Data Augmentation Method for Fact Verification Using GPT-3

Stanford CS224N Custom Project

Jaehwan Jeong
Department of Computer Science
Stanford University
jaehwanj@stanford.edu

Patrick Ryan
Department of Mathematics
Stanford University
pmryan@stanford.edu

Harry Shin
Department of Computer Science
Stanford University
ds816@stanford.edu

Abstract

Fact verification is an NLP task that tries to verify a claim based on a retrieved evidence. Previous work in curating a dataset for fact verification mainly relied on manual labor [1]. Machine learning-based attempts to automatically generate datasets for the Recognizing Textual Entailment (RTE) step of fact verification also carry a limitation in that they adopt rule-based heuristics to create simple `NON-ENTAILMENT` samples [1, 2, 3]. We propose a novel data augmentation method for the RTE task where we generate an evidence-claim pair from a question-answer pair by exploiting GPT-3's pretrained knowledge to create non-entailing RTE samples [4]. We observe that our method generates sophisticated non-entailed samples as the evidence generated by GPT-3 often elaborates on only a portion of the input claim. That is, while both the evidence and the claim describe a similar phenomenon, their focuses are different to the extent that the evidence does not entail the claim as a whole. We also consider various methods such as false-claim generation and perplexity-based filtering to ensure that the augmented samples are non-entailing. Our experiment results show that augmenting FEVER and ANLI [5] with our proposed method improved the RTE classifier's performance on the dev sets.

1 Key Information to include

TA mentor: Anna Goldie. External collaborators: No. Sharing project: No.

2 Introduction

Along with the rapidly growing amount of information available comes the problem of verifying whether such information is reliable or not. This problem is most pronounced for textual information and central to discussions on "fake news", misinformation, disinformation, etc [6]. Thus, it is crucial to develop methods of *automatically* detecting whether a piece of information is reliable. Fact verification is an NLP task involving verification of a claim by retrieving evidence. This task can be broken down into two sub-tasks: information retrieval from a database and recognizing textual entailment (RTE). Conventionally, RTE is the task of labelling an evidence-claim pair according to whether the set of evidence entails, contradicts, or does not provide enough information to verify the claim. Standard datasets for RTE are created manually, an expensive and time-consuming operation.

In order to ensure that RTE datasets can be efficiently created across a wide array of domains, it is paramount to develop methods of automatically generating and labelling evidence-claim pairs. Prior machine-based data augmentation methods for fact verification rely on rule-based heuristics for generating contradiction examples, such as swapping entities in the claim of an entailing evidence-claim pair (E, C) to achieve a contradicting example (E, C'). We propose a novel data augmentation method, in which we feed question-answer pairs from a pre-existing dataset into GPT-3 to generate challenging contradicting evidence-claim pairs, relying on GPT-3's pretrained knowledge[1]. We find that ELECTRA Small achieves a boost of 3 percentage points in accuracy and 6 percentage points in F1 score against the ANLI dev set when trained using data generated from our proposed method.

3 Related Work

Dataset development for fact verification is an actively ongoing field of research. Fact Extraction and VERification dataset (FEVER) [1] continues to serve as a primary baseline for the fact verification task. The dataset evaluates the entire pipeline of fact verification where the model, given an input claim, retrieves evidence from a preprocessed Wikipedia dump and then performs a three-way RTE (`CONTRADICTION`, `ENTAILMENT`, `NEUTRAL`) between the evidence and the claim. The recently introduced dataset Fact Extraction and VERification over Unstructured and Structured Information (FEVEROUS) [2] also incorporates information present in a tabular format in its examples to build a more robust fact verification module. Adversarial Natural Language Inference [5] is a challenging RTE dataset created via an adversarial human-and-model-in-the-loop process that holds examples that were able to deceive the RTE classifier in the loop. Although all of the three datasets laid a crucial foundation in the field of fact verification, their main limitation is that they were manually created, and generating more fact verification samples following such methods will be a time-consuming and expensive task.

