710-1007

Lines 1008-1250

Lines 1251–1491

Lines 1492–1924

Lines 1925–2210

```
I

Now Beowulf bode in the burg of the Scyldings,
leader beloved, and long he ruled
in fame with all folk, since his father had gone
away from the world, till awoke an heir,
haughty Healfdene, who held through life,
sage and sturdy, the Scyldings glad.
Then, one after one, there woke to him,
to the chieftain of clansmen, children four:
Heorogar, then Hrothgar, then Halga brave;
and I heard that -- was -- 's queen,
the Heathoscylfing's helpmate dear.
To Hrothgar was given such glory of war,
such honor of combat, that all his kin
obeyed him gladly till great grew his band
of youthful comrades. It came in his mind
to bid his henchmen a hall uprear,
a master mead-house, mightier far
than ever was seen by the sons of earth,
and within it, then, to old and young
he would all allot that the Lord had sent him,
save only the land and the lives of his men.
Wide, I heard, was the work commanded,
for many a tribe this mid-earth round,
to fashion the folkstead. It fell, as he ordered,
in rapid achievement that ready it stood there,
of halls the noblest:  Heorot {1a} he named it
```

Figure 3.9 On the left, Beowulf as split up in Sparknotes - by line number. On the right, Beowulf as split up by Gutenberg - by section.

Thus, while we made an attempt at automating the chapterizing process, we ultimately found that this task must be done by hand. Instead, we constructed a version of the dataset where chapters are ignored and the model is fed the full novel for each question. Summary statistics for the dataset can be found in Table 3.3 and more comprehensive statistics can be found in Appendix B.

# books	136
# questions	5333

Table 3.3 Summary statistics for Gutenberg dataset

Chapter 4

Sparknotes Baseline Model

4.1 Overview

In this chapter we discuss our models for selecting an answer from the Sparknotes multiple choice questions using Sparknotes summaries. While these summaries are shorter than the full-length novels, they are still too long to be processed directly by BERT, since BERT is a feedforward neural network that accepts a maximum input sequence length of 512 subword units, and most summaries contain more than 512 words. So, we first implement a pipeline that begins a paragraph extraction step that selects the relevant part of the summary before trying to answer the question. Paragraph extraction takes in a summary, a question and a list of answers – without knowing the correct answer – and outputs the paragraph it thinks contains the information relevant to the question. Once we have a single paragraph for each question, we can feed this paragraph into a neural network to identify the correct answer.

4.2 Paragraph Extraction

4.2.1 Overview

Our first step is to extract the most relevant paragraph from the input summary to pair with the question, answer pair.

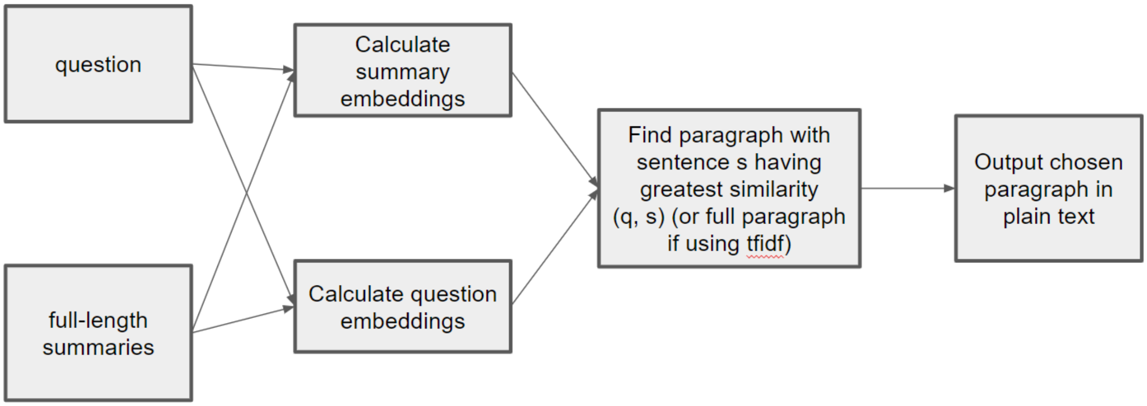

question
full-length summaries
Calculate summary embeddings
Calculate question embeddings
Find paragraph with sentence s having greatest similarity (q, s) (or full paragraph if using tfidf)
Output chosen paragraph in plain text

Figure 4.1 Paragraph extraction flow diagram

One standard method for paragraph extraction is to use TF-IDF with cosine similarity and Tay et al. (2019) use a variant of this approach on full-length novels. TF-IDF is a technique in information retrieval used to embed texts based on the frequency of the words that appear in the text. The equation is composed of two terms: the term frequency and the inverse document frequency. The term frequency is defined as:

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}}$$

where $f_{t,d}$ is the number of times the term t appears in the document d. The inverse document frequency is defined as:

$\mathrm{idf}(t,D) = \log\frac{N}{|\{d \in D : t \in d\}|}$ where D is the set of all documents in the corpus and $N = |D|$.

Finally, we put this together to get: $\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$

We then apply TF-IDF to the words in each paragraph and to the question to get vector representations of each chunk of text. To generate the vector representations using TF-IDF, we use the sklearn TF-IDFVectorizer. This takes in a collection of raw documents and outputs their vector representations, each with dimensionality |D|. Given these vector representations, we can perform cosine similarity. From here, the cosine similarity between each paragraph and the given question is calculated. The paragraph with the greatest score is considered to be the most similar to the question and thus the most informative.

Question	**Chosen Paragraph**
What can Winston's role in the Party best be described as? Cosine similarity with chosen paragraph: 0.2711 Does the chosen paragraph answer the question? No	From a drawer in a little alcove hidden from the telescreen, Winston pulls out a small diary he recently purchased. He found the diary in a secondhand store in the proletarian district, where the very poor live relatively unimpeded by Party monitoring. The proles, as they are called, are so impoverished and insignificant that the Party does not consider them a threat to its power. Winston begins to write in his diary, although he realizes that this constitutes an act of rebellion against the Party. He describes the films he watched the night before. He thinks about his lust and hatred for a dark-haired girl who works in the Fiction Department at the Ministry of Truth, and about an important Inner Party member named O'Brien—a man he is sure is an enemy of the Party. Winston remembers the moment before that day's Two Minutes Hate, an assembly during which Party orators whip the populace into a frenzy of hatred against the enemies of Oceania. Just before the Hate began, Winston knew he hated Big Brother, and saw the same loathing in O'Brien's eyes.
What is the purpose of the Ministry of Truth? Cosine similarity with chosen paragraph: 0.5803 Does the chosen paragraph answer the question? Yes	Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party's oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party's official version of past events. Winston thinks about the other Ministries that exist as

	part of the Party's governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party's loathsome activities.
What is the relationship between the proles and the Party? Cosine similarity with chosen paragraph: 0.4738 Does the chosen paragraph answer the question? Yes	From a drawer in a little alcove hidden from the telescreen, Winston pulls out a small diary he recently purchased. He found the diary in a secondhand store in the proletarian district, where the very poor live relatively unimpeded by Party monitoring. The proles, as they are called, are so impoverished and insignificant that the Party does not consider them a threat to its power. Winston begins to write in his diary, although he realizes that this constitutes an act of rebellion against the Party. He describes the films he watched the night before. He thinks about his lust and hatred for a dark-haired girl who works in the Fiction Department at the Ministry of Truth, and about an important Inner Party member named O'Brien—a man he is sure is an enemy of the Party. Winston remembers the moment before that day's Two Minutes Hate, an assembly during which Party orators whip the populace into a frenzy of hatred against the enemies of Oceania. Just before the Hate began, Winston knew he hated Big Brother, and saw the same loathing in O'Brien's eyes.

Table 4.1 Example results from paragraph extraction with TF-IDF and cosine similarity

Cosine similarity here is defined as:

$similarity(q,p) = \frac{\sum_{i=1}^{n} q_i p_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} p_i^2}}$ where q is the question and p is the given paragraph.

In addition to calculating embeddings using TF-IDF, we also run experiments using BERT word embeddings (Devlin et al., 2018), Infersent sentence embeddings (Conneau et al., 2017) and SentenceBERT (Reimers and Gurevych, 2019). These are techniques that generate word or sentence embeddings which we use similarly to the dense representation of the sentence vectors generated by TF-IDF that we previously discussed.

Instead of |D| dimensions in TF-IDF, the BERT-based embeddings have far fewer dimensions.

Our BERT word embeddings are calculated using the basic uncased BERT which contains 12 layers. The output of running the BERT word embeddings over a sentence is a matrix of three dimensions:

1) The number of layers (12 in this case)
2) The number of words / subword tokens (the length of the sentence)
3) The number of features (768 in our case)

To get a final array of word embeddings, we take the mean of the output of the final four layers and discard the previous eight[5]. This gives us a matrix of dimension sentence_length * 768. When we do this same process to embed the question, we will get a question embedding of dimension question_length * 768.

The process is the same for SentenceBERT (Reimers and Gurevych, 2019) and Infersent (Conneau et al., 2017) –named for the Stanford Natural Language Inference Datasets it was trained on– except that they output an embedding of dimension length * feature size since they perform pooling on the output layers for you. While BERT is trained to generate embeddings for single words in context, both SentenceBERT and Infersent are trained specifically on tasks where they learn to embed whole sentences. Both Infersent and SentenceBERT are trained on the Natural Language Inference task,

[5] Method was taken from https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/

another name for the Recognizing Textual Entailment Task which we explore in-depth in Section 5.2. Thus, we expect their sentence embeddings to be of higher quality than BERT word embeddings.

In all of BERT, Infersent, and SentenceBERT, once we have the list of sentence embeddings for each summary, to calculate cosine similarity between the paragraph and the question. We can think of each paragraph as having a matrix which has a size of the number of sentences in the paragraph by the sentence embedding size. We need to collapse each paragraph matrix into a single vector, so that we can compute the angle between the paragraph embedding and question embedding. There are a few ways to compute a single vector to represent a paragraph using its sentence embeddings:

1) Let the paragraph embedding be the sum of its sentence embeddings. In other words, given sentence embedding vectors $e_1 \cdots e_n$, the paragraph embedding e_p is:

$$e_p = \sum_{i=1}^{n} e_i$$

2) Let the paragraph embedding be the mean of its sentence embeddings. In other words, given sentence embedding vectors $e_1 \cdots e_n$, the paragraph embedding e_p is:

$$e_p = \frac{1}{n} \sum_{i=1}^{n} e_i$$

3) Calculate the cosine similarity between each sentence in the paragraph and the question. Then, for each paragraph choose the sentence most similar to the question and choose the paragraph containing the best sentence.

We try all 3 options and find that option 3 performs the best (see Table 4.9).

Question	**Chosen Paragraph**
What can Winston's role in the Party best be described as? Cosine similarity with chosen paragraph: 0.3664 Does the chosen paragraph answer the question? No	Winston looks down and realizes that he has written "DOWN WITH BIG BROTHER" over and over again in his diary. He has committed thoughtcrime—the most unpardonable crime—and he knows that the Thought Police will seize him sooner or later. Just then, there is a knock at the door.
What is the purpose of the Ministry of Truth? Cosine similarity with chosen paragraph: 0.3512 Does the chosen paragraph answer the question? No	On a cold day in April of 1984, a man named Winston Smith returns to his home, a dilapidated apartment building called Victory Mansions. Thin, frail, and thirty-nine years old, it is painful for him to trudge up the stairs because he has a varicose ulcer above his right ankle. The elevator is always out of service so he does not try to use it. As he climbs the staircase, he is greeted on each landing by a poster depicting an enormous face, underscored by the words "BIG BROTHER IS WATCHING YOU."
What is the relationship between the proles and the Party? Cosine similarity with chosen paragraph: 0.4090 Does the chosen paragraph answer the question? No	On a cold day in April of 1984, a man named Winston Smith returns to his home, a dilapidated apartment building called Victory Mansions. Thin, frail, and thirty-nine years old, it is painful for him to trudge up the stairs because he has a varicose ulcer above his right ankle. The elevator is always out of service so he does not try to use it. As he climbs the staircase, he is greeted on each landing by a poster depicting an enormous face, underscored by the words "BIG BROTHER IS WATCHING YOU."

