Gemma

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large language models, available in English, with open weights for both pre-trained variants and instruction-tuned variants. Gemma models are well-suited for a variety of text generation tasks, including question answering, summarization, and reasoning. Their relatively small size makes it possible to deploy them in environments with limited resources such as a laptop, desktop or your own cloud infrastructure, democratizing access to state of the art AI models and helping foster innovation for everyone.

Gemma-2-2b-it:

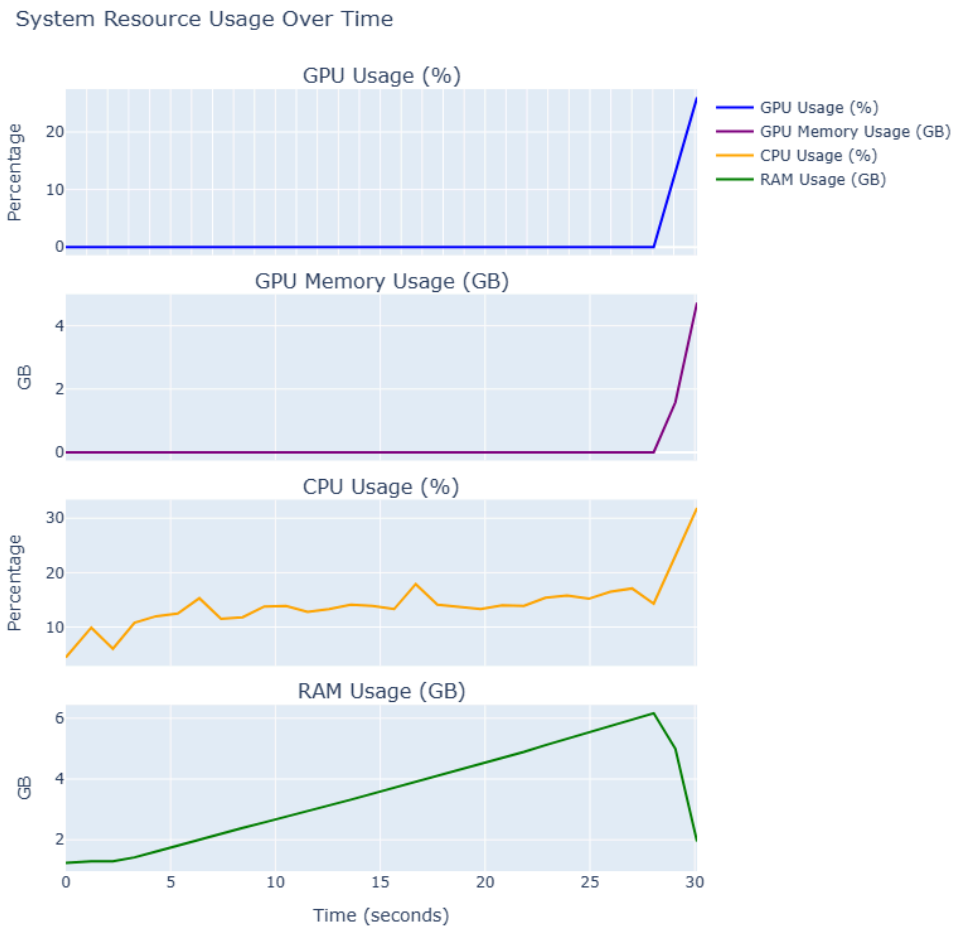
A PyTorch implementation of Gemma 2B instruction tuned model. It is a 2B parameter base model that has been instruction-tuned so it can respond to prompts in a conversation manner.

The Gemma-2-2b architecture leverages advanced model compression and distillation techniques to achieve its superior performance despite its compact size. These methods enable the model to distill knowledge from larger predecessors, resulting in a highly efficient yet powerful AI system.

Gemma-2-2b was trained on a substantial dataset comprising 2 trillion tokens, utilizing Google's state-of-the-art TPU v5e hardware. This allows for rapid and effective training, ensuring the model can handle diverse and complex tasks across multiple languages.

Compared to other models in the Gemma family, such as the 9 billion (9B) and 27 billion (27B) parameter variants, Gemma 2 2b stands out for its balance between size and efficiency. Its architecture is designed to perform exceptionally well on a wide range of hardware, from laptops to cloud deployments, making it a versatile choice for both researchers and developers.