There have been attempts to develop an RTE dataset based on pre-existing datasets in the field of question answering as well. For instance, [7] proposes a method that transforms a given question-answer pair into a declarative claim and uses the passage given in parallel as the claim's evidence. [8] takes a step further and takes advantage of a question answering dataset that was originally meant for disambiguation. That is, given an ambiguous question (e.g. "When was the movie created?" might be asking either for the year the movie was filmed or released) and its disambiguated question-answer candidates, the authors propose to swap the questions and answers across different samples to generate challenging non-entailing samples. However, such methods' biggest shortcoming is that they also assume the availability of a manually gathered dataset where question-answer pairs and their relevant passages are given in parallel.

The authors of [3] propose a machine-based RTE data generation method that first extracts a question-answer pair from a given passage using a natural language generator and then repeats the claim generation process as outlined by [7]. Although the module is completely automated and therefore can produce RTE samples in a cost-efficient manner, the quality of its non-entailing samples may not be as challenging as those created by a human, due to the heuristics used to generate these samples (e.g. Named-Entity-Recognition based entity swapping and introduction of out-of-scope information).

Perplexity-based fact checking proposed by [9] serves as a useful algorithm for computing a rough numerical estimate of a given evidence-claim pair's credibility. Their research is based on the hypothesis that if we condition the perplexity score of a sample on its evidence, a non-entailing sample will naturally have a higher perplexity score than an entailing sample. Table 1 contains FEVER examples as well as their perplexity scores that we computed. We observe that `NON-ENTAILMENT` samples in FEVER yield higher perplexity scores than `ENTAILMENT` samples. We use this method to heuristically filter out potential noise from our generated dataset.

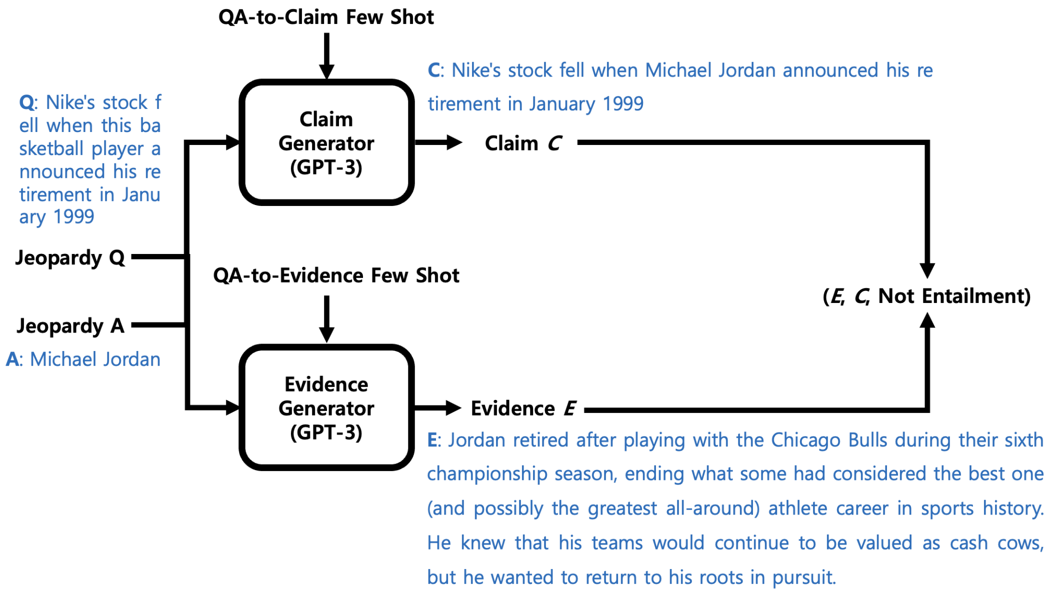
[1] Some might suggest that we should directly perform RTE using GPT-3, instead of creating a dataset. However, notice that GPT-3 comes at the cost of sizable computational resources and inference time. Therefore, given realistic limitations of industry RTE applications, it is much more feasible to train a smaller language model like BERT that performs RTE with less overhead, and this paper aims to use GPT-3 to generate a higher-quality RTE dataset that can be reused.

