Fine-tuning DistilBERT for Energy-Efficient Machine Reading Comprehension: Performance Analysis on SQuAD 2.0

Submitted: 08Dec2024

Tracy Volz, volz.tracy@gmail.com

Abstract

This paper explores the fine-tuning of DistilBERT on the SQuAD 2.0 dataset to develop an energy-efficient machine reading comprehension (MRC) system. Through experimentation with 20 model variations, the research investigated different hyperparameters, optimizers, and model architectures while utilizing two distinct dataloader approaches: variable-length and fixed-length truncation. The best-performing model achieved normalized scores of 64.4% for Exact Match (EM) and 71.4% for F1, falling short of the target metrics derived from BERT's performance (68.43% EM and 75.47% F1). While more complex architectures incorporating multi-head attention mechanisms were tested, they showed decreased performance compared to simpler models. The research revealed significant challenges with "Why" questions and highlighted the importance of balanced question type distribution in training data.

1. Introduction

Machine reading comprehension (MRC) is a subset of Natural Language Processing (NLP) tasks in the realm of Question Answering Systems. The goal of a simple MRC model is to read a body of text (ex: a paragraph of information), then accurately answer a context-based question (V, 2023). In 2016, Stanford introduced the "Stanford Question Answering Dataset" version 1.1 (i.e. SQuAD 1.1), This dataset includes context paragraphs and ~100K question-answer pairs that correspond to included context (Rajpurkar, 2016). Two years later, SQuAD 2.0 was introduced; it included the original 100K question answer-pairs, along with ~50K unanswerable questions (Rajpurkar, 2018). With SQuAD 2.0, accuracy is more difficult to achieve because the model must answer questions when information is represented in the context or abstain from answering when no answer is represented in the context paragraph. The SQuAD datasets are considered the gold standard for MRC experiments because they offer a large corpus of high-quality data that has been verified by humans (Rajpurkar, 2017).

BERT is a Bidirectional Encoder designed to pre-train deep bidirectional representations using a randomized masking technique (Devlin, 2019). Despite having a reputation for producing state-of-the-art results for NLP tasks such as classification and machine comprehension, BERT is computationally expensive. In 2019, approximately 1 year after BERT was released, Huggingface introduced DistilBERT, a BERT-based model that aims to retain 97% of BERT performance with a model that runs 60% faster and is 40% smaller than BERT (Sanh, 2020). Authors accredit the performance of the compressed model to the process called knowledge distillation. With this process, the student -- DistilBERT with 40% smaller transformer -- is pre-trained by distilling supervision of a teacher -- BERT with a full-sized transformer. As such, the DistilBERT model was pre-trained on the same body of text as the BERT model; however, the DistilBERT model contains six layers (rather than 12), and it does not have a token-type embeddings or poolers. Huggingface reports the DistilBERT model is small enough to run on mobile devices (Sanh, 2020).

There are several goals for this paper: 1) Explore fine-tuning of the energy-efficient DistilBERT on the SQuAD 2.0 dataset for the purpose of developing an MRC system, 2) Meet or exceed target performance metrics, 3) Explore the balance between energy-efficient fine-tuning and model accuracy – this research will explore performance metrics (EM and F1 scores) of simple models and determine if complex models (like those containing multi-head attention layers) are necessary to achieve target results.

2. Background

In a 2024 paper titled, "Comparative Analysis of State-of-the-Art Q&A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset", author Cem Özkurt from the Sakarya University of Applied Sciences captured architecture and performance differences of many BERT based models. Özkurt did not make suggestions for model design, but they present performance metrics and highlighted DistilBERT's value for situations with computational resource limitations. Performance metrics were reported as EM = 64.89% and F1 = 68.18% (Özkurt, 2024).

In 2023, the International Journal on Recent and Innovation Trends in Computing and Communication accepted a paper titled, "The DistilBERT Model: A Promising Approach to Improve Machine Reading Comprehension Models." Authors present hyperparameters and suggest the model design should include a custom question-answering head that sits on top of the pre-trained model. Performance metrics were reported as EM = 88.2% and F1 = 84.13% (V, 2023). One observation waws thier approach to splitting the data – they used portions of SQuAD 1.1 for test and validation, and used SQuAD 2.0 for training. Because content in SQuAD 1.1 and 2.0 overlap, this means the reported EM and F1 scores would have been exaggerated because they aren't representative of model performance on unseen data.

In a recent paper titled, "BERT-Based Model for Reading Comprehension Question Answering", authors describe fine-tuning a BERT model on the SQuAD 2.0 dataset, and reporting performance metrics as EM = 61.1% and F1 = 72.5% (Mosaed, 2023). Similarly, the BERT release paper from 2019 reported SQuAD 2.0 performance results on the BERT -Large model as EM = 80.0% and F1 = 83.1% (Devlin, 2019). The success of models presented in this paper will be measured using these metrics.

3. Methods

3.1. Dataset Split and Pre-Processing

Fine-tuning and evaluations were performed using SQuAD 2.0, accessed through Huggingface's dataset download library. Rather than using Huggingface's pre-defined split, most models utilized an 80/10/10 split for training, validation, and testing, respectively. To evaluate the impact of using a smaller dataset during training, some models utilized a 40/10/10 split. Both split approaches ensure final test metrics reflect realistic model performance on unseen data.

Data quality verification tasks identified 315 question-answer pairs where the expected answer did not match the true answer. These pairs were removed as they constitute an insignificant portion (~0.2%) of the dataset (see Appendix A - Initial Data Quality Verification).

One of two dataloader classes was utilized prior to running the models: one with a variable-length truncation approach, and another with a fixed-length truncation approach and reduced max tokens. Using chi-square analysis, the distributions of question types in training, validation, and test groups were analyzed to determine if significant differences existed between the "Variable-Length Trunc" and "Fixed-Length Trunc" groups. Each group produced a very low chi-squared statistic (between 0.0058 and 0.0374) and a p-value of 1.0 with 7 degrees of freedom. These results indicate that differences between the groups are due to chance, confirming no significant difference between the groups (see Appendix B - Data Loaders). The following table describes sample counts for each dataloader class:

Dataloader Name	Max Tokens	Training Samples	Validation Samples	Test Samples	Total Samples
"Variable-Length Trunc"	512	113,000	14,000	14,000	141,000
"Fixed-Length Trunc"	384	113,506	14,181	14,021	140,190

3.2. Performance Metrics

Exact match (EM) and F1 scores are common performance metrics used in machine reading comprehension because together they indicate both perfectly accurate predictions and partially correct predictions. These metrics appear in the reviewed literature and are presented for the models evaluated in this paper. However, the literature did not clearly indicate whether models were normalized prior to metric calculation. This paper reports normalized EM and F1 scores for all model experiments. For details on normalization parameters, exact match (EM) definition, and F1 calculation, see Appendix C - Performance Metrics.

3.3. Measure of Success

Huggingface claims that DistilBERT can retain 97% of BERT's accuracy. Based on this information, the average performance of the BERT models previously described in this paper establishes a measure of success. Therefore, the DistilBERT models presented in this paper will be considered successful if the EM meets or exceeds 0.97*(61.1+80.0)/2 = 68.43% and if the F1 meets or exceeds 0.97*(72.5+83.1)/2 = 75.47%.

3.4. Experimentation

Prior to fine-tuning, baseline results showed that the pre-trained DistilBERT model produced an EM score of 14.3% and an F1 score of 16.8% when tested on the SQuAD 2.0 test data. These results indicated that fine-tuning would be necessary to achieve the previously mentioned EM and F1 goals.

All fine-tuning was performed using Google Colab with a High-RAM A100 GPU. Training utilized a simple neural network framework with configurable dropout layers and a linear classifier head to capture the start and end positions of the predicted text. As previously stated, one of two dataloaders was utilized prior to running the model: one with a variable-length truncation approach, and the other with a fixed-length truncation approach with reduced max tokens. This research examined 20 model variations, documented in Appendix D - Model Hyperparameters and Results.

