KPMG Trusted AI:
Energy Usage and Sustainability

Data science, machine learning, and sustainability applied to derive and optimize energy insights.

Break Through Tech x Cornell Tech Fellowship 2025-26
KPMG 1C

Our Team

Krithika Kondapalli
Computer Science +
Sustainability Studies
University of Florida ‘27

Athena Tian
Biomedical Engineering +
Legal Studies
Northwestern University ‘28

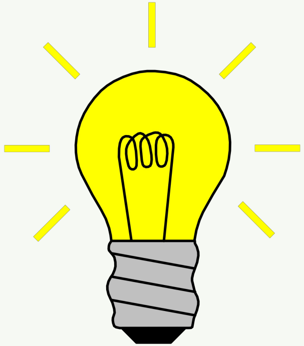
Sandy Zheng
Computer Science +
Entrepreneurship &
Innovation
University of Chicago ‘27

Mercy Ifiegbu
Computer Science
Bucknell University ‘26

Shreyosee Chowdhury
Computer Science +
Administration and
Management
City College of New York ‘27

Agenda

AI Energy Consumption



67.5 kWh

Daily energy consumption of AI queries by KPMG audit agents

#4

Finance has fourth-highest energy burden to market capitalization ratio

40%

Microsoft reduced the Water Usage Effectiveness (WUE) from 0.49L/kWh (2021) to 0.3 L/KWh (2024)

KPMG: Trusted AI Initiative

Values Driven

Human Centric

Trustworthy

Explainability

Data Integrity

Transparency

Fairness

Accountability

Reliability

Security

Sustainability

Safety

Privacy

Project Overview

Initial Hypothesis: Energy usage primarily depends on the model type.
Research Question: Within each model type, which specific features drive energy usage?

Project Goals:

- ❑ Analyze AI systems energy consumption with Random Forest Model.
- ❑ Identify top drivers of energy usage.

Outcomes:

- ❑ Suggest operational strategies for reducing energy consumption.
- ❑ Strategize to improve efficiency and sustainability

01

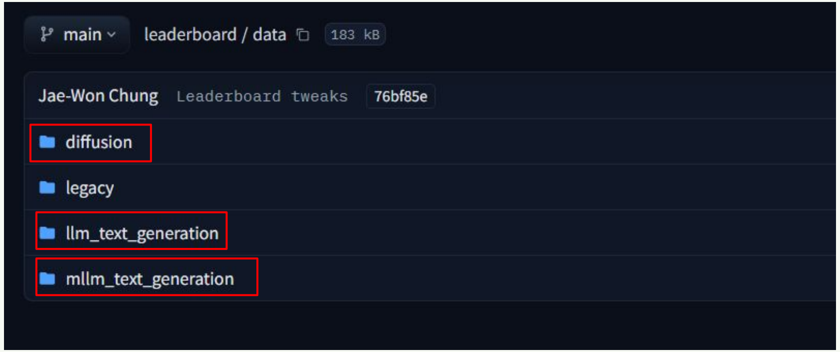
Data Preparation & Understanding

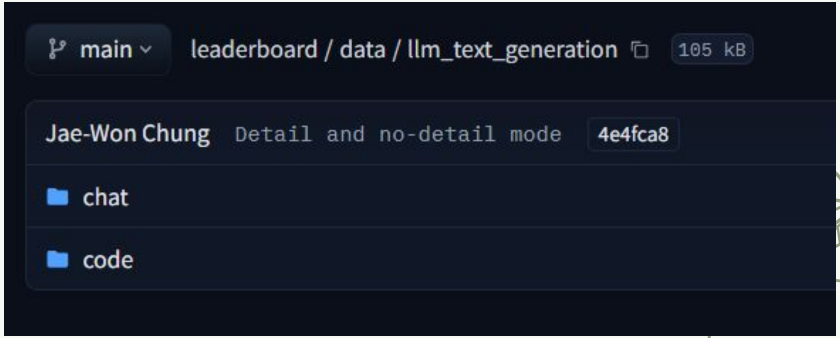
Our Data Set

This data set tracks energy use of GenAI models across tasks to help evaluate sustainable energy usage.