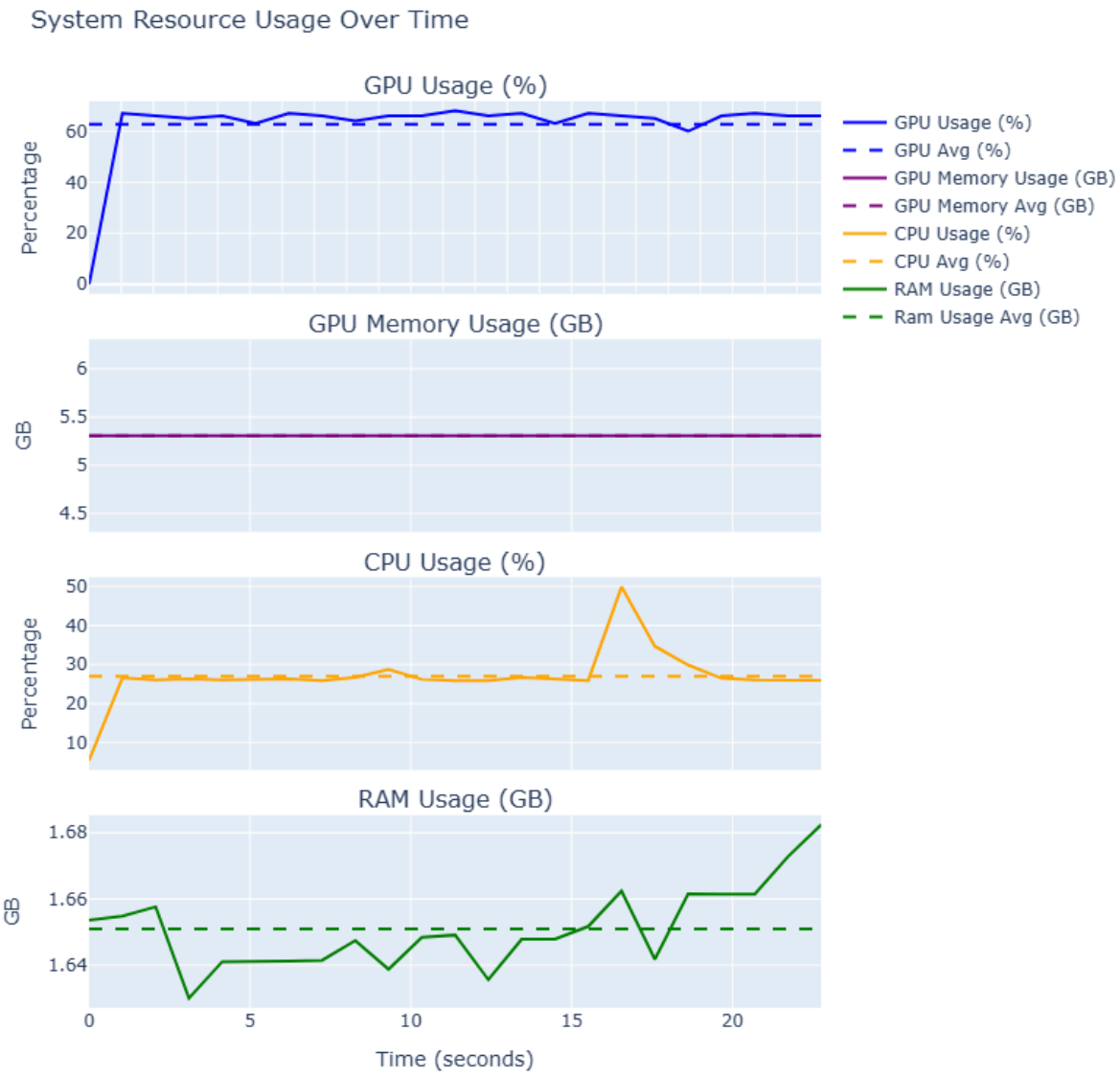
Loading the Model to Kaggle:



Time taken to load model to Kaggle T4 GPU: 30.1109 seconds
Maximum GPU Usage: 26 %
GPU memory used : 4.729 GB
Maximum CPU usage : 31.9 %
Maximum RAM usage: 6.164 GB

Performing Single Inference:

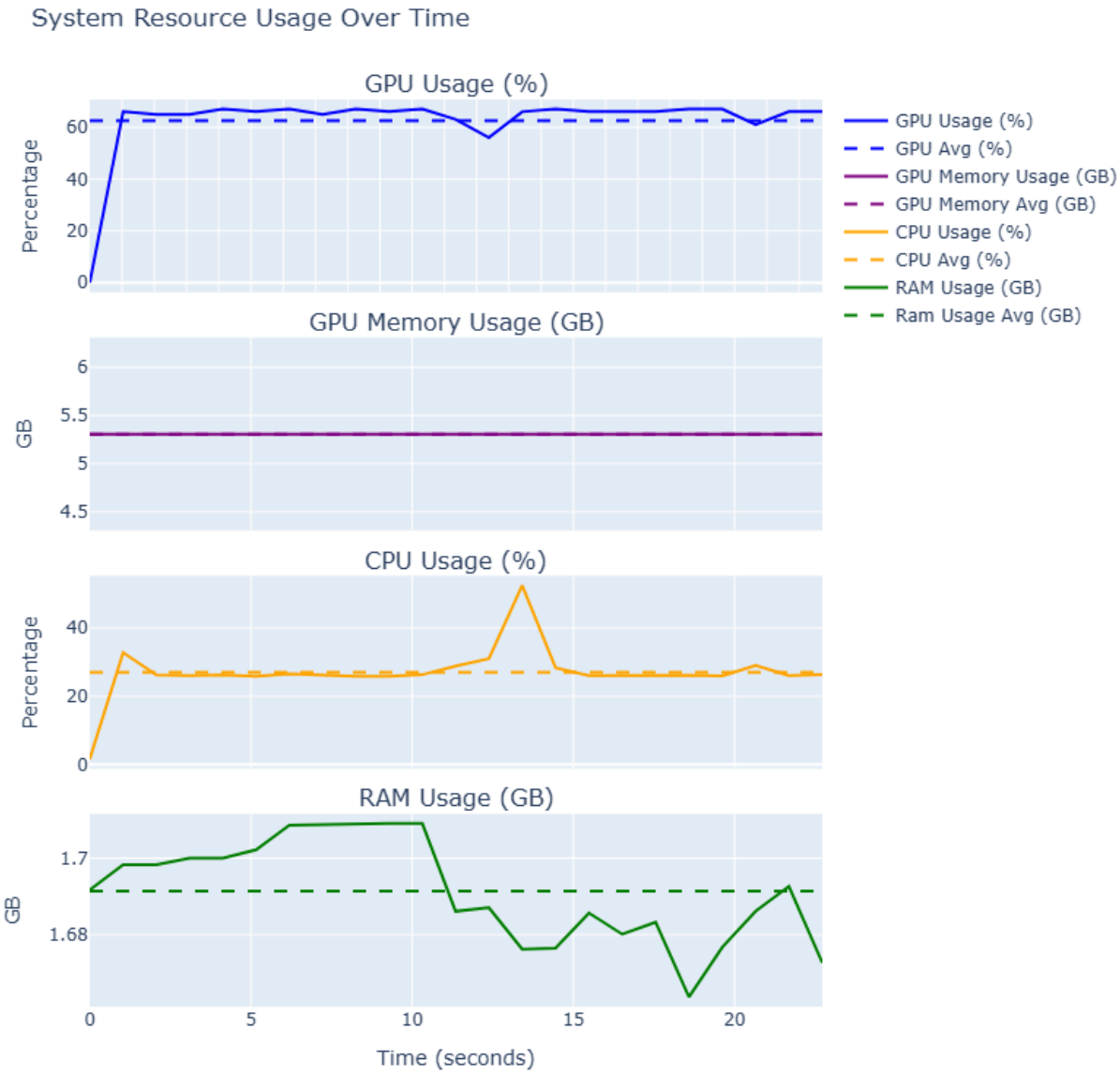
Input Data = “What is the capital of France?”
Output = “The capital of France is **Paris**. 🇫🇷”
Output Length = 500 tokens



Time Elapsed: 22.75566 seconds
Average GPU Usage: 62.73 %
GPU Memory Usage: 5.307617 GB
Average CPU Usage : 26.986 %
Average Ram Usage: 1.65095 GB

Single Sentence Translation:

