**Training Palm2 on the MakerSuite Platform**

***Strategies for AI-Driven Essay Recognition***

**James Lo**

**James Zhou**

**Jinny Kim**

1. **Introduction**

In the era of advanced Artificial Intelligence (AI) tools, such as ChatGPT, the ethical implications of using AI-generated content in academic settings have become a pressing concern. The challenge of distinguishing between human and machine-generated text raises questions about academic integrity, especially in the context of college application essays. This research, for the Google Hackathon, aims to address this issue through the development of Palm2 LLM, aiming to discover the performance by creating our tuning model to analyze text inputs based on writing style, structure, and content, to evaluate whether a passage is AI-generated or human-written.

The need for Palm2 is underscored by the growing prevalence of AI tools in the college application process, college assignments, exams, etc., and the ethical dilemmas they pose. (Zinshteyn, M., &amp; Jones, C., 2023, October 23) Educational institutions exhibit varying stances on AI-generated content, ranging from outright rejection to cautious acceptance. Existing detection tools face limitations, necessitating a more nuanced approach. Palm2, designed for this hackathon, seeks to navigate these ethical complexities by discerning subtle differences in writing styles, thereby contributing to upholding academic integrity while fostering a conversation about the ethical use of AI in education.

1. **Data**

The data used in this research to tune and train the model are mainly from a recent Kaggle competition (Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, Maggie Demkin. 2023), which consists of essays written from students with a given prompt. The dataset is formatted as follows:

* Id - A unique identifier for each essay.
* Prompt\_id - Identifies the prompt the essay was written in response to.
* Text - The essay text itself.
* Generated - Whether the essay was written by a student (0) or generated by an LLM (1). This field is the target and is not present in test\_essays.csv.

Each essay in the dataset corresponds to a specific writing prompt, which are "car-free cities" and "does the electoral college work" and are marked as 0 and 1 in the "Prompt-id" column. Both prompts required students to write an explanatory essay for the target audience to explain the reasoning behind the existing situation. In the first dataset, 1378 original human writings were provided, of which 708 of the essays are "Topic 0" and 670 of the essays are "Topic 1."; Since this is used for training, AI generated none of the passages, and all "Generated" columns are stored as 0. The Average Length of the passage in the dataset is approximately 550, which is calculated by the string length with estimation, based on the study by Ilker Cingillioglu, “Detecting AI-generated essays: the ChatGPT challenge”(Cingillioglu, I. 2023), for AI detection purpose, the training set is recommended contain with essays no less than 400 words, else it would be hard for the model to find the useful pattern to distinguish the major difference between AI writings and Human Writings.

In order to provide sufficient AI writing examples to tuning our own model for Palm2 to use on the makersuit platform, we applied GPT-4 with Super-analytic mode to help us alter the passages from the previous dataset. The GPT-4 were asked to rephrase each passage inside the text column and set to be academic writing style, limited to not expanding much on its generated passage length. In this step, all of the 1378 passages from previous were transmuted by GPT-4. Next, the original writing dataset is merged with the AI-writings by GPT-4, and randomly drawn with passages to create three new dataset. Each dataset consists of 200 writings of student writings and 300 writings of GPT-4 generated, all three datasets do not contain duplicate passages, and are marked as “Training Set”, “Test Set-1”, “Test-set 2” for training purpose and later validation process.

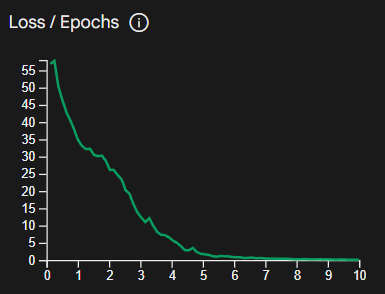
1. **Model Tuning Process and Analysis Result**

We use the Makersuit platform to access the Palm2 model and test with the text prompt and data prompt session on the platform. First, we create multiple tune models from one prompt we created and tried with different key parameters provided by Makersuit to adjust the training process. The prompt entered to train the model as following:

*“Please review the text provided below. Based on analysis of the writing style, structure, and content, estimate the likelihood of whether the passage was generated by an AI or written by a human.”*

And the corresponding example input with output from our training set formatted as following:   
 *Input: “Passage from GPT-4 generated/Human Writing”*

*Output:”AI Generated”/”Human Writing”*

The main adjustable key factors differentiate the models are: Batch Size, Epochs and Learning rate. The based model used for training is Text Bision and the limited example pass to the training process is limited up to 500. By trying up more than 12 combinations and going through different parameters, the best tuned model we created is trained under 10 Epochs, 64 Batch size and 0.005 Learning rate. The loss curve of the model is presented below:  
 

In which we could see from the graph that the model is having a down sloping from steep to smooth after each step of epochs, or the training times upon dataset. The loss is decrease rapidly during the first epochs and showing an outward curving downward from first epoch to third epoch, which we think the model is distinguish the difference between the pattern of two different writings and each stopping on the curve indicates improvement found by the model to reduced the loss. Later, during the fifth epoch toward tenth epoch, the curve is slowly closing to the x-axis, which means that the existing data can’t make more significant improvements from the current dataset and suggest around seven epochs through the training should be enough to produce nearly the same result.