ML.Energy Leaderboard Data Set ~ 470

Diffusion ~ 118	llm ~ 314	mllm ~42
image-to-video	chat	chat
text-to-image	code	
text-to-video		





Our Data's Features

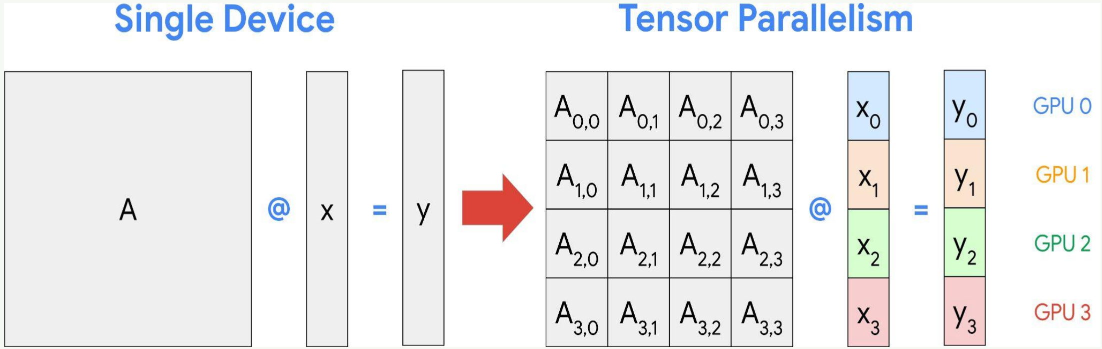
- Model
- GPU
- TP: Tensor Parallelism
- PP: Pipeline Parallelism
- Energy/req (J)
- Avg TPOT
- Token tput (tok/s)
- Avg Batch Size (reqs)
- Max Batch Size (reqs)

Individual Raw Entry:

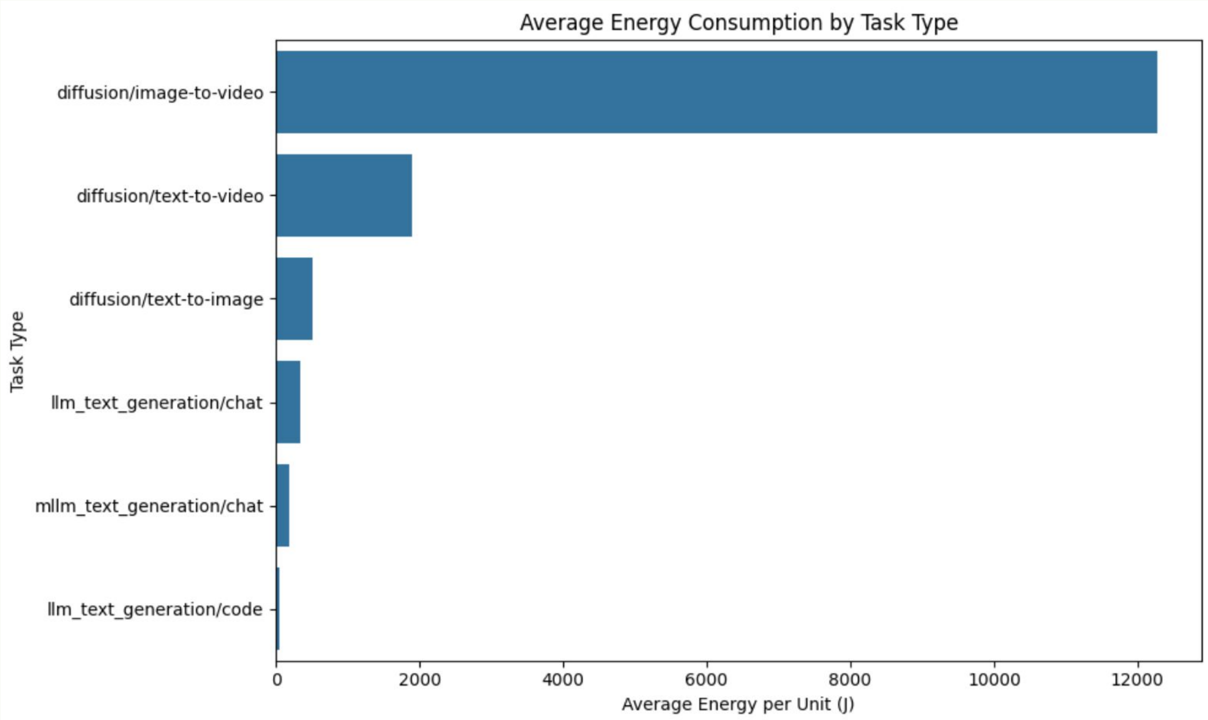

```json
{
  "Model": "google/gemma-2-27b-it",
  "GPU": "NVIDIA A100-SXM4-40GB",
  "TP": 4,
  "PP": 1,
  "Energy/req (J)": 230.644433524673,
  "Avg TPOT (s)": 0.11220224938943191,
  "Token tput (tok/s)": 1051.3354539260472,
  "Avg Output Tokens": 392.47133333333335,
  "Avg BS (reqs)": 127.72361537073786,
  "Max BS (reqs)": 128
}
```

