Decoding Sherlock Holmes: A Comparative Analysis of BERT Models' Performance

Team 2

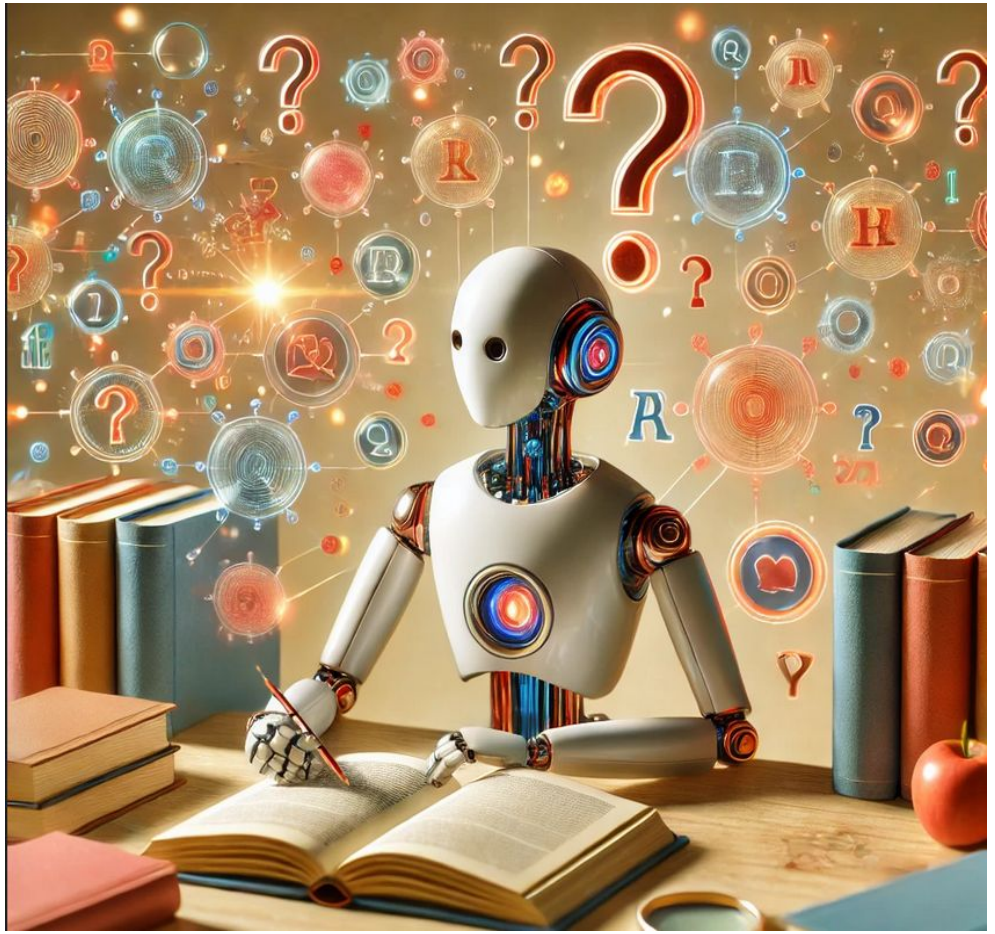
Ria Patel, Devanshi Patel, Sai Dasari, Amy An, Jihun Kim

Motivation

- Addressing the unique challenges of literary texts:
  - Rich and complex contexts.
  - Subtle relationships between characters and events.

- Evaluating current models to identify limitations in understanding and answering questions from literary narratives.

- Literature holds deep, complex stories that require careful understanding—can machines keep up?



Introduction

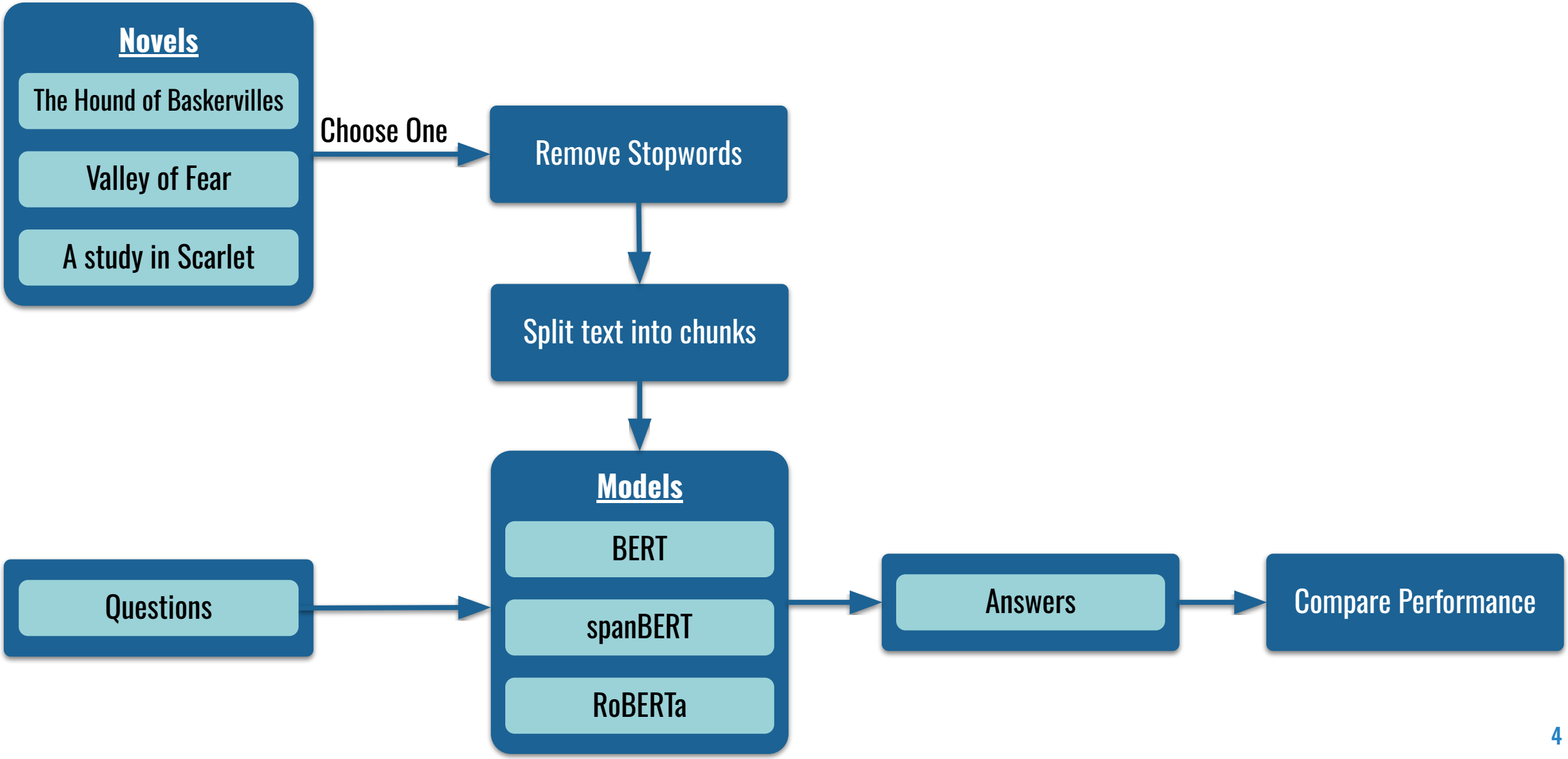
Objective

- Compare performance of BERT-based models for question-answering tasks focusing on plot analysis of literary texts.
- Analyze how well these models understand and interpret complex narrative contexts.

Why BERT ?

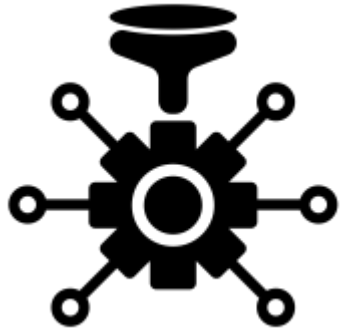
- Handles context-sensitive language understanding.
- Excels in tasks like question answering, making it ideal for analyzing complex plots.