For initial models, hyperparameters and the dataloader ("variable-length truncated") were selected based on recommendations from "Question Answering with DistilBERT" (Herbst, 2023). These hyperparameters included 5 epochs, a learning rate of $4x10^{-5}$, a dropout rate of 0.15, and RMSprop optimization. To reduce overfitting, the research explored combinations of dropout rates (0.15-0.225), epochs (3-5), and training set completeness (80% and 40%). The average model achieved EM≈62% and F1≈69%. Comparable experiments were then conducted using an AdamW optimizer, as recommended in the March edition of the International Journal of Informatics Visualization (Ahda, 2024). Results from models using the AdamW optimizer were similar to those using the RMSprop optimizer, averaging EM≈64% and F1≈71%. Time to complete fine-tuning activities ranged between 45 and 107 minutes, with an average being 76 minutes. Review of sample questions showed these models retained accurate "Truth" content and successfully produced exact matches for both answerable and unanswerable questions. See Appendix E - Sample Questions from "variable-length truncated" Data Loader. Results for this group are captured in Appendix D, Table A.

Upon review, the original dataloader ("variable-length truncated") lacked several aspects that could have impacted model accuracy: 1) a method to ensure consistent padding/truncation, as recommended in "BERT Fine-tuning Tutorial with PyTorch" (McCormick, 2019); 2) removal of empty lines; and 3) confirmation that all critical components were present (context, question, answer, and start position). Additionally, reducing the max tokens to 384 could improve model efficiency without compromising practical performance (Chowdhury, 2024). Multiple models were processed with the new dataloader ("fixed-length truncated"), which included these components. Despite promising results (EM≈76% and F1≈82%), review of sample questions revealed issues with how the dataloader handled the span of the "Truth" content. These issues significantly impacted the accuracy of "Truth" content for some samples; therefore, these results are not reported as successful models. See Appendix F - Sample Questions from "fixed-length truncated" Data Loader. Results for this group are captured in Appendix D, Table B.

To determine if increased complexity would improve performance, additional experiments were conducted using a model that included a multi-head attention mechanism, frozen DistilBERT parameters, and additional transformer layers. The experimental variations included attention head counts of 2, 8, and 12; dropout rates of ~0.175; epochs between 4-19; and training set completeness of 80% and 40%. Average model results showed EM≈45% and F1≈50%, which represented decreased performance compared to the previously described "variable-length truncated" model. Time to complete fine-tuning activities ranged between 56 and 475 minutes, with an average being 269 minutes. Results for this group are captured in Appendix D, Table C.

4. Results and Discussion

As previously mentioned, models using the "Fixed-Length Trunc" loader exhibited issues with the accuracy of "Truth" content; consequently, these results are not reported as successful models. Additionally, results from the complex model showed no improvement compared to the simple model that utilized the "Variable-Length Trunc" loader and were therefore omitted. The following table presents results from both the pre-trained DistilBERT model before fine-tuning (i.e., baseline) and the best-performing models from the "Variable-Length Trunc" loader group:

Model Description	Test EM (Norm.)	Test F-1 (Norm.)
Pre-trained DistilBERT -- prior to performing fine-tuning (ie: baseline)	14.3%	16.8%
Fine-tune results -- “Variable-Length Trunc” loader, simple model, 80/10/10 split, dropout = 0.18, RMSprop optimizer, learning rate = 4x10^{-5}, epochs = 3	64.4%	71.4%

Fine-tuning was completed in ~64 minutes, which was on the low end of possible times for the “Variable-Length Trunc" loader with simple model group (45 to 107minutes). Further analysis of the question type distribution for this model showed that "Why" questions were the worst-performing category of the group (EM=0.389 and F1=0.52). This is not surprising considering the "Why" question type has the lowest representation in the training data (~1.5%). In reviewing the sample questions, it appears that "Why" questions are often less straightforward than "Who" and "When" questions because their answers tend to be multi-faceted. A future model would ideally incorporate more robust training data with evenly distributed question types. Additionally, because many samples received low F1 scores despite being semantically equivalent to the correct answers, a future model should incorporate additional metrics to better capture semantic similarity (e.g., cosine similarity).

5. Conclusion

This research demonstrates that fine-tuning DistilBERT for machine reading comprehension can significantly improve performance over the baseline pre-trained model, though it fell short of achieving the target metrics of EM=68.43% and F1=75.47% (representing 97% of BERT's reported performance on SQuAD 2.0). The best results were achieved using a simple model architecture with the variable-length truncation dataloader, achieving normalized scores of 64.4% EM and 71.4% F1, while more complex architectures incorporating multi-head attention mechanisms showed decreased performance. The research identified a notable weakness in handling "Why" questions, which can be attributed to their underrepresentation in the training data and their inherently more complex, multi-faceted nature. Data quality and processing proved to be crucial factors, as evidenced by the challenges encountered with the fixed-length truncation dataloader despite its initially promising metrics.

Future work should focus on three key areas: developing more balanced training datasets across question types, incorporating semantic similarity metrics to better capture partially correct answers, and exploring alternative approaches to knowledge distillation that might better preserve BERT's performance characteristics. Additionally, this study suggests that while DistilBERT offers computational efficiency advantages, achieving performance parity with BERT on complex MRC tasks may require additional optimization strategies.

References

Ahda , F., Prasetya Wibawa , A., Dwi Prasetya, D., & Arbian Sulistyo, D. (2024, March). Comparison of Adam Optimization and RMSprop in Minangkabau- Indonesian Bidirectional Translation with Neural Machine Translation. https://www.joiv.org/index.php/joiv/article/view/1818

Chowdhury, S. K. (2024, September 16). *Token optimization: The backbone of effective prompt engineering*. IBM Developer. https://developer.ibm.com/articles/awb-token-optimization-backbone-of-effective-prompt-engineering/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. arXiv.org. https://arxiv.org/abs/1810.04805

Herbst, S. (2023, March 5). *Question answering with Distilbert*. Medium. https://medium.com/@sabrinaherbst/question-answering-with-distilbert-ba3e178fdf3d

McCormick, C., & Ryan, N. (2019, July 22). *Bert fine-tuning tutorial with pytorch*. BERT Fine-Tuning Tutorial with PyTorch · Chris McCormick. https://mccormickml.com/2019/07/22/BERT-fine-tuning/

Mosaed, A. A., H. Hindy and M. Aref, "BERT-Based Model for Reading Comprehension Question Answering," *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, Egypt, 2023, pp. 52-57, doi: 10.1109/ICICIS58388.2023.10391167.

Namuth-Covert, D., Haines, C., & Merk, H. (2024). *Chi-square test for goodness of fit in a plant breeding example*. passel. https://passel2.unl.edu/view/lesson/9beaa382bf7e

Özkurt, C. (2024a, February 19). *Comparative analysis of state-of-the-art Q&A models: Bert, Roberta, Distilbert, and Albert on Squad V2 Dataset*. Chaos and Fractals. https://journals.adbascientific.com/chf/article/view/17

Rajpurkar, P. (2017). *The stanford question answering dataset: Background, Challenges, Progress*. mlx. https://shorturl.at/SUcVB

Rajpurkar, P., Jia, R., & Liang, P. (2018, June 11). *Know what you don't know: Unanswerable questions for squad*. arXiv.org. https://arxiv.org/abs/1806.03822

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, October 11). *Squad: 100,000+ questions for machine comprehension of text*. arXiv.org. https://arxiv.org/abs/1606.05250

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, March 1). *Distilbert, a distilled version of Bert: Smaller, faster, cheaper and lighter*. arXiv.org. https://arxiv.org/abs/1910.01108