Simplified Terms

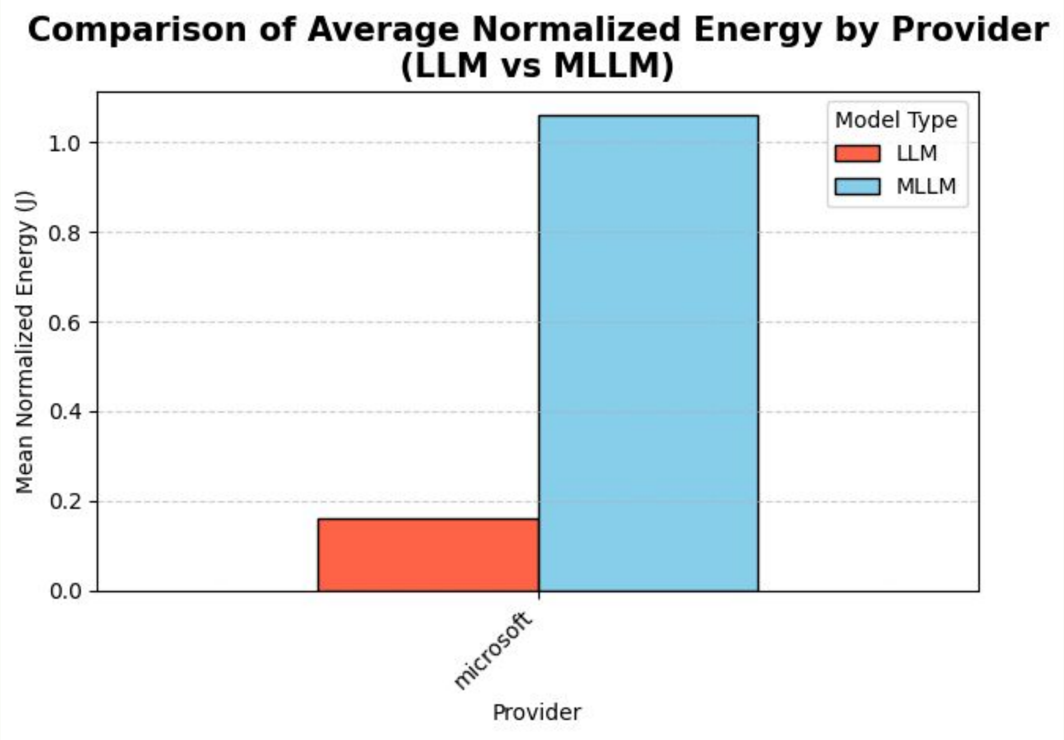
Tensor Parallelism	Splitting models across devices for speed
Energy/req (J)	Energy per request or query
Avg TPOT	Average response time
Token tput (tok/s)	Generation speed (words or images per second)
Token	Basic processing unit