Input Data = “म पुस्तक पढिरहेको छु। Translate to English.”
Output = “I am reading a book.”
Output length = 500 tokens



Time Elapsed : 22.7204 seconds
GPU Usage : 62.52174 %
GPU Memory Used: 5.307617 GB
CPU Usage : 27.0087 %
Average RAM usage : 1.6914 GB

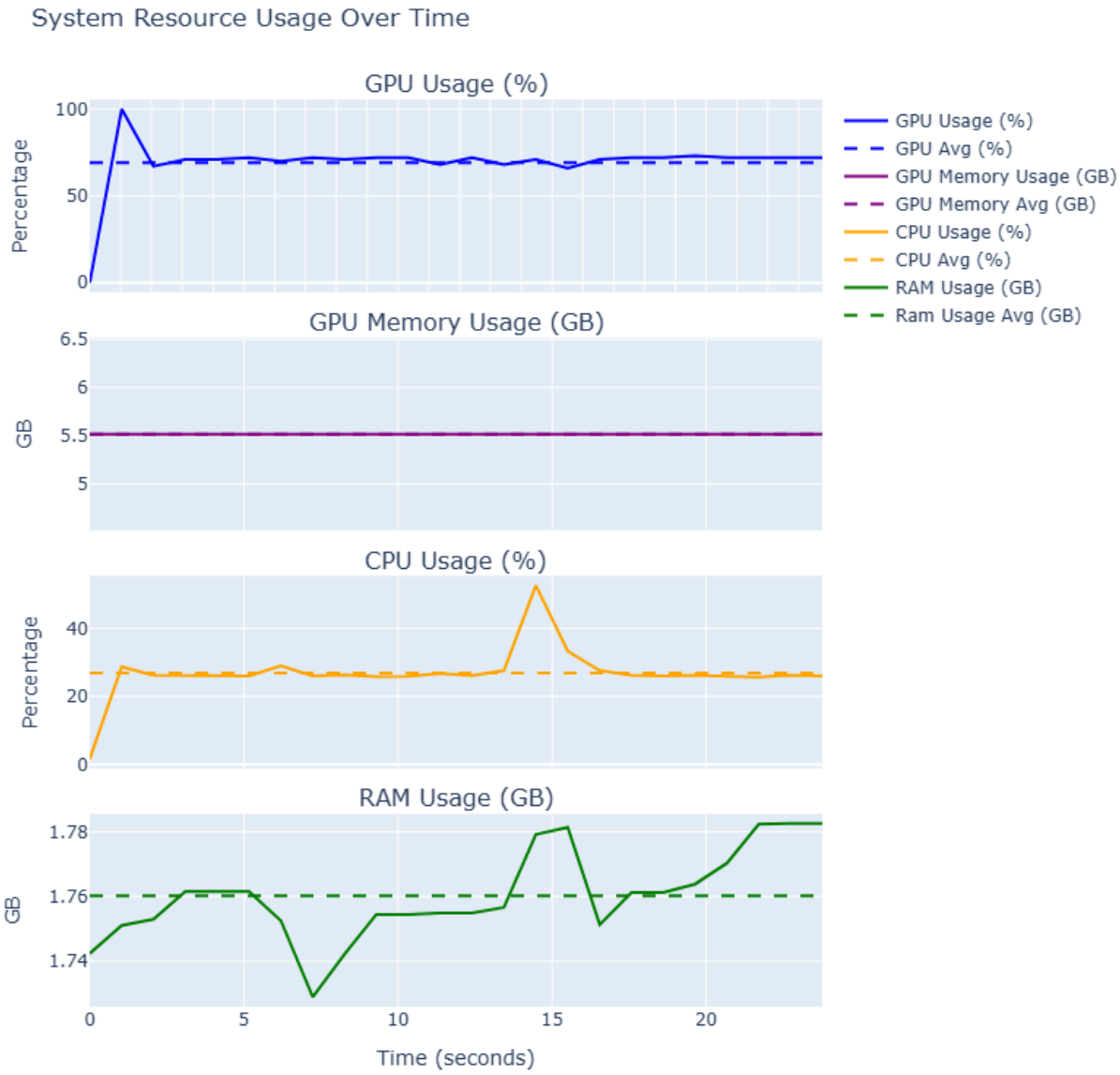
Calculating average inference time for 10 identical inputs.

