Detecting Counterfeit Amazon Products Using Deep Learning

Jaegook Alex Kim
Georgia Institute of Technology
Atlanta, Georgia, USA

Max Immanuel Hill
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Counterfeit products have become a serious problem in online marketplaces, especially on Amazon, one of the biggest e-commerce platforms. Consumers shop for products solely depending on the seller's pictures, reputation, and other buyer's reviews. Buyers have difficulties verifying the product before purchasing as products are often sold by a wide variety of third party sellers [1]. Existing methods to detect counterfeit products currently either require implementation of potential solutions at the production level [2], or rely on non-comprehensive machine learning algorithms to perform basic analysis. In this paper, we utilize a dataset of amazon products and reviews from Amazon to approach the identification of counterfeit products as a deep learning binary classification task. We compare the performance of BERT and Llama2, two deep learning models, to traditional machine learning models such as Random Forest, Naive Bayes, and Logistic Regression on this classification task. We test the performance of these models both in a balanced setting to determine how accurately these models can identify "likely counterfeit" products, and in an imbalanced setting to test model performance in a realistic setting. The results indicate that the deep learning models provide a significant advantage when trying to correctly classify counterfeit products, and that the performance of all models significantly deteriorates in a more realistic, unbalanced setting.

1 INTRODUCTION

In recent decades, e-commerce has increasingly come to shape the way people shop for products all across the world. Amazon, one of the most recognizable online retailing platforms, plays a large role in connecting buyers to sellers, reportedly allowing 4.1 billion item to be shipped by US-based sellers, and shipping products to customers in more than 100 countries and regions across the world [6]. However, with the rise of these online sales, counterfeit products have become a serious problem in online marketplace. Oftentimes for customers, there is little assurance that the products they are buying are legitimate, with common reports of products that are either of lower quality than the product order or an entirely different product altogether [18,19].

These counterfeit sales also have significant, negative impacts on manufacturers across the globe. With smaller manufacturers unable to police the counterfeits listed on these sites, third party sellers can continue to substitute fraudulent products in legitimate listings [19]. Larger companies are also significantly impacted by these counterfeit sales, with luxury brands estimated to lose over $50 billion USD in sales every year [7]. These counterfeit products can be responsible for irreparable damage to brand reputation as customers lose trust with the inauthentic products. This paper attempts to address the counterfeiting problem on e-commerce websites by devising a method to flag potential counterfeit products, in this case those specifically sold on Amazon.

We look to apply deep learning models to detect "counterfeit" products, or inauthentic products online. Our definition of counterfeit covers both products that are described correctly, but not are authentic (ie. Gucci bag, that is a counterfeit), and products that does not match the seller's descriptions (ie. seller says they are selling blankets, but are selling pillows). We define our problem in this paper as a binary classification task of inauthentic products, in which we plan on implementing BERT and Llama2 language models to analyze the corresponding list of reviews and descriptive features. The model produces a classification of whether or not a product is counterfeit.

Using deep learning models to identify counterfeit products comes with multiple challenges. To start, it is incredibly difficult to verify any ground truth information regarding whether a product sold is counterfeit or not, as we do not have direct access to sold products. Because smaller manufacturers may not have sufficient resources to coordinate directly with online retailers and directly ensure the quality of the products being sold, the deep learning methods implemented must prevent counterfeit sales without direct access to the product. While reviews provide a great source of information regarding whether or not a product may be counterfeit, there still may be difficult to determine whether or not a product is counterfeit depending upon the customer's assessment. Some genuine products may in general be poorly made and be considered by buyers to be "fake", while other counterfeit products may closely resemble and function as their original counterparts and may fall under the radar, and therefore not be detected as fake. Another issue with identifying counterfeit products is that products on Amazon regularly changes sellers depending upon whether a product is still in production or how in demand the product is for buyers. It is difficult to accurately determine whether a product is counterfeit at the listing level, as the authenticity of a product varies from seller to seller. Furthermore, while Amazon allows multiple sellers to vend the same product, the reviews for purchases do not specify any information regarding which seller sold them the product.