Table 4.2 Example results from paragraph extraction with Infersent and cosine similarity

Question	**Chosen Paragraph**
What can Winston's role in the Party best be described as?	Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is

Cosine similarity with chosen paragraph: 0.5370533 Does the chosen paragraph answer the question? Yes	technically a member of the ruling class, his life is still under the Party’s oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party’s official version of past events. Winston thinks about the other Ministries that exist as part of the Party’s governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party’s loathsome activities.
What is the purpose of the Ministry of Truth? Cosine similarity with chosen paragraph: 0.4710 Does the chosen paragraph answer the question? Yes	Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party’s oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party’s official version of past events. Winston thinks about the other Ministries that exist as part of the Party’s governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party’s loathsome activities.
What is the relationship between the proles and the Party? Cosine similarity with chosen paragraph: 0.4845	Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party’s oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the

Does the chosen paragraph answer the question? No	Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party's official version of past events. Winston thinks about the other Ministries that exist as part of the Party's governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party's loathsome activities.

Table 4.3 Example results from paragraph extraction with SentenceBERT and cosine similarity

4.2.2 Paragraph Extraction Improvements

In the paragraph extraction process described above, we compare the embedding of each paragraph with the embedding of the question without the multiple choice answers. We consider a variant on the paragraph extraction process that incorporates the multiple choice answers. We can do this by incorporating the answers: either by concatenating the question with all of the answers or by concatenating the question with each individual answer in turn. We then compare the question concatenated with all of the answers or one of the answers to the paragraphs in the summary as before. For the training and validation sets, we can also compare each paragraph embedding to the embedding of the question concatenated with the correct answer. For the test set, we cannot consider only the correct answer since the correct answer is not known to the model during testing.

The purpose of using the correct answer alone concatenated with the question is that the correct answer is exactly what we are trying to identify in the text. Thus, we improve the accuracy of the paragraph extraction process by feeding in the answer we are looking for. Having the correct paragraph in the train and dev sets is important because the model we teach to identify the correct answer given the paragraph needs correct paragraphs in order to learn to identify correct answers. Otherwise, it is simply learning to guess. We explain the answer selection model in detail in Section 4.3.

4.2.3 Evaluation

To evaluate the different paragraph extraction methods, we create a testset of 100 questions taken from 20 summaries. We manually label the correct paragraph in the summary for each question. We can then evaluate how each extraction method performs on the test set. For evaluation, we want to understand not only how well our methods do at identifying the correct paragraph, but, if they are wrong, just how wrong they are. In other words, was the correct paragraph the second choice or the last choice for the system. To evaluate this, we use mean reciprocal rank (MRR). Given a list of questions Q and a list of summaries P, we say that:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Where we define $rank_i$ to be the rank position of the correct paragraph.

As an example, consider a list of three questions and three one sentence paragraphs where P_i contains the answer to Q_i:

Q1: *What can Winston’s role in the Party best be described as?*

Q2: *What is the purpose of the Ministry of Truth?*

Q3: *What does Winston tell Julia about his wife in Chapter 3?*

P1: *Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One.*

P2: *From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party's official version of past events.*

P3: *Winston tells Julia about a walk he once took with his ex-wife Katherine, during which he thought about pushing her off of a cliff.*

We can construct the following table to calculate the MRR.

Question	Ranking	Correct Paragraph	Rank	Reciprocal rank
What can Winston's role in the Party best be described as?	P3, P2, **P1**	P1	3	1/3
What is the purpose of the Ministry of Truth?	P1, **P2,** P3	P2	2	1/2
What does Winston tell Julia about his wife in Chapter 3?	**P3,** P2, P1	P3	1	1/1

Table 4.4 An example MMR calculation

The MRR would then be,

$$(1/3 + 1/2 + 1)/3 = 0.611$$

Table 4.5 shows the results for the paragraph extraction methods described in this section:

What is compared with	Similarity	MRR

each paragraph?		
Choose a paragraph at random	None	0.457
The question alone	TFIDF + cosine	0.785
The question together with all four answers	TFIDF + cosine	0.815
The question with each answer individually and choosing the score of the best question, answer pair	TFIDF + cosine	0.785
The question alone	SentenceBERT + cosine	0.820
The question together with all four answers	**SentenceBERT + cosine**	**0.863**
The question with each answer individually and choosing the score of the best question, answer pair	SentenceBERT + cosine	0.835

Table 4.5 Results from the MMR evaluation

There appears to be a meaningful difference between the different methods. The best performing method is SentenceBERT and cosine using the question concatenated with all of the four answers.

Question	**Chosen Paragraph (Correct)**
What can Winston's role in the Party best be described as?	Winston is an **insignificant official** in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party's oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party's official version of past events. Winston thinks about the other Ministries that exist as part of the Party's governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded

	Ministry of Love, the center of the Inner Party's loathsome activities.
What is the purpose of the Ministry of Truth?	Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party's oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer **altering historical records to match the Party's official version of past events**. Winston thinks about the other Ministries that exist as part of the Party's governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party's loathsome activities.
What does Winston tell Julia about his wife in Chapter 3?	**Winston tells Julia about a walk he once took with his ex-wife Katherine, during which he thought about pushing her off of a cliff.** He says that it would not have mattered whether he pushed her or not, because it is impossible to win against the forces of oppression that govern their lives.
What does Winston trace on the café table?	Winston, now free, sits at the Chestnut Tree Café, where dismissed Party members go to drink. He enjoys a glass of Victory Gin and watches the telescreen. He accepts everything the Party says and does. Without acknowledging it to himself, he can still smell the rats. **On the table, Winston traces "2 + 2 = 5" in the dust.** He remembers seeing Julia on a bitter-cold day that March. She had thickened and stiffened, and he now found the thought of sex with her repulsive. They acknowledged that they had betrayed one another, and agreed to meet again, though neither is truly interested in continuing their relationship. Winston thinks he hears the song lyrics "Under the spreading chestnut tree / I sold you and you sold me," which he heard when he saw the political prisoners there many years earlier. He begins to cry. He remembers a moment of happiness with his mother and sister, but thinks it must be a false memory. He looks up and sees a picture of Big Brother on the telescreen, making him feel happy and safe. As he listens to the war news, he reassures himself of both the great victory he has won over himself and his newfound love for Big Brother.
Who comes to Winston's door while he is writing in his diary?	Winston opens the door fearfully, assuming that the Thought Police have arrived to arrest him for writing in the diary. However, it is only **Mrs. Parsons**, a neighbor in his apartment building, needing help with the plumbing while her husband is away. In Mrs. Parsons's apartment, Winston is tormented by the fervent Parsons children, who, being Junior Spies, accuse him of thoughtcrime. The Junior Spies is an organization of children who monitor adults for disloyalty to the Party, and frequently succeed in catching them—Mrs. Parsons herself seems afraid of her zealous children. The children are very agitated because their mother won't let them go to a public hanging of some of the Party's political enemies in the park that evening. Back in his apartment, Winston remembers a dream in which a man's voice—O'Brien's, he thinks—said to him, "We shall meet in the place where there is no darkness." Winston writes in his diary that his

	thoughtcrime makes him a dead man, then he hides the book.

Table 4.6 Examples of correct paragraphs chosen using SentenceBERT paragraph extraction

Question	**Chosen Paragraph (Incorrect)**
How does the image of Big Brother make Winston feel at the end of the novel?	One day, in a sudden, passionate fit of misery, Winston screams out Julia's name many times, terrifying himself. Though he knows that crying out in this way will lead O'Brien to torture him, he realizes his deep desire to continue hating the Party. He tries to bottle up his hatred so that even he will not recognize it. Therefore, when the Party kills him, he will die hating Big Brother—a personal victory. But he cannot hide his feelings. When O'Brien arrives with the guards, Winston tells him that he hates Big Brother. O'Brien replies that obeying Big Brother is not sufficient—Winston must learn to love him. O'Brien then instructs the guards to take Winston to Room 101.
What is in Room 101?	One day, in a sudden, passionate fit of misery, Winston screams out Julia's name many times, terrifying himself. Though he knows that crying out in this way will lead O'Brien to torture him, he realizes his deep desire to continue hating the Party. He tries to bottle up his hatred so that even he will not recognize it. Therefore, when the Party kills him, he will die hating Big Brother—a personal victory. But he cannot hide his feelings. When O'Brien arrives with the guards, Winston tells him that he hates Big Brother. O'Brien replies that obeying Big Brother is not sufficient—Winston must learn to love him. O'Brien then instructs the guards to take Winston to Room 101.
What personal victory does Winston hope to achieve over the Party?	After some time, Winston is transferred to a more comfortable room and the torture eases. He dreams contently of Julia, his mother, and O'Brien in the Golden Country. He gains weight and is allowed to write on a small slate. He comes to the conclusion that he was foolish to oppose the Party alone, and tries to make himself believe in Party slogans. He writes on his slate "FREEDOM IS SLAVERY," "TWO AND TWO MAKE FIVE," and "GOD IS POWER."
What is the relationship between the proles and the Party?	Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party's oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party's official version of past events. Winston thinks about the other Ministries that exist as part of the Party's governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party's loathsome activities.
Why does Winston consider suicide?	"Winston remembers an occasion when he caught the Party in a lie. In the mid-1960s, a cultural backlash caused the original leaders of the Revolution to be arrested. One day, Winston saw a few of these deposed leaders sitting at the Chestnut Tree Café, a gathering place for out-of-favor Party members.

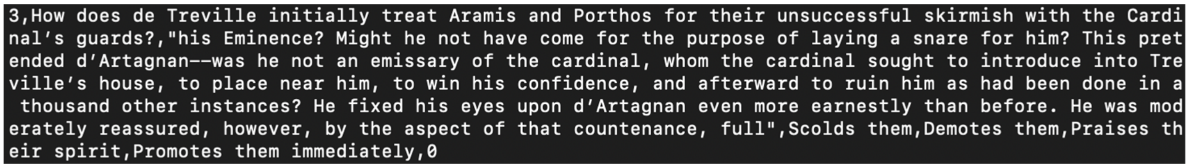
	A song played—"Under the spreading chestnut tree / I sold you and you sold me"—and one of the Party members, Rutherford, began to weep. Winston never forgot the incident, and one day came upon a photograph that proved that the Party members had been in New York at the time that they were allegedly committing treason in Eurasia. Terrified, Winston destroyed the photograph, but it remains embedded in his memory as a concrete example of Party dishonesty.

Table 4.7 Examples of incorrect paragraphs chosen using SentenceBERT paragraph extraction

4.2.4 Paragraph Extraction with Project Gutenberg

In addition to performing paragraph extraction on the Sparknotes summaries, it is also worth seeing how these methods perform on full-novel texts from Project Gutenberg. To perform paragraph extraction on entire novels, we use one of the methods previously mentioned: SentenceBERT embeddings with cosine similarity. We create a sliding window of 80 tokens and choose the sequence of 80 tokens that is most similar to the question. Figure 4.2 shows an example question with its context.

Once the paragraph extraction is complete, we can evaluate its performance. We take a sample of 100 question, context pairs and evaluate how many contexts contain the answer to their corresponding question. In the 100 pairs evaluated, only 5 appeared to contain the answer to the question.



```
3,How does de Treville initially treat Aramis and Porthos for their unsuccessful skirmish with the Cardi
nal's guards?,"his Eminence? Might he not have come for the purpose of laying a snare for him? This pret
ended d'Artagnan--was he not an emissary of the cardinal, whom the cardinal sought to introduce into Tre
ville's house, to place near him, to win his confidence, and afterward to ruin him as had been done in a
 thousand other instances? He fixed his eyes upon d'Artagnan even more earnestly than before. He was mod
erately reassured, however, by the aspect of that countenance, full",Scolds them,Demotes them,Praises th
eir spirit,Promotes them immediately,0
```

Figure 4.2 An example question and its context

4.3 Multiple Choice Model

4.3.1 Overview

Once a paragraph has been chosen, the data, now containing the single chosen paragraph, the question, the answers, and the label is passed into RoBERTa (Liu et al., 2019), which is a model that improves on BERT by training on additional data for a longer period of time. Using this improved training procedure, RoBERTa has been shown to improve performance on tasks such as MNLI (Multi-Genre Natural Language Inference), QA, and RTE. We use RoBERTa to fine-tune on the multiple choice task (see Section 2.6 for an introduction to fine-tuning for question answering). For running the RoBERTa model, we use the Hugging Face[6] library. Hugging Face is a popular NLP framework for using state-of-the-art transformers such as BERT and RoBERTa. Hugging face contains thousands of pre-trained models for use in either Tensorflow or Pytorch - we use Tensorflow.