Then, we step on trying our models on both data prompt and text prompt with the model-10.0005.64 which expects to have higher accuracy based on the loss curve. Firstly, we tried with the data prompt and entered multiple example inputs from “Test Set-1” and “Test Set-2”, each time with 20 inputs and evaluated the output. However, the data prompt gives an unsatisfactory answer by reporting the output as "generated", which is not the intended output that we have trained and therefore cannot count for accuracy at all. Unlike the data prompt, we used the same 20 inputs in the data prompt session, entered one at a time and asked the tuned model, the model correctly given output with either “Human Writing” and “AI generated”.

By getting the model running correctly, we test with different results on the text prompt session from both test-set with 50 results, The accuracy of the model is not considered to provide an accurate answer as result. Upon evaluating the false-positive result, which is the GPT-4 generated passages falsely detected by the model as human writings, the model only provides around 50% accuracy, which is close to the accuracy of making random guesses and not able to say it catches the deciding pattern of it. For the false-negative result, the model concludes around 75% of the passages correct as human writings, which is quite acceptable on this side of detection. Also, in some extreme cases, the model will produce "no content" output, which doesn't seem to have any connection or pattern between the normal and problem input. As a result, the model has a rate of around 62.5% accuracy on detecting a new passage outside the training dataset, and needs to look for more improvement, whether using a wider range of example training or providing more details on the tuning process to better adjust the model on catching patterns.

1. **Discussion**
2. **Problems Faced**

* **Inadequate Output Data for Training and Lack of Method for Re-training purpose**

We attempted to train Palm2 by providing text input and expecting varied output with freedom by model to answer the prompt question like: “Provide a percentage score on the text” or “mark the sentence that is most likely to be AI-generated”. However, this approach encountered challenges as not all of the 500 data samples had corresponding output data other than given values, and experiencing failure to produce more creative output from the model. On this scope of the model tuning, it might be fixed if the platform can provide a method or way for users to re-training their existing model and adjust the prompt used, this should provide the model with a higher ability upon providing creative content and adjust the precision.

* **Inconsistent Results and Overfitting**

Initial tests with the model produced undesirable results. The model seemed to rely heavily on the format of example inputs and outputs to identify patterns, resulting in responses that closely mimicked the provided format and were often incorrect. After the validation stages, we also try with some internet sources data from other topics, since the training set is limited within two explanatory topics, the outcome of the model performed even worse compared to the current test sets. Consider the limitations and difficulties on gathering real-human-writing and use for training, and also the Palm2 only allows for up to 500 examples, the models trained under this method don't do quite well to achieve its goal.

* **Issues with Lane Spacing**

We observed that the presence of lane spacing, particularly in AI-generated text, affected the model’s predictions. Additionally, certain human writings that resembled poems and contained similar lane spacing were incorrectly classified as AI-generated text. The problem can sometimes be fixed after deleting the lane spacing and getting the correct output but requires manually adjusting the input data.

* **Difficulties Implementing API**

The tune model is strictly limited access by the Palm2 API and requires Oauth 2.0 authentication. The API for the text prompt is a lot easier to access without the authentication. Considering the tune model is provided for a wider range of users, the API authentication should be set easier to access and provide convenience for multiple operating environment users.

1. **Suggestions to MakerSuite**

* **Incorporate Recursive Learning and Example Limitation**

We recommend that MakerSuite enhance Palm2’s capabilities by allowing for recursive learning and provide larger capacity for the example training dataset. This would enable the model to iteratively learn and improve its predictions with varieties and could be further optimized based on user input and corrections.

* **Improve Handling of Data Prompt**

The data prompt session doesn’t seem to work well with multiple answers and produce false or non-sense results. Given that line spacing affects the model’s accuracy, we suggest that MakerSuite should refine Palm2’s ability to analyze and interpret text formatting elements, such as lane spacing, to avoid misclassifications.

* **Pattern Recognition Enhancements**

We propose that improvements be made to allow Palm2 to recognize subtle nuances and patterns in the text, ensuring more accurate identification of AI-generated content even when small changes are made to the input.

**References**

Zinshteyn, M., &amp; Jones, C. (2023, October 23). Academic Officials Struggle to Find out If AI Wrote Students’ College Essays.

Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, Maggie Demkin. (2023). LLM - Detects AI Generated Text. Kaggle. <https://kaggle.com/competitions/llm-detect-ai-generated-text>

Cingillioglu, I. (2023). Detecting AI-generated essays: the ChatGPT challenge. International Journal of Information and Learning Technology, 2056-4880. <https://doi.org/10.1108/IJILT-03-2023-0043>

**Appendix**

Our original focus was on Digital Twin; however, having engaged in this project, we are keen on pivoting our research towards LLM for our EL-2 final projects.

All three of us – James Lo, James Zhou, and Jinny Kim – actively participated and contributed equally to the project.