We conduct our study to answer two primary research questions:
RQ1) How accurately can deep learning models classify counterfeit products on Amazon?
RQ2) How well do these model perform on an imbalanced dataset (ie. in a realistic setting)?
The dataset we use for this study is from the Amazon Review dataset collected by McCauley et al. [5]. The dataset was manually labeled to indicate whether a product sold had a high likelihood of being counterfeit, with around 8,500 product reviews labeled from three distinct categories and an eventual 587 items that consisted of 202 "likely counterfeit" products constituting the final balanced dataset.

We trained a supervised fine-tuned Llama2 on the curated dataset, and found that the model had an F1-score of 0.50 and precision of 0.67 for the balanced dataset and an F1-score of 0.26 and recall of 0.65 for the unbalanced dataset. These results compare favorably to the other machine learning models and similarly to BERT. While there is little research involving preventing the circulation of counterfeit products exclusively at the retail level, we believe these experiments not only take a first step at evaluating the potential use of employing deep learning models to detect counterfeit sales, but also demonstrates that these models offer an advantage in identifying counterfeit products on e-commerce platforms.

2 LITERATURE SURVEY

The paper *A Survey Of Counterfeit Product Detection* [2] conducts a systematic review regarding existing literature involving the identification of counterfeit products. Existing methods involved avoiding counterfeits at the production level, such as using QR Codes, RFID tags, and Holographic barcodes. It also covers identification at the retail and consumer levels, using methods such as semi-automatic workflow analysis. The study then rates the methods it has identified in terms of processing information related to counterfeits real-time, predictive capabilities, the methods security, and its trustworthiness. Most of the methods alluded to in the paper require intervention at the production level and the retail level, while implementing a deep learning model identifying counterfeit products only would require collecting data regarding products at the retail level.

Identification and Impact of Online Deceptive Counterfeit Products: Evidence from Amazon [1] is the most similar to our research project. The researchers focus on identifying and analyzing the impact of online counterfeit products on Amazon. They use common machine learning models to classify counterfeit products. We wish to use this related work and improve on it significantly. We want to take on a similar approach but data clean the amazon review dataset to take out fake reviews using neural fake news detection methods before applying deep learning language models to try classifying whether a product is counterfeit or not.

Creating and detecting fake reviews of online products [3] focuses on two main topics : creating fake reviews with language models and detecting fake reviews using language models. The paper discusses the performance between GPT2 and BERT and it concludes that GPT2 performs better in all relevant fields. Humans are able to detect fake reviews with 55.36% accuracy while fakeRoBERT has an accuracy of 96.64%. Our research can use this related work and clean them out of the dataset to train our model more accurately in detecting counterfeit products.

"Do not deceive me anymore!" interpretation through model design and visualization for Instagram counterfeit seller account detection[4] is a study that used Instaloader API to flag counterfeit sale posts on Instagram. The paper analyzed the sellers who posted the flagged sale posts and manually labeled whether the seller was a potential counterfeit seller or not. The proposed model had an account detection accuracy of 100%. The paper comes with limitations as they focus solely on the seller and do not analyze whether the product itself is counterfeit. Additionally, the paper assumes that the sellers have static behavior in posting counterfeit sales posts.

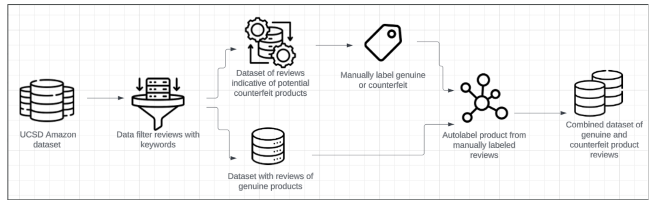

Figure 1: Data Pipeline

Our research will focus on the product rather than the seller as we believe this methodology will be better in resolving the overall issue of counterfeit products. Sellers in online platforms can create new accounts, but the product itself will not change.