To fine-tune on the multiple choice problem, the model needs to be set up correctly. The context is concatenated with the question and then each possible answer is concatenated with the context and question, with a delimiter separating them. Each context-question-answer string is then fed into the transformer model to get the final embeddings. We then run everything through an initially untrained, final linear layer and apply a softmax activation function to the four answer choices to get probabilities. We take the maximum probability to be the predicted label. Figures 4.3 and 4.4 provide a graphical overview of

[6] https://huggingface.co/

the entire architecture. In Figure 4.4, the blue and purple rectangles represent the concatenated context, question and answer which are passed into the transformer, RoBERTa, before being passed into a final linear layer and converted to probabilities with a softmax over the four answer choices..

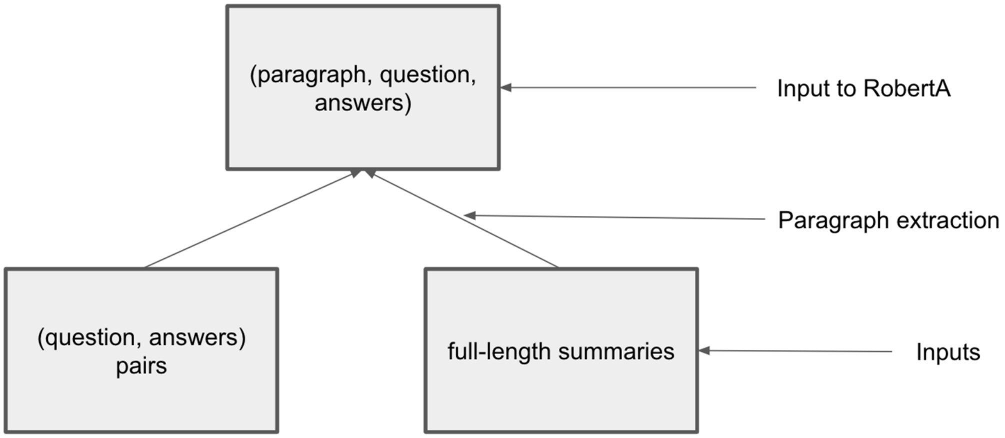


Figure 4.3 Paragraph extraction overview

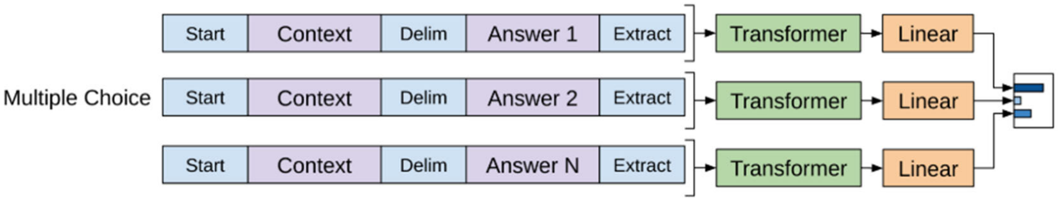


Figure 4.4 Illustration of how we set up the multiple choice task for RoBERTa (Radford et al., 2018)

4.3.2 Experiments

We first run the paragraph extraction on the entire Sparknotes dataset. Once the supporting paragraphs have been chosen, the dataset is split up into a training set (80% of the entire dataset), a validation set (10%), and a test set (10%). Using sklearn, the data is randomly shuffled prior to being split so as to ensure the test set is composed of questions from different types of texts - novels, plays, speeches, etc. We chose the hyperparameters to be the hyperparameters used for fine-tuning RoBERTa on a different multiple choice dataset, SWAG (Zellers et al., 2018) - see Table 4.8 for the details of the hyperparameters used. Note that while RoBERTa can accept sequences of length up to 512, we had to keep the maximum sequence length at 80 to ensure we do not run into memory issues. Finally, per Figure 4.4, the experiment is set up as a multiclass classification problem with four labels representing the four possible answers.

Learning Rate	5e-5
Batch Size	16
# Epochs	3
Maximum total input sequence length after tokenization	80

Table 4.8 Model hyperparameters

4.3.3 Results (With Truncation)

Table 4.9 shows the results of the four baseline experiments. All four experiments achieved relatively similar results. Our baseline performance is relatively weak for all four models and TF-IDF performs similarly to the other more sophisticated embedding methods.

Paragraph Extraction Method	**Model**	**Evaluation Accuracy (higher is better)**	**Evaluation Loss (lower is better)**
N/A	Random guessing	.25	N/A
BERT word embedding + cosine	RoBERTa	0.460	1.2

Infersent embedding + cosine	RoBERTa	0.463	1.15
TF-IDF + cosine	RoBERTa	0.440	1.18
SentenceBERT + cosine using best sentence	RoBERTa	0.461	1.14
SentenceBERT + cosine summing over sentences in paragraph	RoBERTa	0.436	1.20
SentenceBERT + cosine averaging over sentences in paragraph	RoBERTa	0.437	1.22

Table 4.9 Results of baseline experiments on the sparknotes dataset. The evaluation accuracy measures only the models accuracy on the quiz and does not measure our success at choosing the correct paragraph.

We then consider the modification offered in Section 4.2.3. Specifically, concatenating the answer to the question during the paragraph extraction process for training and concatenating all answers to the question for paragraph extraction during inference. Table 4.10 shows the results.

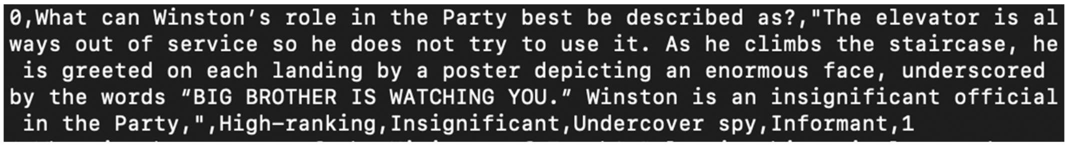
Paragraph Extraction Method - Train + Val	**Paragraph Extraction Method - Test**	**Model**	**Evaluation Accuracy (higher is better)**	**Evaluation Loss (lower is better)**
Sentence BERT embeddings and comparing the answer to each paragraph.	Sentence BERT embeddings and comparing the question to each paragraph.	RoBERTa	**0.471**	**1.185**
Sentence BERT embeddings and comparing the answer concatenated with the question to each paragraph.	Sentence BERT embeddings and comparing the question to each paragraph.	RoBERTa	0.469	1.126

Sentence BERT embeddings and comparing the answer to each paragraph.	Sentence BERT embeddings and comparing the question concatenated with all four answers to each paragraph.	RoBERTa	0.469	1.126

Table 4.10 Results of re-running the baseline with the modified paragraph extraction process. We use the same hyperparameters as before.

4.3.4 Results (On Subparagraph Chunks Without Truncation)

A significant challenge with training the fine-tuning model is managing memory limitations. Our GPUs are limited such that the input to the model - containing the question, context, and answer - must be no more than 80 tokens long. In these cases, we truncate the paragraph and only give the model the first 80 tokens. Thus, for long paragraphs, we lose information when the model cuts off the input. To address this, we can extract smaller sized, non-overlapping chunks as context, rather than extracting a full paragraph. Figure 4.5 depicts one such example with a chunk size of 50 tokens..



```
0,What can Winston’s role in the Party best be described as?,"The elevator is al
ways out of service so he does not try to use it. As he climbs the staircase, he
 is greeted on each landing by a poster depicting an enormous face, underscored
by the words “BIG BROTHER IS WATCHING YOU.” Winston is an insignificant official
 in the Party,",High-ranking,Insignificant,Undercover spy,Informant,1
```

Figure 4.5 An example where the context is a chunk of length 50

Below are the results when running the best performing model - SentenceBERT using answers concatenated with questions and RoBERTa to fine-tune - using different sized contexts chunks. Note the improvement! With smaller chunks, our model is still able to identify relevant chunks and less information is lost during fine-tuning.

Context Size	Evaluation Accuracy	Evaluation Loss
Full Paragraph Truncated at 80 words (Average paragraph length = 104.36 words)	0.471	1.185
80 Words	0.499	1.175
60 Words	0.545	1.085
40 Words	**0.580**	1.098
30 Words	0.579	1.099

Table 4.11 Results using modified context lengths

4.4 Working Through an Example

In order to better understand the flow of the models described in this section, we work through an example here. We can consider the example shown in Figure 3.5 & 4.6. The JSON as shown is exactly the input to our model, except that we have removed the HTML for clarity (the HTML is not used in the model).

We begin by extracting the question:

Q: *What can Winston's role in the Party best be described as?*

We also extract the four answer choices:

A1: *High-ranking*

A2: *Insignificant* **(the correct answer)**

A3: *Undercover spy*

A4: *Informant*

And we extract the context - the summary:

P1: *On a cold day in April of 1984, a man named Winston Smith returns to his home, a dilapidated apartment building called Victory Mansions. Thin, frail, and thirty-nine years old, it is painful for him to trudge up the stairs because he has a varicose ulcer above his right ankle. The elevator is always out of service so he does not try to use it. As he climbs the staircase, he is greeted on each landing by a poster depicting an enormous face, underscored by the words "BIG BROTHER IS WATCHING YOU."*

P2: *Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party's oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer altering historical records to match the Party's official version of past events. Winston thinks about the other Ministries that exist as part of the Party's governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans economic shortages; and the dreaded Ministry of Love, the center of the Inner Party's loathsome activities.*

Given this summary, we want to identify which of the 40-word chunks contains the answer to the given question. We begin by splitting the summary into 40-word chunks. We use the SentenceBERT embeddings to identify the most similar paragraph. Specifically, we go through each of the chunks and calculate the embedding for every sentence in the chunk. For each sentence, we then calculate the cosine similarity between it and the question. For example, the sentence As he climbs the staircase, he is greeted on each landing by a poster depicting an enormous face, underscored by the words "BIG BROTHER IS WATCHING YOU." has a similarity score of 0.307 with the question. Once we have the similarity scores of each sentence, we set the similarity score of each 40-word chunk to be the maximum of all of its sentence scores. Table 4.12 shows the results.

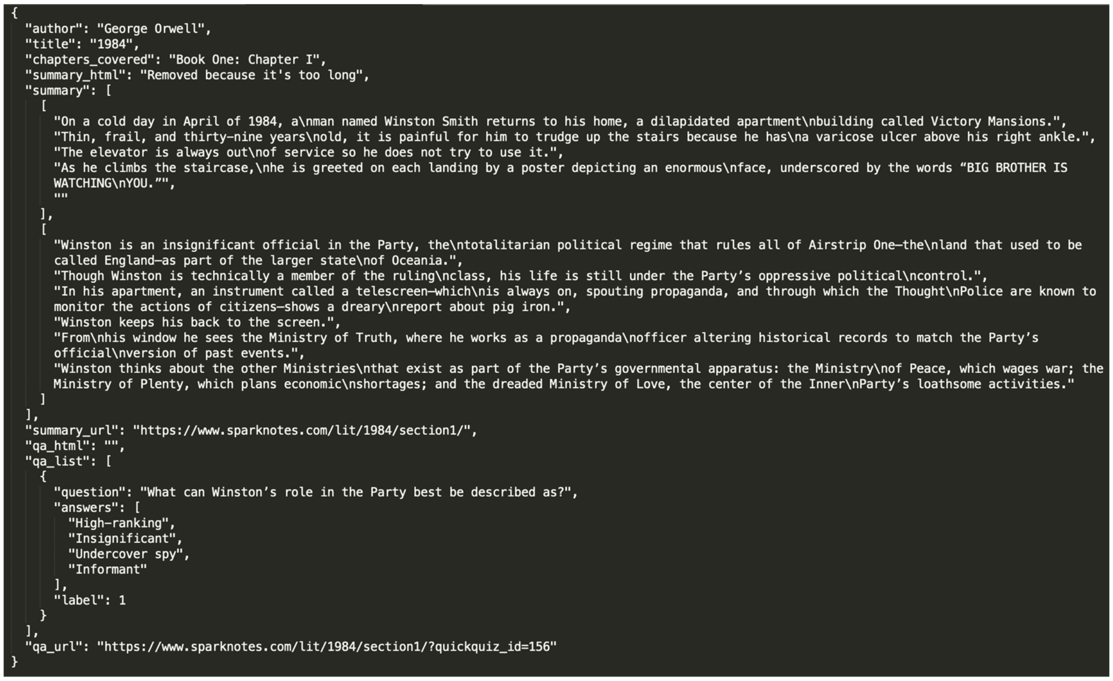
Index	Chunk	Similarity Score

1	On a cold day in April of 1984, a man named Winston Smith returns to his home, a dilapidated apartment building called Victory Mansions. Thin, frail, and thirty-nine years old, it is painful for him to trudge up the stairs	0.562
2	because he has a varicose ulcer above his right ankle. The elevator is always out of service so he does not try to use it. As he climbs the staircase, he is greeted on each landing by a poster depicting	0.447
3	an enormous face, underscored by the words “BIG BROTHER IS WATCHING YOU.” Winston is an insignificant official in the Party, the totalitarian political regime that rules all of Airstrip One—the land that used to be called England—as part of the	0.617
4	**larger state of Oceania. Though Winston is technically a member of the ruling class, his life is still under the Party’s oppressive political control. In his apartment, an instrument called a telescreen—which is always on, spouting propaganda, and through which**	0.648
5	the Thought Police are known to monitor the actions of citizens—shows a dreary report about pig iron. Winston keeps his back to the screen. From his window he sees the Ministry of Truth, where he works as a propaganda officer	0.562
6	altering historical records to match the Party’s official version of past events. Winston thinks about the other Ministries that exist as part of the Party’s governmental apparatus: the Ministry of Peace, which wages war; the Ministry of Plenty, which plans	0.647
7	economic shortages; and the dreaded Ministry of Love, the center of the Inner Party’s loathsome activities.	0.647