Type	Evidence	Claim	Perplexity
ENTAILMENT 1	Born in Saint James , Trinidad and Tobago and raised in South Jamaica , Queens , New York , Minaj earned public attention...	Nicki Minaj was born in Trinidad and Tobago.	3.82
ENTAILMENT 2	Sierra Leone Sierra Leone became an independent Nation on 27 April 1961 from Britain...	Sierra Leone gained sovereignty from Britain in 1961.	14.76
NON-ENTAILMENT 1	Samsung Life Insurance is a South Korean multinational insurance company headquartered in Seoul...	Samsung Life Insurance is a multinational boy band.	55.85
NON-ENTAILMENT 2	Buzz Aldrin Buzz Aldrin (born Edwin Eugene Aldrin Jr. , January 20 , 1930) is an American engineer and former astronaut.	Buzz Aldrin failed astronaut training.	82.73

Table 1: Perplexity Scores for FEVER samples

4 Approach

We take the reverse approach of [3]: instead of generating a question-answer pair from a set of retrieved evidence and then generating a claim from the pair, we instead take a pre-existing question answer pair from the Jeopardy! dataset [10] and generate separately the evidence and claim by providing GPT-3 with a few shot samples of Q-A-Evidence and Q-A-Claim, respectively (see appendix for the few-shot prompts). The Q-A-Claim few-shot samples convert a question-answer pair into a declarative claim, following the method proposed by [7]. On the other hand, Q-A-Evidence few-shot samples explain the relationship between the given question and answer using external information (see Figure 1).



Q: Nike's stock fell when this basketball player announced his retirement in January 1999

A: Michael Jordan

C: Nike's stock fell when Michael Jordan announced his retirement in January 1999

E: Jordan retired after playing with the Chicago Bulls during their sixth championship season, ending what some had considered the best one (and possibly the greatest all-around) athlete career in sports history. He knew that his teams would continue to be valued as cash cows, but he wanted to return to his roots in pursuit.

Figure 1: Diagram of our proposed data augmentation procedure

We observe that even when we instruct GPT-3 to generate an entailing evidence-claim pair by providing several entailing samples as its few-shot prompt, oftentimes, the evidence generated at inference time does not entail the claim it generated. In fact, we have empirically observed that roughly 90% of samples created based on our method are non-entailing samples, due to GPT-3's behavior in the evidence generation step. The few-shot examples for evidence generation introduces out-of-scope information in the evidence, and we expect GPT-3 to incorporate its massive pretrained knowledge to generate an evidence in a similar fashion. Although GPT-3 indeed generates a high-quality narrative of the key idea or event described in the input, its output often does not preserve the semantic content of the original question-answer pair. In other words, GPT-3 is mistakenly leaving out some information present in the input during generation. Considering how GPT-3 creates claims that fully preserve the entire information of the input question-answer pair, the evidence naturally cannot entail the claim. We exploit this behavior to create semantically similar non-entailing samples.

For example, given the question "Nike's stock fell when this basketball player announced his retirement in January 1999" and answer "Michael Jordan" from the Jeopardy! dataset, GPT-3 generated a corresponding claim "Nike's stock fell when Michael Jordan announced his retirement in January 1999" and an evidence "Jordan retired after playing with the Chicago Bulls during their sixth championship season, ending what some had considered the best one (and possibly the greatest all-around) athlete career in sports history...." Note that the GPT-3 generated evidence does not address the "Nike's stock" in the original claim and instead elaborates on Jordan's retirement, thereby failing to entail the claim. Nonetheless, both the claim and the evidence describe the same subject, making this a challenging NON-ENTAILMENT sample. See appendix for additional examples GPT-3 generated.

Furthermore, we follow the perplexity calculation method described in [9] using an off-the-shelf GPT-2 model to compute the perplexity scores of the GPT-3 generated samples and filter out those with low perplexity, as they may potentially be entailing samples. Formally, let $X = \{x_{e_0}, \ldots, x_{e_E}, x_{c_0}, \ldots, x_{c_C}\}$, where E and C denote the number of evidence tokens and claim tokens, respectively. Then we calculate the perplexity of X by the formula

$$PPL\left(X\right) = \sqrt[C]{\prod_{i=1}^{C} \frac{1}{p\left(x_{c_i} \mid x_{c_0}, \ldots, x_{c_E}, \ldots, x_{c_{i-1}}\right)}}.$$

Note that this differs from typical calculations of complexity of a given tuple of tokens in that the conditional probabilities of the evidence tokens $p\left(x_i \mid x_{e_0}, \ldots, x_{e_{i-1}}\right)$ are not used in the calculation of the perplexity of X. See appendix for some low perplexity examples that we have filtered out.

