Towards a Better Understanding of AI

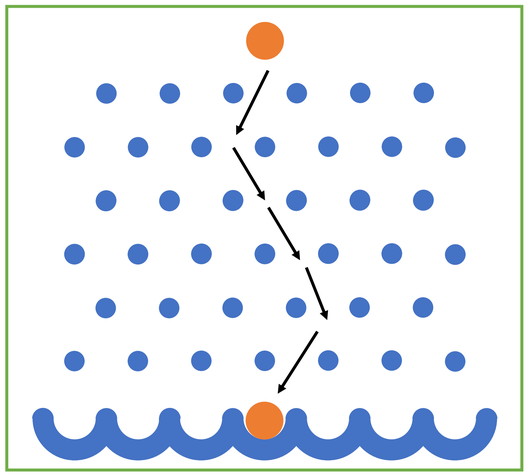
Ian Hastings[1], Oscar Hernandez[2], Aaron Welch[3]

[1]Next Generation Pathways to Computing, [2]Computational Sciences and Engineering, [3]Computer Science and Mathematics

1. Abstract

Artificial intelligence is everywhere. It's been implemented into graphic design, advertisements, search engines, and even teaching. Tech companies are in an arms race to develop the most accurate and highest performing AI, paying top dollar for power large-scale data centers. But as we use AI, it uses energy. But what determines the amount of energy used? And how do AIs work in the first place? My name is Ian Hastings, and I analyzed different sizes of AI models and compared their energy and power usage.

2. Background



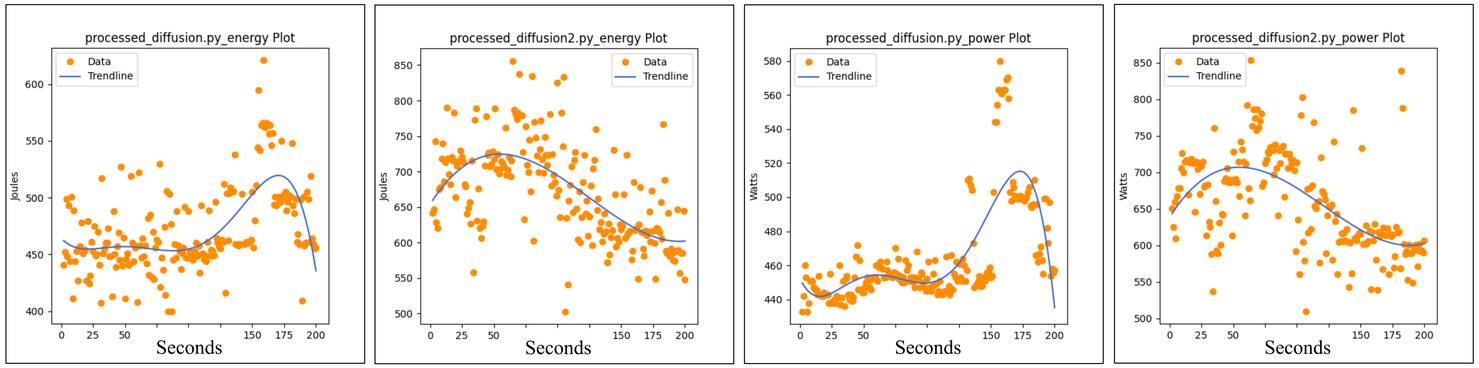
Plinko Diagram

AI Process:
- First: The input text is tokenized
- Then: Relationships between tokens are computed
- Finally: Output is decoded

- Like plinko, the input is shifted based on the parameters and training of the AI

3. Method and Results

Base Model Energy | Refined Model Energy | Base Model Power | Refined Model Power



The energy and power consumed by the refined model is much higher than the base model.

4. Results

- Prompt: "A hyper realistic photograph of a dew-covered spider web sparkling in the morning sun. Each strand is individually discernible, with water droplets reflecting the vibrant colors of a nearby flower."

- Precision vs Consumption

Refined Diffusion Model

Base Diffusion Model

5. Next Steps

- Run more models to compare performance across different sizes
- Compare very different model structures and their consumption of power
- Use a performance monitor, like Vampir, to analyze every detail

6. References and Acknowledgements

I acknowledge my mentors Aaron Welch and Oscar Hernandez for their support and contributions to this project. This research was supported in part by an appointment to the RENEW Pathway Summer Schools Program at Oak Ridge National Laboratory, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education.
This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.