Table 4.12 Results of running SentenceBERT with cosine on each of the chunks

We can see that paragraph 4 has the greater score and so we extract that paragraph to be the context. Given the extracted paragraph, we can create a new training example. The example is a row in a csv file of the form (idx, question, context, ending0, ending1, ending2, ending3, label) where the number indicates the answer choice index and *label* marks the index of the correct answer.

We now finish by feeding this training example into our RoBERTa model and checking the label it outputs against the correct label.



```
{
  "author": "George Orwell",
  "title": "1984",
  "chapters_covered": "Book One: Chapter I",
  "summary_html": "Removed because it's too long",
  "summary": [
    [
      "On a cold day in April of 1984, a\nman named Winston Smith returns to his home, a dilapidated apartment\nbuilding called Victory Mansions.",
      "Thin, frail, and thirty-nine years\nold, it is painful for him to trudge up the stairs because he has\na varicose ulcer above his right ankle.",
      "The elevator is always out\nof service so he does not try to use it.",
      "As he climbs the staircase,\nhe is greeted on each landing by a poster depicting an enormous\nface, underscored by the words “BIG BROTHER IS
      WATCHING\nYOU.”",
      ""
    ],
    [
      "Winston is an insignificant official in the Party, the\ntotalitarian political regime that rules all of Airstrip One—the\nland that used to be
      called England—as part of the larger state\nof Oceania.",
      "Though Winston is technically a member of the ruling\nclass, his life is still under the Party’s oppressive political\ncontrol.",
      "In his apartment, an instrument called a telescreen—which\nis always on, spouting propaganda, and through which the Thought\nPolice are known to
      monitor the actions of citizens—shows a dreary\nreport about pig iron.",
      "Winston keeps his back to the screen.",
      "From\nhis window he sees the Ministry of Truth, where he works as a propaganda\nofficer altering historical records to match the Party’s
      official\nversion of past events.",
      "Winston thinks about the other Ministries\nthat exist as part of the Party’s governmental apparatus: the Ministry\nof Peace, which wages war; the
      Ministry of Plenty, which plans economic\nshortages; and the dreaded Ministry of Love, the center of the Inner\nParty’s loathsome activities."
    ]
  ],
  "summary_url": "https://www.sparknotes.com/lit/1984/section1/",
  "qa_html": "",
  "qa_list": [
    {
      "question": "What can Winston’s role in the Party best be described as?",
      "answers": [
        "High-ranking",
        "Insignificant",
        "Undercover spy",
        "Informant"
      ],
      "label": 1
    }
  ],
  "qa_url": "https://www.sparknotes.com/lit/1984/section1/?quickquiz_id=156"
}
```

Figure 4.6 Our training example from Chapter 3

4.5 Using Coreference Resolution to Improve Baseline Performance

Many of the answers to Sparknotes questions are written using pronouns rather than naming the entity being referred to. For example, in one quiz on The Three Musketeers, we have the following question and answers:

Q: *Why does Kitty help D’Artagnan intercept Milady’s correspondence with the Comte de Wardes?*

A1: *She loves him,*

A2: She hates Milady,

A3: She hates the Cardinal,

A4: She likes all intrigue

Identifying the correct answer requires first identifying that “She” refers to the character *Kitty* in each of the answers. This is in itself a non-trivial task. In some cases, the pronoun is completely removed from the answer. For example, there is no pronoun corresponding to *Jogona* in the answers to this question:

Q: *What does Jogona do with the account of his testimony that the narrator types up?*

A1: *Buries it,*

A2: *Carries it in a pouch,*

A3:*Frames it,*

A4: *Puts it under his pillow*

4.5.1 Experiment and Results

We hypothesize that, if the pronouns in the Sparknotes questions are replaced with the actual entity name - and entities are added where they are completely missing - then the task should become more manageable. To test this, we take a sample of 75 questions and manually replace all pronouns with their named entities. Appendix D contains a list of all of the questions before and after coreference resolution. As an example, the second example above would become:

Q: *What does Jogona do with the account of his testimony that the narrator types up?*

A1: *Jogona buries the account of his testimony that the narrator types up,*

A2: *Jogona carries the account of his testimony that the narrator types up in a pouch,*

A3: *Jogona frames the account of his testimony that the narrator types up,*

A4: *Jogona puts the account of his testimony that the narrator types up under his pillow*

We then run inference using our baseline model with SentenceBERT to see what performance would be like on this sample. We found that accuracy and loss both improved. Results are summarized in Table 4.13.

Model	**Test set**	**Evaluation Accuracy**	**Evaluation Loss**
Baseline Model	Normal	0.58	1.098
Baseline Model	New samples	0.721	0.684

Table 4.13 Results of experiment using manual coreference resolution

Chapter 5

Transfer Learning Approaches

Another method of tackling this problem is using transfer learning from models pre-trained on other QA tasks. Transfer learning is a technique in machine learning where you use knowledge gained from solving one task to solve a different, but related task. We focus on a specific transfer learning task using other NLP tasks that relate closely to question answering. Transfer learning is appealing for a couple of reasons:

1) As mentioned in Chapter 3, the sparknotes dataset is relatively small. Training only on the sparknotes dataset is thus less effective than training on a much larger dataset.
2) Evaluating other models on our dataset allows us to test the robustness of these datasets and see whether they generalize well to other domains.

5.1 RTE Model

5.1.1 RTE Problem Setup

We model the QA task as a Recognizing Textual Entailment (RTE) task (Dagan et al., 2013). The RTE task takes in a context and a hypothesis sentence. The system is then asked to decide whether the hypothesis is entailed by the context or not entailed by the context. Non-entailment can mean either the hypothesis is contradicted by the context or that the context is related to the hypothesis but doesn’t entail it. Thus, we can model the QA task as a RTE task as follows: For each (context, question, answers) tuple, convert this into a set of (context, hypothesis) pairs where the context is a paragraph taken from the Sparknotes summary associated with the question. For example, consider the context, question, and answer:

C: *The general manager of the Central Station had taken the boat out two days before under the charge of a volunteer skipper, and they had torn the bottom out on some rocks. In light of what he later learns, Marlow suspects the damage to the steamer may have been intentional, to keep him from reaching Kurtz. Marlow soon meets with the general manager, who strikes him as an altogether average man who leads by inspiring an odd uneasiness in those around him and* ***whose authority derives merely from his resistance to tropical disease.***

Q: *Where does Marlow believe the general manager's authority comes from?*

A1: *Resistance to tropical disease* **(correct answer)**

A2: *Family connections*

A3: *Above average intelligence*

A4: *Kindness*

We can manually convert this into (context, hypothesis) pairs by turning the question and answer into a hypothesis:

H1: *Marlow believes the general manager's authority comes from resistance to tropical disease.*

H2: *Marlow believes the general manager's authority comes from family connections.*

H3: *Marlow believes the general manager's authority comes from above average intelligence.*

H4: *Marlow believes the general manager's authority comes from kindness.*

Here, H1 would be entailed since it is confirmed as true by the context. Since the rest of the answers are incorrect, H2, H3, and H4 would not be entailed by the context paragraph.

Finally, there are also instances where a hypothesis is neither supported nor contradicted by the context - where the context is simply irrelevant. For example, given the same context above, we could consider the hypothesis:

H5: *Emily likes to code in Python*

This statement is neither supported nor contradicted by the context. Moving forward, we will consider such hypotheses to be not entailed.

Note that, while this example is simple and could be done automatically, this kind of conversion can be more subtle and complex. For example, consider the question and answer:

Q: *When Marlow arrives at the Inner Station, what explanation does the Russian trader give for the natives attacking the steamer?*

A: *They were hungry*

After conversion this becomes:

When Marlow arrives at the Inner Station, the explanation the Russian trader gives for the natives attacking the steamer is that they were hungry.

In this case, we converted “what” into “the”, “give” to “gives” and added “is that” before tacking on the answer.

5.1.2 RTE and QA

RTE and QA have been tackled together before. Specifically, Wang and Manning (2010) used a tree-edit model to perform competitively on both RTE and QA tasks. Wang and Manning used the same setup as described above for RTE. For QA, they consider the setup is similar: the system is given a question and a possible answer and returns true if the sentence correctly answers the question. For example, they consider the question:

Q: W*ho beat Floyd Patterson to take the title away?*

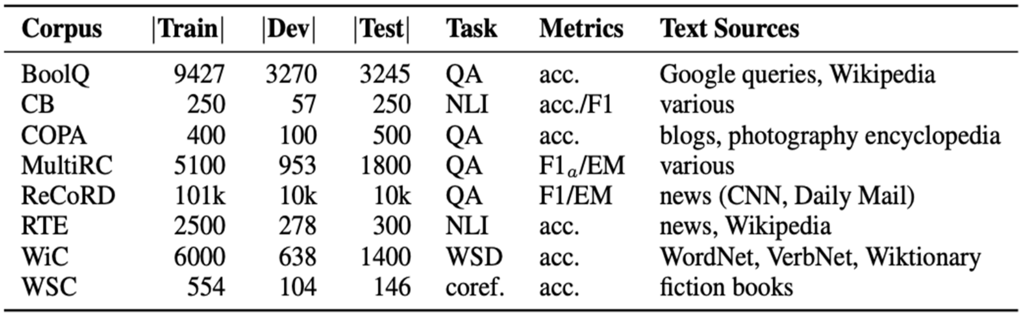
With the answer:

A: *He saw Ingemar Johansson knock down Floyd Patterson seven times there in winning the heavyweight title.*

Note that this problem formulation mirrors the RTE problem formulation and is similar to the adaption described in Section 5.2.1.

5.1.3 SuperGLUE

SuperGLUE (Wang et al., 2019) is a set of NLP tasks created to serve as a benchmark for measuring the success of new methods across multiple tasks. Figure 5.3 shows the summary statistics for each of the datasets included in SuperGLUE and Appendix C



Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

Figure 5.3 : The tasks included in SuperGLUE (Wang et al., 2019). For MultiRC, the number of total answers is listed for 456/83/166 train/dev/test questions.

Contains descriptions of each task and examples taken from each dataset.

SuperGLUE is more than just a set of benchmarks, it can be used to create multi-head models, pre-trained on different tasks and fine-tuned as a single or multi-head model. Multi-head models are models that are trained on multiple tasks concurrently. The tasks share a set of neural layers and then each task has a different final layer and activation function. All tasks perform backpropagation down to the shared layers such that the shared layers learn from all of the different tasks. Figure 5.4 shows an example of a multi-head architecture with three tasks. Below, we run experiments testing how a model pre-trained on the SuperGLUE tasks performs on our dataset.