Another method we use to ensure that the examples we generate are non-entailing is to transform the original question-answer pair QA into QA' where A' is an incorrect answer. We rely on the fact that Jeopardy! questions are created so that there is only one answer, ensuring a high likelihood that the claim created based on a different entity A' is incorrect. Here, we first generate an A' that is similar, but not identical, to the original A using GPT-3 (see appendix for the few-shot prompts). Then, we feed QA' into GPT-3 to generate both a claim and evidence. We observe that claim-extraction of QA' usually results in a false claim. However, the evidence generated from QA' rarely entails the claim. While it is true that GPT-3 can sometimes *hallucinate* non-existent or incorrect information [11], it is rarely able to hallucinate reliably to generate text that properly entails the false claim it generated. In some sense, GPT-3 can lie, but not well. We exploit this behavior to reliably generate non-entailing samples.

5 Experiments

5.1 Data

The baseline of our experiment will be two pre-existing fact verification datasets: FEVER and ANLI [1, 5]. We balance the number of ENTAILMENT and NON-ENTAILMENT samples in the datasets by removing ENTAILMENT examples for FEVER and NON-ENTAILMENT examples from ANLI. Since ANLI maintains a non-entailing NEUTRAL class in addition to FEVER's ENTAILMENT/CONTRADICTION split, we collapse the two labels into NON-ENTAILMENT for the ANLI train set. We also maintained an extra version of the ANLI dev set, E/N, containing entailing and neutral samples to see if our data augmentation methods help distinguish such samples as well. We decided to keep the dev sets separate so that we can better compare the model's performance against ANLI dev set to that against FEVER dev set, which only contains ENTAILMENT/CONTRADICTION. The full dataset distribution is collated in table 2.

We experiment how different augmentation methods improve model performance. We have two main data augmentation methods of interest: one is our proposed method using GPT-3 and the other is a

Class	FEVER	ANLI	GPT-2	GPT-3	FEVER DEV	ANLI E/C	ANLI E/N
ENTAILMENT	20,000	20,000	0	0	6,666	1,000	1,000
NON-ENTAILMENT	20,000	20,000	10,000	10,000	6,666	1,000	1,000
Total	40,000	40,000	10,000	10,000	13,332	2,000	2,000

Table 2: Dataset Distribution

"baseline" augmentation method where we use a Natural Language Generation model to generate non-entailing claims from evidence as a parallel experiment. For our proposed method, we use question-answer pairs from the Jeopardy! dataset [10] as the seed data to generate training samples as described in the previous section. For the conventional augmentation method, we use GPT-2 fine-tuned on the CONTRADICTION samples in FEVER to generate false claims from evidence and compare the augmentation method's impact on the model performance to ours. We generate 10, 000 augment NON-ENTAILMENT samples using each method.

In addition, we applied three filtering methods to our GPT-3-augmented data and trained them as well: mid perplexity-scored, high perplexity-scored, and false-answered-based filtering. For high perplexity-scored pairs, we only keep 5,000 GPT-3 samples with highest perplexity scores, and for mid perplexity scores, we only keep the middle 5,000 samples. For false-answered pairs, we use 5,000 samples whose claims were generated based on question and false-answer pairs.

[12].

5.2 Evaluation method

To test our data augmentation technique, we fine-tune the pretrained ELECTRA Small model on the various datasets described in the previous section. Note that the baseline datasets originally have three classes (CONTRADICTION, ENTAILMENT, and NEUTRAL) as does the conventional RTE dataset, but in our experiment we merge the NEUTRAL and CONTRADICTION classes to form a NON-ENTAILMENT class to perform an ENTAILMENT / NON-ENTAILMENT binary classification. We make this simplification for two reasons: the unavailability of evidence for the NEUTRAL pairs in FEVER and the realistic limitation of distinguishing NEUTRAL and CONTRADICTION samples generated by GPT-3.

We evaluate the models on three different test sets: FEVER, ANLI E/C, and ANLI E/N. The evaluation metrics we use are accuracy, precision, recall, and F1 on ENTAILMENT. In addition to model performance, we also evaluated the mean and standard deviation of perplexity scores of the datasets to see the types of data our proposed GPT-3 method generated.