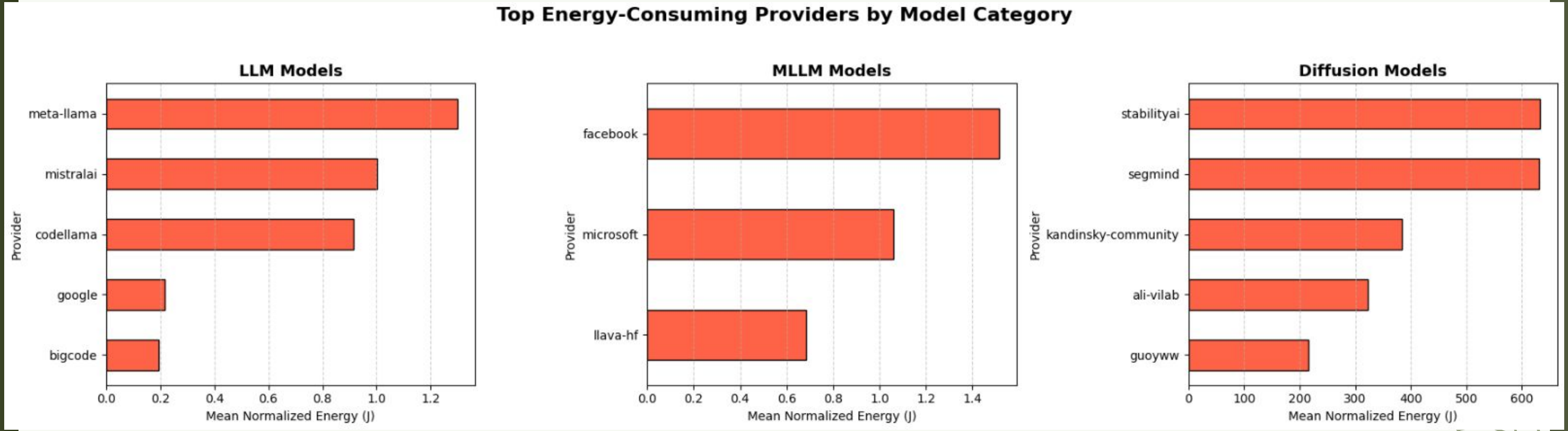
Data Insights: Task Type



Data Insights: LLM vs MLLM



Data Insights: Top Providers



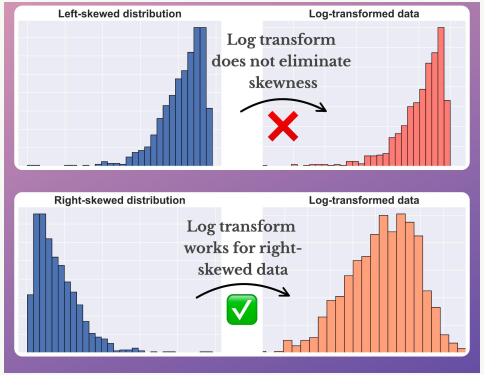
Larger models = more parameters → higher compute → higher energy

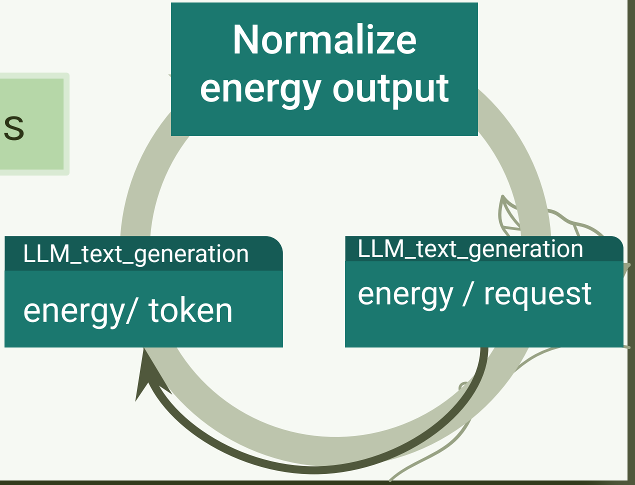
Data Preparation

Standardized energy across model types

Handle missing/ non-applicable features

Log-transformed energy values





02

Modeling & Evaluation

Modeling Approach: Which Model?

Random Forest was chosen due to the **complexity** and **nuance** of sustainability, this dataset, and real-world applications.

