Enhancing SQuAD QA Model Generalization Through Quoref Coreference Training

Ramin Mohammadi
UT Austin
ramin.mohammadi@utexas.edu

Abstract

Models trained on the SQuAD reading comprehension dataset often rely on lexical overlap between questions and context, limiting their ability to generalize to QA tasks that require more challenging or different reasoning skills. Such behavior can be defined by dataset artifacts which is the reliance on patterned lexical cues to make predictions, hindering generalization capabilities. This work examines whether training on SQuAD mixed with Quoref—a dataset explicitly designed to require multi-sentence and coreferential reasoning, with minimal lexical shortcut cues—improves generalization for the ELECTRA-small model. I train three configurations (SQuAD-only, SQuAD+Quoref, and SQuAD+ContrastOnlyQuoref) for 3 epochs and evaluate them on SQuAD, Quoref, Quoref contrast samples, and a subset of Quoref's test set. Using data cartography based on per sample F1 mean and variance scores across 5 training epochs, SQuAD validation samples are labeled as *easy*, *ambiguous*, or *hard* to perform an analysis on per-label accuracy under each training configuration. While EM (Exact Match) and F1 scores on SQuAD remained practically unchanged, incorporating Quoref increases accuracy on ambiguous and hard SQuAD samples but reduces accuracy on easy SQuAD samples. In addition, EM and F1 scores on Quoref samples largely increased, expanding the model's generalization capabilities, but slightly lower performance on Quoref contrast samples compared to non contrast Quoref indicates dataset artifacts remain present. I show that training on Quoref samples requiring coreferential reasoning improves performance on hard and ambiguous SQuAD questions.

1 Introduction

The QA (Question-Answering) problem in NLP (Natural Language Processing) is given a context and question, predict the answer using information in the context. The baseline model in this experimentation was the ELECTRA model (Clark et al., 2020) trained on the SQuAD dataset (Rajpurkar et al., 2016), and the proposed approach is training the model on a combined dataset of SQuAD and Quoref samples.

Quoref is a reading comprehension dataset that requires resolving coreferences across multiple sentences and avoids including simple lexical-overlap cues that SQuAD frequently relies on. This design forces models to perform more complex reasoning rather than exploiting surface-level shortcuts (Roa et al., 2020). Due to Quoref having the same JSON format as SQuAD, but structured slightly differently, and both regarding the same QA task, Quoref was viewed as a "harder" subset of QA problems to enhance the performance on the SQuAD validation set.

The original intent was to find premade contrast samples for SQuAD to test the presence of dataset artifacts. To avoid having to manually create the contrast samples, exploration of the contrast set paper (Gardner et al., 2020) led to coming across the Quoref dataset, which the authors made a set of contrast samples for. The baseline SQuAD model had poor performance on the Quoref samples, as seen in Table 5, which reflected the flaw in the model's generalizability to other kinds of QA problems, representing that the baseline model performing well on its validation set did not imply that its a well trained QA model. QA models have to learn various kinds of analysis for reading comprehension questions, just as a human would, which motivated the idea of training on Quoref samples along with the SQuAD set to gain further reading comprehension skills on the QA task.

[*] Quoref Dataset: https://quoref-dataset.s3-us-west-2.amazonaws.com/train_and_dev/quoref-train-dev-v0.1.zip

[†] Quoref Contrast Set: https://github.com/allenai/contrast-sets/tree/main/quoref

Context:
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion ***Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title***. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Question:
Which NFL team won Super Bowl 50?

Prediction:
Carolina Panthers

References:
`['Denver Broncos', 'Denver Broncos', 'Denver Broncos']`

Figure 1: Baseline model incorrect prediction a SQuAD validation sample. Can see "Super Bowl 50" is stated at the beginning of the context but the lexical overlap of the question is in the next sentence and have to infer the answer as it is not directly restating the question in the context. Baseline is an ELECTRA model trained only on SQuAD's training set. Meanwhile, the model trained on SQuAD + QuoRef samples predicted this problem correctly reflecting improved reasoning on some SQuAD samples by training on Quoref samples.

