SCI Foundation Models Preliminary Results

1 Evaluation Metrics and Results

To fully understand the quality and carbon impact of various foundation models at the inference stage, it is essential to comprehensively evaluate them from different key perspectives such as response time, power/energy, output quality, and carbon intensity. In this report, we present our evaluation metrics and results from these important perspectives.

1.1 Response Time, Power and Energy

Table 1 shows the average total power used by each output produced by the respective models. The results are broken down as follows:

- Time - The average amount of time in seconds taken for a prompt to be evaluated and an output generated. In the case of GPT-2, for instance, each prompt took roughly 27.5 seconds to be evaluated before the model generated and recorded a reply.
- Power - Average power consumed in watts over the total time each prompt was evaluated until terminating in an output.
- Total Energy - The amount of joules consumed during the entire prompt evaluation process. This is calculated by the formula $Watts * Seconds = Joules$.

From Table II, we can observe that the number of parameters is directly related to the response time and total energy consumed by each model. For example, GPT-J 6B is nearly twice as power-hungry per run than the next closest model (GPT-Neo 2.7B) when deployed in CPUs, using over 6,800 joules of energy to generate a single response. Further, GPT-J 6B takes nearly 18x as long to complete a single output as compared to GPT-Neo 125M. The power consumed by the models is fairly consistent. However, GPT-2 (with 1.5 billion parameters) appears to be an exception by using $\sim$13% more power than the next closest model (GPT-Neo 1.3B). However, GPT-2 takes less time to run thus still saves roughly 12% energy than GPT-Neo 1.3B.

Table 2 summarizes the results when deploying these models on GPUs, from which we can observe that both the response time and energy efficiency have

Table 1: Average Foundation Model Power Usage - CPU Only

Model	Time (Seconds)	Power (Watts)	Total Energy (Joules)
GPT-J 6B	139.1	48.9	6802.9
GPT-Neo 2.7B	68.6	51.8	3553.5
GPT-Neo 1.3B	35.7	54.2	1934.9
GPT-Neo 125M	7.8	50.3	392.3
GPT-2	27.5	61.7	1696.8

Table 2: Average Foundation Model Power Usage - GPU Accelerated

Model	Time (Seconds)	CPU Power (Watts)	GPU Power (Watts)	Total Energy (Joules)
GPT-Neo 1.3B	2.1	42.6	36.8	166.7
GPT-Neo 125M	1.0	42.9	21.7	64.6
GPT-2	2.4	43.2	30.5	176.9

been significantly increased for all three foundation models. While the CPU consumes between 15%-30% less power, the GPU utilizes a considerable amount, resulting in roughly 25% more total power on average when compared to the CPU only results. However, the time taken to produce each output is drastically reduced. For example, we see a 17x speedup for GPT-Neo 1.3B, which contributes to 91% overall energy reduction. Contrasting GPT-Neo 1.3B and GPT-2 again, their response time and energy efficiency are fairly comparable when deploying on GPUs. It can also be observed that the smallest model GPT-Neo 125M is 2 times faster and consumes much less energy but the quality of its output is a concern.

1.2 Writing Performance

We consider the quality of NLP outputs is as important as their energy and carbon efficiency. In fact, in some cases it might be even more important, depending on the circumstances. For example, an AI system utilizing NLP that guides a doctor through a complex medical procedure would need to be much more accurate than an NLP system that does general translation. Thus, a rigid evaluation on model quality is required. According to the Program for the International Assessment of Adult Competencies (PIAAC) in a 2012 study, the average adult in the United States reads at approximately an 8th grade level [1] [2]. As this work is concerned with NLP outputs of a more general nature, we target the 8th reading level as a baseline for output evaluation. In 2012, the Texas Education Agency adopted The State of Texas Assessments of Academic Readiness (STAAR) [3]. As part of this adoption, evaluative measures were made available to all Texas teachers to ensure a broadly applicable model that could assess various skills expected of students within the public school system, including reading and writing. Via STAAR, a rubric [4] and scoring guide [5] are available for English I, a class taken by all freshman-level (9th grade) students in Texas. STAAR rubrics are not available to assess 8th grade reading and writing, hence the selection of this rubric instead.