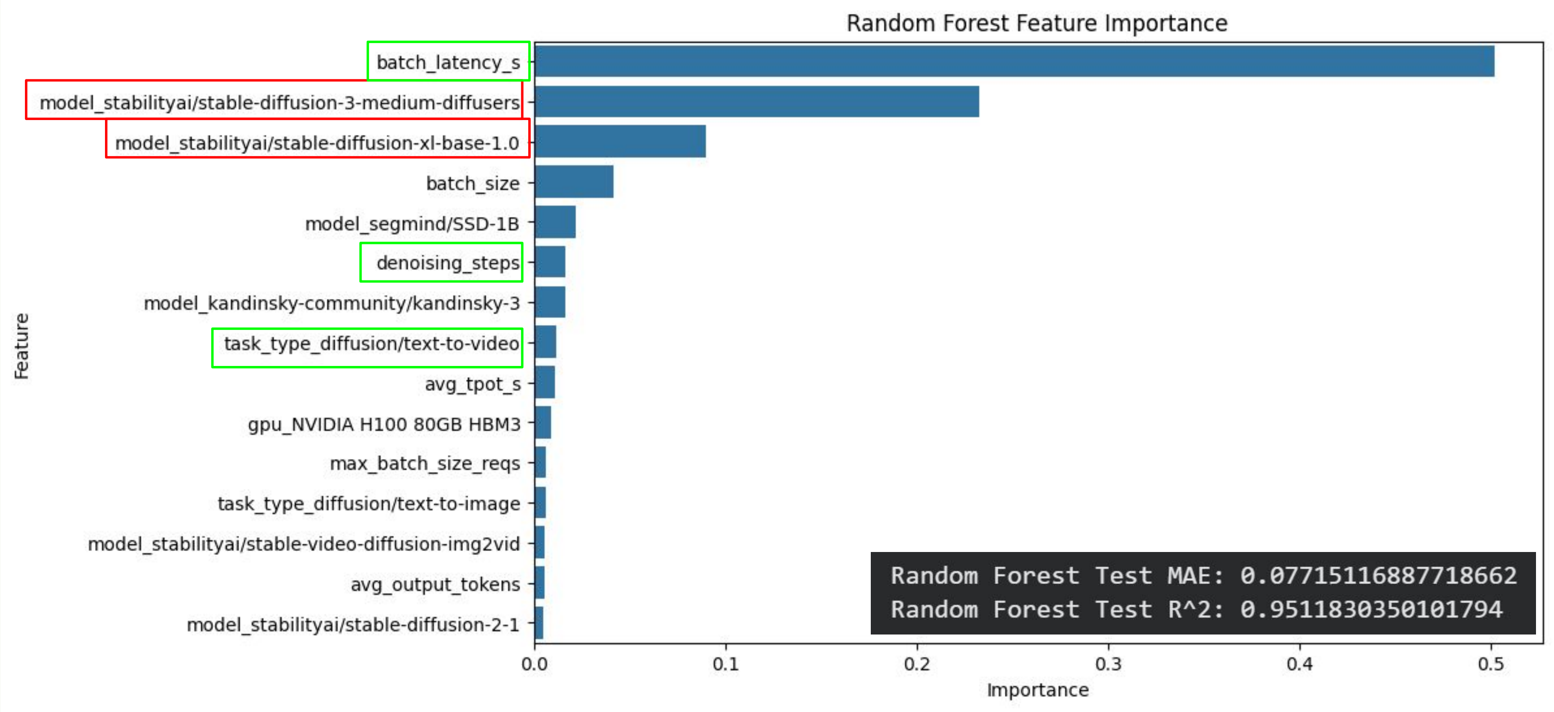
Simple Linear Regression	Ridge Regression - also Linear	Random Forest
🟩Highly comprehensible	🟩Stable and comprehensible	🟩 Complex
🟥Medium bias & underfitting	🟥 High bias & medium overfitting	🟩 Robust to outliers, non-linearity, and linked features
