Automatic Grading For Essay Questions Using Deep Learning Models

Mostafa Abdelsalam Mohamed Emam Abdelmonem Mansour
Mohamed Bahgat Mariem Ahmed

June 2024

Abstract

The increasing prevalence of online education has intensified the demand for efficient and scalable assessment methods, particularly for essay-based questions, which are time-consuming and subjective to grade manually. This paper presents an innovative approach to automate essay grading using fine-tuned large language models (LLMs) based on transformer architectures. By integrating techniques such as Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA), our system evaluates essay responses in both English and Arabic, leveraging semantic understanding to compare student answers against reference responses. A custom dataset, constructed through synthetic generation, real-world collection, and reformation of existing corpora like SQuAD and QuAC, supports robust training and evaluation. The proposed framework achieves high accuracy (up to 97%) in classifying factual correctness, offering instant, consistent, and reliable feedback. This work addresses challenges such as grading fairness, multilingual support, and educator workload, contributing to the advancement of AI-driven educational tools.

1 Introduction

The rapid evolution of online education has transformed traditional assessment methodologies, creating an urgent need for innovative solutions to enhance scalability, efficiency, and accuracy in evaluating student performance. Among these challenges, essay-based question grading is particularly labor-intensive and subjective, often demanding extensive time and expertise from educators. This challenge is amplified in languages such as Arabic, where intricate linguistic nuances and cultural context further complicate the evaluation process.

Advancements in artificial intelligence (AI), particularly within natural language processing (NLP) and deep learning, provide promising avenues to automate essay grading. By leveraging these technologies, it becomes possible to not only alleviate the workload on educators but also to deliver more consistent, immediate, and accurate feedback to students. This paper presents a senior project conducted at the Faculty of Computers and Artificial Intelligence, Benha University, entitled “Automatic Grading For Essay Questions Using Deep Learning Models.” Submitted in partial fulfillment of the requirements for a Bachelor’s degree in Computers and Artificial Intelligence in June 2024, this project aims to revolutionize the assessment paradigm through the integration of advanced transformer-based large language models (LLMs) into an online examination system.

Our approach exploits the powerful capabilities of LLMs, which are fine-tuned using techniques such as Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA), to automatically evaluate essay responses in both Arabic and English. By comparing student answers against reference responses within a specific context, the system aspires to deliver instant, accurate, and reliable feedback. This feedback not only reduces the manual grading burden on educators but also assists students in identifying areas for improvement in real time.

The significance of this work lies in its potential to overcome key limitations of traditional examination formats—including time constraints, grading fairness, and the dependence on extensive human oversight. By incorporating advanced autocorrect functionalities alongside sophisticated language models, our system enhances grading efficiency and supports multilingual assessment, with a particular emphasis on Arabic—a language that has historically been underrepresented in automated evaluation tools.

This paper outlines the complete methodology comprising dataset creation, model fine-tuning, and the development of a web-based grading platform. Furthermore, it examines the benefits and challenges associated with deploying such a system in educational environments. Ultimately, this work seeks to contribute to the expanding field of AI-driven education by offering a scalable and adaptable solution that empowers both students and educators.

2 Related Work

Automatic essay grading has been a long-standing research challenge in educational technology and natural language processing (NLP). Traditional systems, such as e-rater, relied on manually engineered features, syntactic parsing, and statistical models to evaluate writing quality, but lacked the ability to understand semantics or generalize across topics.

With the rise of deep learning, neural architectures brought significant improvements to language understanding. Encoder-based models like BERT and RoBERTa enabled deeper semantic analysis, and Siamese networks using models such as SBERT and ST5 were widely adopted for paraphrase detection and similarity-based grading tasks [7, 8].

Recent efforts have focused on applying large language models (LLMs), including encoder-decoder (e.g., T5, BART) and decoder-only architectures (e.g., GPT, LLaMA, Bloomz), for open-ended educational tasks such as essay scoring, answer generation, and question answering. These models offer stronger reasoning capabilities but often require significant resources to fine-tune.

To address the cost and efficiency challenges, parameter-efficient fine-tuning techniques like Low-Rank Adaptation (LoRA) [3] and Quantized LoRA (QLoRA) [2] have been proposed. These approaches enable effective fine-tuning of large models on commodity hardware while maintaining performance.

Several benchmark datasets support progress in question-answering and educational NLP, including SQuAD [5], QuAC [1], and PAWS [8]. However, these datasets are not specifically designed for essay grading, particularly in multilingual contexts. Our work reformulates QA datasets to support essay correctness classification and introduces enhancements such as ground-truth-informed labels and multilingual extensions.