Table 3: Average Foundation Model Writing Performance - CPU Only

Model	Writing Performance Score
GPT-J 6B	5.83
GPT-Neo 2.7B	6.63
GPT-Neo 1.3B	6.17
GPT-Neo 125M	3.57
GPT-2	5.83

Table 4: Average Foundation Model Writing Performance - GPU Accelerated

Model	Writing Performance Score
GPT-Neo 1.3B	5.87
GPT-Neo 125M	3.5
GPT-2	5.83

The rubric is divided into three portions:

- Organization/Progression - the clarity of organization, focus, and responsiveness of the output
- Development of Ideas - the output contains details and examples appropriate to the prompt, and also how engaging and interesting it is to read
- Use of Language/Conventions - precision of word choice, clear, correct grammar, and purposeful sentence structure

For each output generated by selected foundation models, the portions of the rubric are evaluated separately and assigned a numeric score between 1 and 4. These three portions are then added together, resulting in a total score between 3 and 12, with 12 being the highest possible score, i.e., the output is an example of superior writing. 3 denotes a nonsensical output, with poor grammar, little acknowledgement of the prompt, and lack of overall coherence. To eliminate as much bias as possible, all outputs are scored randomly, i.e., selected and scored with no indicator as to the model employed.

All 240 outputs are scored using the aforementioned rubric, and the final results are tallied and averaged in tables 3 and 4 for comparisons. The average writing performance for each model is denoted, separated between CPU Only and GPU Accelerated results.

Comparing the CPU Only and GPU Accelerated results, it is immediately apparent that there is no statistical significance between the quality of output generated by either method. Thus, when possible, it would be best practice to always use the GPU if speed and power savings are valued. However, a very significant difference can be seen in the score differential amongst the more power-hungry models. As shown in table 1, GPT-J 6B uses vastly more energy during the production of each output, but performs roughly on par with GPT-Neo 2.7B, GPT-Neo 1.3B, and GPT-2. GPT-Neo 2.7B has the highest

writing performance score (12% better than GPT-J 6B) but consumes only 52% of power. There could be many reasons for this disparity, including the quality of the prompt selected, researcher bias when reading the outputs, or pure randomness. It is entirely possible that some models are simply better at certain sorts of outputs than others.

Interestingly, GPT-Neo 1.3B and GPT-2 are very similar in writing performance, which aligns closely with their very similar energy consumption. When comparing these two models to GPT-Neo 2.7B, there is roughly a 2X energy increase for the at-best aforementioned 12% better score, and thus an important decision must be made: how valuable is that 12% increase in quality?

Finally, it is worth noting that judging the actual quality of output from the various foundation models revealed the difficulty in reproducing natural language within a synthetic environment. Even though strict rules for English exist, there are countless instances where those rules might be broken for sake of clarity, emphasis, or even stylistic choice. What separates good English from bad English might be a matter of taste, and thus the criteria for tailoring output changes from domain to domain. Indeed, the "mere" 12% increase in quality of output might be extremely significant if an expert system is needed vs. an online chatbot that can be expected to spout gibberish in some instances.

1.3 Software Carbon Intensity (SCI)

SCI is a rate of carbon emissions per one unit of R, and is represented by the following equation:

$$SCI = ((E * I) + M) \text{ per } R \tag{1}$$

Where:

- E = Energy consumed by a software system
- I = Location-based marginal carbon emissions
- M = Embodied emissions of a software system
- R = Functional unit (e.g. carbon per additional user, API-call, ML job, etc.)

R in this case refers to each individual run of the AI, i.e. evaluating a single prompt and producing an output. E per each R has already been calculated in tables 1 and 2. As the test server is physically located in San Marcos, TX, power is supplied by Electric Reliability Council of Texas (ERCOT). Per the 2019 Grid Electricity Emissions Factors report, systems utilizing the ERCOT power grid generate 0.4784 $kgCO_2e$ per kWh [6], and is assigned as the value of I. M is calculated as follows:

$$M = TE * (TR/EL) * (RR/TR) \tag{2}$$

Where:

- TE = Total Embodied Emissions, the sum of LCA emissions for all hardware components
- TR = Time Reserved, the length of time the hardware is reserved for use by the software
- EL = Expected Lifespan, the anticipated time that the equipment will be installed
- RR = Resources Reserved, the number of resources reserved for use by the software
- TR = Total Resources, the total number of resources available

TE is somewhat difficult to calculate, and thus we use the Boavizta dataset to estimate the TE for the test server utilized in these experiments [7] [8]. Per AMD, the Ryzen 2950x processor has a TDP of 180W [9]. The closest system in the Boavizta database with a similar processor TDP has a value of is a Dell PowerEdge R540 with a processor TDP of 165W. As the PowerEdge is a very similar high-end system, this is a good sample for approximating embodied emissions. gwp_total for the PowerEdge is 8230 kgCO_2e, and thus this value is used for TE in these calculations. Due to the lack of information available on carbon emissions during the GPU manufacturing process, TE (and subsequently SCI) is calculated under the assumption that the server only utilized the CPU when running models.