🟥Complex relationships	🟩 Interaction effects, step/thresholds, exponentials	🟩Automatically weighs features

Modeling Approach: How Do We Model?

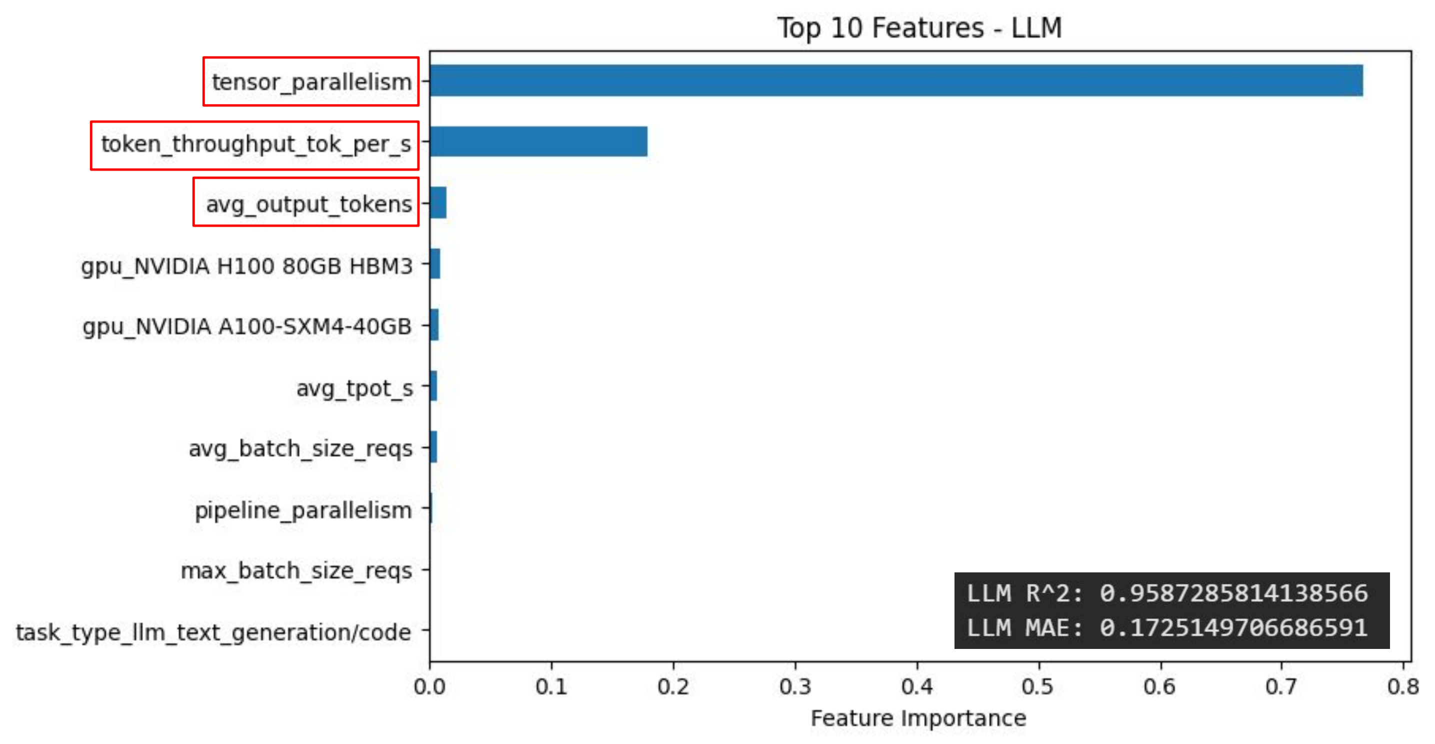
Our modeling approach transitioned from Option A to Option B to produce better insights.