Context:
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was ***played on February 7, 2016***, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Question:
What day was the Super Bowl played on?

Prediction:
February 7, 2016

References:
`['February 7, 2016', 'February 7', 'February 7, 2016']`

Figure 2: Baseline model correct prediction on a SQuAD validation sample. The answer is stated directly in the context (strong lexical overlap), requiring little reasoning effort beyond span extraction.

2 Related Work

2.1 Contrast Sets

Contrast sets (Gardner et al., 2020) reveal the presence of dataset artifacts. Contrast samples are manual perturbations to a test set which are small enough changes to a data sample that alter its ground truth label but preserves the lexical/syntactic artifacts that were present in the original example. A model may create a decision boundary that does not reflect the true decision boundary, so using enough test samples, you want to perturb every sample (in your subset) to fill in local gaps to the data distribution so the model's decision boundary can be evaluated. It's impossible to fill in all of the gaps to the data distribution so the idea is to perturb samples in multiple ways to acquire multiple variations of a sample to challenge and test the model's learned decision boundary in local spots of the distribution. If a model performs worse on the contrast set, it indicates that dataset artifacts are present in the model's weights.

The authors of the contrast set paper created contrast samples for various datasets including QA, sentiment analysis, multiple choice questions, etc. But, among the QA datasets: DROP, QUOREF, ROPES, BOOL, and MC-TACO, Quoref was the only one that had the same JSON format as SQuAD, including an answer_start label being the first index in the context where the answer is present. In addition, Quoref samples require understanding/references across multiple parts in a context passage vs. SQuAD which are more localized lexical overlap, reasoning questions. So, as seen in Table 5, the baseline model (ELECTRA trained on only SQuAD) samples performs poorly on Quoref eval sets. The challenging aspect of having to go from local reasoning (perform well on SQuAD samples) to across context reasoning (perform well on Quoref samples) motivated the idea of training on Quoref samples in addition to SQuAD, training on a harder subset of samples. Performance was evaluated training on SQuAD, SQuAD & Quoref, and SQuAD & contrast Quoref samples.

2.2 Data Cartography

Data cartography is the characterization of samples in a dataset with the following 3 labels: ambiguous, easy to learn, and hard to learn. The labels can be acquired by observing performance on a set of samples across multiple training epochs and then aggregating the results per sample by taking their mean and variance across the epochs (Swayamdipta et al., 2020). Data cartography here was utilized to analyze performance as seen in Tables 2 and 3. Further details on cartography are explained in the next section. Note, the cartography paper presents the findings that training on ambiguous examples promotes generalization, easy samples are important for convergence, and find that hard to learn samples are labeling errors, but it was mentioned that it their finding was model specific and to their NLI classification task, whereas here we're performing the QA task which is not a classification task and the metric used for cartography is F1 which is not as sensitive to a human mislabel error compared to a classification task since it measures at a token level. Also, samples in the SQuAD set are presented with multiple possible ground truth answers to help prevent the human mislabeling issue.

3 Method

3.1 Datasets

SQuAD and Quoref QA datasets were utilized. SQuAD was the baseline dataset and Quoref was the additional dataset that the ELECTRA model was trained on. In addition, from the contrast sets paper (Gardner et al., 2020), the authors handcrafted contrast samples on a subset of Quoref's test set. Both the contrast Quoref samples and the subset of Quoref test samples were utilized in this experimentation for training and evaluation.

The model was trained on 3 different training sets: (1) SQuAD, (2) both SQuAD and Quoref training samples, (3) SQuAD and Quoref contrast samples, each for 3 epochs. The SQuAD and Quoref mixes were concatenated to the same training set, so as seen in Table 1, (1) consisted of 87599 samples, (2) consisted of 87599 + 19399 samples, and (3) consisted of 87599 + 700 samples.