TR is selected as three years due to the length of actual usefulness of the system; this is a typical length of time a system of this sort is deployed before replacement is necessary to keep up with advancements in technology. EL, however, is set at five years as this is the average length of time a server is expected to be deployed before replacement. Essentially it is assumed that although a system of this sort can be expected to perform without hardware failure for five years, after three years the hardware will be insufficient to run contemporary jobs.

RR and TR are both set to 1. Here we assume most of the system resources will be utilized when executing large tasks submitted to these foundation models. Using the values above, the value of M would be calculated as:

$$M = 8230 kgCO_2e * (3/5) * (1/1) = 4938 kgCO_2e \tag{3}$$

For purposes of this calculation, joules of energy must be converted to kWh, with 3,600,000 joules equal to one kWh. Using GPT-J 6B, the total energy per model run would thus be 6802.9/3600000 = 0.00189 kWh. Further, we assume the foundation model service will be available for the entire lifespan of the server (3 years) and thus we convert the time for each run into the total number of AI model runs over that time span: 3 years * 31,536,000 seconds/years / 139.1 seconds = 680,143.78 total runs. Hence, the total amount of energy utilized by GPT-J 6B over the lifespan of the server would be 680,143.78 * 0.00189 kWh

Table 5: Calculated SCI for All Foundation Models

	GPT-J 6B	GPT-Neo 2.7B	GPT-Neo1.3B	GPT-Neo 125M	GPT-2
Total kWh - CPU	1285.26	1361.31	1424.35	1321.75	722.72
Total Runs - CPU	680,143	1,379,125	2,650,084	12,129,230	3,440,290
Total Emissions - CPU	5552.97	5589.25	5619.41	5570.32	5283.75
SCI - CPU	8.16	4.05	2.12	0.46	3.45
Total kWh - GPU	-	-	2086.13	1697.69	1937.06
Total Runs - GPU	-	-	45,051,428	94,608,000	39,420,000
Total Emissions - GPU	-	-	5936.01	5750.17	5864.69
SCI - GPU	-	-	0.13	0.06	0.15

= 1285.47 kWh. SCI for GPT-J 6B is then calculated as:

$$TotalEmissions = (1285.47kWh \times 0.4784kgCO_2e/kWh) + 4938kgCO_2e = 5552.97kgCO_2e (5)$$

SCI is per run of each instance of the GPT-J 6B model and would be calculated simply as 5552.97/680,143.78 = 0.0082 $kgCO_2e$, or 8.2 gCO_2e.

We calculate the SCI of all foundation models and list the results in Table 5. It is apparent that the carbon expenditure during creation of the server itself has a massive impact on the overall SCI value. This indicates that selecting the device to run a foundation model is extremely important when considering carbon waste. Further, while GPT-J 6B is the most power hungry model for each run, the fact that the smaller GPT-Neo 2.7B and 125M models can run significantly more times during the server deployment time, which means those models could actually use more power and produce more total carbon waste if run continuously. Another significant finding is that using the GPU produces a much lower carbon impact from the ability to run the model many more times. SCI is substantially lower, but overall power usage is still higher.

References

[1] M. G. (ETS), R. F. (ETS), L. M. (Westat), T. K. (Westat), and J. H. (Westat), "Literacy, numeracy, and problem solving in technology-rich environments among u.s. adults: Results from the program for the international assessment of adult competencies 2012," https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014008, 2012.

[2] A. Wylie, "What's the latest u.s. literacy rate?" https://www.wyliecomm.com/2021/08/whats-the-latest-u-s-literacy-rate/, 2021.

[3] T. E. Agency, "Staar english i and english ii resources," https://tea.texas.gov/student-assessment/testing/staar/staar-english-i-and-english-ii-resources, 2021.

[4] ——, "English i expository writing rubric," https://tea.texas.gov/sites/default/files/Rubric-EOC-Eng1-WrtgExpository.pdf, 2011.

[5] ——, "English i expository scoring guide," https://tea.texas.gov/sites/default/files/2021-staar-english-1-scoring%20guide-tagged.pdf, 2021.

[6] Carbonfootprint.com, "2019 grid electricity emissions factors," https://www.carbonfootprint.com/docs/2019_06_emissions_factors_sources_for_2019_electricity.pdf, 2019.

[7] Boavizta, "Digital and environment: How to evaluate server manufacturing footprint, beyond greenhouse gas emissions?" https://www.boavizta.org/en/blog/empreinte-de-la-fabrication-d-un-serveur, 2021.

[8] ——, "Environmental footprint data," https://github.com/Boavizta/environmental-footprint-data/blob/main/boavizta-data-us.csv, 2022.

[9] AMD, "Amd ryzen threadripper 2950x processor," https://www.amd.com/en/products/cpu/amd-ryzen-threadripper-2950x, 2022.