5.3 Experimental details

We use GPT-3 Davinci for both claim and evidence generation with max token length of 80 for both tasks. For false answer generation, we use GPT-3 Curie with max token length of 10. We train ELECTRA Small for each dataset with batch size 32 and learning rate of 5e-5 for 15 epochs on a single GeForce RTX 2080GPU, taking about 2 hours to train on each. For the GPT-2 augmentation method, we trained GPT-2 with batch size 2 and learning rate of 5e-4 for 10 epochs on a single GeForce RTX2080GPU, taking about 2 hours to train. For perplexity calculation, we use a pre-trained GPT-2 model without any fine-tuning, following the method outlined by [9].

5.4 Results

Table 3 shows the recall, precision, accuracy, and F1 scores of ELECTRA Small models fine-tuned on the baseline FEVER and ANLI train sets as well as their GPT-2/GPT-3 augmented counterparts and evaluated on the FEVER, ANLI E/C, and ANLI E/N dev sets. In addition, the replaced GPT3 model is fine-tunened on the baseline train sets with half of their non-entailing samples replaced with the 10,000 GPT3 generated samples and therefore has the same amount of samples as the original baseline. This model's performance captures the quality of our GPT3-generated data compared to human-annotated data from FEVER and ANLI.

We observe that dataset augmentation with GPT-2/GPT-3 improved accuracy and F1 score across all three dev sets, with pronounced improvement of 3% accuracy on the ANLI dev set with GPT-3 augmented samples. It is worthy of note that our augmented models achieved a boost in F1 score by improving on recall while maintaining the same precision on predicting `ENTAILMENT`. This suggests that the data augmentation methods help the models better identify challenging entailing evidence-claim pairs. This is quite unexpected given that we provided more challenging non-entailing evidence-claim pairs.

Data	FEVER DEV	ANLI DEV E/C	ANLI DEV E/N
Baseline	0.922/**0.925**/0.924/0.923	0.598/0.547/0.552/0.571	0.602/0.614/0.612/0.608
Replaced GPT3	**0.947**/0.884/0.912/0.915	0.729/0.533/0.545/0.616	0.730/0.595/0.617/0.656
Baseline+GPT2	0.937/0.915/**0.925/0.926**	0.683/0.553/0.566/0.611	0.684/0.610/0.623/0.645
Baseline+GPT3	0.943/0.910/**0.925/0.926**	**0.738/0.554/0.572/0.633**	**0.736/0.618/0.641/ 0.672**

Table 3: Baseline vs Baseline with data augmentation (Recall / Precision / Accuracy / F1)

Table 4 shows the recall, precision, accuracy, and f1 scores of ELECTRA Small models fine-tuned on the baseline FEVER AND ANLI train sets that are augmented with three filtered versions of the GPT-3 augmented data: mid perplexity scored, high perplexity scored, and false-answered ones.

Among the GPT-3 generated samples, we see that samples with high perplexity scores yielded greater improvement on both accuracy and F1 score than those with average perplexity scores. In addition, we observe that the false-answer method on our GPT3-generated data, where we perform entity swapping on claims such as "Nitrogen makes up around 78% of the atmosphere, oxygen only about 20%" to turn it into "Nitrogen makes up around 78% of the atmosphere, hydrogen only about 20%," yields a boost in accuracy and F1 score on the ANLI E/C dev set over the two perplexity filtered data sets, which may be due to reduced noise from GPT-3 generated `ENTAILMENT` samples.

Data	FEVER DEV	ANLI DEV E/C	ANLI DEV E/N
Mid-PPL	0.933/0.914/0.923/0.923	**0.690**/0.548/0.560/0.611	0.685/0.608/0.622/0.644
High-PPL	**0.934/0.919/0.926/ 0.926**	0.686/0.550/0.563/0.611	**0.690**/0.610/0.625/**0.648**
False-Answer	0.931/0.914/0.922/0.922	0.678/**0.559/0.572/0.613**	0.671/**0.618/0.629**/0.643

Table 4: Different filtering methods for GPT-3 generated samples (Recall / Precision / Accuracy / F1)

Table 5 shows the perplexity scores of the various datasets on which we fine-tune ELECTRA Small. Note that for both FEVER and ANLI, the perplexity scores for the `ENTAILMENT` samples are lower than those for the `NON-ENTAILMENT` samples, as expected. The gap between the perplexity scores for the ANLI is small, suggesting how the dataset is a lot more challenging than FEVER. We can also observe how the perplexity scores for the GPT-3 generated samples are big in general, implying that the samples are more likely to be non-entailing. Note that our perplexity scores are computed using GPT-2, so GPT-2 augmented samples show relatively low perplexity scores as expected.