For evaluation, the following sets were used: (4) SQuAD validation set, (5) Quoref validation set, (6) contrast Quoref set, and (7) the subset of Quoref's test set that the contrast samples were based on. For the tables in this paper, (5) is denoted by "QuoRef" and (7) is referred to as "QuoRefOriginal". The ContrastOnlyQuoref and QuorefOriginal do not have a training/validation split as these are just a subset of samples from the Quoref dataset.

SQuAD and Quoref are the same kind of QA task which consists of a given context (some kind of passage), a question, an array of possible answers, and an answer_start attribute which indicates the starting index in the context where the answer first appears. But, their JSON files have a different structure, so the Quoref JSON training and validation set files were re-written in the same structure as SQuAD's using a Python script in order for both datasets to be handled similarly during training and evaluation.

3.2 Cartography Labeling SQuAD Validation

The SQuAD validation set was categorized into 3 labels: ambiguous, easy, and hard. The metric for determining these classifications was the per sample F1 mean and variance across 5 training epochs.

More specifically, using the baseline model which is trained only on SQuAD training samples, after every epoch, the model makes predictions on the SQuAD validation set and the F1 score is computed for every sample using the model's predicted answer to the question and the possible ground truth answers. Each question can have multiple possible correct answers so the F1 score for a sample is the max among the correct answers. After every epoch, the scores were written to a CSV file with the sample's id, epoch, and F1 score. Then, after training for 5 epochs, using the CSV data, the mean F1 score and F1 variance was computed for each sample where the population was the individual sample's F1 score in each epoch (compute the F1 mean and variance across the epochs for each SQuAD validation sample).

The cartography paper (Swayamdipta et al., 2020) analyzed datasets corresponding to the NLI classification task and used the metric "confidence" which captured how confidently the learner assigns the true label to the observation based on its probability distribution, being the mean probability of the true label across epochs during training. Here,the F1 score was used instead of confidence to suit the QA task which is not a classification task. The F1 score is used in QA tasks to measure the overlap between the predicted tokens and ground truth tokens. So, just like the cartography paper but in the perspective of F1 as the metric, ambiguous is defined as samples with high F1 variance, easy to learn are samples with high mean F1 scores and low F1 variance, and hard to learn are samples with low F1 means and low F1 variance. But the threshold for classifying these labels were not specified in the paper.

Two approaches were used for classifying the

Dataset	Train	Validation
SQuAD	87,599	10,570
Quoref	19,399	2,418
ContrastOnlyQuoref	–	700
QuorefOriginal	–	415

Table 1: Dataset splits used for training and validation across SQuAD and Quoref variants. ContrastOnlyQuoref and QuorefOriginal are subsets of the Quoref dataset so they do not have a data split. ContrastOnlyQuoref was used for training and evaluation.

Training Set	Ambiguous	Easy	Hard
SQuAD (Baseline)	893/2331 (0.3831)	6904/7571 (0.9119)	18/668 (0.0269)
SQuAD + Quoref	894/2331 (0.3835)	6848/7571 (0.9045)	44/668 (0.0659)
SQuAD + ContrastOnlyQuoref	928/2331 (0.3981)	6858/7571 (0.9058)	47/668 (0.0704)

Table 2: Manual Cartography Threshold Labeling: Per-label accuracy (fraction and proportion) on ambiguous, easy, and hard subsets of the SQuAD validation set across different training dataset configurations. Fractions represent the number of correct samples / total samples for that label. A prediction was "correct" if the predicted answer was an exact match with one of the possible ground truth answers.

samples using their F1 mean and variance which can be seen in Figure 5. The first approach was manually setting thresholds: mean >= 0.8 and <= 0.01 variance was easy, mean <= 0.2 and variance <= 0.01 was hard, and else ambiguous. The second approach was using K-means k=3 to partition the samples where the center with the max F1 mean coordinate was easy (coordinates were F1 mean and variance), min F1 mean coordinate was hard, and else the last centroid was assigned as ambiguous. The labels for each approach were written to a CSV file with the sample's id and label so during evaluation, the labels could be acquired. The results for the number of samples correct per label using each approach (manual thresholding and K-means) can be seen in Tables 2 and 3. This cartography mapping was only analyzed on the SQuAD validation set and labeling was in the perspective of the baseline (ELECTRA trained only on SQuAD).