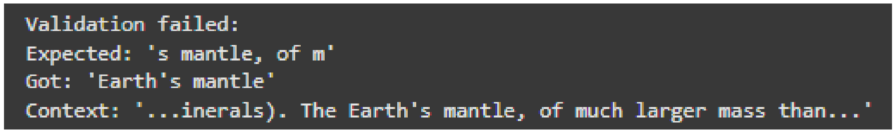
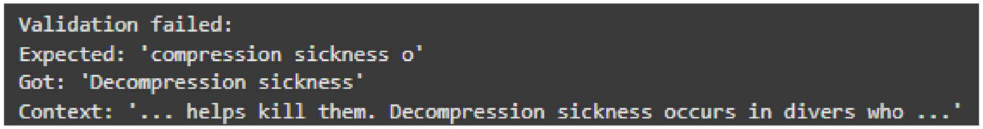
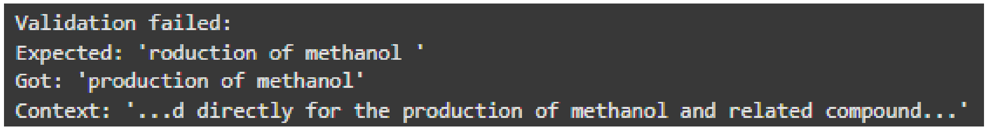
V, S., Shahapure, N. H., PM, R., B, N., Khandelwal, P., Anand, A., Agrawal, P., & Srivastava, V. (2023, August 2). *The distilbert model: A promising approach to improve machine reading comprehension models*. International Journal on Recent and Innovation Trends in Computing and Communication. https://ijritcc.org/index.php/ijritcc/article/view/7957

Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020, June 14). *Linformer: Self-attention with linear complexity*. arXiv.org. https://arxiv.org/abs/2006.04768

Appendix A: Initial Data Quality Verification

Data quality verification tasks identified 315 question-answer pairs where the expected answer did not match the true answer. The validation check used the start_answer index as a starting point, then counted forward using the length of the answer. The string of this output was compared to the actual answer to ensure the content was an exact match. If the content was not an exact match, the question-answer pair was removed. Removing these 315 question-answer pairs is not expected to negatively impact model results due to lack of training data, as they constitute an insignificant portion of the dataset (~0.2%). Furthermore, their removal is expected to have a positive impact on model results since these samples would have reduced the overall EM and F1 scores due to the mismatch between expected and true answers.

The following table provides a few examples of samples that failed this validation check:

Validation Failed Example	Comment
 Validation failed: Expected: 's mantle, of m' Got: 'Earth's mantle' Context: '...inerals). The Earth's mantle, of much larger mass than...'	Dropped 5 letters from beginning, everything before apostrophe
 Validation failed: Expected: 'compression sickness o' Got: 'Decompression sickness' Context: '... helps kill them. Decompression sickness occurs in divers who ...'	Dropped two letters from beginning
 Validation failed: Expected: 'roduction of methanol ' Got: 'production of methanol' Context: '...d directly for the production of methanol and related compound...'	Dropped one letter from beginning

Due to inconsistencies in the failures, no root-cause was identified. The data was re-loaded from Hugginface several times. Results remain unchanged for this validation check.

Appendix B: Data Loaders

After removing 315 samples from the initial data quality verification step, one of two dataloader classes was utilized prior to running the models – one designed for masking and one designed to handle triplets of information (context-question-answer) with additional validation checks.

Summary of aspects from the “Variable-Length Trunc” dataloader:

Additional Aspects and Validation	**Observation**
Removes leading / trailing whitespace (Note: doesn’t remove unanswerable questions)	---
Removes samples where tokens exceed 512 (Note: ensures only complete question-answers are present)	None were found after removing 315 samples from initial data verification
Removes the last file of each dataset – looses a maximum of 1000 rows for each split	---

Note: The Variable-Length Trunc code was obtained from Github through a Medium post from S. Herbst titled, “*Question answering with Distilbert*. Medium” (Herbst, 2023).

Summary of aspects from the “Fixed-Length Trunc” data loader:

Additional Aspects and Validation	**Observation**
Removes empty lines (Note: doesn’t remove unanswerable questions)	No samples were removed after 315 samples from initial data verification
Confirms all critical components are present – context, question, answer and start position. (Note: unanswerable questions show a “-1”)	No samples were removed after 315 samples from initial data verification
Max tokens set to 90% of 512 (Note: tokens reduced to 384 to improve memory and time efficiency (Wang, 2020)	No samples were removed after 315 samples from initial data verification
Ensures consistent padding / truncation	---

Summary of aspects that overlap

Both dataloaders output a dictionary of tensors: input_ids, attention_mask, start_positions, and end_positions. Both dataloaders add padding for situations where less than the allowable tokens are present. Both set the Sets the start and end positions to 0 for unanswerable questions.

Summary of differences

The “Fixed-Length Trunc” dataloader ensures consistent padding / truncation. The “Variable-Length Trunc” dataloader does not truncate the sequences to a fixed length. Both set the Sets the start and end positions to 0 for unanswerable questions; however the Dataset dataloader does this in a more complex way which could be contributing to errors that led to reduced accuracy of the model.

Appendix B: Data Loaders

The following table describes sample counts and distribution of question types for the “Variable-Length Trunc” class with Max Tokens set to 512:

Training Samples	Training – Question Type Distribution	Validation Samples	Validation– Question Type Distribution	Test Samples	Test– Question Type Distribution
113,000	what (64,957 | 57.5%) who (11,786 | 10.4%) how (11,616 | 10.3%) which (8,399 | 7.4%) when (7,570 | 6.7%) where (4,679 | 4.1%) other (2,320 | 2.1%) why (1,673 | 1.5%)	14,000	what (8,019 | 57.3%) who (1,458 | 10.4%) how (1,469 | 10.5%) which (1,011 | 7.2%) when (911 | 6.5%) where (565 | 4.0%) other (316 | 2.3%) why (251 | 1.8%)	14,000	what (8,084 | 57.7%) who (1,453 | 10.4%) how (1,400 | 10.0%) which (1,045 | 7.5%) when (946 | 6.8%) where (619 | 4.4%) other (278 | 2.0%) why (175 | 1.2%)

The following table describes sample counts and distribution of question types for the “Fixed-Length Trunc” class with Max Tokens set to 384:

Training Samples	Training – Question Type Distribution	Validation Samples	Validation– Question Type Distribution	Test Samples	Test– Question Type Distribution
113,506	what (65,250 | 57.5%) who (11,832 | 10.4%) how (11,677 | 10.3%) which (8,435 | 7.4%) when (7,603 | 6.7%) where (4,699 | 4.1%) other (2,330 | 2.1%) why (1,680 | 1.5%)	14,181	what (8,123 | 57.3%) who (1,477 | 10.4%) how (1,484 | 10.5%) which (1,021 | 7.2%) when (925 | 6.5%) where (572 | 4.0%) other (324 | 2.3%) why (255 | 1.8%)	14,190	what (8,188 | 57.7%) who (1,469 | 10.4%) how (1,422 | 10.0%) which (1,064 | 7.5%) when (958 | 6.8%) where (627 | 4.4%) other (283 | 2.0%) why (179 | 1.3%)

Using a Chi-square analysis, the distribution of each group – training, validation and test – were analyzed to determine if significant differences were present between the “Variable-Length Trunc” and “Fixed-Length Trunc” groups. Each group produced a very low chi-squared statistic (between 0.0058 and 0.0374) with a high p-value of 1.0 and 7 degrees of freedom. These results indicate the differences between the groups are due to chance, therefore there is no significant difference between the groups (Namuth-Covert, 2024)

Appendix C: Performance Metrics

All Exact Match (EM) and F-1 Scores reported in Appendix D were performed on normalized inputs (predicted and actual text). Normalization was performed to ensure the comparison captures the most meaningful elements of the text.