3 DATASET

This paper elects to use the Amazon Review dataset collected by McCauley et al [5] from 2018, which comprises over 233.1 million unique reviews from 20 million users over a time span of 12 years. The products reviewed are from a broad spectrum of categories, for example: Appliances, Books, Electronics, Luxury Beauty, Office Products, and Toys & Games. The data included features such as ratings, text, and helpfulness votes for reviews, along with related product information such as descriptions, category information, price, brand and image features. In this paper, publicly available review and metadata json files were downloaded and used from the following categories: Luxury Beauty, Appliances, and Industrial/Scientific. The choice of these three categories was motivated by the belief that these categories adequately represent the vast assortment of Amazon products. According to a recent audit of Amazon by the U.S. Government Accountability Office, there is a demonstrably high percentage of counterfeit cosmetic products circulating in the Amazon marketplace [8]. Thus, the inclusion of the Luxury Beauty also suggests that the curated dataset will consist of a relatively large number of likely counterfeit products. The initial dataset we downloaded consisted of 2,935,738 reviews and 210,291 products. A large dataset is necessary in order to extract a sufficient amount of counterfeit products for the deep learning models to run on. Due to the inherent imbalance between counterfeit and genuine products on the marketplace, this dataset was later reduced down to a significantly smaller number of labeled product, with an overview of this data processing detailed in Figure 1.

The initial dataset does not include ground truth labels indicating whether or not a product is counterfeit. Therefore, we must rely on the wisdom of the crowd to determine the authenticity of a product, in this case the buyers. Because buyers not only have direct access to a product but also are likely knowledgeable about the product they are buying, we can utilize their reviews to determine the ground truth label of whether or not a product is counterfeit.

To gather the most relevant reviews related to counterfeit products, we first cleaned the data by attempting removing duplicate reviews, converted all words to lowercase, removed htmls and links from the review texts, and removed typos. Next, we separated the reviews into two different files, where one included keywords indicating that its associated product might potentially be counterfeit, and the other did not include any suspected keywords. The keywords we filtered for were "fake", "counterfeit", "knock off", "imitation", "fraudulent", "copycat", "forged", "imitation", "phony", and "inauthentic". The products that did not have any of these suspected keywords were automatically labeled "likely genuine".

After this filtering, there were a total of 8,531 reviews that contained the aforementioned keywords. We acted as annotators, manually went through all reviews to determine whether the review text clearly indicated whether the product was counterfeit. To maintain consistency, we created strict rules that guided the annotation process. An annotator would label a review as "likely counterfeit" or a value of "0" if one of the following conditions was met regarding the product reviewed:

- The customer clearly states the product is a counterfeit or fake.
- The customer indicates a strong suspicion based on usage that the product they received is counterfeit.
- The customer states that at least one of their purchases of the associated item returned a counterfeit product.

Otherwise, we labeled the review as "likely genuine" or a value of "1". As a note of clarification, the labeling process focuses exclusively on identifying counterfeit products, and while related, is not necessarily dependent upon the quality of the product.

The labeled data was then used to determine whether a product sold was "likely counterfeit", or had a high incidence of counterfeit sales. The frequency of counterfeit labeled reviews product ranged from 1 to 78. We set a threshold of 3 "likely counterfeit" reviews to label a product with the ground truth value of "likely counterfeit" or a value of "0". This data process resulted in 202 "likely counterfeit" labeled products, which is less than 0.01% of the initial number of products we started off with. Our team decided it was best to create a balanced dataset to train our model, as the number of likely counterfeit products were significantly less to likely genuine products. We randomly sampled 385 "likely genuine" products that did not have any suspected keywords and finalized with a dataset of 587 total items.

Our final balanced dataset has an average of 144.7 reviews per product, and we flagged 454 Luxury Beauty product reviews as "likely counterfeit", while 104 reviews for Industrial & Scientific, and 28 reviews for Appliances. The high amount of Luxury Beauty products was the result of a high incidence of counterfeit products labeled for the indicated category. We chose 8 features in total: product number, title, features of the product, description, price, brand, category, and average rating. Textual reviews for products were not included both because they were used to label ground truth values of products, and because review information is not necessarily available before consumers purchase the product, as required by the models.