4 Results

Training ELECTRA on SQuAD and Quoref samples, rather than SQuAD alone, led to higher accuracy on ambiguous and hard samples, but reduced accuracy on easy samples for the SQuAD validation set. This trend becomes even more pronounced, with even fewer incorrect predictions on easy samples, when the model is trained on only the 700 Quoref contrast samples in addition to SQuAD. This can be seen in Tables 2 and 3. Due to the cross-sentence and multi-entity reasoning required by Quoref, the model exhibits reduced performance on easy SQuAD samples, which primarily rely on simpler lexical-overlap cues. The model appeared to have exchanged some of the behaviors it learned from SQuAD for improved performance on Quoref-style reasoning. As shown in Table 5, SQuAD+Quoref training leads to significantly higher scores on Quoref evaluations, while performance on the SQuAD validation set decreased very slightly. The SQuAD+ContrastOnlyQuoref model even answers 18 more SQuAD validation questions correctly than the baseline where “correct” means the predicted answer exactly matches one of the ground truth answers.

The nature of the easy samples being simple lexical overlap of the question and answer in the context conflicts with Quoref samples. But, in return, the model showed improved accuracy on the more complex SQuAD samples, particularly those labeled ambiguous or hard, which aligns with the proposed hypothesis of incorporating Quoref. Though, training on the SQuAD+Quoref sets did not result in universally beneficial reasoning for the QA task as both the baseline and Quoref models had the same incorrect predictions for many samples in the SQuAD validation set, as seen by their identical EM and F1 scores in Table 4.

Its worth noting that for some samples, the predictions were the right intended answer but differed by some tokens such as punctuation, so its possible the cartography per label accuracies in Tables 2 and 3 are hallucinations.

Overall, although EM and F1 scores on the SQuAD validation set remain relatively unchanged, the model exhibits substantial gains on the Quoref

Training Set	**Ambiguous**	**Easy**	**Hard**
SQuAD (Baseline)	415/848 (0.4894)	7352/8576 (0.8573)	48/1146 (0.0419)
SQuAD + Quoref	415/848 (0.4894)	7280/8576 (0.8489)	91/1146 (0.0794)
SQuAD + ContrastOnlyQuoref	419/848 (0.4941)	7302/8576 (0.8514)	112/1146 (0.0977)

Table 3: Kmeans k=3 Cartography Labelling: Per-label accuracy (fraction and proportion) on ambiguous, easy, and hard subsets of the SQuAD validation set across different training dataset configurations. Fractions represent the number of correct samples / total samples for that label. A prediction was "correct" if the predicted answer was an exact match with one of the possible ground truth answers.

Training Set	**EM**	**F1**	**Correct**	**Incorrect**	**Accuracy**
SQuAD (Baseline)	77.4	85.4	7815	2755	0.7394
SQuAD + Quoref	77.1	85.2	7786	2784	0.7366
SQuAD + ContrastOnlyQuoref	77.4	85.3	7833	2737	0.7411

Table 4: Exact Match (EM), F1, and correctness statistics (number of correct & incorrect samples and accuracy) on the SQuAD validation set for models trained on three different dataset configurations.

evaluations. This indicates a broader improvement in QA capability: beyond the short-span, lexical-matching behavior encouraged by SQuAD, the model acquires stronger cross-sentence and coreference reasoning skills from Quoref. However, the SQuAD+Quoref model performs worse on the Quoref contrast set than on the original Quoref samples, suggesting that certain dataset artifacts are still present.