Normalization -- The following bullets capture aspects of normalization utilized in this report:

- Convert prediction and answer to lower case
- Remove string punctuation from prediction and answer (!"#$%&'()*+,-./:;<=>?@[]^_`{|}~)
- Remove articles "a", "an", and "the"
- Remove leading / training spaces, standardize to single space between words

Exact Match (Normalized) – Compares the normalized predicted text to the normalized truth text (i.e: answer). For an exact match, the function returns True; If not an exact match, the function returns False. The average EM score for all test data was reported as the “Test EM (Norm.)” for each model in Appendix D – Model Hyperparameters and Results

F-1 Score (Normalized) – Normalizes predicted text and truth text, then computes common tokens between the two. Computes precision ratio, recall ratio and F1 score with the following relationships:

$$\text{Precision} = \frac{\text{count of common tokens}}{\text{count of predicted tokens}} \qquad \text{Recall} = \frac{\text{count of common tokens}}{\text{count of truth tokens}} \qquad \text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The average F1 score for all test data was reported as the “Test F-1 (Norm.)” for each model in Appendix D

Exact match (EM) and F-1 scores are common performance metrics used in machine reading comprehension because together, they paint a picture of the model’s rate of producing perfectly accurate predictions and partial correct predictions. These metrics were presented in the literature examined in this paper, and they are presented for the models evaluated in this paper. However, one item that was not clear from the literature was whether or not models were normalized prior to running metrics. In this paper, we report normalized EM and F-1 scores for the model experiments. The best models (captured in the results section) also included non-normalized exact match and F-1 scores to give readers a baseline of these performance results.

Appendix D – Model Hyperparameters and Results

Table A: All models with “Variable-Length Trunc” dataloader using a Simple Model (ie: no added attention mechanism)

Model #	Split	Dropout	Optimizer	Learning Rate	Epochs	Test EM (Norm.)	Test F-1 (Norm.)	Training Time
2	Traditional (80/10/10)	0.15	RMSprop	$4x10^{-5}$	5	0.617	0.693	~ 107 minutes
3	Traditional (80/10/10)	0.25	RMSprop	$4x10^{-5}$	4	0.625	0.702	~86 minutes
4	Traditional (80/10/10)	0.225	RMSprop	$4x10^{-5}$	4	0.638	0.711	~86 minutes
5	Traditional (80/10/10)	0.18	RMSprop	$4x10^{-5}$	3	0.644	0.714	~64 minutes
6	Traditional-Half (40/10/10)	0.225	RMSprop	$4x10^{-5}$	4	0.594	0.671	~45 minutes
7	Traditional-Half (40/10/10)	0.18	RMSprop	$4x10^{-5}$	4	0.597	0.673	~45 minutes
13	Traditional (80/10/10)	0.225	AdamW	$4x10^{-5}$	4	0.643	0.713	~92 minutes
16	Traditional (80/10/10)	0.14	AdamW	$4x10^{-5}$	4	0.635	0.706	~88 minutes
22	Traditional (80/10/10)	0.18	AdamW	$4x10^{-5}$	3	0.636	0.704	~54 minutes
20	Traditional (80/10/10)	0.13	AdamW	4x10-5	4	0.600	0.685	~94 minutes

Table B: All models with “Fixed-Length Trunc” dataloader using a Simple Model (ie: no added attention mechanism)

Model #	Split	Dropout	Optimizer	Learning Rate	Epochs	Test EM (Norm.)	Test F-1 (Norm.)	Training Time
12	Traditional (80/10/10)	0.18	AdamW	4x10-5	3	0.766	0.817	~48 minutes
14	Traditional (80/10/10)	0.18	AdamW	$4x10^{-5}$	4	0.769	0.821	~64 minutes
15	Traditional (80/10/10)	0.16	AdamW	$4x10^{-5}$	4	0.764	0.817	~64 minutes
17	Traditional (80/10/10)	0.13	AdamW	$4x10^{-5}$	4	0.767	0.822	~64 minutes
19	Traditional (80/10/10)	0.13	AdamW	$4x10^{-5}$	4	0.762	0.810	~84 minutes

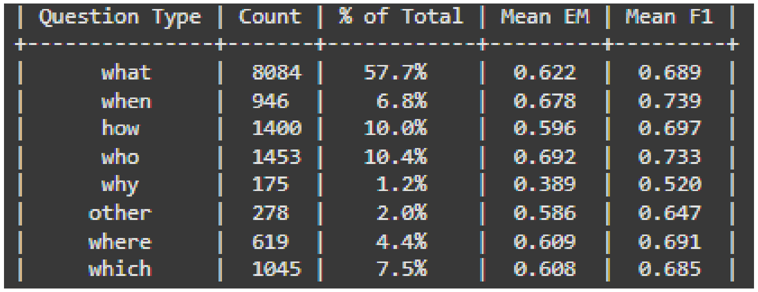
Appendix D – Model Hyperparameters and Results

Table C: All models with "Variable-Length Trunc" dataloader using a Complex model (ie: added attention mechanism)

Model #	Split	Number of heads	Dropout	Optimizer	Learning Rate	Epochs	Test EM (Norm.)	Test F-1 (Norm.)	Training Time
8	Traditional-Half (40/10/10)	8	0.18	RMSprop	$4x10^{-5}$	4	0.351	0.381	~ 56 minutes
9	Traditional (80/10/10)	12	0.18	RMSprop	$4x10^{-5}$	10	0.492	0.554	~250 minutes
10	Traditional (80/10/10)	12	0.17	RMSprop	$4x10^{-5}$	19	0.539	0.600	~475 minutes
23	Traditional (80/10/10)	2	0.18	AdamW	$4x10^{-5}$	15	0.410	0.446	~293 minutes

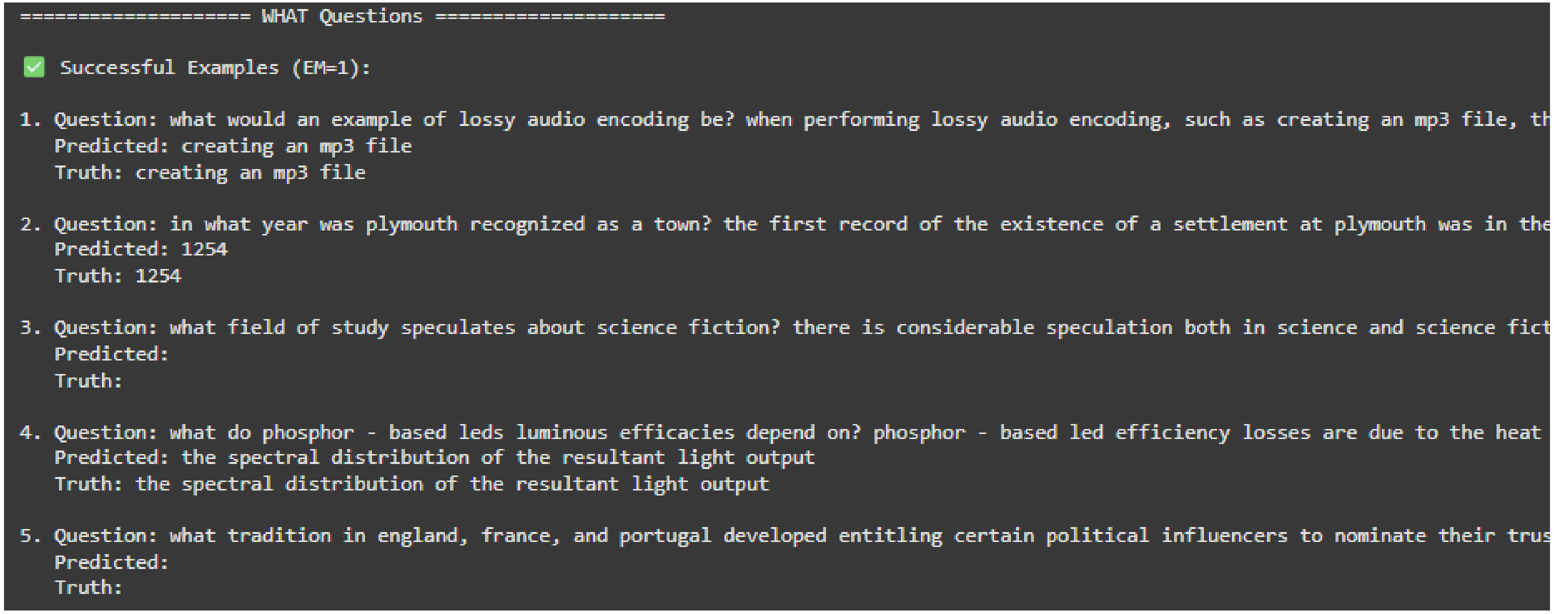
Appendix E - Sample Questions from “variable-length truncated” Data Loader