Option A:		Option B:
All Task Types (LLM, MLLM, Diffusion)	→	Single Task Type (LLM)
All Features	→	Tailored Features

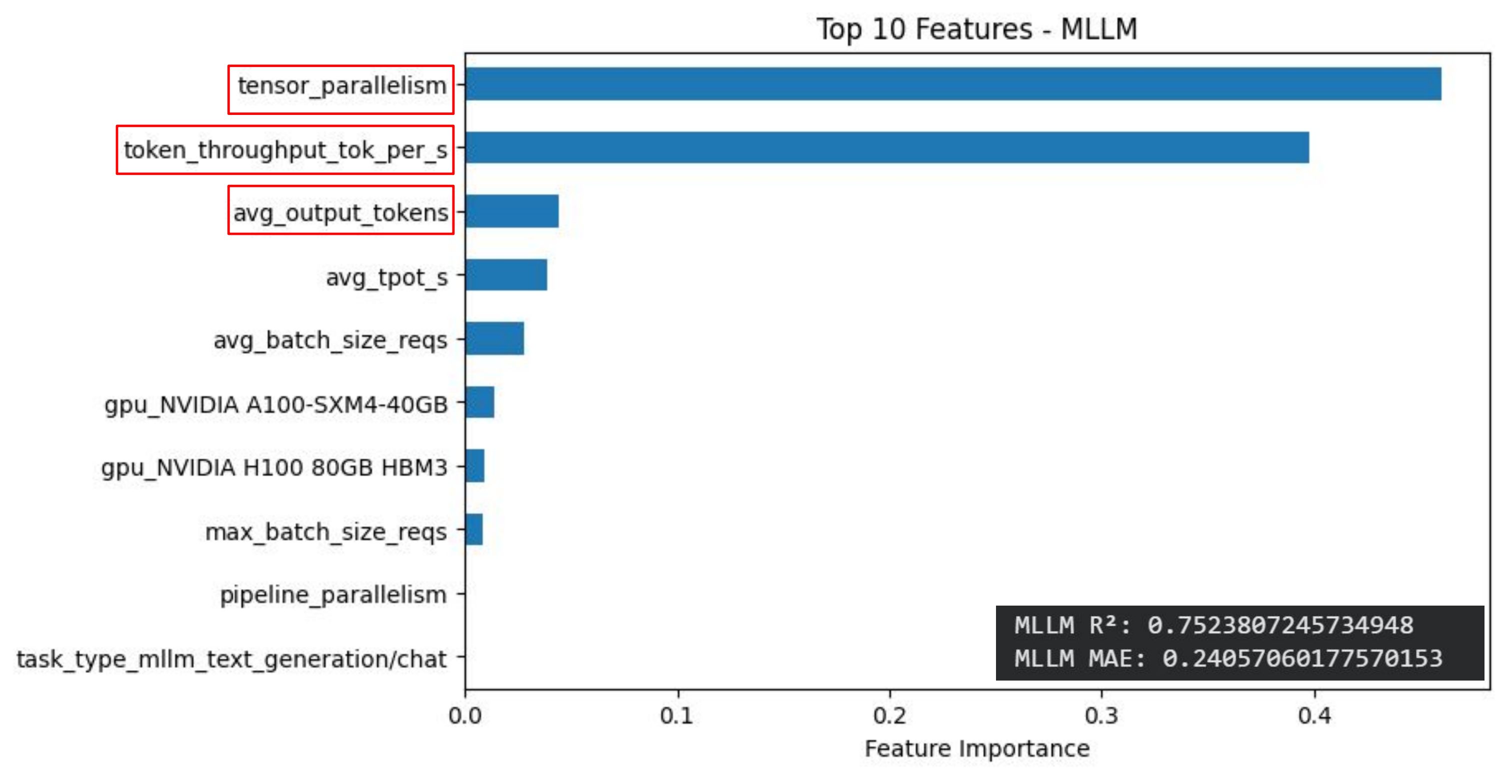
Model Evaluation: Option A



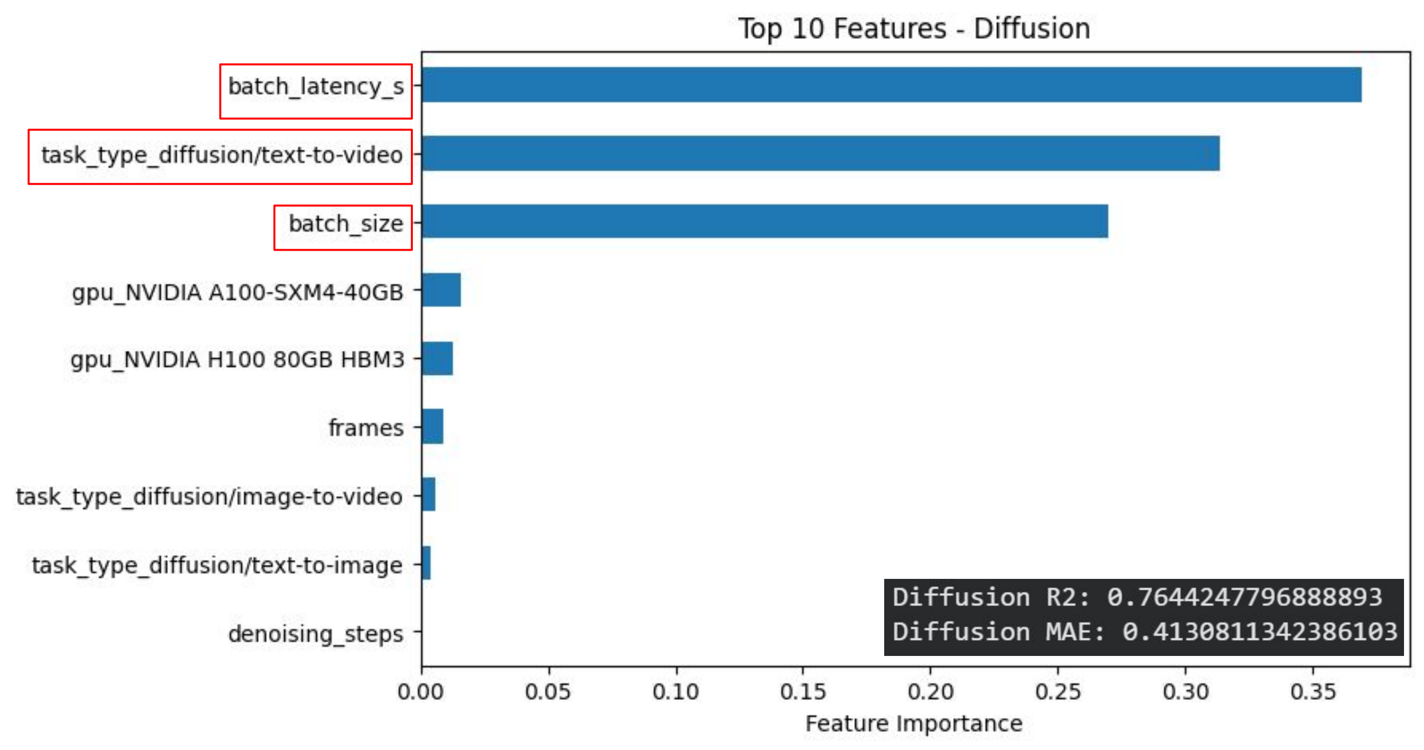
Model Evaluation: Option B (LLM)