Moreover, our use of OpenOrca-derived samples [4] enables scalable evaluation, while preserving quality through expert-reviewed reformations. Compared to prior approaches, our pipeline uniquely combines fine-tuned LLMs, enhanced labeling, multilingual corpora, and confidence-based scoring to deliver high-performance automatic grading in both English and Arabic.

3 Methodology

In this section, we present the end-to-end design of our automatic essay grading framework, covering model selection, dataset construction (including reforming and augmenting existing corpora), label engineering, multilingual extension, and training/evaluation protocols.

3.1 Semantic Similarity Baseline

- **Models:** SBERT, ST5 in a Siamese-network configuration (Figure 9).
- **Procedure:** Compute a continuous similarity score between student and reference answer spans.
- **Limitations:**
    1. Decoder-only LLMs lack the bidirectional encoder required by Siamese architectures.
    2. Reducing open-ended responses to a scalar score underutilizes large-model reasoning and world knowledge.
- **Conclusion:** We transitioned to a direct binary classification approach using a 7 B-parameter decoder-only LLM.

3.2 LLM-Based True/False Classification

- **Input:** Triplet of (Question, Student Answer, Context).
- **Objective:** Fine-tune the LLM to output a single token—True or False—indicating factual correctness.
- **Prompt template:**

```
Instruction: Determine whether the student’s answer is true or
false given the context below.

Context: ...

Question: ...

Student Answer: "..."

Response:
```

3.3 Dataset Construction

We synthesize a comprehensive training corpus via four complementary streams (Figure 10):

3.3.1 Dataset Preparation

A robust dataset forms the backbone of any machine learning system, particularly for tasks involving language understanding and generation. Given the absence of a pre-existing dataset tailored to our specific needs—evaluating essay responses in Arabic and English against reference answers—we employed a three-pronged approach to dataset creation: synthetic generation, collection, and reformation of existing datasets.

- **Synthetic Dataset Generation Using ChatGPT:** Leveraging the generative capabilities of large language models (LLMs) like ChatGPT, we created synthetic datasets by providing example prompts and extracting question-answer pairs. To ensure diversity, model parameters such as temperature (set to 1) and frequency penalty (set to 1) were adjusted to maximize randomness while maintaining relevance. This method, while efficient, incurred a significant cost (approximately $600 for 52,000 samples), prompting exploration of alternative strategies.
- **Data Collection:** We collaborated with educational institutions to gather real student essays, adhering to ethical guidelines and anonymization protocols to protect

privacy. Additionally, online platforms and crowdsourcing were considered to expand the dataset, ensuring diversity in topics, writing styles, and student backgrounds. This approach, however, proved time-consuming and resource-intensive given our team size of six.

- **Reformation of Existing Datasets:** To supplement our data, we reformed existing datasets such as SQuAD and QuAC, which provide context-question-answer structures. These datasets were adapted to fit our format (context, question, student answer, and truthfulness label), focusing on dialogues and open-ended responses. This reformation process involved extracting relevant features and aligning them with our grading objectives, balancing quality with scalability.

The final dataset was curated to include a mix of synthetic and real-world samples, with input lengths capped at 512 tokens to prevent truncation issues during tokenization, as detailed in the input length check process.