4 EXPERIMENTAL SETTING

Counterfeit product identification is a binary classification task. To evaluate the baseline and our model's performance, we decided to use precision, recall, F-1 score, and accuracy. Precision will measure how many counterfeit products that each model predicted are actually counterfeit. This is important as flagging authentic products as counterfeit can harm sellers that are doing honest business. Recall measures how many counterfeit products we correctly predict out of all the counterfeit products in the validation dataset. F-1 score will provide us insight into both precision and recall as having the harmonic mean between the two values will give insight into how well our model is performing in detecting counterfeit products. We decided to also include accuracy as our initial training dataset is relatively balanced. Although the e-commerce product corpus is very imbalanced because of the small number of counterfeit products, our team decided accuracy can provide insight on how well our model performs overall in correctly predicting counterfeit products. For all baselines and models, the datasets have an 80-20 split for training and validating. All models are also run on Google Colab Pro, using a NVIDIA T4/V100 GPU containing 15GB GPU RAM.

Our baseline models are from Cao et. al [1], where the paper implemented random forest, naive bayes, and logistic regression machine learning algorithms to classify counterfeit products. We attempt to replicate their work with the datasets to see how well common machine learning algorithms perform. The paper does not specify their parameters, so 250 decision trees were used for the random forest algorithm.

Our second baseline is Bidirectional Encoder Representations from Transformers (BERT), a commonly used deep learning model in research when doing text classification tasks. There exists no previous or related work that incorporated BERT to try classifying counterfeit products. Our team believed using a deep learning model would be best fitting in comparing with our proposed method of incorporating Llama2. We implemented bert-base-uncased, an open source pre-trained model available on Hugging Face [9]. BERT needs the input to be sequential data, unlike random forest or logistic regression that use one-hot encoding. We put our data features into a list of strings, and utilize the BertTokenizer class to get tokenized outputs. Then, we used the pytorch dataset[10] to create a Dataset object that would be used to train the pre-trained BERT model.

5 PROPOSED METHOD

Our team created a model based of Llama2, Meta AI's open source large language model (LLM) that was released in July 2023. Llama2 was pre-trained with 2 trillion tokens and a context length of 4096. There are three different Llama2 models of 7 billion, 13 billion, and 70 billion parameters. Our team worked with the 7 billion parameter Llama2 because of GPU limitations. 7B Llama2 takes up 14GB of GPU RAM, and Google Colab Pro has 15GB of GPU RAM.

Our team believed that Llama2 was the latest and best performing state of the art language model that would outperform all baselines. Prior benchmarkings have also shown that Llama2 outperform BERT in classifying complex tasks [16].

Our team first split the dataset 80-20 training and validation and uploaded it onto Hugging Face. We then used Llama2, Hugging Face Format [11], that would allow us to leverage *AutoTrain*, a Hugging Face tool that automatically trains and deploys state of the art LLMs [12].

Fully fine-tuning Llama2 is expensive and requires a lot of storage as a fully fine-tuned model will be the same size as the original pretrained model. It was infeasible given the time frame and limited computing power accessible. Our solution was to utilize *parameter efficient fine tuning* (PEFT). We only fine-tuned a small number of the model's parameters while freezing most parameters of Llama2, which results in greatly decreasing computational and storage costs. PEFT also overcomes a phenomenon observed when fully fine-tuning pre-trained LLMs called catastrophic forgetting. Catastrophic forgetting is when a language model is trained to task A, and then trained to do task B. However, in the process of learning task B, the fine-tuned model "forgets" how to perform task A [13].

After researching the benchmarks between 4-bit and 8-bit quantization of Llama2, we concluded that losing precision from using 4-bit would not significantly impact our model's overall performance [14]. 4-bit quantization runs about 1.5 times faster than 8-bit [15], thus our team believed it would be optimal to save on computing costs while slightly sacrificing precision.

We used a training batch size of 2, which is relatively small. Unfortunately, our team did not have a choice as we attempted to run a batch size of 4, but we ran out of memory in the midst of training our model.