Input Data = नेपालको राजधानी के हो?
Output Length = 500 tokens

Average GPU Usage = 67.5656%
GPU Memory Usage = 5.3076 GB
Average CPU Usage = 27.35 %
Average RAM Usage = 1.7358 GB
Total Time Taken = 227.1612 seconds
Average Time for a single inference = 227.1612 / 10
= 22.71612 seconds

Summarization Test:

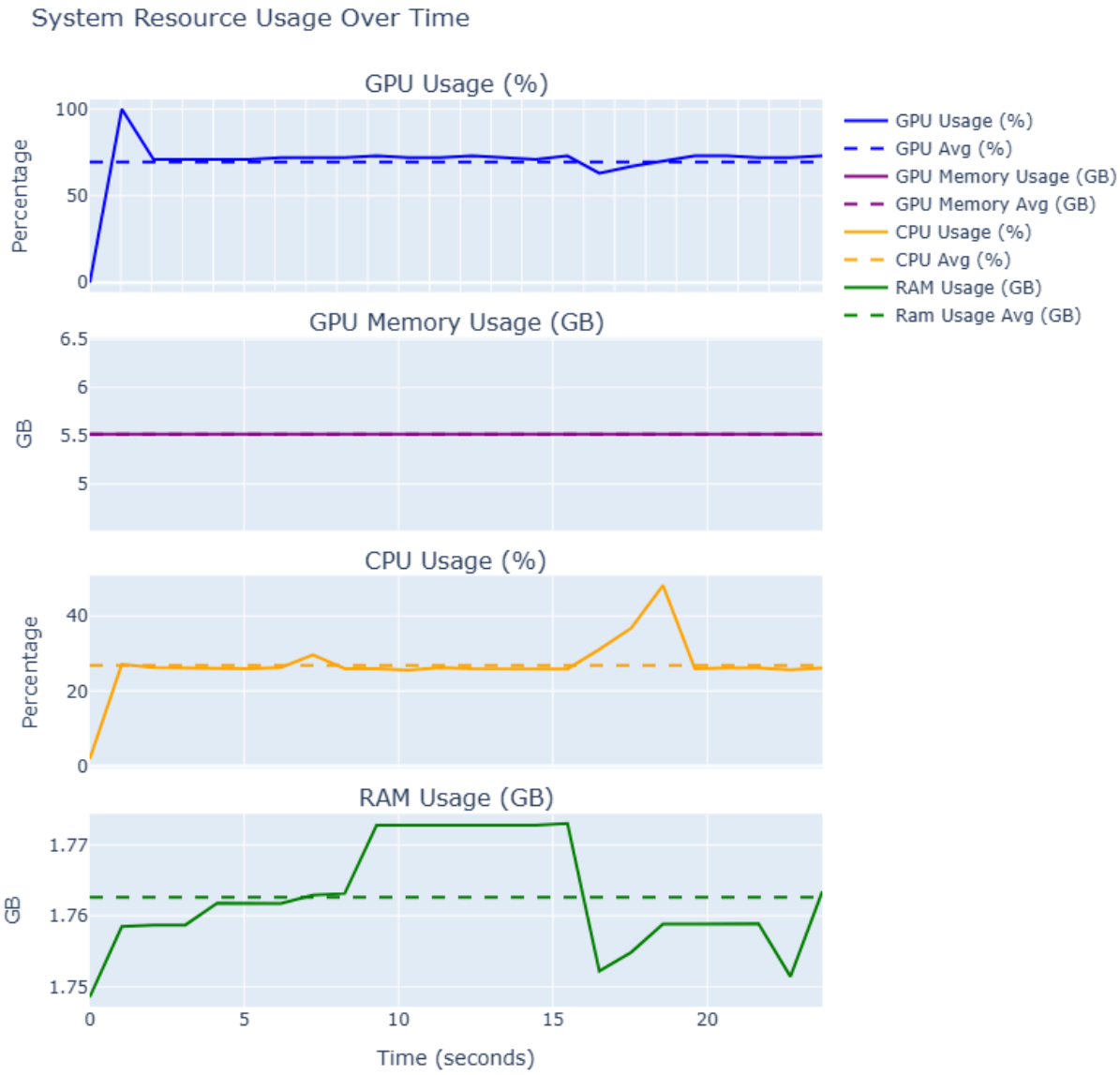
Input Data Length = 788 tokens
Output Length = 500 tokens