Workflow



Data Preparation and Data Engineering

- **Source:** Text extracted from Project Gutenberg.
  - 3 Books of Arthur Conan Doyle:
    - The Hound of Baskervilles
    - Valley of Fear
    - A Study in Scarlet

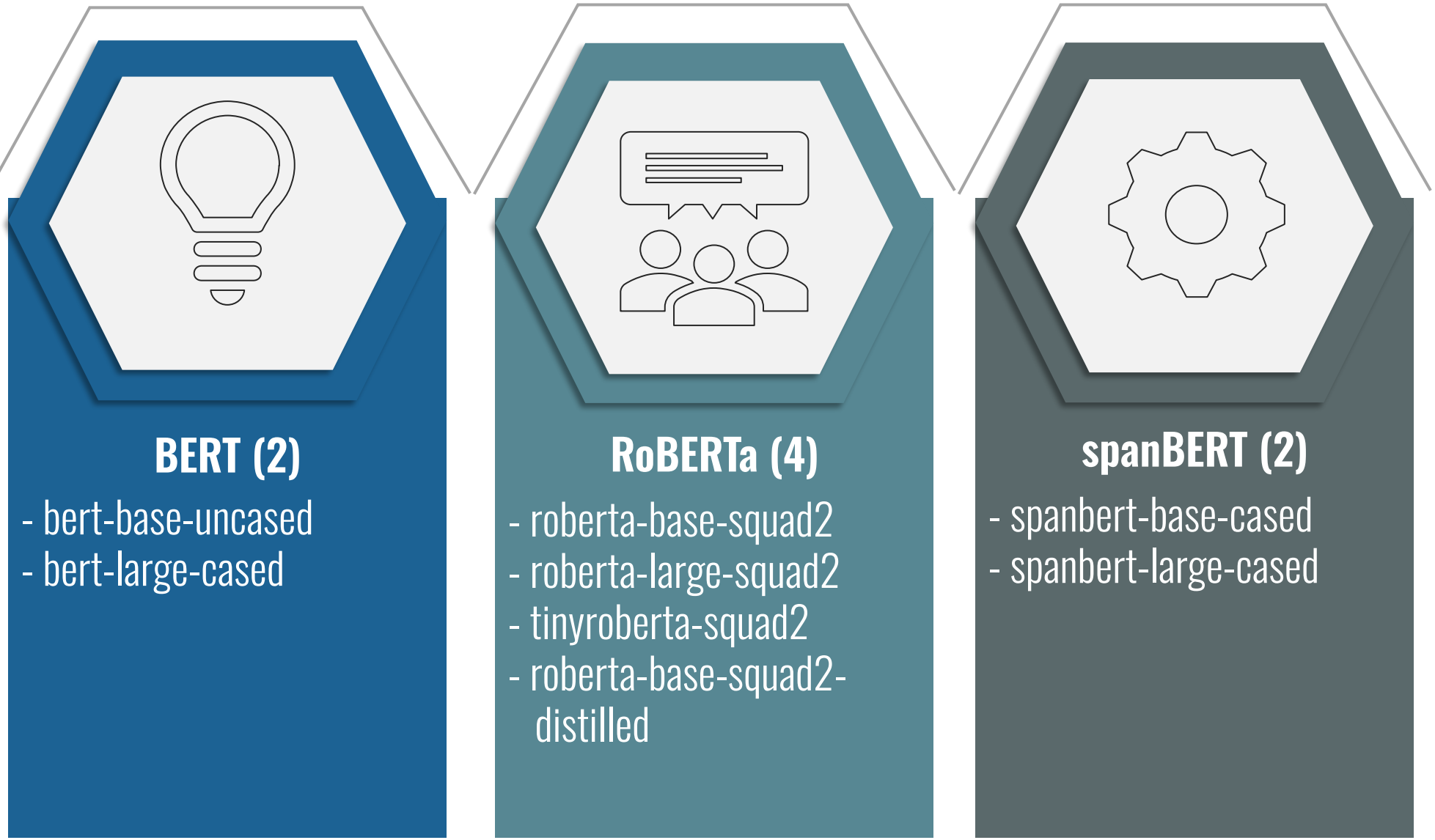


- **Preprocessing:**
  - Removed headers, footers, special characters; standardized spacing.
  - Removed stopwords and lowercase text.
- **Chunking:** Divided text into 512-word chunks to fit model constraints.
- **Pipeline:** Configured BERT model for question-answering tasks.
- **Input Formatting:** Paired questions with relevant chunks for processing.

Questions Chosen

1. Who supports Sherlock Holmes on the investigation?
2. Who is the victim?
3. Who killed the victim?
4. Where does the murder take place?
5. What is the murder weapon?
6. When does Sherlock Holmes begin to unravel the details that lead to solving the murder?
7. How does Sherlock find the murderer?
8. What is the motive for the murder?
9. What is the evidence that led Sherlock Holmes to the murderer?
10. What is the plot twist of the story?

Models



BERT (2)

- bert-base-uncased
- bert-large-cased

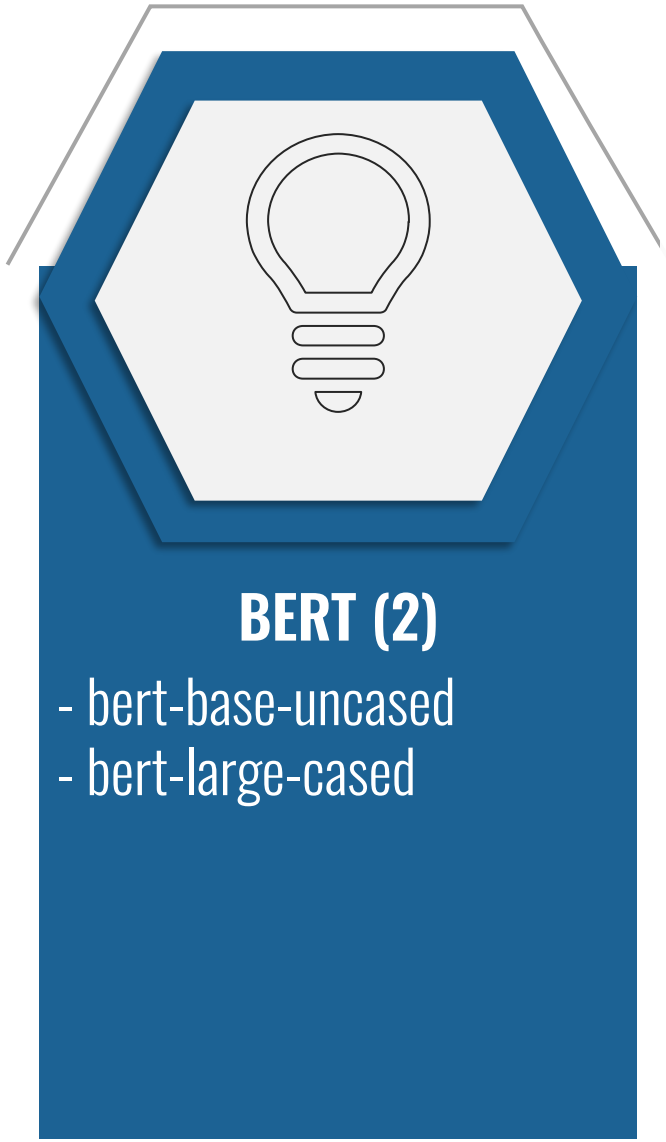
RoBERTa (4)

- roberta-base-squad2
- roberta-large-squad2
- tinyroberta-squad2
- roberta-base-squad2-distilled

spanBERT (2)

- spanbert-base-cased
- spanbert-large-cased

Models



bert-base-uncased:

- 110M parameters
- No differentiation between capitalized and uncapitalized words
- Strips accent markers in words

bert-large-cased:

- 340M parameters
- Distinguishes between capitalized and uncapitalized words

Training objectives:

1. **Masked Language Modeling (MLM)** - learns bidirectional representation of sentences
2. **Next Sentence Prediction (NSP)** - learns sequential order of sentences

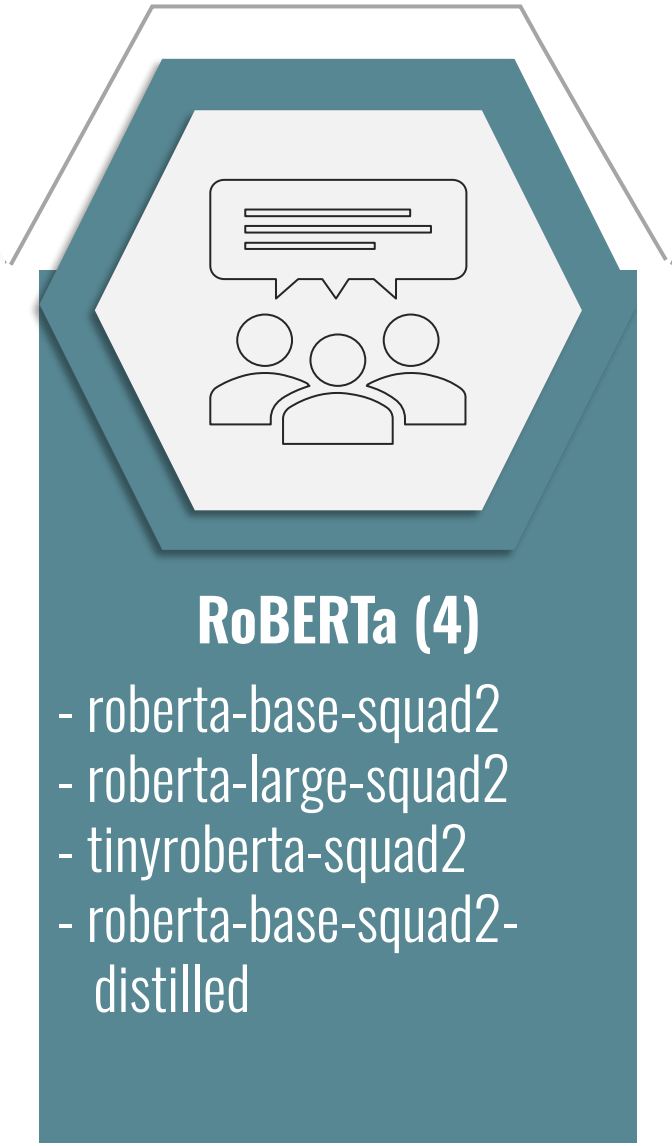
Models

roberta-base-squad2:
- 124M parameters

roberta-large-squad2:
- 354M parameters

tinyroberta-squad2:
- Distilled version of base model using TinyBERT approach
- Teacher: roberta-base-squad2
- ~62.5M parameters*

roberta-base-squad2-distilled:
- Distilled version of base model using Hinton approach
- Teacher: roberta-large-squad2
- ~178M parameters*



SQuAD2.0: Stanford Question Answering Dataset
- Consists of 100,000 questions with 50,000 unanswerable questions that look similar to answerable questions

Distillation:
- Prediction layer distillation - minimize difference between the outputs of the prediction layer between student & teacher
  - Hinton approach
- Intermediate layer distillation - minimize differences between hidden states & attentions of student/teacher
  - Before prediction layer distillation
  - TinyBERT approach

* rough estimate according to some articles - no explicit number found

Models

spanbert-base-cased:

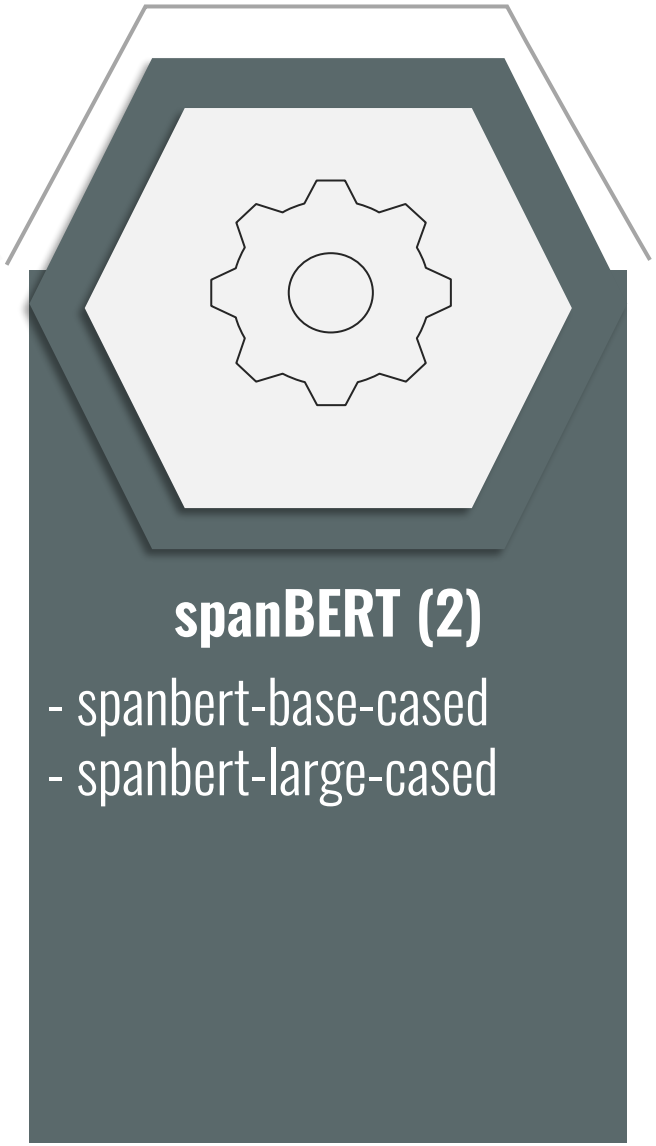
- 110M parameters

spanbert-large-cased:

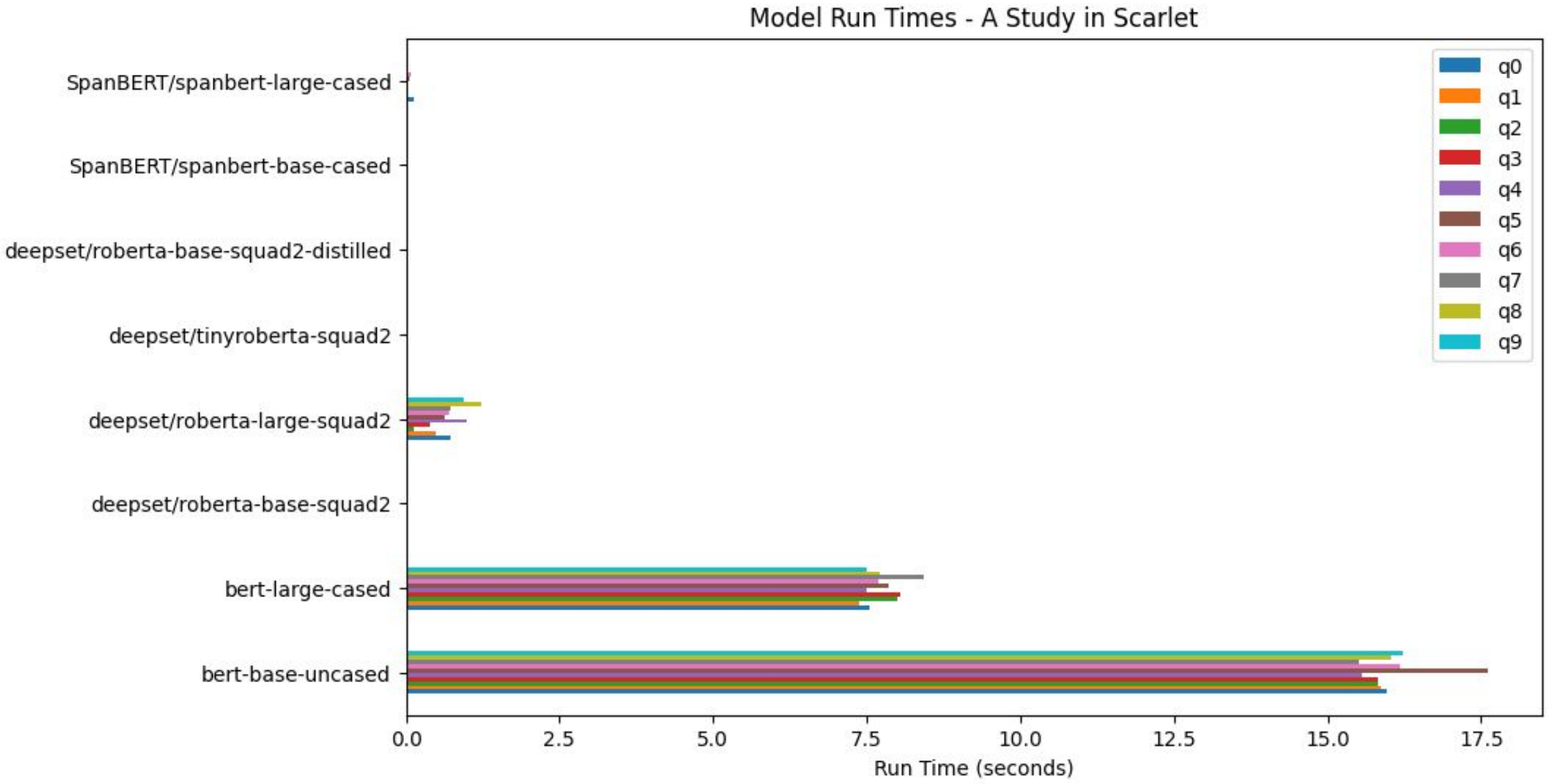
- 340M parameters

SpanBERT has same model configuration as BERT, but differs in:

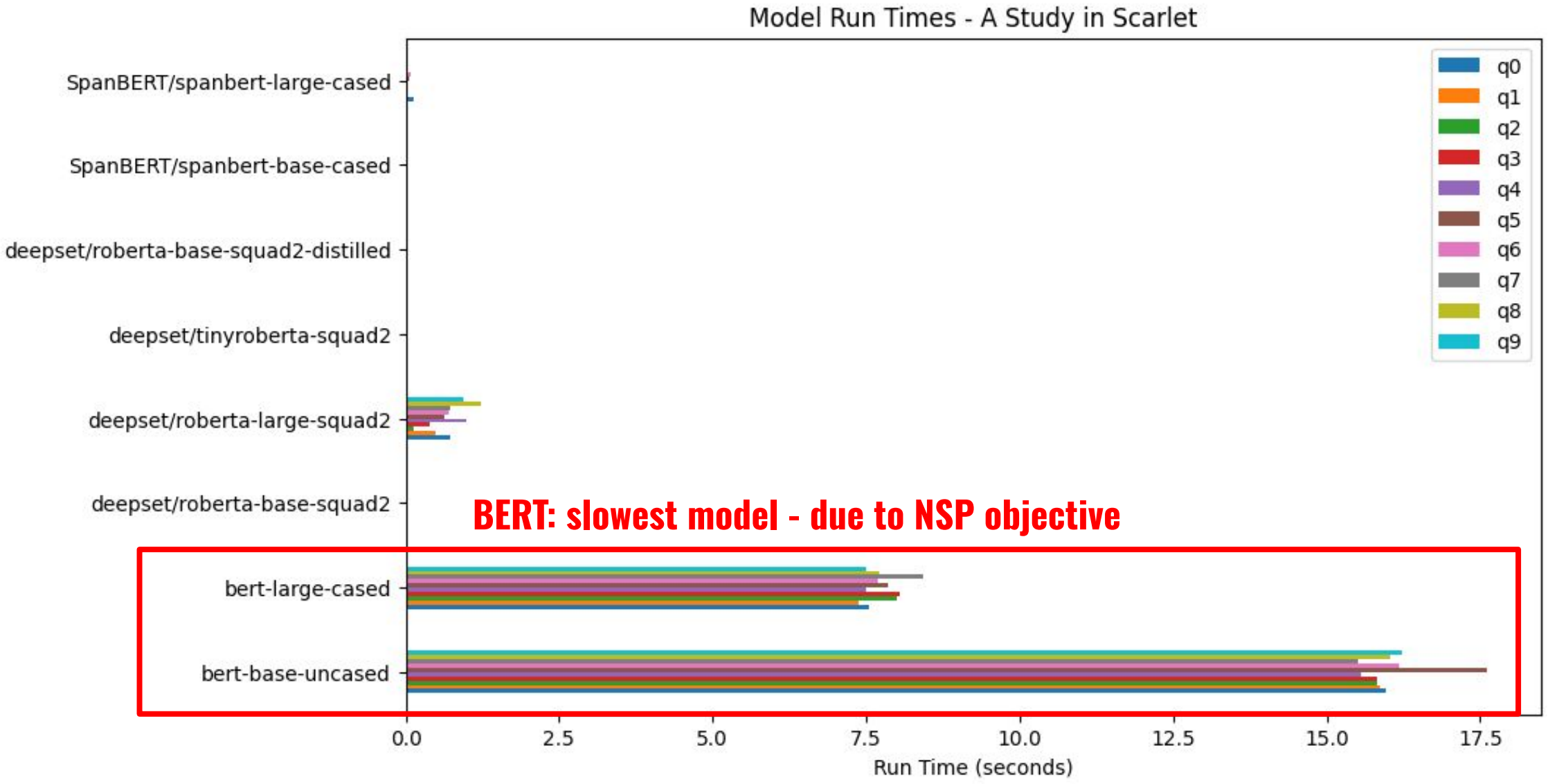
- Masking contiguous random spans (not random tokens)
- **Span-boundary objective (SBO)** - learns to predict content of masked span without relying on individual tokens within the span



Results - A Study in Scarlet



Results - A Study in Scarlet



Results - A Study in Scarlet

	q0	q1	q2	q3	q4	q5	q6	q7	q8	q9
bert-base-uncased	221B	221B	221B	221B	221B	221B	221B	221B	221B	221B
bert-large-cased	dull	dull	dull	dull	dull	odd pounds it, nothing taken. Whatever motives	dull	dull	century lawyer, suppose. writing legal twist i...	dull
deepset/roberta-base-squad2	Lestrade	Drebber	Drebber	Salt Lake City	rifle	Echo_ day	drunk sort o' man	writer going put female name Rachel	sharp needles	that fool Lestrade, thinks smart, gone upon wr...
deepset/roberta-large-squad2	Lestrade	Lucy Ferrier	Brigham Young	Brixton Road	knife	'	'	robbery	Brigham Young	Brigham Young
deepset/tinyroberta-squad2	Lestrade	Joseph Stangerson	Sherlock Holmes	Scotland Yard	rifle	hunter'	wound upon dead man's person	despotism hatred Liberalism	dark plain sky	thickens
deepset/roberta-base-squad2-distilled	Jefferson Hope	father	Drebber	Brixton Road	rifle	mountains	Sherlock Holmes yet finished breakfast	Brigham Young	Knowledge Literature	Sherlock Holmes
SpanBERT/spanbert-base-cased	Latin character, may	see that?" cried. "It seems knows deal should.	Latin character, may	Latin character, may	see that?" cried. "It seems knows deal should.	mantelpiece—a red wax one—and light saw	gloomy	see that?" cried. "It seems knows deal should.	Latin character, may	Latin character, may
SpanBERT/spanbert-large-cased	Again, absurd suppose sane man would carry del...	shall see me." tore	shall see me." tore	shall see me." tore	Again, absurd suppose sane man would carry del...	second man to-day used	second man to-day used	Again, absurd suppose sane man would carry del...	intended kill cold blood. would rigid justice ...	Again, absurd suppose sane man would carry del...