3.3.2 Reforming Existing QA Datasets

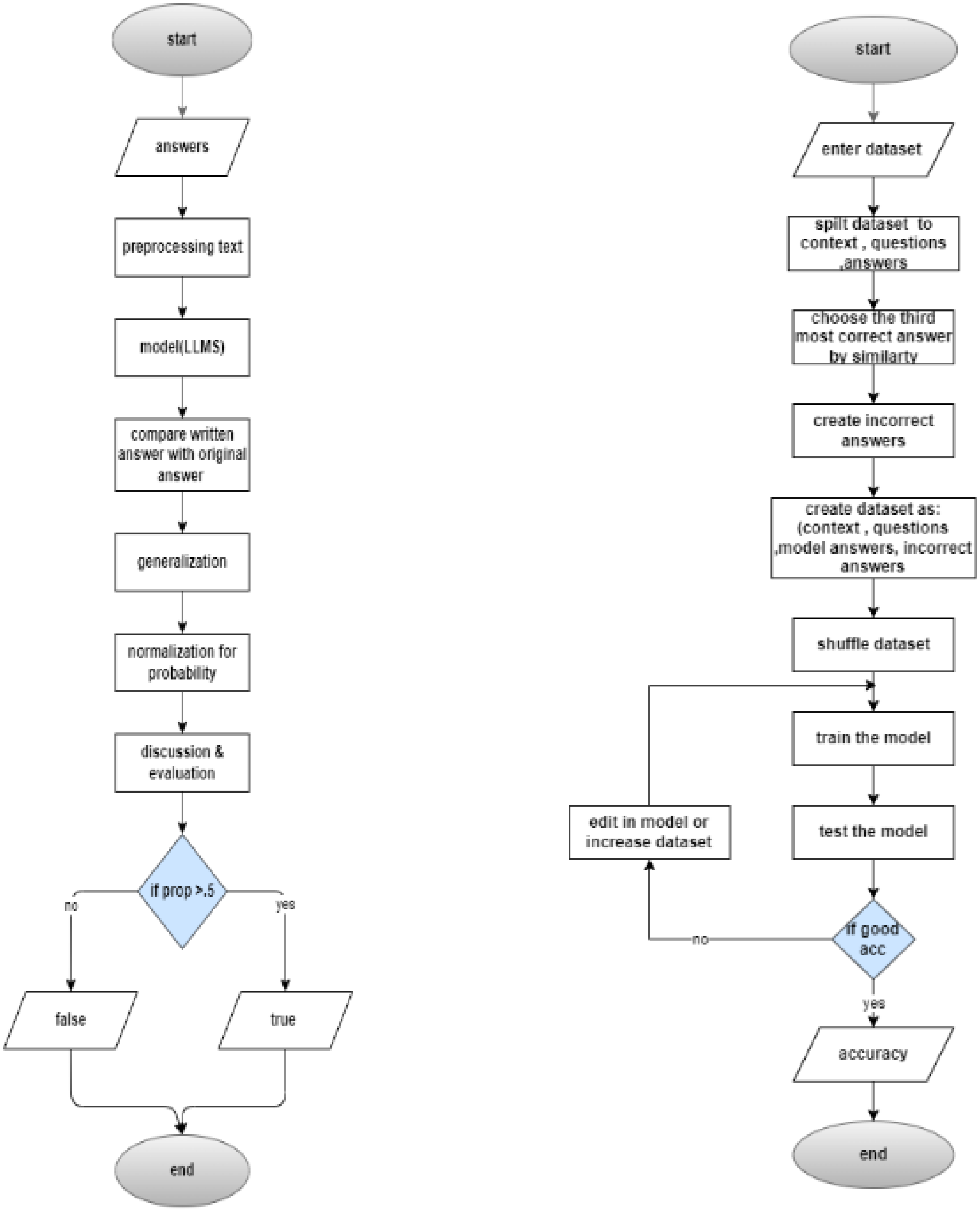
start
answers
preprocessing text
model(LLMS)
compare written answer with original answer
generalization
normalization for probability
discussion & evaluation
if prop >.5
no
yes
false
true
end
start
enter dataset
spilt dataset to context , questions ,answers
choose the third most correct answer by similarty
create incorrect answers
create dataset as: (context , questions ,model answers, incorrect answers
shuffle dataset
train the model
test the model
edit in model or increase dataset
if good acc
no
yes
accuracy
end

Figure 31 : dataset making flowchart

Figure 1: Dataset reformation process

- **Targets:** QuAC and SQuAD 2.0.
- **QuAC Selection:** Multi-turn Wikipedia dialogs with paragraph-length answers (Figure 13).
- **Format conversion:**
  - **True examples:** Directly map each teacher span to Label = True.
  - **False examples:**
    - **Random sampling:** Pair with unrelated spans (baseline).
    - **RAG-based distractors:**
      1. Index all spans.
      2. Retrieve top-k semantically similar spans to the true answer.
      3. Exclude exact matches.
      4. Select the **third** most similar span for maximal challenge.
      5. Back-translate for lexical variation.
      6. Enforce a $1:1$ True : False ratio.
- **Preliminary reforming results:**
  - Random baseline → 40% accuracy on held-out QuAC
  - RAG distractors → 8̃0% accuracy, confirming increased difficulty

3.3.3 Label Enhancement via Ground-Truth Inclusion

- **Rationale:** Binary labels alone insufficiently guide LLM reasoning.
- **Enhanced label format:**
  - **True:**
    صحيح أن الجواب هو ⟨category⟩.