Time Elapsed = 23.76627 seconds
Average GPU Usage = 69.125 %
GPU Memory Usage = 5.518555 GB
Average CPU Usage = 26.84583 %
Average RAM Usage = 1.76026 GB

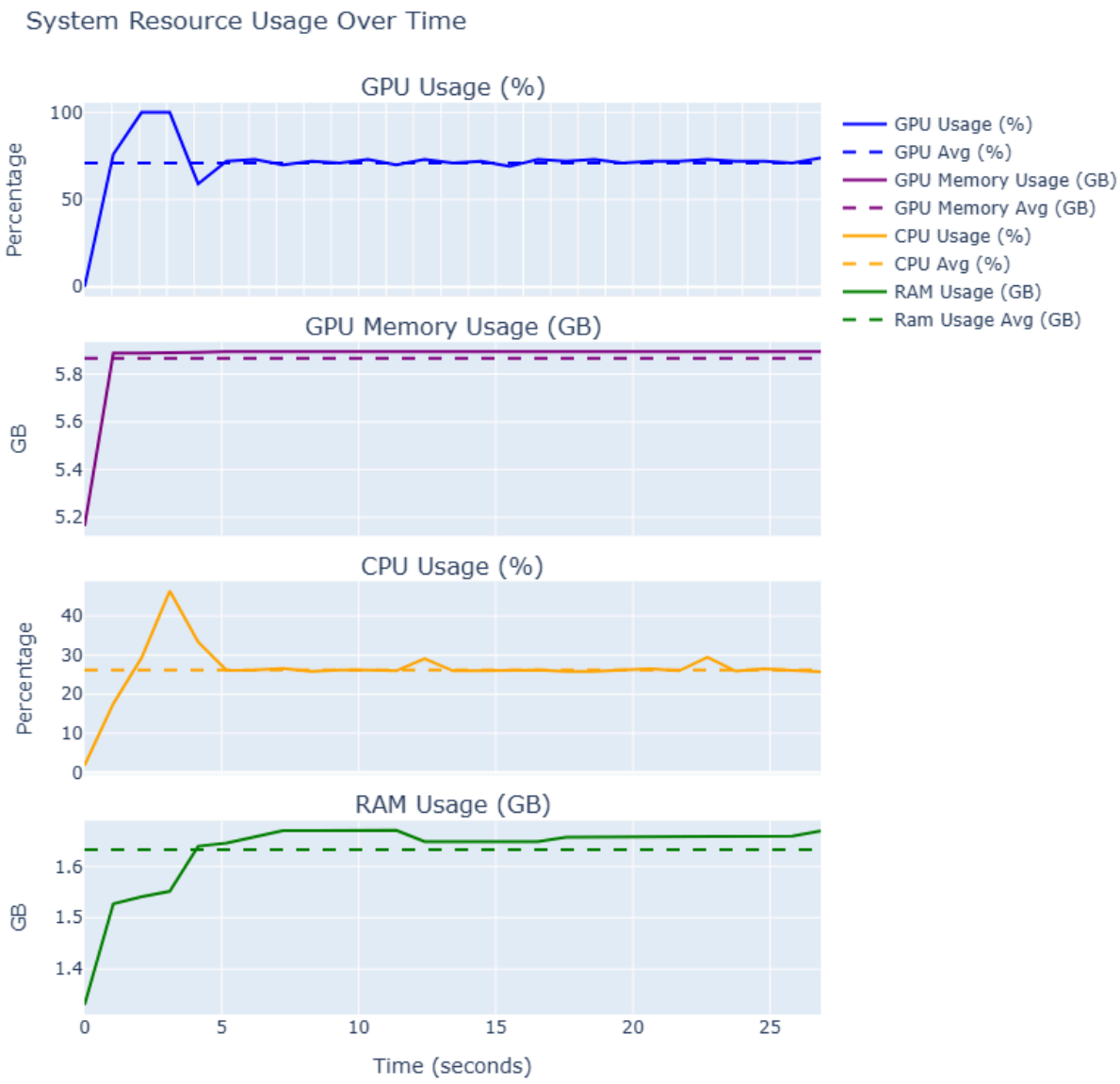
Translation Test on Longer Input Tokens:

Input Data Length = 788 tokens.
Output Length = 500 tokens



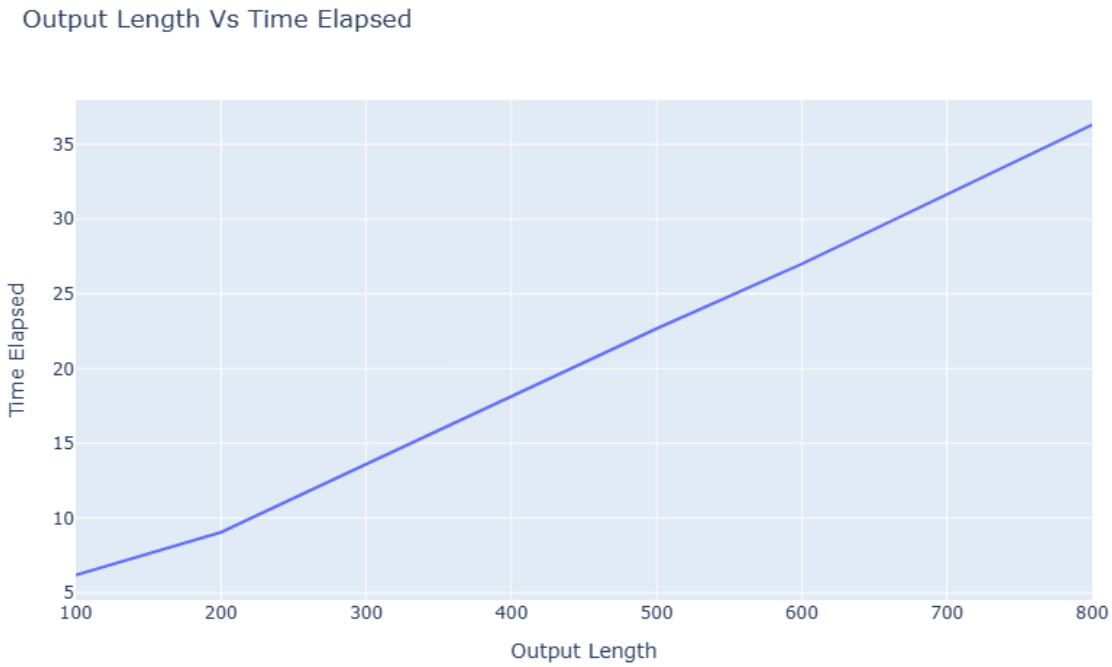
Time Elapsed = 23.72506 seconds
Average GPU Usage = 69.54167 %
GPU Memory Usage = 5.518555 GB
Average CPU Usage = 26.81667 %
Average RAM Usage = 1.762645 GB

Summarization Test :
Input Data Length = 1578 tokens
Ouput Length = 500 tokens



GPU Average Usage = 70.9629 %
GPU Memory Usage = 5.8935 GB
Average CPU Usage = 26.26 %
Average RAM Usage = 1.6323 GB
Time Elapsed = 26.8527

Output Length VS Time Elapsed for Constant Input Data:



The plot shows the time taken to generate output of certain length. It shows that output length and time taken has a linear relationship.

Summary:

All tests performed with output length = 500 tokens.

Input	Time Elapsed (seconds)	Average GPU Usage (%)	Max GPU Usage (%)	Average CPU Usage (%)	Max CPU Usage (%)	Average RAM Usage (GB)	Max RAM Usage (GB)	GPU Memory Usage (GB)
Loading the gemma-2-2b-it model	30.1109	-	26	-	31.9	-	6.1640	4.7290
What is the capital of France?	22.7557	62.7300	67	26.9860	49.9	1.65095	1.6820	5.3076
म पुस्तक पढिरहेको छु। Translate to English.	22.7204	62.5217	67	27.0087	52.5	1.69140	1.7090	5.3076
नेपालको राजधानी के हो?	22.7161	67.5656	71	27.3500	55.8	1.73580	1.7747	5.3076
Summarization: Input Length = 788 tokens	23.7662	69.1250	100	26.8458	52.6	1.76026	1.7825	5.5185
Translation: Input Length = 788 tokens	23.7251	69.5454	100	26.8167	48.2	1.76260	1.7728	5.5186
Summarization: Input Length = 1578 tokens	26.8527	70.9629	100	26.2556	46.4	1.63235	1.6694	5.8935