Results - A Study in Scarlet

BERT struggles to answer any questions given

	q0	q1	q2	q3	q4	q5	q6	q7	q8	q9
bert-base-uncased	221B	221B	221B	221B	221B	221B	221B	221B	221B	221B
bert-large-cased	dull	dull	dull	dull	dull	odd pounds it, nothing taken. Whatever motives	dull	dull	century lawyer, suppose. writing legal twist i...	dull
deepset/roberta-base-squad2	Lestrade	Drebber	Drebber	Salt Lake City	rifle	Echo_ day	drunk sort o' man	writer going put female name Rachel	sharp needles	that fool Lestrade, thinks smart, gone upon wr...
deepset/roberta-large-squad2	Lestrade	Lucy Ferrier	Brigham Young	Brixton Road	knife	'	'	robbery	Brigham Young	Brigham Young
deepset/tinyroberta-squad2	Lestrade	Joseph Stangerson	Sherlock Holmes	Scotland Yard	rifle	hunter'	wound upon dead man's person	despotism hatred Liberalism	dark plain sky	thickens
deepset/roberta-base-squad2-distilled	Jefferson Hope	father	Drebber	Brixton Road	rifle	mountains	Sherlock Holmes yet finished breakfast	Brigham Young	Knowledge Literature	Sherlock Holmes
SpanBERT/spanbert-base-cased	Latin character, may	see that?" cried. "It seems knows deal should.	Latin character, may	Latin character, may	see that?" cried. "It seems knows deal should.	mantelpiece—a red wax one—and light saw	gloomy	see that?" cried. "It seems knows deal should.	Latin character, may	Latin character, may
SpanBERT/spanbert-large-cased	Again, absurd suppose sane man would carry del...	shall see me." tore	shall see me." tore	shall see me." tore	Again, absurd suppose sane man would carry del...	second man to-day used	second man to-day used	Again, absurd suppose sane man would carry del...	intended kill cold blood. would rigid justice ...	Again, absurd suppose sane man would carry del...

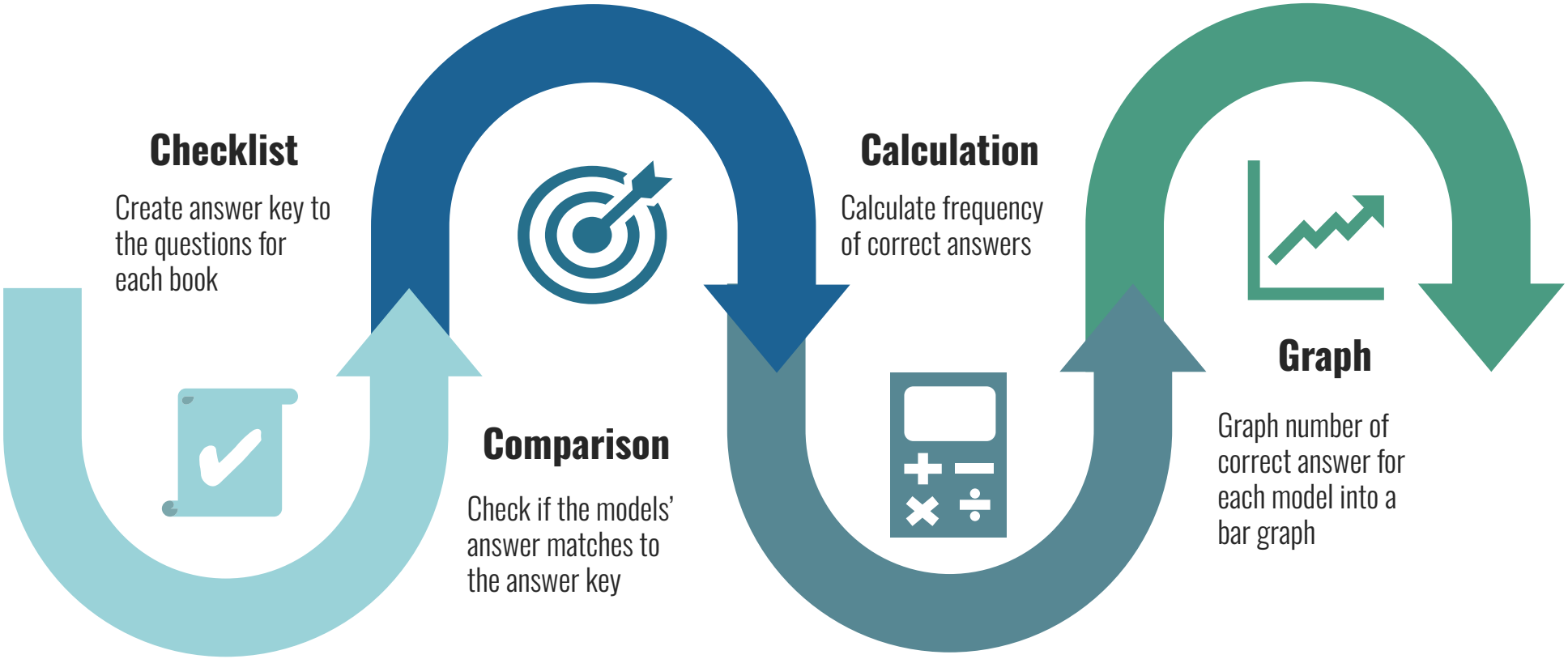
SpanBERT is just as bad as BERT, with the exception of the spanbert-large-cased model

Results - A Study in Scarlet

Roberta models are the best performing - the most correct sensical answers to given questions

	q0	q1	q2	q3	q4	q5	q6	q7	q8	q9
bert-base-uncased	221B	221B	221B	221B	221B	221B	221B	221B	221B	221B
bert-large-cased	dull	dull	dull	dull	dull	odd pounds it, nothing taken. Whatever motives	dull	dull	century lawyer, suppose. writing legal twist i...	dull
deepset/roberta-base-squad2	Lestrade	Drebber	Drebber	Salt Lake City	rifle	Echo_ day	drunk sort o' man	writer going put female name Rachel	sharp needles	that fool Lestrade, thinks smart, gone upon wr...
deepset/roberta-large-squad2	Lestrade	Lucy Ferrier	Brigham Young	Brixton Road	knife	'	'	robbery	Brigham Young	Brigham Young
deepset/tinyroberta-squad2	Lestrade	Joseph Stangerson	Sherlock Holmes	Scotland Yard	rifle	hunter'	wound upon dead man's person	despotism hatred Liberalism	dark plain sky	thickens
deepset/roberta-base-squad2-distilled	Jefferson Hope	father	Drebber	Brixton Road	rifle	mountains	Sherlock Holmes yet finished breakfast	Brigham Young	Knowledge Literature	Sherlock Holmes
SpanBERT/spanbert-base-cased	Latin character, may	see that?" cried. "It seems knows deal should.	Latin character, may	Latin character, may	see that?" cried. "It seems knows deal should.	mantelpiece—a red wax one—and light saw	gloomy	see that?" cried. "It seems knows deal should.	Latin character, may	Latin character, may
SpanBERT/spanbert-large-cased	Again, absurd suppose sane man would carry del...	shall see me." tore	shall see me." tore	shall see me." tore	Again, absurd suppose sane man would carry del...	second man to-day used	second man to-day used	Again, absurd suppose sane man would carry del...	intended kill cold blood. would rigid justice ...	Again, absurd suppose sane man would carry del...