We ran multiple tests and training and found that setting the learning rate to 2e-4 gave us the best results, while we ran 5 epochs as a compromise to not train our model for too short or too long. Finally, although Llama2 was pre-trained with a context length of 4096 tokens, our dataset did not contain any input tokens longer than 2000 tokens. Therefore, our team set the max length to 2048 to further reduce computational time and cost.

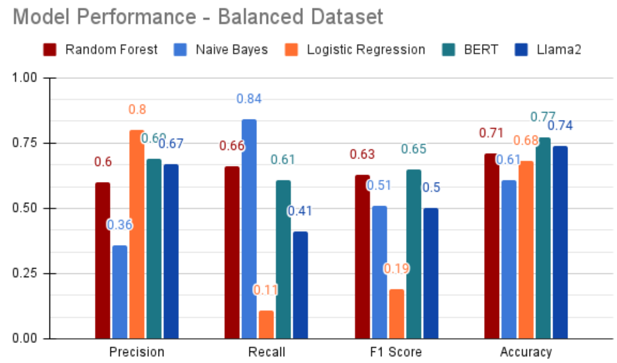

Figure 2: Model Results for Balanced Dataset

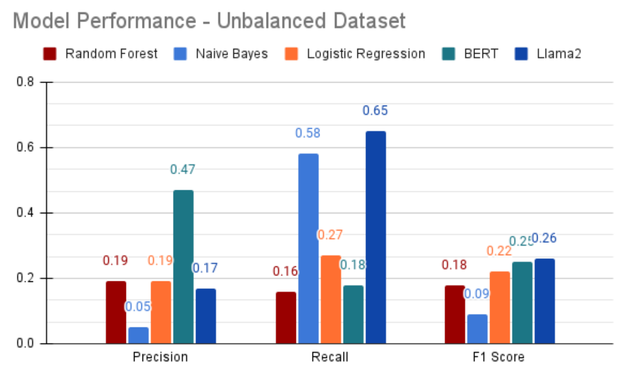

Figure 3: Model Results for Unbalanced Dataset

6 EXPERIMENTS

RQ1: How accurately can we classify "likely counterfeit" products on Amazon?‘

With our first question, we determine how well the models perform in a balanced setting, or when the dataset has a similar number of genuine and counterfeit products. As seen in Figure 2, we find that Llama2 has a F1-score of 0.50 and accuracy of 0.74, with a high precision and low recall value suggesting higher confidence when predicting counterfeit products as opposed to identifying large amounts of counterfeit products. While Llama2 performs more favorably than the Logistic regression and Naive Bayes models for F1-score and accuracy, Llama2 performs comparably to the baseline models with these two metrics. BERT appears to be the top performing model in the balanced dataset setting, marginally outperforming all of the other models in both F1-score (0.65) and Accuracy (0.77). Random Forest (with a F1-score of 0.63 and accuracy of 0.71) performs comparably to its deep learning counterparts, while the Logistic Regression performs poorly in F1-score (0.19) and Naive Bayes performs poorly for accuracy (0.61).

RQ2: How does changing the classification setting (balanced versus unbalanced) affect model performance?

In this next question, our models are tested on our realistic, unbalanced dataset, where the number of genuine products is significantly larger than the number of counterfeit products. The unbalanced dataset has a 1:18 ratio of likely counterfeit to genuine products. Accuracy has been removed as an evaluation metric due to the bias of the metric when operating on an unbalanced dataset. While the performance of all models suffers as a result of the greater imbalance, as seen in Figure 3, Llama2 improves relative to the other baselines. Llama2 achieves the highest F1-Score (0.26) with high recall. This means that Llama2 is able to correctly classify 60.5 of all likely counterfeit products flagged. On the other hand, it has relatively low precision (0.17), which suggests that Llama2 flags a lot of genuine products as likely counterfeit.