5 Conclusion

In this work, I explored how training on both SQuAD and Quoref samples can improve accuracy on ambiguous and hard to learn SQuAD samples, while simultaneously trading off accuracy on easy samples. In addition, performance on the SQuAD set minimally decreased while gaining substantial increased performance on Quoref coreference reasoning QA problems, enhancing generalization capabilities. Data cartography was performed on the SQuAD validation set by computing each sample's F1 mean and variance across training epochs with the baseline SQuAD training set, providing insight to how different types of questions were affected by the additional Quoref sample training.

Although the baseline SQuAD-trained ELECTRA achieved strong SQuAD validation performance, it failed on many Quoref QA problems, demonstrating reliance on dataset artifacts such as lexical proximity between question and answer spans which are minimally present in Quoref samples. The SQuAD-Quoref trained models gaining increased accuracy on ambiguous and hard to learn samples but decreased accuracy on easy samples may imply that the model just learned a different set of dataset artifacts due to no net improvements. In addition, the SQuAD-Quoref trained model's lower performance on the Quoref contrast samples compared to non contrast Quoref samples indicates that dataset artifacts remain present.

While Quoref training improved reasoning on harder samples, persistent dataset artifacts continue to constrain the model's ability to generalize reliably. Future work may explore training on ambiguous Quoref or SQuAD samples, since ambiguous samples can lead to increased generalization and exploring whether the SQuAD samples labeled as hard to learn were actually human mislabeled errors as inferred in the cartography paper for the NLI dataset that the authors analyzed (Swayamdipta et al., 2020).

References

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah A Smith. 2020. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Pradeep Dasigi Roa, Matt Gardner Singh, and 1 others. 2020. Quoref: A reading comprehension dataset with

	Validation Sets							
Training Set	**SQuAD**		**QuoRef**		**QuoRefContrast**		**QuoRefOriginal**	
	EM	F1	EM	F1	EM	F1	EM	F1
SQuAD (Baseline)	77.4	85.4	18.0	27.4	21.3	31.2	18.8	28.6
SQuAD + Quoref	77.1	85.2	60.9	65.5	47.4	53.1	63.9	68.7

Table 5: EM and F1 scores for ELECTRA trained on 2 different training set configurations, evaluated across SQuAD, QuoRef, QuoRefContrast, and QuoRefOriginal validation sets. Note the state of the art F1 score on the Quoref dataset is 70.5 (Roa et al., 2020).

Training Set	**Ambiguous**	**Easy**	**Hard**	**EM**	**F1**
SQuAD (3 epochs)	893/2331 (0.3831)	6904/7571 (0.9119)	18/668 (0.0269)	77.4	85.4
SQuAD (5 epochs)	870/2331 (0.3732)	6943/7571 (0.9171)	0/668 (0.0000)	77.6	85.8

Table 6: Effect of training duration (3 vs. 5 epochs) on per-label accuracy (ambiguous, easy, hard) on manual cartography threshold labels and SQuAD validation performance (Exact Match and F1). Demonstrates that training for too many epochs leads to overfitting, characterized by increased accuracy on easy samples but reduced performance on ambiguous and hard samples. Cartography labeling here corresponds to manual thresholding.

questions requiring coreferential reasoning. *arXiv preprint arXiv:1908.05803*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