The following output describes EM and F-1 metrics by question type for the “Variable-Length truncated” data loader:



```
| Question Type | Count | % of Total | Mean EM | Mean F1 |
+---------------+-------+------------+---------+---------+
|      what     | 8084 |   57.7%    |  0.622  |  0.689  |
|      when     | 946  |    6.8%    |  0.678  |  0.739  |
|      how      | 1400 |   10.0%    |  0.596  |  0.697  |
|      who      | 1453 |   10.4%    |  0.692  |  0.733  |
|      why      | 175  |    1.2%    |  0.389  |  0.520  |
|     other     | 278  |    2.0%    |  0.586  |  0.647  |
|     where     | 619  |    4.4%    |  0.609  |  0.691  |
|     which     | 1045 |    7.5%    |  0.608  |  0.685  |
```

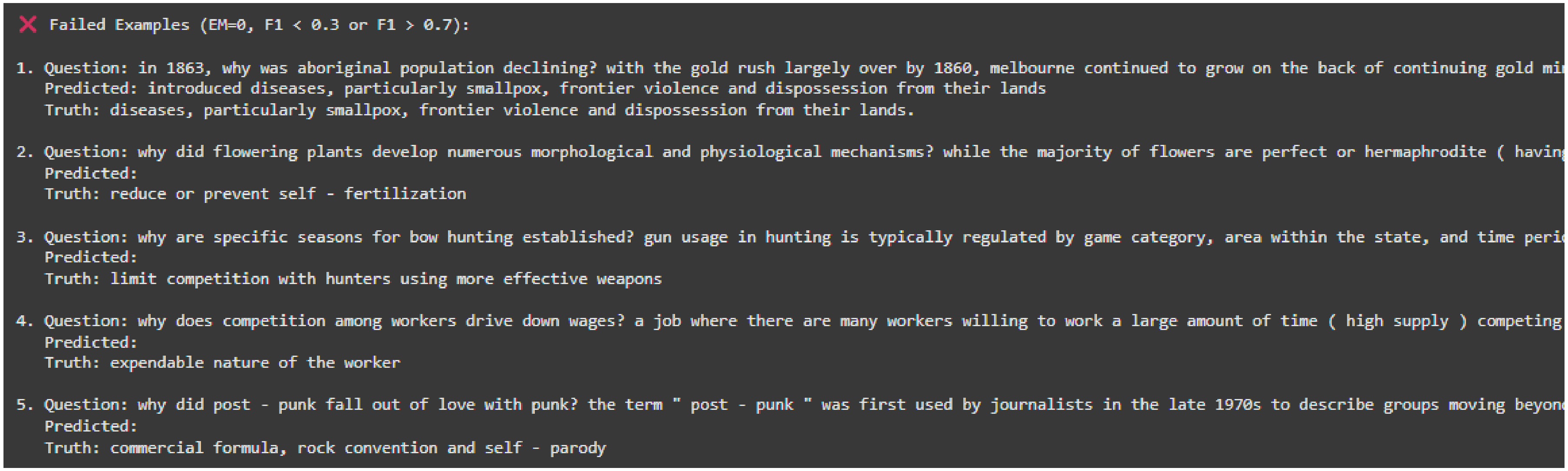
Review of the sample questions show these models retain accurate “Truth” content, and the models are capable of producing exact matches for answerable and unanswerable questions.



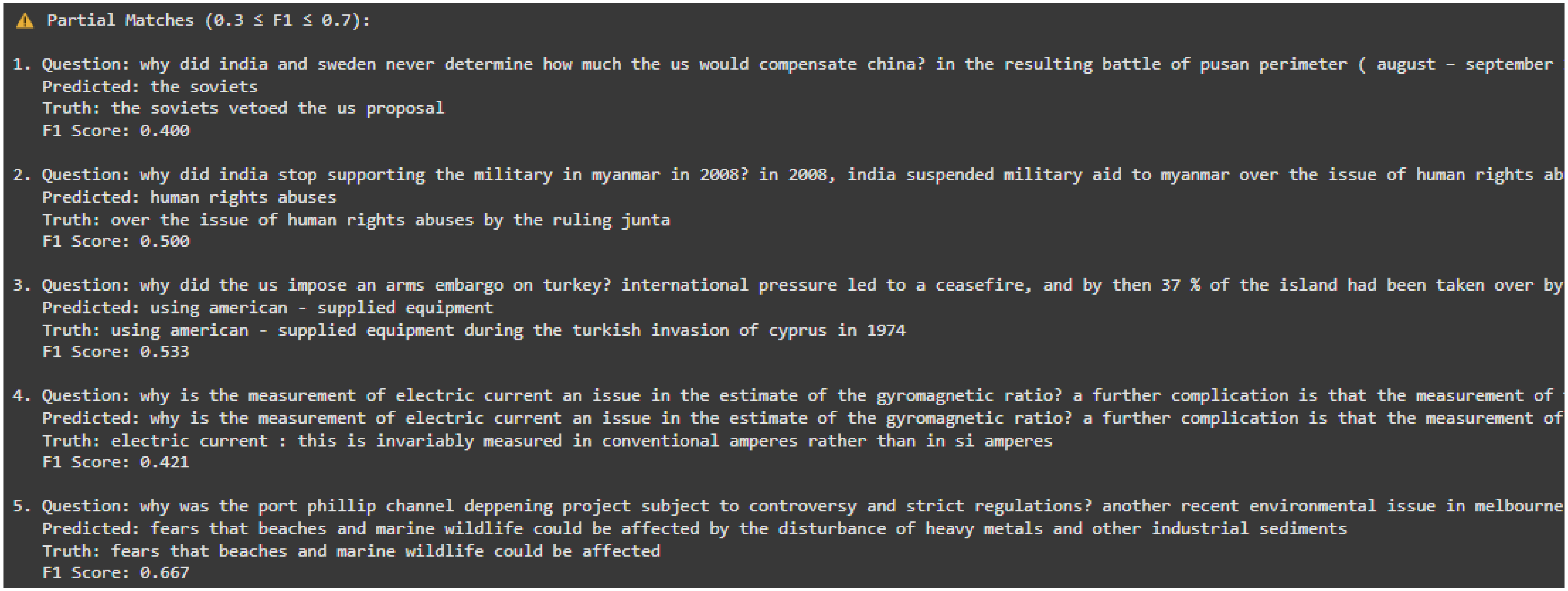
```
==================== WHAT Questions ====================

✅ Successful Examples (EM=1):

1. Question: what would an example of lossy audio encoding be? when performing lossy audio encoding, such as creating an mp3 file, th
   Predicted: creating an mp3 file
   Truth: creating an mp3 file

2. Question: in what year was plymouth recognized as a town? the first record of the existence of a settlement at plymouth was in the
   Predicted: 1254
   Truth: 1254

3. Question: what field of study speculates about science fiction? there is considerable speculation both in science and science fict
   Predicted:
   Truth:

4. Question: what do phosphor - based leds luminous efficacies depend on? phosphor - based led efficiency losses are due to the heat
   Predicted: the spectral distribution of the resultant light output
   Truth: the spectral distribution of the resultant light output

5. Question: what tradition in england, france, and portugal developed entitling certain political influencers to nominate their trus
   Predicted:
   Truth:
```

Based on the metrics for each question type, the “When” questions had the best performance (EM = 67.8% and F1 = 73.9%), and the “Why” questions had the worst performance (EM = 38.9% and F1 = 52.0%). A review of the failed examples show that 4 out of 5 “Why” responses produced no prediction and 4 out of 5 partial matches were either missing the tail end of the response or contained two much information on the tail end of the response. Poor performance of “Why” questions could have stemmed from the tiny fraction of “Why” questions available in the training data -- ~1.5%, the smallest percentage of all question types.