BERT, while still appearing to be one of the top performing models, achieves the opposite results. BERT achieves relatively high precision (0.47), which means it performs better in correctly classifying the counterfeit products, while having low false positives. Machine learning models such as Random Forest, and Logistic Regression, perform comparably to their deep learning counterparts, while Naive Bayes now performs poorly.

While Llama2 performs comparably to BERT in both experimental settings and to Random Forest in a balanced setting, there is evidence to suggest why Llama2 didn't perform as well as predicted relative to the other baselines. While it has been documented that Llama models have outperformed BERT models on complex datasets in the past, this observation had primarily been noted for larger datasets [16]. This could also explain why Llama2 was seen to have improved in performance compared to BERT for the larger, unbalanced dataset. Llama2 could have also been limited by computational constraints associated with working in Google Colab Pro. Because of limited GPU RAM, we were forced to limit the size of the model and parameters such as batch size to conserve memory, which has been known to introduce more noise into the model [17].

Note that there is little established precedent regarding the level at which to expect machine and deep learning models to identify counterfeit products. While the F1 and accuracy scores for deep learning may not be high in a balanced setting (0.5 for Llama2, 0.65 for BERT), these values indicate that the models perform better than random and provide users with a tool that can help a user identify whether or not a product listing is counterfeit. This is further supported by the accuracy of these models (0.74 for Llama2, 0.77 for BERT), indicating a certain amount of confidence in correctly predicting the authenticity of products. Overall, Llama2 and BERT can provide a significant advantage in helping to identify counterfeit products on Amazon. We also find that while every model exhibits reduced performance in a more unbalanced, realistic setting, Llama2 exhibits the greatest performance at identifying counterfeit products among them.

7 CONCLUSION

While our work is an important first step towards using deep learning methods to address counterfeit sales on e-commerce platforms, it is necessary to recognize shortcomings of our model and implementation. First, while the model includes a variety of features related to products being sold, it does not consider any information related to the seller of the product. On e-commerce platforms such as Amazon, many of the products shipped are vended to customers by third-party sellers, these platforms even allowing multiple sellers to distribute the same product. Oftentimes, the difference between a customer receiving a genuine and a counterfeit product is often down to the seller, some of whom decide to sell counterfeits to lower the price of their individual listing [18]. Without seller information, the model is unable to utilize crucial information which may be directly linked to whether a product may be counterfeit. The dataset size also potentially limits the performance of the model. With an inherent imbalance in the marketplace between genuine and counterfeit products, it was difficult to confidently identify a large amount of counterfeit products, even among the over two-hundred thousand product listings within the McCauley dataset. Finally, our model is limited to the lack of computing power. Because we were limited to working with 51GB of RAM and 15GB of GPU RAM on Google Colab Pro to train and test our fine-tuned Llama2 model, we were forced to artificially constrain key parameters during our experiments, using the smallest Llama2 model and batch size possible to work within these limits. Future work could utilize the 13B or 70B parameter Llama2 model as benchmarks show they perform significantly better than the 7B parameter Llama2 model [15].

There are multiple areas research into counterfeit product identification could address in the future. Primarily, more data regarding the circulation of counterfeit products on online retail websites needs to be created or publicized. There are very few publicly available datasets that qualitatively describe the sales of products on e-commerce websites, with those existing providing little information to indicate whether a product may be counterfeit. With the publication of labeled datasets indicating counterfeit product sales, researchers would not only be able to confirm the findings of our paper but could also identify key hallmarks of counterfeit products and variations of the characteristics across different platforms. Finally, researchers could also utilize the inherent network structure of online retail websites to help identify counterfeit items. By modeling the vending relationship between products and third-party sellers, researchers would not only be provided with more contextual information to discern whether a product being sold is counterfeit, but could also identify malevolent sellers who are attempting to sell counterfeits in these online marketplaces.

8 CONTRIBUTIONS

All team members have contributed a similar amount of effort.

9 REFERENCES

[1] Ziyi Cao, Sanjeev Dewan, and Jinan Lin. 2022. *Identification and Impact of Online Deceptive Counterfeit Products: Evidence from Amazon.* SSRN.