Context:
At Montgomery Advertising in New York City, Duke Crawford is having trouble handling the account of cosmetics manufacturer ***Michele Bennett***, one of the company's most important clients—and his former fiancée. Still determined to win him back, Michele refuses to sign a contract until Duke reciprocates her affection. When Duke threatens to quit Michele's account, his boss James Montgomery assigns him to do the book promotion for a new client, a nerve psychologist named J.O. Loring.
While taking a taxi to the psychologist's office, Duke shaves with an electric razor he invented, but his nervousness and stress result in leaving half his mustache intact. When he arrives at the client's office, Duke discovers that J.O. Loring is in fact an attractive woman named Jo. Staring at the half a mustache, Jo mistakes him for one of her mentally disturbed patients. Determines to prove to himself that he is anesthetized from women, he kisses the doctor. Jo reacts by recommending that he read her book on stress relief titled *Let's Live a Little*. Later that night, Duke is unable to fall asleep.
The next morning, after *Duke* makes an appointment to see Jo as her patient, *Jo advises him* that if he wants his former fiancée to sign the contract, he must ***wine and dine her***. Following her advice, *Duke* arranges ***a date with Michele*** at a nightclub. Wanting to observe the encounter for scientific reasons, Jo arrives at the nightclub with her stuffy surgeon boyfriend, Dr. Richard Field. When Michele notices that Duke and Jo are falling in love, and when she is served a cake with an advertising contract inside instead of a marriage license, she throws his drink at him and storms out of the nightclub. Duke is reduced to a nerve-wracked state—repeating ad slogans over and over.

Question:
What is the full name of the person Jo recommends that Duke must wine and dine?

Prediction:
J.O. Loring

References:
`['Michele Bennett']`

Figure 3: Baseline model incorrect prediction on a Quoref validation example requiring cross-sentence reasoning and coreference resolution. The question refers to the person Jo recommends Duke should *wine and dine her*, which must be linked to *a date with Michele* in the following sentence and then back to *Michele Bennett* in the opening sentence to recover the full name. The baseline instead predicts the closer but incorrect mention *J.O. Loring*.

Context:
After the December 7, 1941, attack on Pearl Harbor, the United States Government swiftly moved to begin solving the "Japanese Problem" on the West Coast of the United States. In the evening hours of that same day, the Federal Bureau of Investigation (FBI) arrested selected "enemy" aliens, including more than 5,500 Issei men. The California government pressed for action by the national government, as many citizens were alarmed about potential activities by people of Japanese descent.
On February 19, 1942, President Franklin D. Roosevelt signed ***Executive Order 9066***, which authorized the Secretary of War to designate military commanders to prescribe military areas and to exclude "any or all persons" from such areas. The order also authorized the construction of what would later be called "relocation centers" by the War Relocation Authority (WRA) to house those who were to be excluded. This *order* resulted in the ***forced relocation of over 120,000 Japanese Americans***, two-thirds of whom were native-born American citizens. The rest had been prevented from becoming citizens by federal law. Over 110,000 were incarcerated in the ten concentration camps located far inland and away from the coast. Manzanar was the first of the ten concentration camps to be established. Initially, it was a temporary "reception center", known as the Owens Valley Reception Center from March 21, 1942, to May 31, 1942. At that time, it was operated by the US Army's Wartime Civilian Control Administration (WCCA). The Owens Valley Reception Center was transferred to the WRA on June 1, 1942, and officially became the "Manzanar War Relocation Center." The first Japanese American incarcerees to arrive at Manzanar were volunteers who helped build the camp. By mid–April, up to 1,000 Japanese Americans were arriving daily, and by July, the population of the camp neared 10,000. Over 90 percent of the incarcerees were from the Los Angeles area, with the rest coming from Stockton, California; and Bainbridge Island, Washington. Many were farmers and fishermen. Manzanar held 10,046 incarcerees at its peak, and a total of 11,070 people were incarcerated there.

Question:
What order resulted in forced relocation of over 120,000 Japanese Americans?

Prediction:
Executive Order 9066

References:
`['Executive Order 9066']`

Figure 4: Correct prediction by the baseline model on a Quoref validation sample. Although Quoref typically requires reasoning across sentences, this example resembles an easy SQuAD-style question in which the answer is stated in the context and requires minimal reasoning beyond span extraction.