Appendix E - Sample Questions from “variable-length truncated” Data Loader



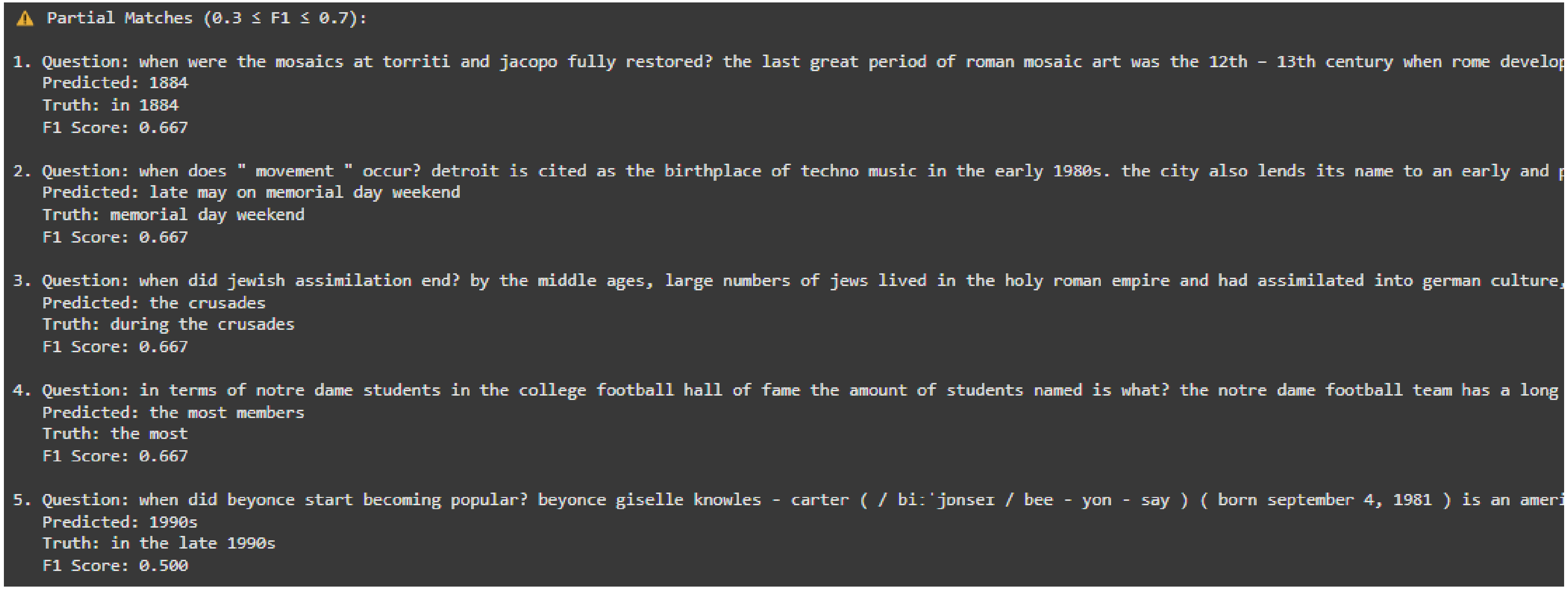
```
❌ Failed Examples (EM=0, F1 < 0.3 or F1 > 0.7):

1. Question: in 1863, why was aboriginal population declining? with the gold rush largely over by 1860, melbourne continued to grow on the back of continuing gold mi
   Predicted: introduced diseases, particularly smallpox, frontier violence and dispossession from their lands
   Truth: diseases, particularly smallpox, frontier violence and dispossession from their lands.

2. Question: why did flowering plants develop numerous morphological and physiological mechanisms? while the majority of flowers are perfect or hermaphrodite ( having
   Predicted:
   Truth: reduce or prevent self - fertilization

3. Question: why are specific seasons for bow hunting established? gun usage in hunting is typically regulated by game category, area within the state, and time peri
   Predicted:
   Truth: limit competition with hunters using more effective weapons

4. Question: why does competition among workers drive down wages? a job where there are many workers willing to work a large amount of time ( high supply ) competing
   Predicted:
   Truth: expendable nature of the worker

5. Question: why did post - punk fall out of love with punk? the term " post - punk " was first used by journalists in the late 1970s to describe groups moving beyon
   Predicted:
   Truth: commercial formula, rock convention and self - parody
```



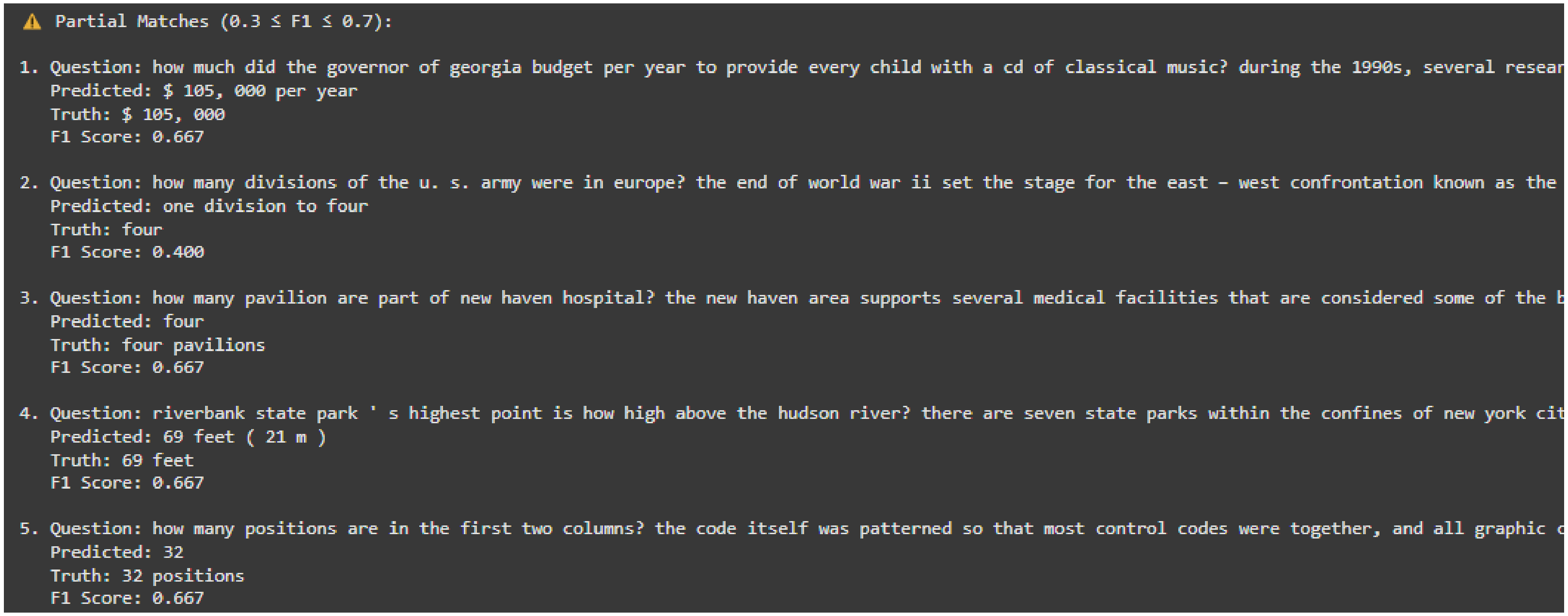
```
⚠ Partial Matches (0.3 ≤ F1 ≤ 0.7):

1. Question: why did india and sweden never determine how much the us would compensate china? in the resulting battle of pusan perimeter ( august – september
   Predicted: the soviets
   Truth: the soviets vetoed the us proposal
   F1 Score: 0.400

2. Question: why did india stop supporting the military in myanmar in 2008? in 2008, india suspended military aid to myanmar over the issue of human rights ab
   Predicted: human rights abuses
   Truth: over the issue of human rights abuses by the ruling junta
   F1 Score: 0.500

3. Question: why did the us impose an arms embargo on turkey? international pressure led to a ceasefire, and by then 37 % of the island had been taken over by
   Predicted: using american - supplied equipment
   Truth: using american - supplied equipment during the turkish invasion of cyprus in 1974
   F1 Score: 0.533

4. Question: why is the measurement of electric current an issue in the estimate of the gyromagnetic ratio? a further complication is that the measurement of
   Predicted: why is the measurement of electric current an issue in the estimate of the gyromagnetic ratio? a further complication is that the measurement of
   Truth: electric current : this is invariably measured in conventional amperes rather than in si amperes
   F1 Score: 0.421

5. Question: why was the port phillip channel deppening project subject to controversy and strict regulations? another recent environmental issue in melbourne
   Predicted: fears that beaches and marine wildlife could be affected by the disturbance of heavy metals and other industrial sediments
   Truth: fears that beaches and marine wildlife could be affected
   F1 Score: 0.667
```