Analysis



Checklist

Create answer key to the questions for each book

Comparison

Check if the models' answer matches to the answer key

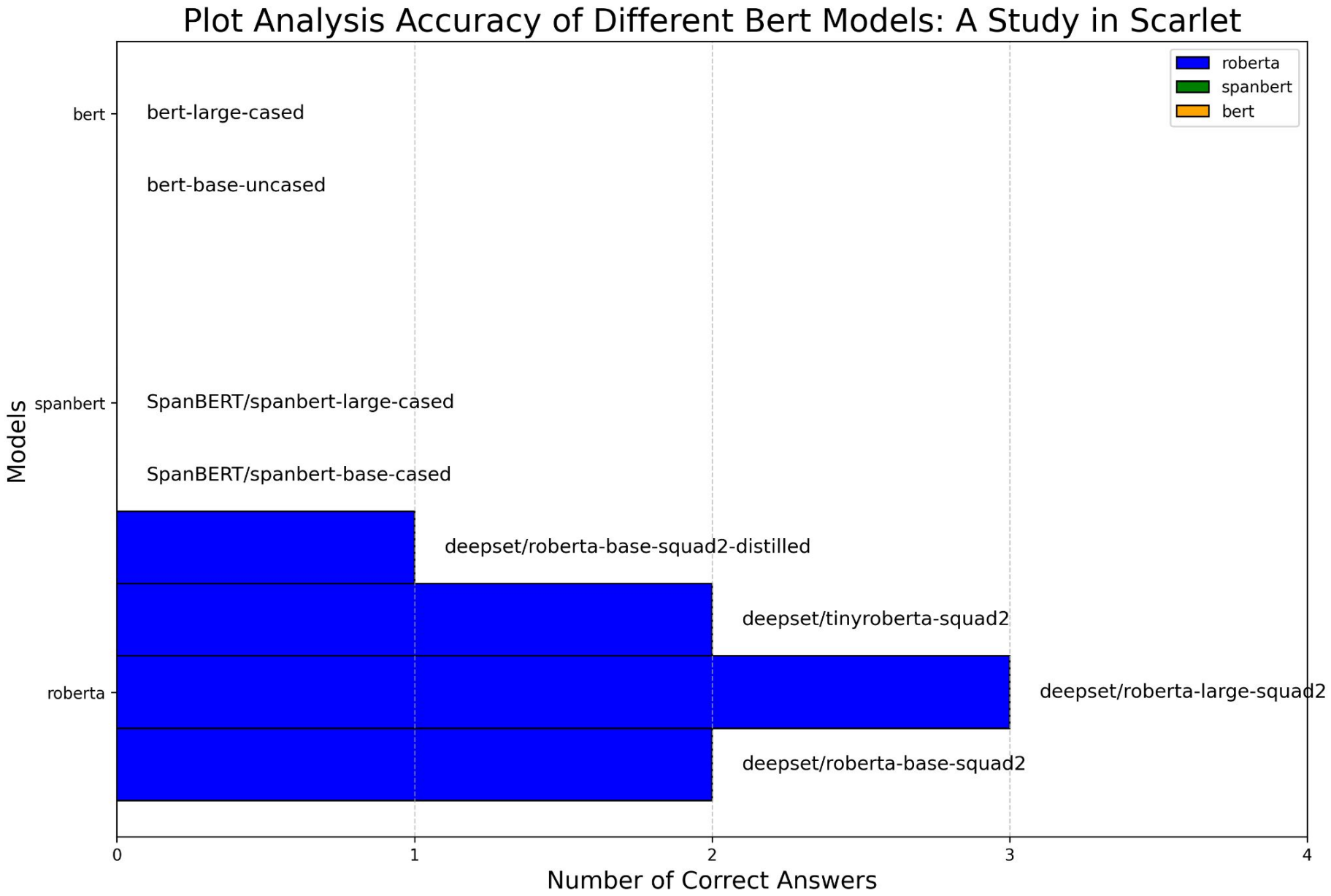
Calculation

Calculate frequency of correct answers

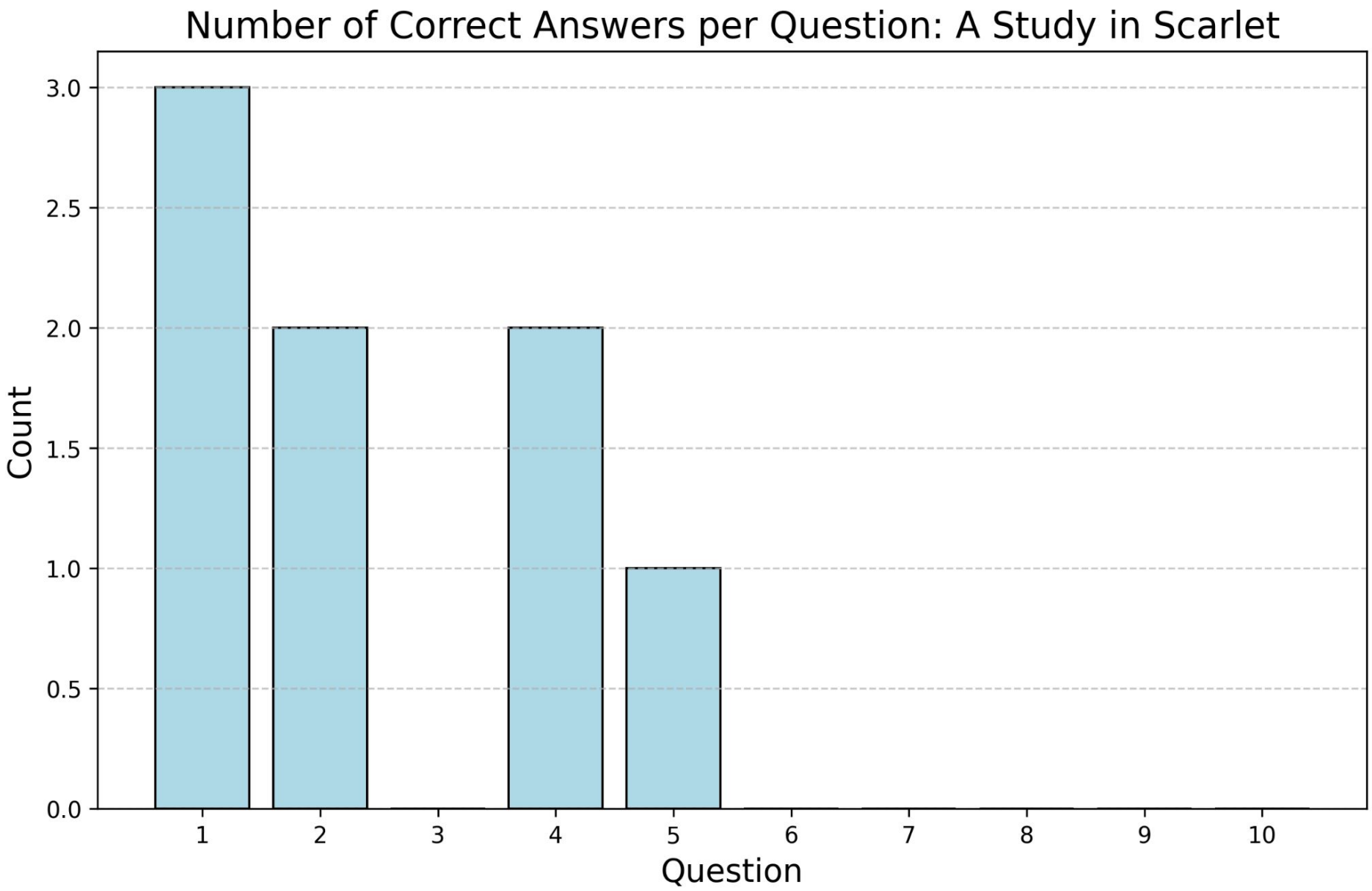
Graph

Graph number of correct answer for each model into a bar graph

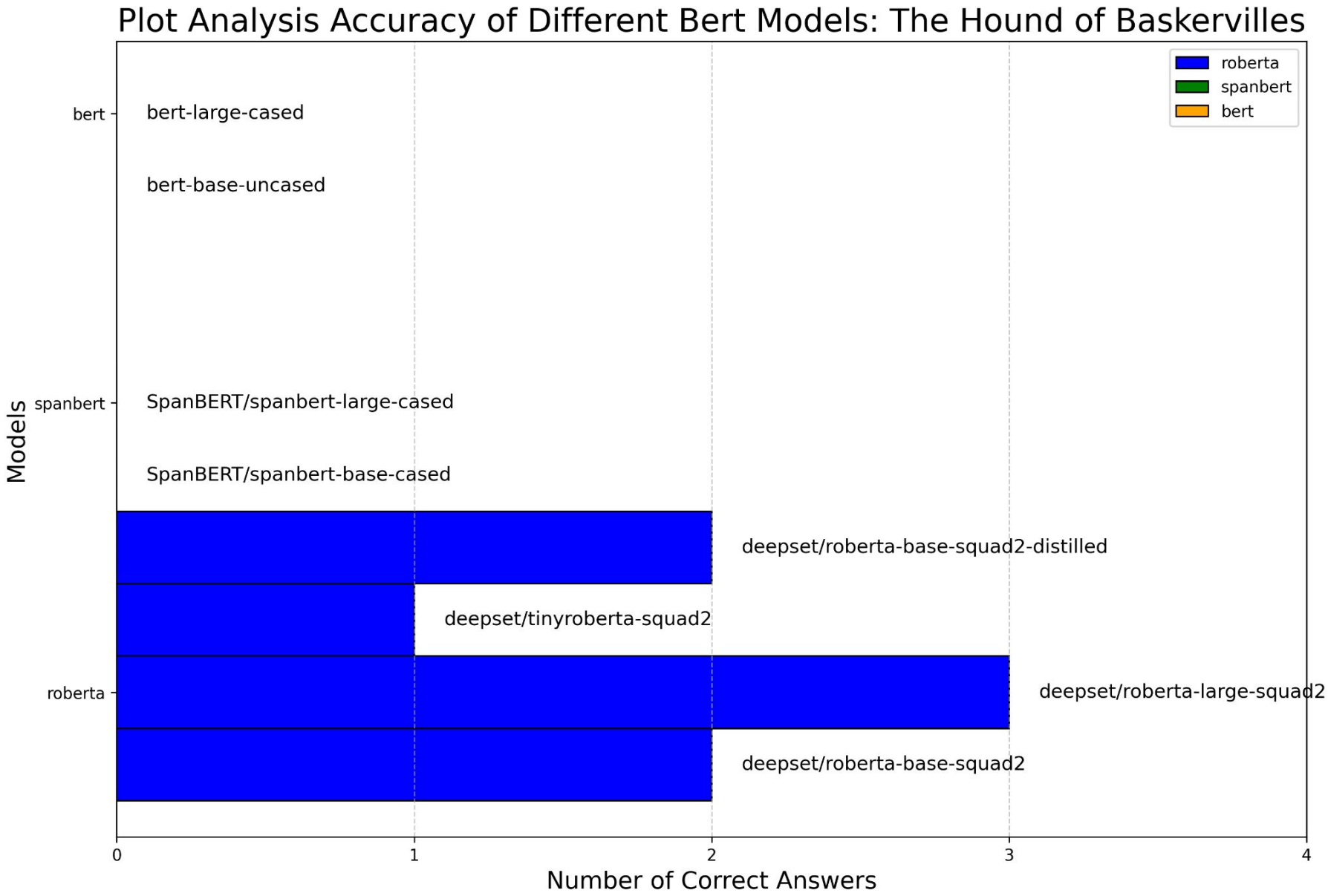
Analysis



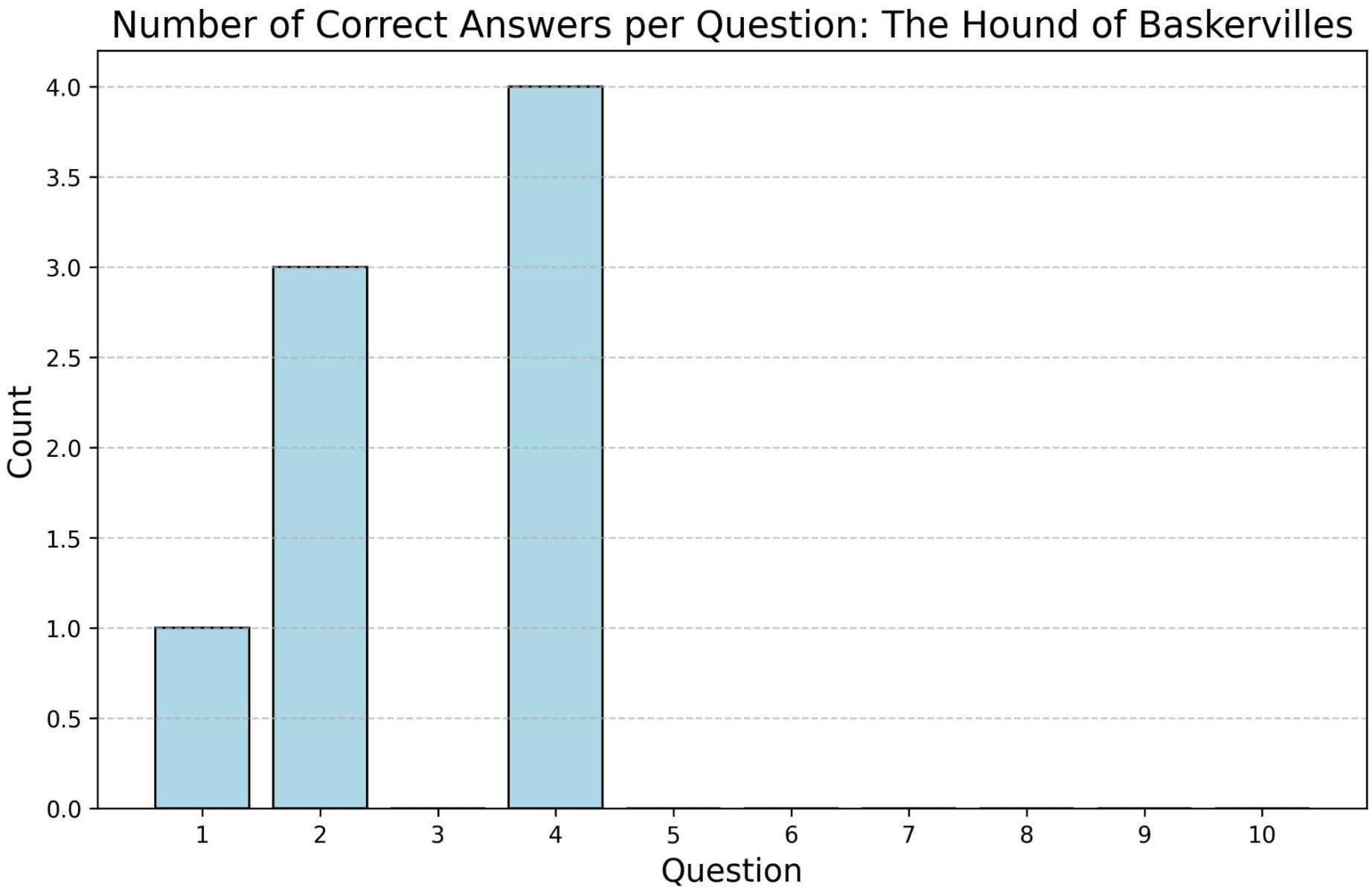
Analysis



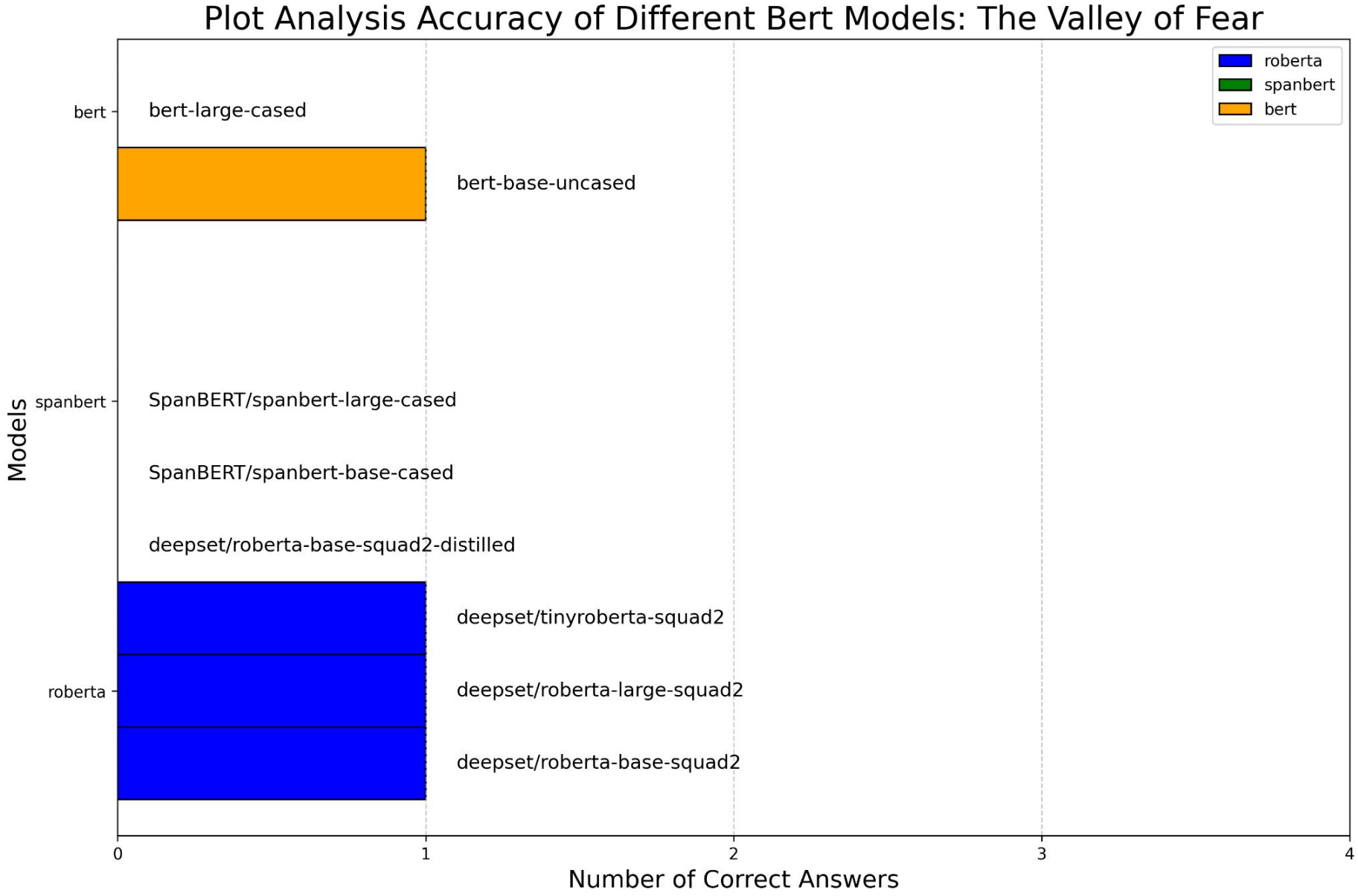
Analysis



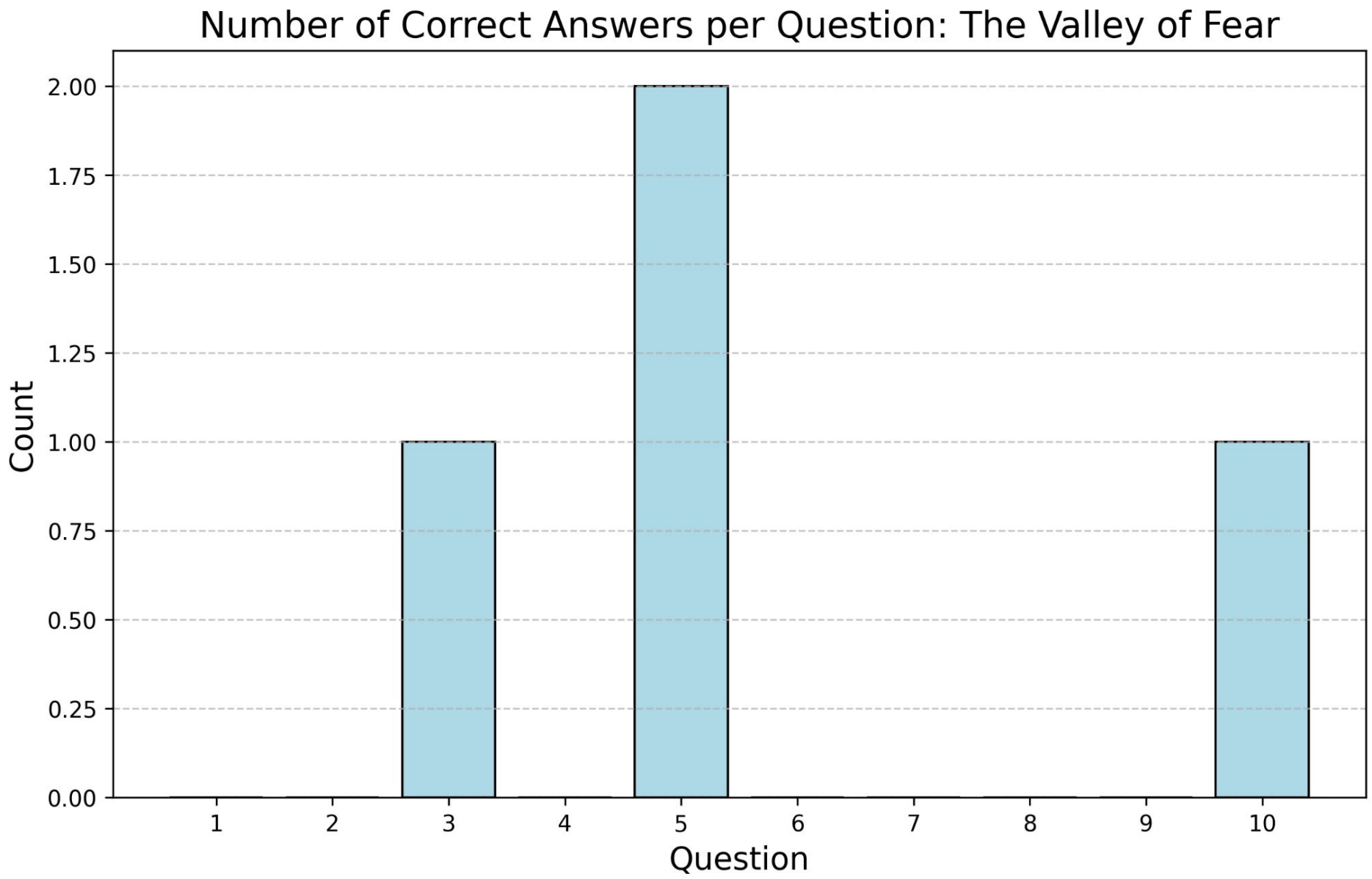
Analysis



Analysis



Analysis



Analysis

- **BERT**
  - Pros:
    - Can be generalized to a variety of tasks, provided that it is fine-tuned
  - Cons:
    - Not able to generate sensical responses - probably due to undertraining
- **RoBERTa**
  - Pros:
    - Surpassed both BERT and SpanBERT - delivered accurate & relevant responses
    - Crucial pretraining methodology: dynamic masking & larger-scale datasets
    - Strong management of complex, context-dependent queries
- **SpanBERT:**
  - Pros:
    - Excelled in span-based extractions
  - Cons:
    - Lacked versatility for different tasks

Conclusion

- BERT is severely under trained, but was meant to be a generalizable model that can easily be fine-tuned to certain tasks - gave parts of words, single words & sentences, etc.
- SpanBERT was not even close to the correct answer - gave long sentences with punctuation.
- Overall, RoBERTa worked the best with the lower-level questions - comprehension achieved
- **All models** incorrectly answered complicated questions.
- **Chunking text** into 512-token chunks added preprocessing complexity, but helped with addressing some model limitations, especially on the compute resources available