[2] Prabhu Shankar and Jayavadivel Ravi. 2019. *A Survey Of Counterfeit Product Detection.* 8, 12 (December 2019).

[3] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. 2022. *Creating and detecting fake reviews of online products.* 64, 12 (December 2022). DOI:https://doi.org/https://doi.org/10.1016/j.jretconser.2021.102771.

[4] Park, J., Gu, J., Kim, H. Y. (2022). *"Do not deceive me anymore!" interpretation through model design and visualization for instagram counterfeit seller account detection.* Computers in Human Behavior, 137, 107418. https://doi.org/10.1016/j.chb.2022.107418

[5] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects.* (November 2019), 188–197. DOI:https://doi.org/10.18653/v1/D19-1018

[6] Quaker, D. (2023, October 10). *Amazon selling stats.* Amazon. https://sell.amazon.com/blog/amazon-stats

[7] *How Much Money lost due to the Counterfeit Fashion Industry?* (2023, April 13). Alpvision.com. https://alpvision.com/counterfeit-fashion-industry/: :text=For%20fashion%20brands%2C%20counterfeits%20represent

[8] Government Accountability Office, Gianopoulos, K.,*INTELLECTUAL PROPERTY Agencies Can Improve Efforts to Address Risks Posed by Changing Counterfeits Market* (2018). Retrieved October 26, 2023, from https://www.gao.gov/products/gao-18-216.

[9] *bert-base-uncased · Hugging Face.* (n.d.). Huggingface.co. https://huggingface .co/ bert-base-uncased

[10] *Datasets DataLoaders — PyTorch Tutorials 1.11.0+cu102 documentation.* (n.d.). Pytorch.org. https://pytorch. org/tutorials/beginner/basics/data_tutorial.html

[11] Thakur, A. (2023, July 19). *abhishek/llama-2-7b-hf-small-shards at main.* Huggingface.co. https://huggingface.co/abhishek/llama-2-7b-hf-small-shards/tree/main

[12] HuggingFace. (n.d.). *AutoTrain.* Huggingface.co. https://huggingface.co/docs/autotrain/index [13] Choi, C. (2022, November 23). *Sleep Can Keep AI From Catastrophic Forgetting - IEEE Spectrum.* Spectrum.ieee.org. https://spectrum.ieee.org/catastrophic-forgetting-deep-learning

[14] Rohling, K. (2023, August 11). *Benchmarking Llama 2 70B inference on AWS's g5.12xlarge vs an A100.* Medium. https://medium.com/@krohling/benchmarking-llama-2-70b-inference-on-awss-g5-12xla

rge-vs-an-a100-9d387d969177: :text=Unsurprisingly%2C%204%2Dbit %20quantized%20models
[15] ggerganov. (n.d.). *llama.cpp/README. md at dc271c52ed65e7c8d fcbaaf84dabb1f788e4f3d0 · ggerganov/llama. cpp.* GitHub. Retrieved December 7, 2023, from https://github.com/ggergan ov/llama.cpp/bl ob/dc271c52ed65e7c8dfcbaaf84dabb1f788e4f3d0/README.mdquanti zation
[16] Bumgardner, C., Mullen, A., Armstrong, S., Hickey, C., Talbert, J. 2023. *Local Large Language Models for Complex Structured Tasks.* https://arxiv.org/pdf/2308.01727.pdf
[17] Shen, K. (2018, June 19). *Effect of batch size on training dynamics.* Medium. https://medium.com/ mini-distill/effect-of-batch-size-on-training-dynamics-21c14f7a716e
[18] Suthivarakom, G. (2020, February 11). *Welcome to the Era of Fake Products.* nytimes.com. Retrieved December 5, 2023, from https://www.nytimes.com/wirecutter/blog/amazon-counterfeit-fake-products/.
[19] Hern, A. (2018, April 27). *Amazon site awash with counterfeit goods despite crackdown.* The Guardian. https://www.theguardian.co m/technology/2018/apr/27/amazon-site-awash-with-counterfeit-go ods-despite-crackdown