Appendix E - Sample Questions from “variable-length truncated” Data Loader

Partial matches for many of the categories were semantically equivalent, and a human reader would have likely accepted the predicted answer as accurate – see samples below for partial matches of the “When” and “How” categories:



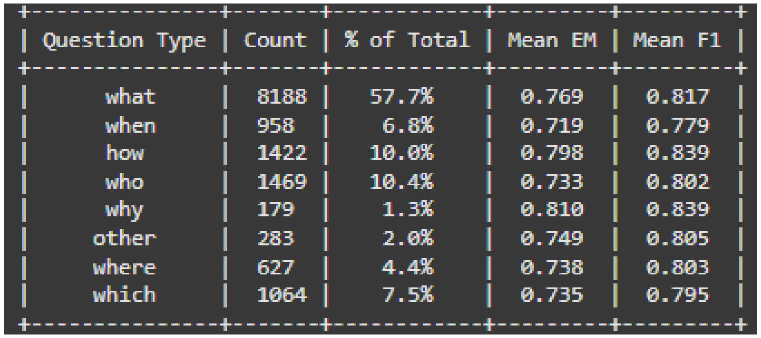
```
⚠ Partial Matches (0.3 ≤ F1 ≤ 0.7):

1. Question: when were the mosaics at torriti and jacopo fully restored? the last great period of roman mosaic art was the 12th – 13th century when rome develop
   Predicted: 1884
   Truth: in 1884
   F1 Score: 0.667

2. Question: when does " movement " occur? detroit is cited as the birthplace of techno music in the early 1980s. the city also lends its name to an early and p
   Predicted: late may on memorial day weekend
   Truth: memorial day weekend
   F1 Score: 0.667

3. Question: when did jewish assimilation end? by the middle ages, large numbers of jews lived in the holy roman empire and had assimilated into german culture,
   Predicted: the crusades
   Truth: during the crusades
   F1 Score: 0.667

4. Question: in terms of notre dame students in the college football hall of fame the amount of students named is what? the notre dame football team has a long
   Predicted: the most members
   Truth: the most
   F1 Score: 0.667

5. Question: when did beyonce start becoming popular? beyonce giselle knowles - carter ( / biːˈjonseɪ / bee - yon - say ) ( born september 4, 1981 ) is an ameri
   Predicted: 1990s
   Truth: in the late 1990s
   F1 Score: 0.500
```



```
⚠ Partial Matches (0.3 ≤ F1 ≤ 0.7):

1. Question: how much did the governor of georgia budget per year to provide every child with a cd of classical music? during the 1990s, several resear
   Predicted: $ 105, 000 per year
   Truth: $ 105, 000
   F1 Score: 0.667

2. Question: how many divisions of the u. s. army were in europe? the end of world war ii set the stage for the east – west confrontation known as the
   Predicted: one division to four
   Truth: four
   F1 Score: 0.400

3. Question: how many pavilion are part of new haven hospital? the new haven area supports several medical facilities that are considered some of the b
   Predicted: four
   Truth: four pavilions
   F1 Score: 0.667

4. Question: riverbank state park ' s highest point is how high above the hudson river? there are seven state parks within the confines of new york cit
   Predicted: 69 feet ( 21 m )
   Truth: 69 feet
   F1 Score: 0.667

5. Question: how many positions are in the first two columns? the code itself was patterned so that most control codes were together, and all graphic c
   Predicted: 32
   Truth: 32 positions
   F1 Score: 0.667
```

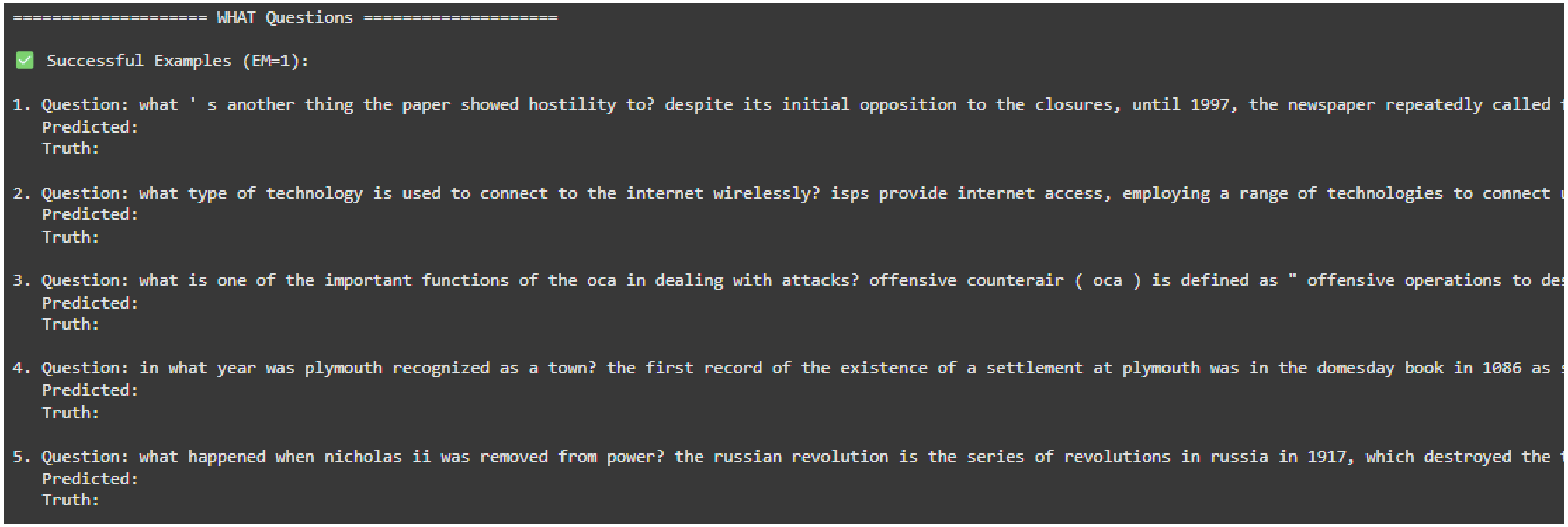
Appendix F - Sample Questions from “fixed-length truncated” Data Loader