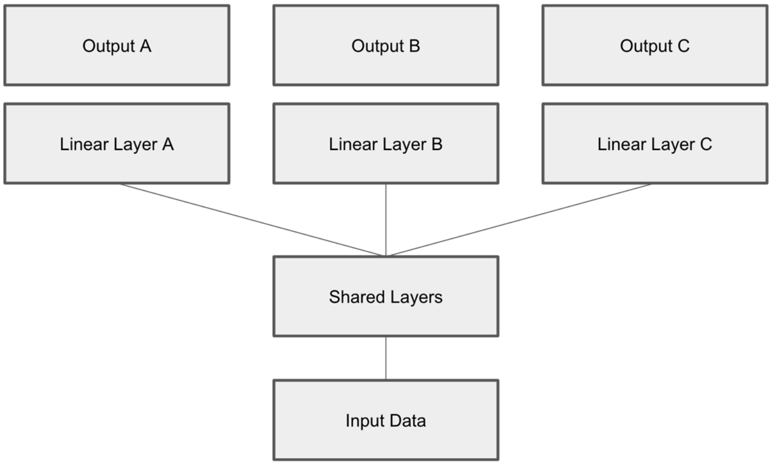
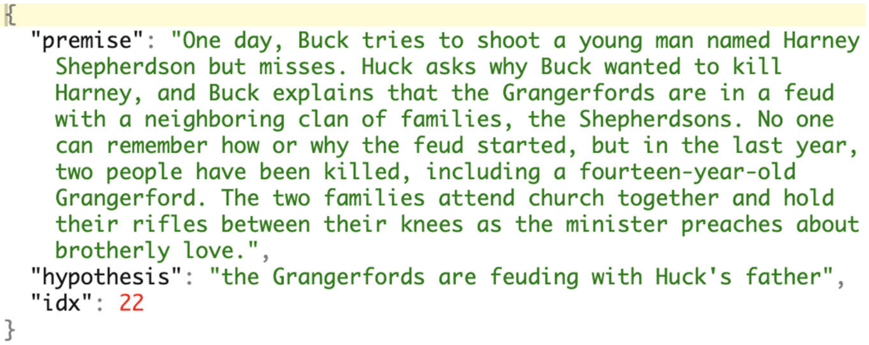


Figure 5.4 An example multi-head architecture with three tasks

5.1.4 RTE processing

Since the conversion can vary across different sentence and answer formats, converting the Sparknotes QA task into a RTE task is challenging automatically and therefore, we did a subset by hand. We created a small sample task to gauge whether, in an idealized setting, an RTE model could perform well on the Sparknotes questions. We converted 25 multiple choice questions, each with four answers into individual RTE examples, where each question-answer pair was a single example. All told, this gives 100 RTE examples of which 25 are entailed. Each of these 100 examples were paired with the paragraph that contained the correct answer to the original question. We then created an additional 25 questions where the paragraph matched with the hypothesis was unrelated such that there are also 100 examples where the model should detect neutral entailment. This sample was then used as the test set for inference. Appendix E contains 25 examples taken from the test set.



```
{
  "premise": "One day, Buck tries to shoot a young man named Harney
    Shepherdson but misses. Huck asks why Buck wanted to kill
    Harney, and Buck explains that the Grangerfords are in a feud
    with a neighboring clan of families, the Shepherdsons. No one
    can remember how or why the feud started, but in the last year,
    two people have been killed, including a fourteen-year-old
    Grangerford. The two families attend church together and hold
    their rifles between their knees as the minister preaches about
    brotherly love.",
  "hypothesis": "the Grangerfords are feuding with Huck's father",
  "idx": 22
}
```

Figure 5.5 An example taken from the RTE experiment

5.1.5 Experiment

For running the model, we used the Jiant toolkit,[7] an NLP software toolkit maintained by the NYU Machine Learning for Language Lab. It is designed for transfer learning and

[7] https://github.com/nyu-mll/jiant/tree/bert-friends-exps

sentence understanding tasks and contains an easy to use implementation of superGLUE. The model was run using hyperparameters recommended by Jiant. We pre-trained on the full superGLUE corpus and then fine-tuned a multi-head model on the RTE task. The model used for training and fine-tuning was RoBERTa, like with the baseline model. Once the model was complete, the RTE results could be evaluated manually. Table 5.2 shows that we were able to achieve accuracy similar to that of the SuperGLUE published results on the full RTE validation set, so we feel confident that it was trained properly.

Our RTE validation accuracy	0.697
SuperGLUE paper RTE validation accuracy	0.716

Table 5.2 Results of training on SuperGLUE tasks and validating on SuperGLUE RTE task

5.1.6 Results

The results of this experiment indicate that the RTE model struggles with the Sparknotes hypotheses. One possible explanation of this is that the context passages that RTE trains on are relatively short and thus it struggles to handle a full paragraph context. Table 5.3 gives the results of the experiment.

Note that, since 50% of the questions are matched with incorrect paragraphs and of the other 50% of the questions, 75% of the answers are incorrect, it ends up that only 12.5% of the examples are entailed. Therefore, it is not sufficient to look only at accuracy, but rather we must consider precision and recall as well.

Total examples	200
Accuracy	85%
Precision	22.2%

Recall	8%

Table 5.3 Results of the RTE experiment

From the results, we can see that the model has trouble identifying entailed sentences. Accuracy looks promising, but precision and recall appear to be relatively low. Specifically, of the 25 entailed hypotheses, our model only correctly identifies two - specified in Table 5.4. Both of these hypotheses are found in the context using very similar language. Thus, it seems our model managed only to identify entailment when the hypothesis, or an important part of it, can be found explicitly in the premise.

The one incongruity in the otherwise drab scene is the rosebush that grows next to the prison door. The narrator suggests that it offers a reminder of Nature's kindness to the condemned; for his tale, he says, it will provide either a "**sweet moral blossom**" or else some relief in the face of unrelenting sorrow and gloom.	the rose outside the prison symbolizes, according to the narrator, A **sweet moral blossom** found within the Hester's story
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester's earlier life flash through her mind: she sees her parents standing before their **home in rural England**, then she sees a "misshapen" scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester reflects on her **childhood home in England** while standing before the crowd

Table 5.4 Correctly identified entailed hypotheses

5.1.7 Working Through A RTE Example

To clarify the RTE flow, we work through an example here. Consider a context paragraph taken from the Sparknotes summary of the opening chapters of The Scarlet Letter:

"The one incongruity in the otherwise drab scene is the rosebush that grows next to the prison door. The narrator suggests that it offers a reminder of Nature's kindness to the

condemned; for his tale, he says, it will provide either a "sweet moral blossom" or else some relief in the face of unrelenting sorrow and gloom."

We are given the multiple choice question:

"What does the rose outside the prison symbolize, according to the narrator?"

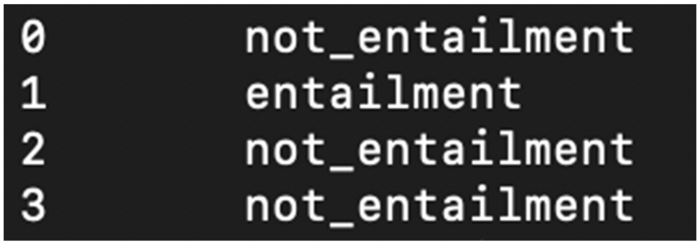
And the answers:

- *"The frailty of mankind"*
- *"A sweet moral blossom found within the Hester's story"*
    - The correct answer
- *"The reader's sympathy for Hester's ordeal"*
- *"The beautiful child that resulted from Hester's affair"*

We begin by converting these answers into hypotheses using the question. In this case, we get four hypotheses with labels:

- *"The rose outside the prison symbolizes, The frailty of mankind"*
    - Not entailed
- *"The rose outside the prison symbolizes, according to the narrator, A sweet moral blossom found within the Hester's story"*
    - Entailed by the context paragraph.
- *"The rose outside the prison symbolizes, according to the narrator, The reader's sympathy for Hester's ordeal"*
    - Not entailed
- *"The rose outside the prison symbolizes, according to the narrator, The beautiful child that resulted from Hester's affair"*
    - Not entailed

These four hypotheses can now be added to the RTE test set. At inference time, the model will try to answer whether each of these four hypotheses are entailed by the context paragraph. We train the model on the superGLUE tasks using the Jiant toolkit and then fine-tune the model on the SuperGLUE RTE task. Finally, the model evaluates this example. In this case, we can see from Figure 5.6 that the model correctly classified the entailment for all four of the hypotheses.



```
0        not_entailment
1        entailment
2        not_entailment
3        not_entailment
```

Figure 5.6 The output of the RTE model where the labels correspond to the four Scarlet Letter hypotheses we have defined here

Chapter 6

Conclusion and Future Work

In this thesis, we introduced a novel question answering dataset using Sparknotes summaries and quizzes and Project Gutenberg novels. Our dataset contains over 20,000 questions written by experts, creating a high-quality question answering dataset focusing on literature. Combining these questions with the Gutenberg novels offers a dataset for evaluating QA systems on novel-length texts.

We also provide a baseline system which, for each question, extracts a chunk of text that it believes contains the information needed to answer the question and then uses the state-of-the-art RoBERTa pre-trained model on the processed data. Our model achieves approximately 58% test accuracy using the Sparknotes summaries and quizzes while random guessing is 25%. We also provide a test set for evaluating paragraph extraction methods and use this to evaluate our own paragraph extraction methods.

Finally, we look to a well-known dataset for transfer learning, SuperGLUE. With SuperGLUE, we convert a subset of the Sparknotes dataset into premise, hypothesis pairs. We then pre-train a BERT model on all SuperGLUE tasks before fine-tuning on the RTE task. Ultimately, this model performs significantly worse on our test set than on the SuperGLUE testset. The RTE task is trained on smaller premises and thus appears to struggle with paragraph-length premises.

Sparknotes quizzes offer a new challenging task for question answering on long texts. We saw some success answering the multiple choice quizzes using the Sparknotes summaries, but the gap between our baseline and our approximation of human performance shows that this is still a challenging task. Future work could include improving on the design of the baseline system to bridge the gap with approximated human performance. Additionally, measuring human performance more carefully using crowd sourcing techniques could give a more accurate idea of the feasibility of the Sparknotes quizzes. Lastly, aligning the Sparknotes chapter summaries with the Project Gutenberg novels has to be done manually, but it would allow for a more reasonable task than trying to answer questions on the full novel and it is worth doing so as to allow for evaluating models on the Project Gutenberg data.

Appendix A

All the code used for experiments can be found here:
https://github.com/ymann/sparknotesqa

Appendix B

List of works in both Project Gutenberg and Sparknotes

Note: The marked books are still being processed by sparknotes (they are on the website, but incomplete) and thus do not have sparknotes summary statistics.