Data	`ENTAILMENT`	`NON-ENTAILMENT`
FEVER	19.75/29.41	28.81/49.33
ANLI	32.64/21.32	33.84/25.89
GPT-2	-	22.99/29.83
GPT-3 (Total)	-	65.42/54.55
GPT-3 (Mid)	-	60.78/24.11
GPT-3 (High)	-	96.77/53.78

Table 5: Perplexity scores of datasets (Mean / Standard Deviation)

6 Analysis

Our data augmentation method on the ANLI train set can incorporate information from multiple clauses to make accurate predictions despite a lack of explicit evidence of the claim. For example, our data augmented models were able to correctly classify the entailing evidence-claim pair "Evil under the sun is a video game released... the pc version was released in 2007, and the wii version one year later." and "The wii version came out in 2008", while the baseline model got it wrong as it failed to combine the information "released in 2007" and "one year later" to deduce that the wii version came out in 2008. Furthermore, we observe the social and ethical impact of our research, with our data augmented models, for instance, correctly classifying the non-entailing evidence-claim pair "Passion play is a 2010 american drama film... executive produced by rebecca wang..." and "The executive producer was male." Although the evidence explicitly states that the film was produced by Rebecca Wang, a female producer, the original model trained on the baseline dataset incorrectly classified the sample to be `NON-ENTAILMENT` because of the gender bias of producers usually being male. We hope that our augmentation method can help address the implicit racial and gender bias that might be present in the datasets we use.

Regarding our perplexity-based filtering method, we observe that roughly 2/3 of generated examples with perplexity score less than 12 are in fact entailing, which is especially significant given that we observed that roughly 90% of generated examples are non-entailing. This provides support for the hypothesis that non-entailing samples have higher perplexity score than entailing samples, and it validates our perplexity-based filtering method. We also find that the examples filtered using this method tend to be very easy evidence-claim pairs, where the claim is typically a restatement of the evidence: "Newsweek's 'Transition' column features birth, marriage, divorce death announcements" and "The magazine's 'Transition' column features birth, marriage, divorce death announcements of celebrities" with a low perplexity score of 3.44. See appendix for more examples.

7 Conclusion

We have provided a novel method for automatically generating non-entailing evidence-claim pairs by transforming a question-answer pair into an evidence-claim pair using GPT-3. We also presented two methods for ensuring the high quality of our examples: one based on filtering samples by perplexity score and the other based on transforming the question-answer pair QA into a false pair QA' to generate a non-entailing evidence-claim pair. We find that our method improves model performance over a variety of baseline datasets by improving model recall while maintaining precision.

Limitations of our work include inability to filter out nonsense examples that GPT-3 generates. One possible method would be to filter by high-perplexity, since nonsense examples are likely to have high perplexity. However, this method is in tension with the hypothesis that high-perplexity examples are also non-entailing. Our preliminary results demonstrate that filtering both low- and high-perplexity examples does not work as well as just filtering low-perplexity examples. More experiments must be done to validate these results.

Another limitation of our research is that we rely on existing question-answer datasets, which may not be readily available in certain domains. In future work, we could do away with the seed question-answer pair, and instead begin with a piece of evidence E (as prior work has done, and which is readily accessible from, say, Wikipedia), extract a question-answer pair from it, and then feed the question-answer pair into our pipeline to generate a non-entailing evidence-claim pair $E'C$. We note the irony of beginning with seed evidence to generate non-entailing evidence. Using GPT-3 to directly generate QA pairs may be an alternate solution as well.

References

[1] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification, 2018.

[2] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information, 2021.

[3] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. Zero-shot fact verification by claim generation. *CoRR*, abs/2105.14682, 2021.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.

Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[5] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020.