The following output describes EM and F-1 metrics by question type for the “Fixed-Length truncated” data loader:



```
+---------------+-------+------------+---------+---------+
| Question Type | Count | % of Total | Mean EM | Mean F1 |
+---------------+-------+------------+---------+---------+
|      what     | 8188  |   57.7%    |  0.769  |  0.817  |
|      when     | 958   |    6.8%    |  0.719  |  0.779  |
|      how      | 1422  |   10.0%    |  0.798  |  0.839  |
|      who      | 1469  |   10.4%    |  0.733  |  0.802  |
|      why      | 179   |    1.3%    |  0.810  |  0.839  |
|     other     | 283   |    2.0%    |  0.749  |  0.805  |
|     where     | 627   |    4.4%    |  0.738  |  0.803  |
|     which     | 1064  |    7.5%    |  0.735  |  0.795  |
+---------------+-------+------------+---------+---------+
```

At first glance, the model appears to perform exceptionally well at identifying unanswerable questions. Below is a screenshot of the 5 sample with EM=1.0 for the “What” questions



```
=================== WHAT Questions ====================

✅ Successful Examples (EM=1):

1. Question: what ' s another thing the paper showed hostility to? despite its initial opposition to the closures, until 1997, the newspaper repeatedly called t
   Predicted:
   Truth:

2. Question: what type of technology is used to connect to the internet wirelessly? isps provide internet access, employing a range of technologies to connect u
   Predicted:
   Truth:

3. Question: what is one of the important functions of the oca in dealing with attacks? offensive counterair ( oca ) is defined as " offensive operations to des
   Predicted:
   Truth:

4. Question: in what year was plymouth recognized as a town? the first record of the existence of a settlement at plymouth was in the domesday book in 1086 as s
   Predicted:
   Truth:

5. Question: what happened when nicholas ii was removed from power? the russian revolution is the series of revolutions in russia in 1917, which destroyed the t
   Predicted:
   Truth:
```

However, in digging a bit deeper, the issue with inaccurate “Truth” answers becomes more obvious.

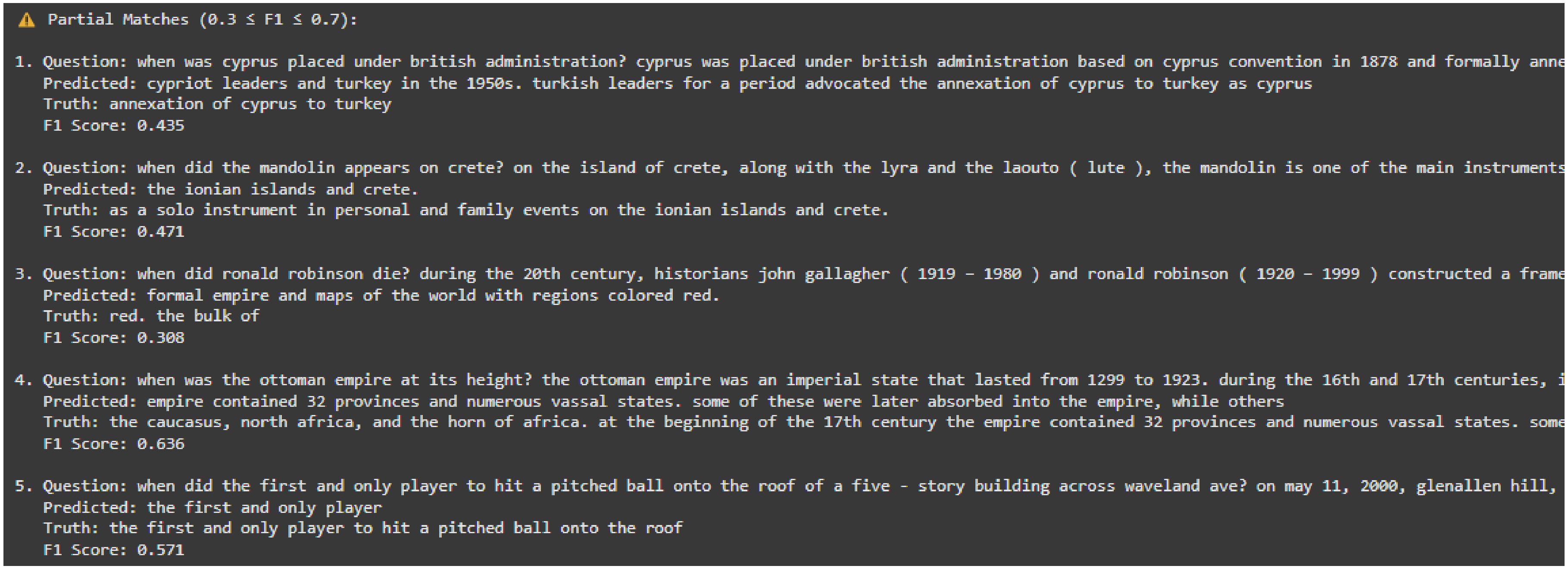
For example, the question: “*when did king louis of hungary approve the privilege of koszyce?*”

From the context we see the correct answer is “1374”: “*in 1374 king louis of hungary approved the privilege of koszyce (polish : " przywilej koszycki " or " ugoda koszycka ") in kosice in order to guarantee the polish throne for his daughter jadwiga. he broadened the definition of who was a member of the nobility and exempted the entire class from all but one tax (łanowy, which was limited to 2*

Appendix F - Sample Questions from “fixed-length truncated” Data Loader

grosze from łan (an old measure of land size)). in addition, the king ' s right to raise taxes was abolished ; no new taxes could be raised without the agreement of the nobility. henceforth, also, district offices (polish : " urzedy ziemskie ") were reserved exclusively for local nobility, as the privilege of koszyce forbade the king to grant official posts and major polish castles to foreign knights. finally, this privilege obliged the king to pay indemnities to nobles injured or taken captive during a war outside polish borders.”

However model predicted “*king louis of hungary approve*” and calculated an EM score of 1.0 because the model thought the Truth was “*king louis of hungary approve*”. The sample analysis shows dozens of examples where the “Truth” is not accurate (see screenshot below); Therefore, despite stellar EM and F-1 scores, it would be considered a failed model. The following screenshot shows more examples of inaccurate reported “Truth”:



```
⚠ Partial Matches (0.3 ≤ F1 ≤ 0.7):

1. Question: when was cyprus placed under british administration? cyprus was placed under british administration based on cyprus convention in 1878 and formally anne
   Predicted: cypriot leaders and turkey in the 1950s. turkish leaders for a period advocated the annexation of cyprus to turkey as cyprus
   Truth: annexation of cyprus to turkey
   F1 Score: 0.435

2. Question: when did the mandolin appears on crete? on the island of crete, along with the lyra and the laouto ( lute ), the mandolin is one of the main instruments
   Predicted: the ionian islands and crete.
   Truth: as a solo instrument in personal and family events on the ionian islands and crete.
   F1 Score: 0.471

3. Question: when did ronald robinson die? during the 20th century, historians john gallagher ( 1919 - 1980 ) and ronald robinson ( 1920 - 1999 ) constructed a frame
   Predicted: formal empire and maps of the world with regions colored red.
   Truth: red. the bulk of
   F1 Score: 0.308

4. Question: when was the ottoman empire at its height? the ottoman empire was an imperial state that lasted from 1299 to 1923. during the 16th and 17th centuries, i
   Predicted: empire contained 32 provinces and numerous vassal states. some of these were later absorbed into the empire, while others
   Truth: the caucasus, north africa, and the horn of africa. at the beginning of the 17th century the empire contained 32 provinces and numerous vassal states. some
   F1 Score: 0.636

5. Question: when did the first and only player to hit a pitched ball onto the roof of a five - story building across waveland ave? on may 11, 2000, glenallen hill,
   Predicted: the first and only player
   Truth: the first and only player to hit a pitched ball onto the roof
   F1 Score: 0.571
```