Title	Number of words in Gutenberg novel	Average number of words in Sparknotes summary	Number of questions in Sparknotes quiz
The Three Musketeers	207530	642.9	65
Adam Bede	196601	619	65
The Aeneid	176777	552.3	60
Agamemnon	20757	309.1	40
The Age of Innocence	94858	519.1	55
The Ambassadors	147793	N/A	N/A
The American	121949	N/A	N/A
Anna Karenina	317922	N/A	N/A
Anne Of Green Gables	94351	942.5	50
Anthem	17741	N/A	N/A
Arms and the Man	23137	N/A	N/A
The Awakening and Selected Short Stories	58312	987.6	40
Babbitt	111536	491.5	60
Beowulf	22740	396.1	50
Bleak House	322146	1484.8	65
Madame Bovary	105244	414.3	60
The Brothers Karamazov	322935	683.6	80
The Call of the Wild	29143	551.4	35
The Mayor of Casterbridge	104847	600.2	55
A Christmas Carol	25504	434.6	25

Common Sense	21032	N/A	N/A
David Copperfield	325648	660.6	70
Crime and Punishment	185649	571.8	65
Robinson Crusoe	114117	667.3	35
Cyrano de Bergerac by Edmond Rostand	37200	N/A	N/A
Daisy Miller	19356	422.8	25
Don Quixote	391784	860.6	90
The Picture of Dorian Gray	76582	499.2	50
Dracula	154817	679.4	50
Dubliners	61984	474.8	75
The Importance of Being Earnest	20740	578.2	30
Emma	144229	486.1	89
An Enemy of the People	28531	700.2	25
White Fang	69383	413	50
Frankenstein	70424	399.5	55
The Autobiography of Benjamin Franklin	62594	711.7	35
The Good Soldier	69041	779.6	45
Great Expectations	168536	447.4	70
Hard Times	99559	430.1	45
Heart of Darkness	34820	450.8	30
Hedda Gabler	26437	399.4	40
Herland	47331	N/A	N/A
Casanova's Homecoming	34646	658.7	24
Howards End	103066	477	50
The Adventures of Huckleberry Finn	101966	662.2	70
The Idiot	220138	N/A	N/A
The Iliad	140704	498.3	65
Incidents in the Life of a Slave Girl	77476	N/A	N/A

Iola Leroy	68409	N/A	N/A
Ivanhoe	176996	419.2	50
Jane Eyre	174503	705.9	50
Jude the Obscure	136237	698	30
The Jungle	137257	665	50
Kidnapped	75879	564.1	28
Little Women	175991	775.7	45
Looking Backwards from 2000 to 1887	71089	646	20
Lord Jim	119683	656.2	60
Far from the Madding Crowd	123666	N/A	N/A
Maggie: A Girl of the Streets	21775	396.8	20
Main Street and Other Poems	7888	689.1	55
Mansfield Park	146271	607.4	60
Middlemarch	301410	554.9	80
The Mill on the Floss	188935	781.8	90
A Modest Proposal	3116	425	20
The Last of the Mohicans	136669	622.3	40
The Count of Monte Cristo	418083	664.7	94
The Moonstone	177765	732.6	90
Northanger Abbey	70538	449	75
The Odyssey	119379	601.1	60
Oliver Twist	147727	542.3	55
O Pioneers!	50531	343.8	30
In Our Town	66890	810.2	20
This Side of Paradise	75962	384.9	50
The Pearl Box	29606	757.2	30
Persuasion	78496	676.2	60
A Journal of the Plague Year	94276	456	40

A Portrait of the Artist as a Young Man	77766	559.5	50
The Power and the Glory	83085	545.5	60
Pride and Prejudice	111016	494.1	59
The Red Badge of Courage	44302	425.2	45
The Red and the Black	182186	454	45
Regeneration	57318	816.9	55
The Return of the Native	129858	556.2	40
A Room With A View	60278	N/A	N/A
The Scarlet Letter	78066	477.2	65
The Turn of the Screw	38602	532.6	40
The Secret Garden	73560	587.8	85
Sense and Sensibility	111591	0	0
The House of the Seven Gables	98802	697.5	55
Siddhartha	37116	407.5	10
Silas Marner	66428	806.4	45
Sister Carrie	141933	628.7	50
Sons and Lovers	144774	527.1	75
The Stranger	20388	519	35
Swann's Way	179349	N/A	N/A
The Alchemist	50165	984.1	60
Through the Looking-Glass	28749	426.8	45
The Time Machine	29726	489.3	30
Treasure Island	64118	591.9	45
Essay on the Trial By Jury	87051	525.2	50
A Tale of Two Cities	124041	585.7	60
Typee	103533	725.1	50
War and Peace	514170	1305.5	50
The Yellow Wallpaper	5523	N/A	N/A

All's Well That Ends Well	21084	N/A	N/A
Antony and Cleopatra	21381	N/A	N/A
As You Like It	19978	485.5	50
Coriolanus	23935	442.6	35
Cymbeline	24653	N/A	N/A
The Comedy of Errors	13834	412.7	30
Hamlet	28910	439.9	75
Henry IV Part 2	22719	477.7	22
Henry VI Part 1	21613	550.8	130
Henry VI Part 2	25575	592.6	77
Henry VI Part 3	24651	502.8	79
King Henry VIII	21838	451.9	55
King John	18417	588.5	40
Love's Labour's Lost	19267	318.7	35
King Lear	23560	418.7	55
Macbeth	15325	555.1	40
Measure for Measure	19037	528.1	50
The Merchant of Venice	19366	498.1	50
The Merry Wives of Windsor	20119	564.4	45
Much Ado About Nothing	19335	473	50
Othello	23423	573	45
Pericles	16800	340	60
King Richard III	25990	469.4	70
Romeo and Juliet	22060	345.6	80
The Taming of the Shrew	18631	401	45
The Tempest	14721	369.1	50
Timon of Athens	16318	538.7	50
The Tragedy of Titus Andronicus	18171	292	40

Troilus and Cressida	23241	N/A	N/A
Twelfth Night	18488	497.5	90
The Two Gentlemen of Verona	18729	310.2	55
The Winter's Tale	30483	426	35

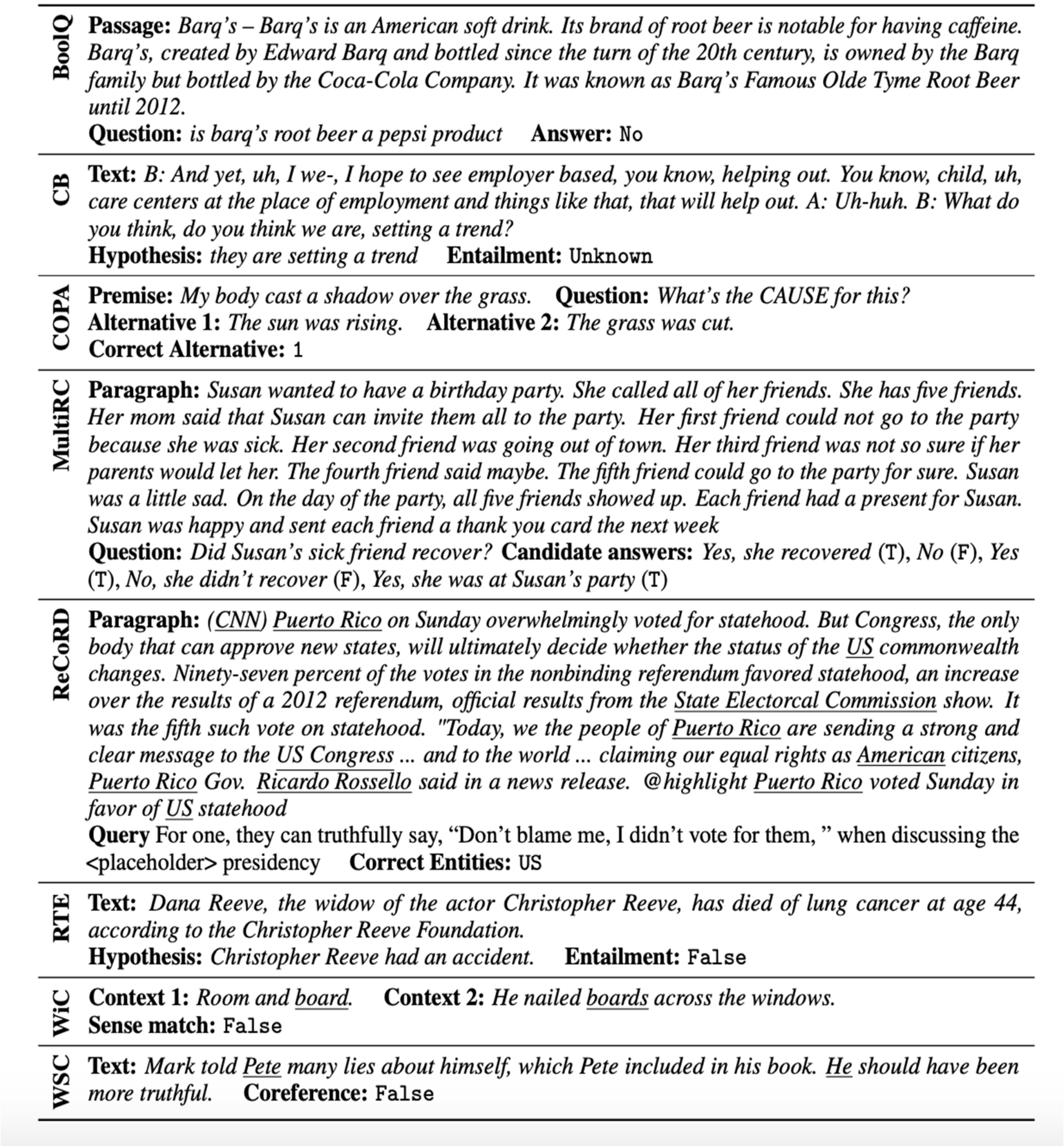
Appendix C

Definitions of tasks in superGLUE datasets

- **BoolQ**: A QA task where each example is a passage together with a yes/ no question.
- **CB:** A three class textual entailment task where each premise contains an embedded clause that is annotated with the degree to which the author believes the clause is true. The hypothesis is then the extractions of that clause.
- **COPA:** COPA is a causal reasoning task where an example is a premise sentence and two possible choices. The system must determine the cause or effect of the premise from two possible choices.
- **MultiRC:** MultiRC is a QA task where each example is a paragraph, a question and possible answers where there can be multiple correct answers.
- **ReCoRD:** A multiple choice QA task where question is a cloze-style fill in the blank and a specific entity is masked out.
- **RTE:** A textual entailment task where each example is a premise and a hypothesis and the system needs to determine if the hypothesis is entailed by the premise.
- **WiC:** Given two sentences and a word that appears in both, the task is to identify if the word is used in the same sense in both sentences.

- **WSC:** Given a sentence with a pronoun and a list of nouns, the system must determine who the pronoun is referring to.

Examples of tasks from SuperGLUE datasets



BoolQ	**Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.* **Question:** *is barq's root beer a pepsi product* **Answer:** `No`
CB	**Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** `Unknown`
COPA	**Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?* **Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** `1`
MultiRC	**Paragraph:** *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week* **Question:** *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (`T`), *No* (`F`), *Yes* (`T`), *No, she didn't recover* (`F`), *Yes, she was at Susan's party* (`T`)
ReCoRD	**Paragraph:** *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood* **Query** For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** `US`
RTE	**Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.* **Hypothesis:** *Christopher Reeve had an accident.* **Entailment:** `False`
WiC	**Context 1:** *Room and board.* **Context 2:** *He nailed boards across the windows.* **Sense match:** `False`
WSC	**Text:** *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** `False`

Appendix D

Coreference Resolution:

Question	Answers before coreference resolution	Answers after coreference resolution
Which statement best reflects Julia’s views toward rebellion?	Like Winston, she believes that rebellion will come from the proles. Unlike Winston, she has no interest in rebellion. She believes that an awareness of sex will lead to mass rebellion within the Party. She believes that the Party is actually a good thing.	Like Winston, Julia believes that rebellion will come from the proles. Unlike Winston, Julia has no interest in rebellion. Julia believes that an awareness of sex will lead to mass rebellion within the Party. Julia believes that the Party is actually a good thing.
What does Winston tell Julia about his wife in Chapter 3?	That he once thought about pushing her over a cliff, That he truly loved her That he feels responsible for her death at the hands of the Party That they were never really married	That Winston once thought about pushing Julia over a cliff That Winston truly loved Julia That Winston feels responsible for Julia That Winston and his wife were never really married.
What has prevented Julia and Winston from meeting more frequently?	She has started to see another man. Winston’s superiors at the ministry have become suspicious of his activities. They have been busy with preparations for Hate Week. Winston is unsure of his feelings toward Julia.	Julia has started to see another man. Winston’s superiors at the ministry have become suspicious of his activities. Winston and Julia have been busy with preparations for Hate Week. Winston is unsure of his feelings toward Julia.
Why does Kitty help D’Artagnan intercept Milady’s correspondence with the Comte de Wardes?	She loves D’Artagnan She hates Milady She hates the Cardinal She likes all intrigue	Kitty loves D’Artagnan Kitty hates Milady Kitty hates the Cardinal Kitty likes all intrigue
What remains ambiguous to D’Artagnan after his brief meeting with Madame Bonacieux on the bridge?	Why he loves her Whether she loves him Whether she is safe What side she supports	Why D’Artagnan loves Madame Bonacieux Whether Madame Bonacieux loves D’Artagnan Whether Madame Bonacieux is safe What side Madame Bonacieux supports
What does Podrick reveal to Tyrion that he did to Mandon?	Exiled him Drowned him Hanged him Tortured him	Exiled Mandon Drowned Mandon Hanged Mandon Tortured Mandon
What does Luwin ask Osha to do for him in the godswood?	End his misery Tend to his wounds Avenge his death Succeed his as maester	End Luwin's misery Tend to Luwin's wounds Avenge Luwin's death Succeed Luwin's as maester
What advantage does Daenerys plan to use in order to retake her family’s throne?	Her enormous army Her illustrious name Her dragons Her sway over men	Daenerys enormous army Daenerys illustrious name Daenerys dragons Daenerys sway over men
What day is Hetty’s execution scheduled for?	Her wedding date Easter Sunday Her birthday Christmas Eve	Hetty’s wedding date Easter Sunday Hetty’s birthday Christmas Eve