[6] Lucas Graves and Alexios Mantzarlis. Amid political spin and online misinformation, fact checking adapts. *The Political Quarterly*, 91(3):585–591, 2020.

[7] Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets, 2018.

[8] Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. Faviq: Fact verification from information-seeking questions. *CoRR*, abs/2107.02153, 2021.

[9] Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. Towards few-shot fact-checking via perplexity, 2021.

[10] J. Wollen. Jeopardy clue dataset. `https://github.com/jwolle1/jeopardy_clue_dataset.git`, 2021.

[11] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*, 2021.

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

A Appendix

Table 6: Example claim, evidence generated based on our method

Jeopardy! Q	Jeopardy! A	GPT-3 Claim	GPT-3 Evidence	Label
4 different species of bears live in Alaska: Kodiak, grizzly, black & this	Polar bears	There are 4 different species of bears in Alaska: Kodiak, grizzly, black and polar bears.	The largest brown bear subspecies native to North America along with being the world's largest land carnivore. The polar bear is found in the Arctic Circle and surrounding arctic regions including Alaska, Canada, parts of Russia, Greenland and Norway.	non-entailment
Nike's stock fell when this basketball player announced his retirement in January 1999	Michael Jordan	Nike's stock fell when Michael Jordan announced his retirement in January 1999	Jordan retired after playing with the Chicago Bulls during their sixth championship season, ending what some had considered the best one (and possibly the greatest all-around) athlete career in sports history. He knew that his teams would continue to be valued as cash cows, but he wanted to return to his roots in pursuit	non-entailment
In 1462 this printer known for movable type had to move out of Mainz	Johannes Gutenberg	In 1462 Johannes Gutenberg, printer known for movable type, had to move out of Mainz	In 1424, Johann Gutenberg was ready to start making copies of his Bible. In Mainz, Germany, he met with the church warden, Johann Fust, to discuss getting his money. Fust said that Gutenberg must pay his debts before starting the Bible.	non-entailment

Table 6: Example claim, evidence generated based on our method

Jeopardy! Q	Jeopardy! A	GPT-3 Claim	GPT-3 Evidence	Label
Vladimir Samsonov is touted as Europe's only hope against China in this game	Ping-pong	Vladimir Samsonov is touted as Europe's only hope against China in the game of Ping-pong	Samsonov, at 14 years old, has been hailed as the next "greatest thing" in table tennis since he was 10, becoming the youngest player in table tennis history to win a World Championship medal.	non-entailment
If you're triskaidekaphobic, you're afraid of this number, & not just on a Friday	thirteen	If you're triskaidekaphobic, you're afraid of the number thirteen, & not just on a Friday	While fear of the number 13 as a bad omen is a common jesting among superstitious people and generally considered imaginary, it arguably has its source in alphanumeric convenience. Of the numbers 1 through 12, the number 13 is the one that stands last	non-entailment
He auditioned for & won the part of Ron Weasley with a rap that he wrote	Rupert Grint	Rupert Grint (actor) auditioned for & won the part of Ron Weasley with a rap that he wrote.	After his first agent told Grint that he would never be good enough to make it as an actor, the actor wrote & rapped on the pavement outside his home in Harrow, testing his own confidence and convincing his mother to send his film & photo negative that had been taken by his father of him	non-entailment

Table 7: Few-shot prompt for claim generation

Few-shot Q	Few-shot A	Claim
For the last 8 years of his life, Galileo was under house arrest for espousing this man's theory	Copernicus	For the last 8 years of his life, Galileo was under house arrest for espousing Copernicus' theory
The city of Yuma in this state has a record average of 4,055 hours of sunshine each year	Arizona	The city of Yuma in Arizona has a record average of 4,055 hours of sunshine each year
Rajaraja I of the Cholas battled to take this Indian Ocean island now known for its tea	Ceylon (or Sri Lanka)	Rajaraja I of the Cholas battled to take Ceylon now known for its tea
Kayak is an example of this, a word that reads the same forwards & backwards	a palindrome	a kayak is a palindrome, a word that reads the same forwards & backwards

Table 8: Few-shot prompt for evidence generation

Few-shot Q	Few-shot A	Evidence
For the last 8 years of his life, Galileo was under house arrest for espousing this man's theory	Copernicus	Galileo's championing of heliocentrism suggested by Copernicus was met with opposition from within the Catholic Church and from some astronomers. Galileo was tried by the Inquisition, found "vehemently suspect of heresy", and forced to recant. He spent the rest of his life under house arrest.