    True: the answer is "⟨ground-truth span⟩."
  - **False:**
    .

    False: the answer is "⟨distractor span⟩."
  - **Impact:**

Model	Before	After
Flan-T5	20%	83%
Bloomz	40%	85%

—˜60 pp average gain demonstrates the efficacy of this strategy (Figure 1).

3.3.4 Multilingual Dataset Construction

- **Automated translation:** Base Arabic corpus via Google Translate (82. 5% general precision).
- **Targeted augmentation & expert review:** LLM-driven post-editing and bilingual expert validation ensure domain fidelity.
- **Randomization for robustness:** Inject nonsensical “answers” (e.g., “asdasjkldamsj,” “112ssa”) to train the model to flag gibberish as False and deter adversarial patterns.
- **Outcome:** Parallel English/Arabic corpora of equal scale and complexity.

3.4 Training, Evaluation & Preliminary Results

- **Fine-tuning:** Binary cross-entropy loss; input format:
  [Instruction] $\langle$SEP$\rangle$[Context] $\langle$SEP$\rangle$[Question] $\langle$SEP$\rangle$[Student Answer]
- **Data split:** 80% train / 10% validation / 10% test (balanced labels).
- **Metrics:** Accuracy, precision
- **Preliminary results:**
  - **English LLM:** 97% accuracy on held-out test set
  - **Arabic LLM:** 90% accuracy on held-out test set

3.5 Confidence-Based Grade Estimation

To ensure that only answers the model is more than “coin-flip” confident about receive credit, we remap the model’s probability P(True) so that:

- Any $P(True) \leq 0.5$ yields zero points.
- Only $P(True) > 0.5$ is linearly scaled into the [0, G] grading range (where G is the maximum points for the question) and then rounded to the nearest 0.5.

Let z_{True} and z_{False} be the unnormalized logits for the tokens "True" and "False", respectively. We first compute:

$$P(True) = \frac{\exp(z_{True})}{\exp(z_{True}) + \exp(z_{False})} \tag{1}$$

Next, define the normalized confidence above the 0.5 threshold:

$$\alpha = \frac{P(True) - 0.5}{0.5} \tag{2}$$

(so that $\alpha = 0$ when $P = 0.5$, and $\alpha = 1$ when $P = 1$).

The final grade is then:

$$\text{Grade} = \begin{cases} 0, & \text{if } P(True) \leq 0.5 \\ \text{round}(\alpha \times G \times 2)/2, & \text{if } P(True) > 0.5 \end{cases} \tag{3}$$

How it works:

1. **Thresholding:** Answers with $P(True) \leq 0.5$ receive 0 points.
2. **Scaling:** For $P < 0.5$, compute $\alpha = (P - 0.5)/0.5$, which lies in [0,1].
3. **Continuous grade:** Multiply α by G to get a raw grade in [0, G].
4. **Discretization:** Round that raw grade to the nearest 0.5 by computing round($\alpha \times G \times 2$) / 2.

Example: If G = 5 and P(True) = 0.7, then:

$$\alpha = (0.7 - 0.5)/0.5 = 0.4 \tag{4}$$

$$\text{raw grade} = 0.4 \times 5 = 2.0 \tag{5}$$

$$\text{Grade} = \text{round}(2.0 \times 2)/2 = 2.0 \tag{6}$$

This ensures that:

- $P \leq 0.5$ always yields 0
- $P = 1.0$ yields the full G points
- All intermediate values are awarded in 0.5-point increments.

4 Experimental Setup

4.1 Datasets:

- **Primary Training/Evaluation:** The custom-built corpus derived from synthetic generation and reformed QuAC/SQuAD datasets, with parallel English and Arabic versions (approx. 52,000 samples total before splitting). Input lengths were capped at 512 tokens.
- **Additional Evaluation:** Models were also evaluated on subsets or variations related to the OpenOrca dataset for generalization assessment.
- **Baseline Comparison:** A Siamese network baseline used the PAWS (Paraphrase Adversaries from Word Scrambling) dataset for semantic similarity classification.
- **Data Split:** Datasets were split into 80% training, 10% validation, and 10% testing, ensuring balanced labels within each split.

