A Performance Analysis on Question and Answering Chatbots

By: Abolfazl Malekahmadi, Mohadese Sheikh Qoraei,

Habibollah Naeimi, Anshul Pundhir

July 2023

neuromatch academy

Pod 509/ Group 1

Overview

- Case Study
- Methodology
    - Dataset
    - Pre-processing
    - BERT
    - GPT-2
- Conclusion

Question Chatbots

- These days chatbots are one of the best solutions for answering FAQs.
- Considering the importance of speed and accurate answering, we wanted to check the possibility of designing a chatbot for new questions.
- We are looking to find out how the additional information and documentation will affect the performance of a chatbot.

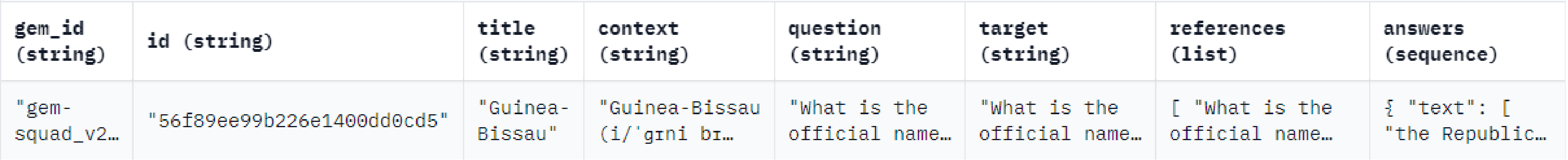
Complexity vs. Knowledge

- In fact, we want to see which is better, a more complex model or a model that is trained with more knowledge and information.
- So we choose GPT-2 and BERT models for this comparison.
- GPT-2 as a more complex model and BERT which is trained not only by questions and answers but also by the context of the questions.

Methodology

Dataset

- SQuAD2.0: The Stanford Question Answering Dataset.
  - A reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles.
  - the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.
  - Contains 100,000+ question-answer pairs on 500+ articles, also over 50,000 unanswerable questions.



gem_id (string)	id (string)	title (string)	context (string)	question (string)	target (string)	references (list)	answers (sequence)
"gem-squad_v2…	"56f89ee99b226e1400dd0cd5"	"Guinea-Bissau"	"Guinea-Bissau (i/ˈɡɪni bɪ…	"What is the official name…	"What is the official name…	["What is the official name…	{ "text": ["the Republic…

Methodology

Pre-processing

- GPT-2
  - Splitting dataset to train and validation.
  - Train and validation ratio = 8 to 2

- BERT
  - Tokenizing using a pre-defined tokenizer.
  - Truncating data to a maximum length of 512 tokens.
  - Searching for rows with no answer and rows with identical answers.

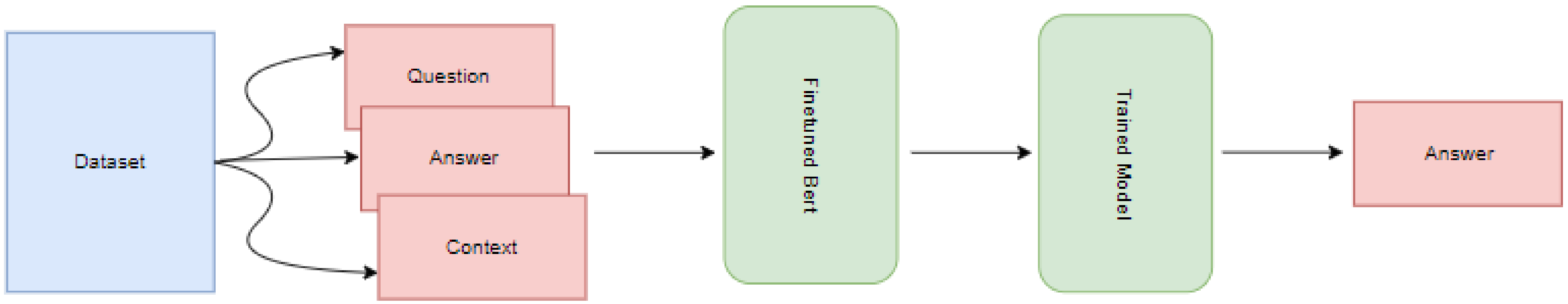
BERT

- BERT: Bidirectional Encoder Representations from Transformers.
- DistilBERT: a small, fast and light model trained by distilling BERT base.
- It has 40% less parameters, runs 60% faster while preserving over 95% of BERT's performances.