Table 8: Few-shot prompt for evidence generation

Few-shot Q	Few-shot A	Evidence
The city of Yuma in this state has a record average of 4,055 hours of sunshine each year	Arizona	The sun is said to shine during about 90% of the daylight hours, making Yuma one of the sunniest places in the world. The city in Arizona receives the most recorded mean sunshine of anywhere on Earth (4,055), although the equipment used by the United States tends to provide higher sunshine estimates than the traditional Campbell–Stokes recorder.
Rajaraja I of the Cholas battled to take this Indian Ocean island now known for its tea	Ceylon (or Sri Lanka)	In 993, Rajaraja I invaded Sri Lanka, which is called Ila-mandalam in the Chola records. This invasion most probably happened during the reign of Mahinda V of Anuradhapura, who according to the Chulavamsa chronicle, had fled to Rohana (Ruhuna) in south-eastern Sri Lanka because of a military uprising.In 1017, Rajaraja's son Rajendra I completed the Chola conquest of Sri Lanka.
Kayak is an example of this, a word that reads the same forwards & backwards	a palindrome	A palindrome is a word, number, phrase, or other sequence of characters which reads the same backward as forward, such as madam, racecar, and kayak.

Table 9: Example GPT-3 generated samples with low perplexity scores

GPT-3 Claim	GPT-3 Evidence	PPL Score
Rivaner is a profitable type of grape planted in Luxembourg's Moselle Valley.	A cooperation in Luxembourg's Moselle Valley is called RJB (Rivaner Joppegauald), and that name is covered by a European Union-wide three-year wine trade mark. Rivaner is a profitable type of grape planted in Luxembourg's Moselle Valley. There are 173,425 hectares of vineyards in Europe (including Moldova.	2.23
The Boy Scouts of America was founded by Baden-Powell in 1907	The Scout movement was formally inaugurated on Brownsea Island in England in 1907 by Baden-Powell, Lieutenant General in the British Army, and founder of the Boy Scouts. In 1910 Baden-Powell wrote the first book on the Scout Movement titled "Scouting for Boys". The movement incorporated in the UK in February 1910, and shortly thereafter Baden-Powell formed.	3.61
In 1989, George H.W. Bush appointed Clarence Thomas to the U.S. Court of Appeals for the District of Columbia.	By 1989, Thomas, a former Department of Education employee, had become the Chairman of the Equal Employment Opportunity Commission. President George H. W. Bush nominated him, as a Republican, to be the second Black American appointed to the United States Court of Appeals for the District of Columbia Circuit, a federal appellate court.	3.84
"A Tale of Two Cities" starts as Dr. Alexander Manette is released after 18 years in the Bastille prison	"A Tale of Two Cities" by Charles Dickens, published in 1859, opens as Dr. Alexander Manette is released after eighteen years of imprisonment in the Bastille by the French revolutionist Marquis de Sade:.	4.41

Table 9: Example GPT-3 generated samples with low perplexity scores

GPT-3 Claim	GPT-3 Evidence	PPL Score
U.S. women working full-time earn 76 cents for every dollar their male counterparts make	The U.S. Department of Labor reported that women earn 76 to 84 cents for every dollar men earn, depending on the education level. Deborah Reed, a women's rights advocate, said, "You would think that we would have closed that gap by now when the laws equalizing opportunities have been in place for four decades."	5.91
Mexico City stands on the ruins of Tenochtitlan, the capital of the Aztec people.	Although the Aztecs were defeated by Cortés and the Spanish in the 16th century, Tenochtitlan was rebuilt into what is now Mexico City.	8.29

Table 10: Few-shot prompt for A' generation

Few-shot A	Few-shot A'
Michael Jordan	Lebron James
Copernicus	Newton
George Washington	Abraham Lincoln
Hamburger	Pizza
Harvard	Yale
Egypt	Morocco
Coffee	Tea
Computer	Television
Leonardo DiCaprio	Brad Pitt
Tom Cruise	George Clooney
Lionel Messi	Cristiano Ronaldo