4.2 Models Evaluated:

A range of transformer architectures and sizes were evaluated:

- **Encoder-Decoder:** Flan-T5 (3B), MT0 (3B), BART (700M)
- **Decoder-Only:** Bloomz (7B), Llama 2 (7B), Marcoroni-7b-DPO-Mergea (7B), Mistral (Original 7B baseline), MistralTri x-v1 (9B)
- **Encoder-Only:** RoBERTa (considered, specific fine-tuning results not fully tabulated in provided summary)
- **Siamese Baseline Models:** BERT, T5-large, BART

4.3 Implementation Details:

- **Framework:** Hugging Face Transformers library.[6]
- **Hardware:** Training primarily conducted on NVIDIA P100 GPUs via Kaggle notebooks.
- **Training Time:** Approximately 50 hours per model on average.

4.4 Evaluation Metrics:

- **Primary:** Accuracy, Precision.
- **Secondary:** F1-Score (reported for select models)

5 Results and Discussion

5.1 Overall Performance:

The fine-tuned LLMs achieved high accuracy on the essay grading classification task.[7] Our best performing model, Marcoroni-7b-DPO-Mergea (a fine-tuned decoder-only model), reached 96% accuracy on the reformed QuAC-derived test set and 97% on the OpenOrca-related test set. Performance varied across architectures and languages, as summarized in Table 1.

Table 1: Model Performance on Essay Grading Classification Task

Model	**QuAC Accuracy**	**OpenOrca Accuracy**	**Epochs**	**Lang**	**Size**	**F1-Score**
Flan-T5	90%	92%	3	en	3B	—
Bloomz	93%	—	4	ar	7B	85.59%
Marcoroni-7b-DPO-Mergea	**96%**	**97%**	2	en	7B	**96.00%**
Llama 2	92%	92%	3	en	7B	—
MT0	—	85%	2	ar	3B	—
MistralTri x-v1	91%	—	2	en	9B	—
Bart	72%	—	1	en	700M	—
Mistral (Original)	*78%*	—	—	*en*	*7B*	—

5.2 Impact of Enhanced Labeling:

The enhanced labeling strategy (Section 2.3) was crucial. As evidenced by preliminary tests (+45-63pp gains) and the high final F1-scores for models trained with it (e.g., Marcoroni: 96%), providing the ground-truth/distractor span within the label significantly improves the model's ability to learn the classification task accurately.

5.3 Architecture Comparison:

Our results challenge the notion that decoder-only models are universally superior for all NLP tasks. While our top performer was a decoder-only model (Marcoroni-7b), other strong results were achieved by encoder-decoder models like Flan-T5. Conversely, some decoder-only models (e.g., original Mistral baseline, Bart) showed lower performance compared to fine-tuned versions or other architectures. This highlights the importance of empirical evaluation and task-specific architecture suitability, rather than relying solely on generative capabilities. Models around the 7B parameter size generally offered a good balance of performance and computational feasibility for this task.

5.4 Multilingual Performance:

The system demonstrated strong performance in both English and Arabic. The English Marcoroni model reached 96-97% accuracy, while the Arabic Bloomz model achieved 93% accuracy (and 85.59% F1-score). The MT0 model also showed reasonable performance (85%) in Arabic. This validates the effectiveness of the automated translation and expert review process for creating the parallel corpus. The slightly lower accuracy in Arabic might reflect the inherent complexity of the language or potential nuances lost in translation, suggesting areas for future refinement.

5.5 Siamese Network Baseline:

For comparison, a Siamese network approach was evaluated on the PAWS dataset for semantic similarity classification. Accuracies were: BERT (84.59%), T5-large (83.99%), and BART (82.59%). While operating on a different task (paraphrase identification) and dataset, these results are substantially lower than the accuracies achieved by our fine-tuned LLMs using direct classification with enhanced labels on the essay grading task, further justifying our chosen methodology.

6 Conclusion

This work successfully demonstrates the application of fine-tuned Large Language Models for automated essay grading in both English and Arabic. By formulating the task as a factual correctness classification against a reference answer and employing an enhanced label engineering strategy that incorporates ground-truth information, we achieved high accuracy (up to 97%). Our findings indicate that while decoder-only models can perform exceptionally well, task-specific empirical evaluation across different architectures (including encoder-decoder) is crucial. The methodology presented, including multi-source dataset construction and multilingual adaptation, provides a robust framework for developing scalable, efficient, and consistent automated grading systems, significantly benefiting both educators and students.

References

[1] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[4] OpenOrca contributors. Openorca: An open reproduction of orca from microsoft research. `https://huggingface.co/datasets/Open-Orca/OpenOrca`, 2023. Accessed: 2024-06-01.

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

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008, 2017.

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:2003.00688*, 2020.

[8] Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308, 2019.