When Dinah is returning to Snowfield, how does she react when Adam says that he knows she will always do what is right?	She laughs She prays She cries She kisses him	Dinah laughs Dinah prays Dinah cries Dinah kisses him
What happens when Robert reaches down Apron's throat to pull out the object that's suffocating her?	He successfully removes it He pushes it deeper He loses his arm Apron drags him around	Robert successfully removes the object Robert pushes it deeper Robert loses his arm Apron drags Robert around
What does Robert want Ira and Haven to do with the badly wounded Hussy?	Abandon her Tend to her wounds Put her up for adoption Kill her	Abandon Hussy Tend to Hussy's wounds Put Hussy up for adoption Kill Hussy
According to Haven, what should Pinky have had long ago?	Her first heat Her spaying Her first birthday party Her vaccines	Pinky's first heat Pinky's spaying Pinky's first birthday party Pinky's vaccines
What does Haven do when Mr. Tanner offers Robert a piglet?	Nothing Takes the piglet himself Thanks him profusely Tries to decline the piglet	Nothing Takes the piglet himself Thanks Mr. Tanner profusely Tries to decline the piglet
What are the Pecks rich in, according to Haven?	Love Livestock Money What they need	Love Livestock Money What the Pecks need
What does Robert make the mistake of showing off after returning home from his last day of school?	His final essay His scraped knee His report card A candy he stole	Robert's final essay Robert's scraped knee Robert's report card A candy Robert stole
What "remedy" does Aunt Matty propose for Robert's "D" in English?	Pulling him out of school Tutoring him personally Buying him a notebook Confiscating his bicycle	Pulling Robert out of school Tutoring Robert personally Buying Robert a notebook Confiscating Robert's bicycle
What does Aeneas do when he learns of Pallas's death?	He breaks down in tears and begs Jupiter to stop the battle. He goes into a murderous rage and cuts down many of the Latins. He challenges Turnus to single combat to avenge his friend. He calls on Venus to turn the Latin soldiers into piles of dust.	Aeneas breaks down in tears and begs Jupiter to stop the battle. Aeneas goes into a murderous rage and cuts down many of the Latins. Aeneas challenges Turnus to single combat to avenge his friend. Aeneas calls on Venus to turn the Latin soldiers into piles of dust.
What wound does Aeneas suffer as the battle begins?	He is stabbed in the arm by a spear. He is shot in the leg with an arrow. He is sliced across the chest by a sword. He is hit in the face by a club.	Aeneas is stabbed in the arm by a spear. Aeneas is shot in the leg with an arrow. Aeneas is sliced across the chest by a sword. Aeneas is hit in the face by a club.
What mistake does Aeneas's father make regarding Apollo's command?	He believes that Apollo is Juno in disguise. He assumes Apollo is telling them to found a city in Greece. He thinks they are supposed to go to Crete. He interprets Apollo's orders as applying to him, not Aeneas.	Aeneas's father believes that Apollo is Juno in disguise. Aeneas's father assumes Apollo is telling them to found a city in Greece. Aeneas's father thinks they are supposed to go to Crete. Aeneas's father interprets Apollo's orders as applying to him, not Aeneas.

What prophecy does the harpy make to Aeneas and his men?	They will not found their city until they are so hungry that they eat their `s. They will not find a place to rest until they have killed the traitor in their midst. They will not make it to Italy until seven of them are sacrificed to the gods. They will not start their empire until they have suffered at Dido's hands	Aeneas and his men will not found their city until they are so hungry that they eat their tables. Aeneas and his men will not find a place to rest until they have killed the traitor in their midst. Aeneas and his men will not make it to Italy until seven of them are sacrificed to the gods. Aeneas and his men will not start their empire until they have suffered at Dido's hands
Why does Dido have mixed feelings about her love for Aeneas?	She knows that he wants to take over her kingdom. She fears that the gods will take him away. She has sworn to never marry again. She is also in love with Anchises.	Dido knows that he wants to take over her kingdom. Dido fears that the gods will take him away. Dido has sworn to never marry again. Dido is also in love with Anchises.
Why does Juno want Aeneas and Dido to get together?	She promised Dido a great love. She hopes to bring peace to the region. She wants to cause the fall of Carthage. She thinks it will hinder Aeneas's quest.	Juno promised Dido a great love. Juno hopes to bring peace to the region. Juno wants to cause the fall of Carthage. Juno thinks it will hinder Aeneas's quest.
What rumors begin to spread about Aeneas and Dido?	They have given up their duties for lust. They are plotting to attack Italy. They have teamed up to invade Greece. They are actually brother and sister.	Aeneas and Dido have given up their duties for lust. Aeneas and Dido are plotting to attack Italy. Aeneas and Dido have teamed up to invade Greece. Aeneas and Dido are actually brother and sister.
Whom does Aeneas leave behind in Sicily?	The men in his band who have fallen in love with Sicilian women, The Trojan women who joined in the burning of the ships "The oldest, weakest, and most unwilling members of his party”, "The fastest, strongest, and most daring members of his group"	The men in Aeneas' band who have fallen in love with Sicilian women, The Trojan women who joined in the burning of the ships "The oldest, weakest, and most unwilling members of Aeneas' party”, "The fastest, strongest, and most daring members of Aeneas' group"
How does Aeneas find the tree with the golden bough?	A golden ram shows him the way., A pair of doves lead him to it., He follows a trickle of blood., He cuts down all of the other trees.	A golden ram shows Aeneas the way., A pair of doves lead Aeneas to it., Aeneas follows a trickle of blood., Aeneas cuts down all of the other trees.
What is the significance of the Trojans' meal of bread and fruit?	They do not offer any to the poor, angering Jupiter. They eat their tables, as the harpy predicted. They consume parts of Carthage in Italy, as Apollo commanded. They pluck the fruit from the trees of Latinus, waking the hydra.	the Trojans do not offer any to the poor, angering Jupiter. the Trojans eat their tables, as the harpy predicted. the Trojans consume parts of Carthage in Italy, as Apollo commanded. the Trojans pluck the fruit from the trees of Latinus, waking the hydra.
What does Latinus offer Aeneas in addition to land?	His daughter His orchards His army His treasury	Latinus' daughter Latinus' orchards Latinus' army Latinus' treasury
How does Ascanius inadvertently start the war?	He seduces a woman who turns out to be Turnus's sister in disguise. He cuts down a tree that Latinus had set aside as sacred. He shoots a favorite stag of Latinus's herdsman. He fishes in Tibernius's sacred river	Ascanius seduces a woman who turns out to be Turnus's sister in disguise. Ascanius cuts down a tree that Latinus had set aside as sacred. Ascanius shoots a favorite stag of Latinus's herdsman. Ascanius fishes in Tibernius's sacred river

What aspect of the narrator's writing process fascinates the native boys?	His fountainpen, His roll-top desk, His typewriter, His leather notebooks	the narrator's fountainpen, the narrator's roll-top desk, the narrator's typewriter, the narrator's leather notebooks
When Lulu doesn't visit for over a week, what does Farah explain has happened to her?	Hunters shot her, She migrated south She broke her leg She got married	Hunters shot Lulu, Lulu migrated south Lulu broke her leg Lulu got married
What does Jogona do with the account of his testimony that the narrator types up?	He buries it He carries it in a pouch Frames it He puts it under his pillow	Jogona buries the account of his testimony Jogona carries the account of his testimony in a pouch Frames the account of his testimony Jogona puts the account of his testimony under his pillow
What do the Somali women find shocking about European marriages?	Spouses share a bed They're are held publicly There's no bride price They require a minister	Spouses share a bed European marriages are held publicly There's no bride price in European marriages European marriages require a minister
What does Agamemnon rebuke his wife for?	Spending too much money Ordering the carpet Cheating on him Sending their son away	Spending too much money Ordering the carpet Cheating on Agamemnon Sending Agamemnon and his wife's son away
According to gossip, Ellen Olenska left her husband because he was ___.	Gay An alcoholic A gambler Unfaithful to her	Gay An alcoholic A gambler Unfaithful to Ellen Olenska
Why does Archer persuade May to leave the opera early?	He wants to confess He feels ill He hates the music He sees Ellen	Archer wants to confess Archer feels ill Archer hates the music Archer sees Ellen
Later in Archer's life, who surprises him by asking if he was ever in love with Ellen Olenska?	His son His sister His servant Ned Winsett	Archer's son Archer's sister Archer's servant Ned Winsett
Why does Gregor's mother pass out when she enters Gregor's room and sees him?	He's eating garbage He's transformed back He's on the wall He hugs her	Gregor's eating garbage Gregor's transformed back Gregor's on the wall Gregor hugs her
Why is Gregor's mobility limited at the beginning of Part 3?	He's in a cage He's in the closet He's injured He's drunk	Gregor's in a cage Gregor's in the closet Gregor's injured Gregor's drunk
When George tries not to let Lennie talk to the boss, why is the boss suspicious?	He worries that George might be taking advantage of Lennie. Because Lennie seems violent and crazy He thinks Lennie is making fun of him. He suspects George and Lennie are planning to rob him	the boss worries that George might be taking advantage of Lennie. Because Lennie seems violent and crazy the boss thinks Lennie is making fun of him. the boss suspects George and Lennie are planning to rob him
With an enormous debt hanging over his head, Lydgate tells Rosamond that they must ___.	Move somewhere smaller Divorce Sell her jewelry Borrow from her father	Move somewhere smaller Divorce Sell Rosamond's jewelry Borrow from Rosamond's father

In Dacca, what does Parvati ask Shiva for?	Directions A bag of marbles His birth certificate A lock of hair	Directions A bag of marbles Shiva's birth certificate A lock of hair
Who does Saleem claim is speaking to him telepathically?	His birthmother Mary Angels Brass Monkey	Saleem's birthmother Mary Angels Brass Monkey
When James attacks Leonard, what does he grab him by?	His torso, His Adam's apple His hair His elbow	Leonard's torso, Leonard's Adam's apple Leonard's hair Leonard's elbow
Who does James sit with at dinner after meeting with his parents?	"Miles, Lilly", No one, "Hank, Leonard", His parents	"Miles, Lilly", No one, "Hank, Leonard", James' parents
What does Philip tell Maggie that he likes, prompting her to kiss him?	Her smile Her hair, Her eyes, Her laugh	Maggie's smile Maggie's hair, Maggie's eyes, Maggie's laugh
What did Mr. Tulliver use as collateral for the loan he took from one of Wakem's clients?	The mill The family furniture His horse The family cottage	The mill The family furniture Mr. Tulliver's horse The family cottage
What is exceptional about the slave Rider?	His social status, His incredible wealth, His great intelligence, His size and strength	Rider's social status, Rider's incredible wealth, Rider's great intelligence, Rider's size and strength
Why is Rider incredibly depressed and moody throughout the story?	He broke his arm, His aunt abandoned him, His young wife died, His owners are morally bankrupt	Rider broke his arm, Rider's aunt abandoned him, Rider's young wife died, Rider's owners are morally bankrupt
What does Rider do when he is put in jail?	Digs an escape hole Resignedly sobs Rips cell door off Kills sheriff and deputies	Rider Digs an escape hole Rider Resignedly sobs Rider Rips cell door off Rider Kills sheriff and deputies
Why is hunting Old Ben such a challenge at the beginning of the story?	He is amazingly fast, He is incredibly stealthy, The men fear him, The hounds fear him	Old Ben is amazingly fast, Old Ben is incredibly stealthy, The men fear Old Ben, The hounds fear Old Ben
How does Boon Hogganbeck eventually kill Old Ben?	Slits his throat Traps him Shoots him Tricks him	Slits Old Ben's throat Traps Old Ben Shoots Old Ben Tricks Old Ben
In reference to what does Boon Hogganbeck yell, "They're mine!" at the end of the story?	The squirrels surrounding him Guns that he's is fixing Gold coins he found His newborn sons	The squirrels surrounding Boon Hogganbeck Guns that Boon Hogganbeck is fixing Gold coins Boon Hogganbeck found Boon Hogganbeck's newborn sons
For what purpose does Carothers give Isaac an envelope?	Apologize for killing doe Reveal his secret identity Pay off Carothers's lover Threaten him discreetly	Apologize for killing doe Reveal Isaac's secret identity Pay off Carothers's lover Threaten Isaac discreetly