- Our model:
  - Uses DistilBertTokenizerFast for tokenizing.
  - Optimized using AdamW with learning rate = 3e-4.
  - Fine-tuned on SQuAD2.0 dataset.
  - Uses the questions, context, and answers columns of the dataset.

BERT

- Structure and sample result:

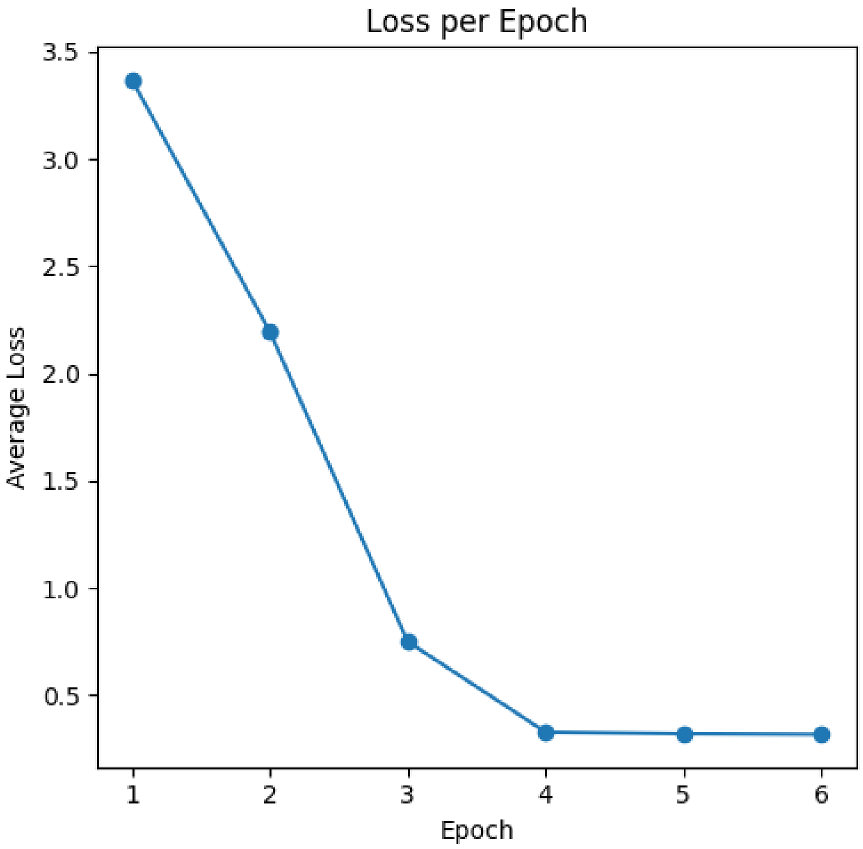


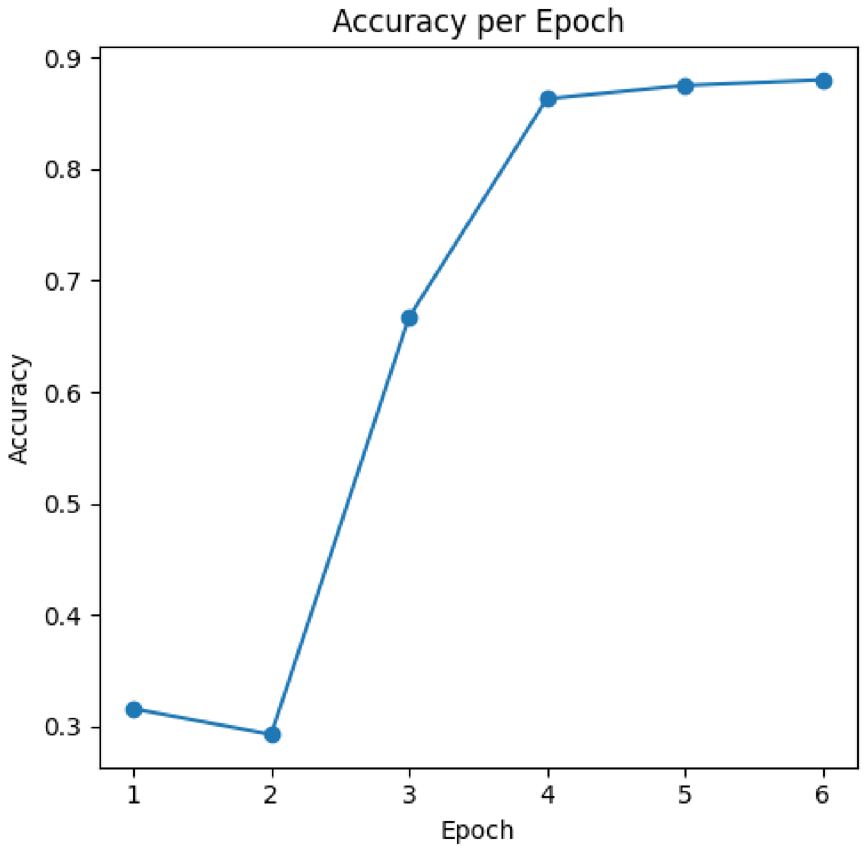
Question: Who is the 44th President of the United States?
Context: The 44th President of the United States is Barack Obama. He served as the President from January 20, 2009, to January 20, 2017.

Answer: barack obama

BERT

- Results:





Average Loss = 0.4028

Accuracy = 85.30%

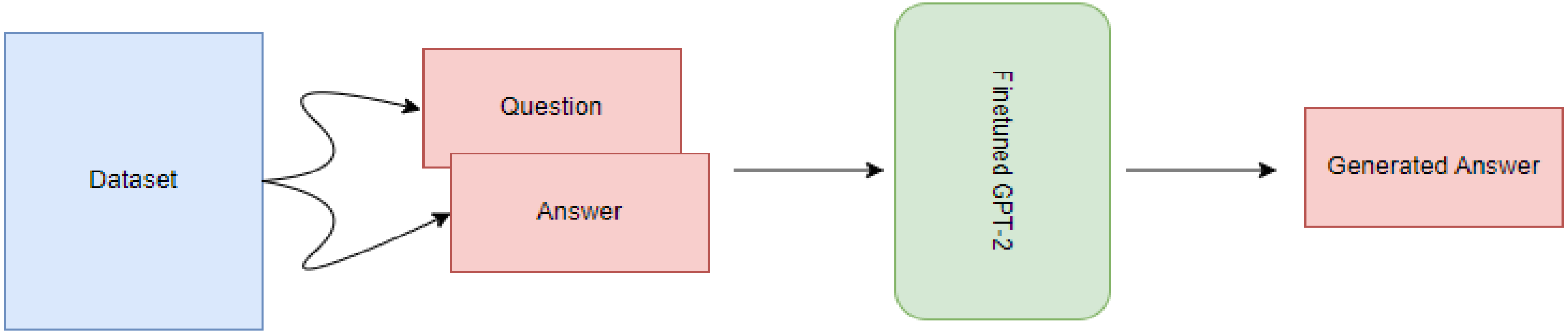
GPT-2

- GPT-2: Generative Pre-trained Transformer 2.
- Pre-trained on BookCorpus, a dataset of over 7,000 unpublished fiction books from various genres, and trained on a dataset of 8 million web pages.

- Our model:
  - Uses pre-trained GPT-2 small with 124M parameters.
  - Fine-tuned on SQuAD2.0 dataset.
  - Uses only the questions and answers columns of the dataset.
  - Accuracy calculated by counting similar words.
  - Steps = 3000 and Learning rate = 0.001

GPT-2

- Structure and sample result:

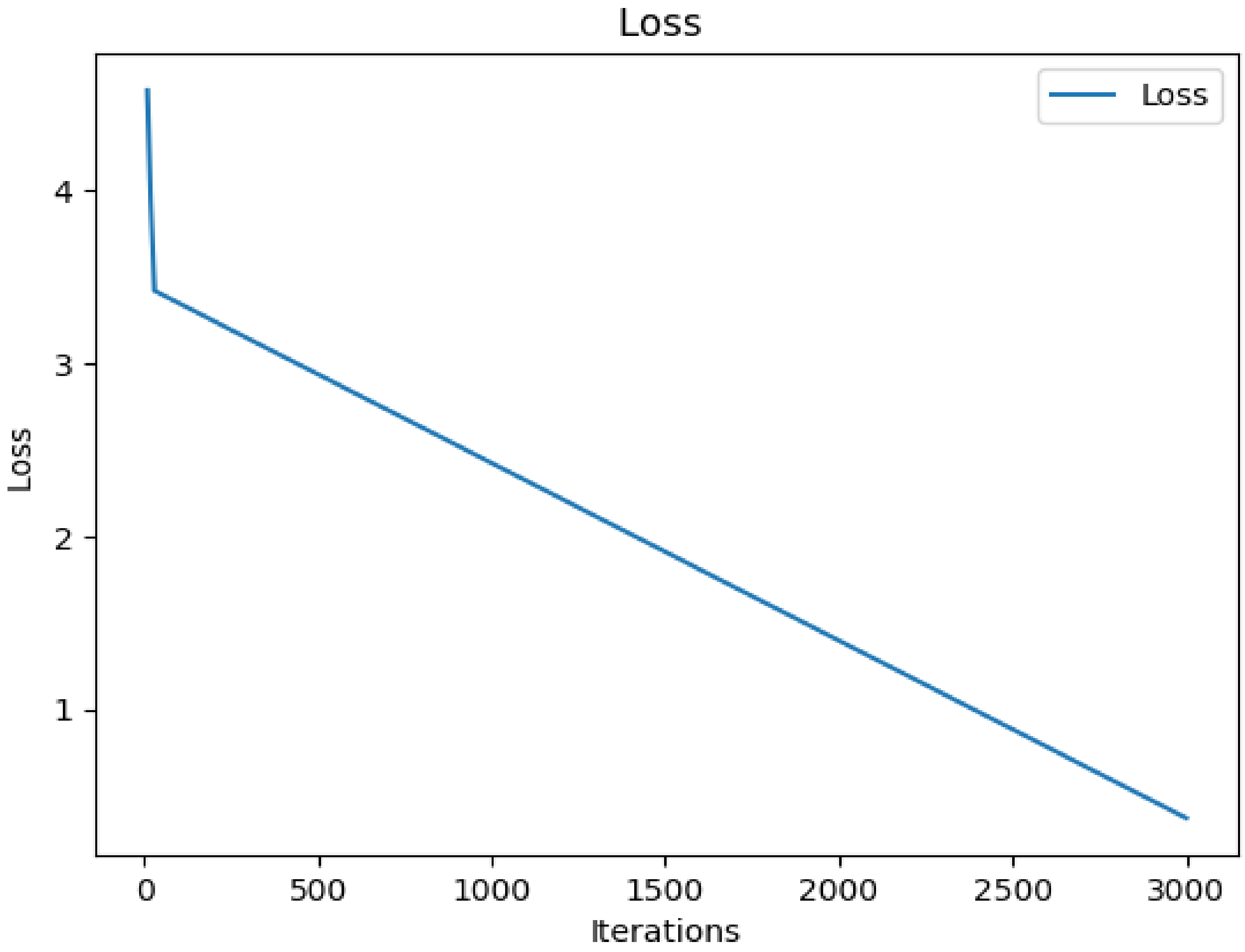


Question: How many people did Carlton have per km2 between 2012 and 2013?
Validation answer: 9,000
Generated answer: 16

Question: Where was there a minute of silence during the relay?
Validation answer: Ruijin, Jiangxi
Generated answer: a moment of silence

GPT-2

- Results:



Precision = 0.11475
Recall = 0.11666
F1 score = 0.11570

Conclusion

The simpler model with richer additional information can outperform the more advanced model with less information related to the question's context.

- Limitations: Mostly Time and Resources
  - Due to the limitations of Colab GPU, we were not able to use the model with more parameters and epochs.
  - Couple times of GPU crashing.
- Future Direction: How can we increase the performance?
  - Much more GPUs
  - Even better datasets
  - Fine-tuning simpler rich model using more complex models like RL or GNN.