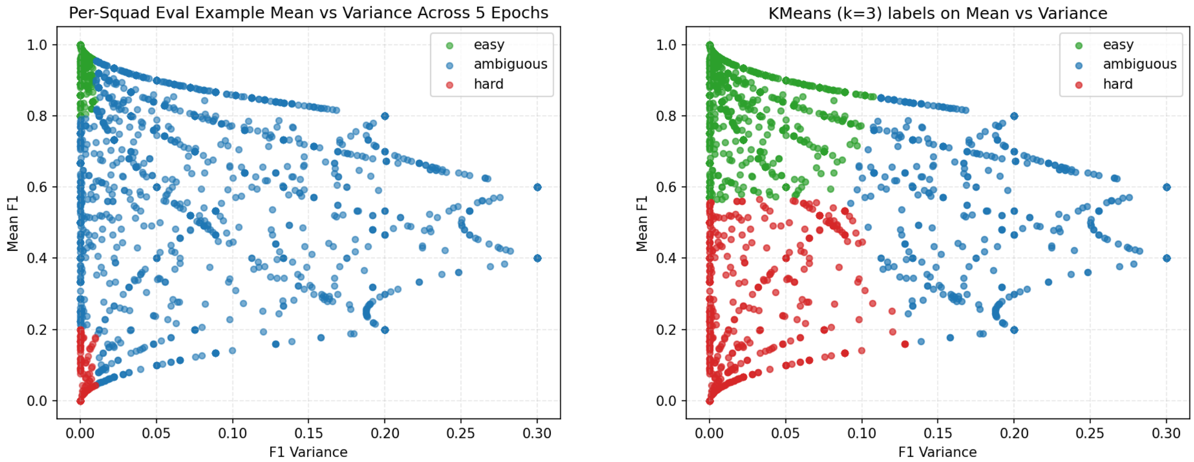

Figure 5: Cartography mapping of SQuAD validation samples based on per-sample F1 mean and F1 variance across five training epochs. Samples with high F1 mean and low variance cluster into the easy region, low-mean/low-variance samples form the hard region, and high-variance samples occupy the ambiguous region. These clusters illustrate the distribution of sample difficulty and model stability during learning. The left scatterplot reflects labeling using manually set thresholds on the mean and variance and the right scatterplot is labeling using K-means clustering and assigning the centroid with the largest F1 mean coordinate as easy, smallest as hard, and else centroid as ambiguous. Both labeling approaches were used for analysis 2 3

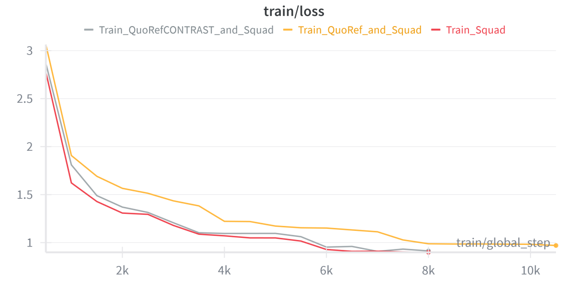

Figure 6: ELECTRA training loss across 3 epochs for model trained on 3 different training sets.

Context:
The game's media day, which was typically *held on* the *Tuesday* afternoon prior to the game, was moved to the ***Monday evening*** and re-branded as *Super Bowl* Opening Night. The event was held on February 1, 2016 at SAP Center in San Jose. Alongside the traditional media availabilities, the event featured an opening ceremony with player introductions on a replica of the Golden Gate Bridge.

Question:
What day of the week was Media Day held on for Super Bowl 50?

Prediction:
Tuesday

References:
`['Monday', 'Monday', 'Monday']`

Figure 7: Incorrect prediction on a SQuAD validation example by the model trained on SQuAD+Quoref which the SQuAD-only baseline model answered correctly. This error suggests that the SQuAD+Quoref model may be overshooting from learned behavior from Quoref samples or still relying on dataset artifacts and surface patterns such as "held on".