How does Isaac react to Carothers giving him the envelope?	He loudly berates him He is silently disappointed He seems completely indifferent He becomes overjoyed	Isaac loudly berates him Isaac is silently disappointed Isaac seems completely indifferent Isaac becomes overjoyed
What does Mollie Beauchamp repeatedly say Carothers Edmonds has done to her son?	Locked him in prison Framed him for murder Sold him to Egypt Beat him viciously	Locked Mollie's son in prison Framed Mollie's son for murder Sold Mollie's son to Egypt Beat Mollie's son viciously
According to Gavin Stevens, why will Beauchamp's son be executed?	He stole money He killed a police officer He ran from the plantation He murdered Carothers Edmonds	Beauchamp's son stole money Beauchamp's son killed a police officer Beauchamp's son ran from the plantation Beauchamp's son murdered Carothers Edmonds
What happens when Sara begins to question Brian on the stand?	"She notices new traits, which she doesn't like or respect." "She begins, then refuses to continue, evoking judicial precedent." She notices all of the qualities that made her fall in love with Brian. She thinks that he has grown less handsome and intelligent over the years	"Sara notices new traits, which she doesn't like or respect." "Sara begins, then refuses to continue, evoking judicial precedent." Sara notices all of the qualities that made her fall in love with Brian. Sara thinks that Brian has grown less handsome and intelligent over the years
What does Sara finally ask Brian?	Why he favors Anna Why he spends so much time looking at the night sky When he is coming home When they can get divorced	Why Brian favors Anna Why Brian spends so much time looking at the night sky When Brian is coming home When Anna and Brian can get divorced
When Sara tells Brian that they're going to lose "her," to which daughter is she referring?	She is referring to both daughters. She isn't sure which daughter she means. Anna Kate	Sara is referring to both daughters. Sara isn't sure which daughter she means. Anna Kate
How does Julia compare Campbell to Anna?	She thinks they're both survivors of troubled families. She thinks they're both stubborn and selfish. She thinks they're both cowards. She thinks they're both impudent and hard-hearted	Julia thinks Campbell and Anna are both survivors of troubled families. Julia thinks Campbell and Anna are both stubborn and selfish. Julia thinks Campbell and Anna are both cowards. Julia thinks Campbell and Anna are both impudent and hard-hearted
How does Campbell respond when his dog begins acting up?	He ignores the dog He gives the dog a biscuit. He asks an officer of the court to take the dog outside. He surreptitiously kicks the dog.,	Campbell ignores the dog Campbell gives the dog a biscuit. Campbell asks an officer of the court to take the dog outside. Campbell surreptitiously kicks the dog.,
Why did Campbell break up with Julia when they were in high school?	Because he was ashamed of his seizures Because they got into a car accident Because he didn't want her to have to deal with his condition Because he didn't really love her	Because Campbell was ashamed of his seizures Because Campbell and Julia got into a car accident Because Campbell didn't want Julia to have to deal with his condition Because Campbell didn't really love Julia

How does Anna respond when Campbell asks if she's willing to take an action that could kill Kate?	She breaks down, tearing her hair and wailing. She says she's willing, even though she knows she""ll feel guilty. She says she's willing, because she knows Kate wants her to. She breaks down, saying that she only wanted to punish Sara.	Anna breaks down, tearing her hair and wailing. Anna says she's willing, even though she knows she""ll feel guilty. Anna says she's willing, because she knows Kate wants her to. Anna breaks down, saying that she only wanted to punish Sara.
Why does Jesse want to get struck by lightning?	To commit suicide To feel alive To make his parents sad To get attention	To commit suicide To feel alive To make Jesse's parents sad To get attention
What does Brian remember about the night of Anna's birth?	That he saw Andromeda That he counted four shooting stars That there were no stars That he and Sara wished for a boy	That Brian saw Andromeda That Brian counted four shooting stars That there were no stars That Brian and Sara wished for a boy
While pregnant, how does Sara think of Anna, her unborn daughter?	As a ray of light and source of joy As a constant source of pleasure As a means of holding her family together As a way to save Kate	As a ray of light and source of joy As a constant source of pleasure As a means of holding Sara's family together As a way to save Kate
Why does Kate clean her room before leaving the house?	In case Anna wants to sleep there In case Sara has guests To prevent her parents from snooping In case she doesn't come back	In case Anna wants to sleep there In case Sara has guests To prevent Kate's parents from snooping In case Kate doesn't come back
Why does Jesse walk into traffic?	To get noticed To see if he’s really invisible To anger his parents To kill himself	To get noticed To see if Jesse's really invisible To anger Jesse's parents To kill himself
Why does Zeus kill Apollo’s son, Aesculapius?	He stole from Zeus He revived the dead He betrayed Apollo He insulted Zeus	Aesculapius stole from Zeus Aesculapius revived the dead Aesculapius betrayed Apollo Aesculapius insulted Zeus
What was Erysichthon’s punishment for cutting down Ceres’ sacred giant oak tree?	He is always starving He cannot die He loses his sight He cannot move	Erysichthon is always starving Erysichthon cannot die Erysichthon loses his sight Erysichthon cannot move
Every winter, Demeter is filled with sorrow: why?	Cold weather Her husband leaves Persephone joins her Persephone goes to Hades	Cold weather Demeter’s husband leaves Persephone joins Demeter Persephone goes to Hades
How does Hercules die?	The Labors of Hercules Hades kills him Zeus kills him He kills himself	The Labors of Hercules Hades kills Hercules Zeus kills Hercules Hercules kills himself
Where does Zeus’s allegiance lie in the Trojan War?	He is neutral With the Greeks With fate With the Trojans	Zeus is neutral With the Greeks With fate With the Trojans

Appendix E

RTE Test Set:

Premise	Hypothesis
The one incongruity in the otherwise drab scene is the rosebush that grows next to the prison door. The narrator suggests that it offers a reminder of Nature’s kindness to the condemned; for his tale, he says, it will provide either a “sweet moral blossom” or else some relief in the face of unrelenting sorrow and gloom.	the rose outside the prison symbolizes, according to the narrator, The frailty of mankind
The one incongruity in the otherwise drab scene is the rosebush that grows next to the prison door. The narrator suggests that it offers a reminder of Nature’s kindness to the condemned; for his tale, he says, it will provide either a “sweet moral blossom” or else some relief in the face of unrelenting sorrow and gloom.	the rose outside the prison symbolizes, according to the narrator, A sweet moral blossom found within the Hester's story
The one incongruity in the otherwise drab scene is the rosebush that grows next to the prison door. The narrator suggests that it offers a reminder of Nature’s kindness to the condemned; for his tale, he says, it will provide either a “sweet moral blossom” or else some relief in the face of unrelenting sorrow and gloom.	the rose outside the prison symbolizes, according to the narrator, The reader's sympathy for Hester's ordeal
The one incongruity in the otherwise drab scene is the rosebush that grows next to the prison door. The narrator suggests that it offers a reminder of Nature’s kindness to the condemned; for his tale, he says, it will provide either a “sweet moral blossom” or else some relief in the face of unrelenting sorrow and gloom.	the rose outside the prison symbolizes, according to the narrator, The beautiful child that resulted from Hester's affair
As the crowd watches, Hester Prynne, a young woman holding an infant, emerges from the prison door and makes her way to a scaffold (a raised platform), where she is to be publicly condemned. The women in the crowd make disparaging comments about Hester; they particularly criticize her for the ornateness of the embroidered badge on her chest—a letter “A” stitched in gold and scarlet. From the women’s conversation and Hester’s reminiscences as she walks through the crowd, we can deduce that she has committed adultery and has borne an illegitimate child, and that the “A” on her dress stands for “Adulterer.”	the women in the crowd criticize Hester's scarlet letter because Its ornate design is inappropriate for a symbol of punishment.
As the crowd watches, Hester Prynne, a young woman holding an infant, emerges from the prison door and makes her way to a scaffold (a raised platform), where she is to be publicly condemned. The women in the crowd make disparaging comments about Hester; they particularly criticize her for the ornateness of the embroidered badge on her chest—a letter “A” stitched in gold and scarlet. From the women’s conversation and Hester’s reminiscences as she walks through the crowd, we can deduce that she has committed adultery and has borne an illegitimate child, and that the “A” on her dress stands for “Adulterer.”	the women in the crowd criticize Hester's scarlet letter because The fine quality fabric should have been saved for a better purpose.

As the crowd watches, Hester Prynne, a young woman holding an infant, emerges from the prison door and makes her way to a scaffold (a raised platform), where she is to be publicly condemned. The women in the crowd make disparaging comments about Hester; they particularly criticize her for the ornateness of the embroidered badge on her chest—a letter “A” stitched in gold and scarlet. From the women’s conversation and Hester’s reminiscences as she walks through the crowd, we can deduce that she has committed adultery and has borne an illegitimate child, and that the “A” on her dress stands for “Adulterer.”	the women in the crowd criticize Hester's scarlet letter because Its crude design is unbecoming and sloppy.
As the crowd watches, Hester Prynne, a young woman holding an infant, emerges from the prison door and makes her way to a scaffold (a raised platform), where she is to be publicly condemned. The women in the crowd make disparaging comments about Hester; they particularly criticize her for the ornateness of the embroidered badge on her chest—a letter “A” stitched in gold and scarlet. From the women’s conversation and Hester’s reminiscences as she walks through the crowd, we can deduce that she has committed adultery and has borne an illegitimate child, and that the “A” on her dress stands for “Adulterer.”	the women in the crowd criticize Hester's scarlet letter because It is misspelled.
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester's infant cries out while she's standing on the scaffold because she is frightened by the shouting of the angry crowd.
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester's infant cries out while she's standing on the scaffold because she's hungry.
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester's infant to cries out while she's standing on the scaffold because a piece of rotten fruit thrown at Hester accidentally hits her.
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester's infant cries out while she's standing on the scaffold because Hester squeezes her too tightly

The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester reflects on her childhood home in England while standing before the crowd
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester reflects on friends from her youth while standing before the crowd
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester reflects on the night of her affair while standing before the crowd
The beadle calls Hester forth. Children taunt her and adults stare. Scenes from Hester’s earlier life flash through her mind: she sees her parents standing before their home in rural England, then she sees a “misshapen” scholar, much older than herself, whom she married and followed to continental Europe. But now the present floods in upon her, and she inadvertently squeezes the infant in her arms, causing it to cry out. She regards her current fate with disbelief.	Hester reflects on her grandfather's farm while standing before the crowd

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Seohyun Back, Seunghak Yu, Sathish Reddy Indurthi, Jihie Kim, and Jaegul Choo. 2018. Memoreader: Large-scale reading comprehension through neural memory controller. In *Proceedings of EMNLP*.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*.

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 the 2018 Conference on Empirical Methods in Natural Language Processing,* pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. From 'F' to 'A' on the NY Regents science exams: An overview of the Aristo project. *ArXiv, abs/1909.01958*, 2019.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 9-11, 2017, pp. 681–691, 2017.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. Recognizing Textual Entailment: Models and Applications. In *Synthesis Lectures on Human Language Technologies,* 2013, 6:4, 1-220.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*.

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 the 28th International Conference on Neural Information Processing Systems, NIPS'15*, Cambridge, MA, USA.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from realworld applications. In *Proceedings of the Workshop on Machine Reading for Question Answering,* pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the International Conference on Learning Representations*.

Sathish Reddy Indurthi, Seunghak Yu, Seohyun Back, and Heriberto Cuayahuitl. 2018. ´ Cut to the chase: A context zoom-in network for reading comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 570–575, Brussels, Belgium. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

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

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, ´ and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, abs/1712.07040.

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Largescale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv, abs/1907.11692*, 2019.

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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

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 Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2230–2235. https://aclweb.org/anthology/D16- 1241.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openaiassets/research-covers/languageunsupervised/language understanding paper*. pdf, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. CoQA: A conversational question answering challenge. *Transactions of the Association of Computational Linguistics.*

Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

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 the 2013*

Conference on Empirical Methods in Natural Language Processing, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018a. Multi-granular sequence encoding via dilated compositional units for reading comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2141– 2151, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537, 2019*.

Mengqiu Wang and Christopher D Manning. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of COLING*, 2010.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive 76 pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.