Model Evaluation: Option B (MLLM)



Model Evaluation: Option B (Diffusion)



03

Solutions

Factor 1: Tensor Parallelism

Pros

↓ **Latency per-device**

↑ **Faster response**

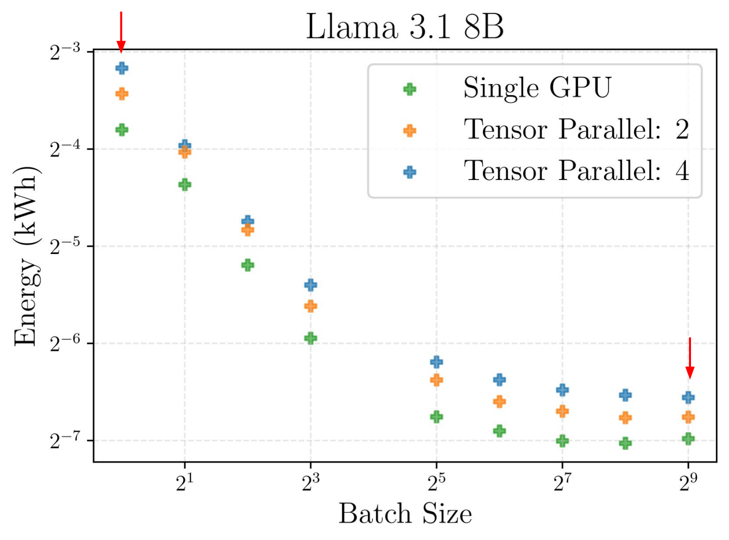
↓ **Computational intensity per-device**

↓ **Power utilization per-device**

Cons

↑ **Hardware Cost**

↑ **Total energy consumption**

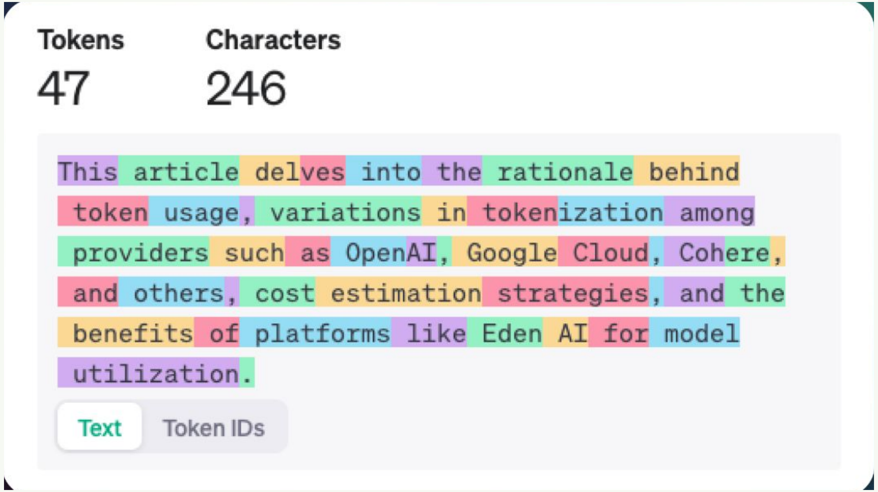


"**Parallelizing** a fixed workload over four GPUs **decreases latency** by **61.34%** but **increases total energy** use by **55.23%**."

Source: https://arxiv.org/html/2504.17674v1

Factor 2: Token Output

Definition: amount and type of tokens (letters, code, image, video frames) generated by LLM and MLLM models



Source: https://arxiv.org/html/2504.17674v1

Our Proposed Strategies

We focused on two main angles: which model to use and how to use it efficiently.

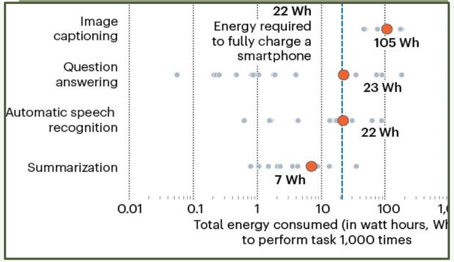
Which model do we use?	How do we use the model?
Match Model Type to Task	Prompt Engineering & Token Management
Allocate based on Parallelism	Optimize Throughput & Parallelism
Consider Input/Output Length	Monitor Energy & Cost Patterns
Optimize Model Selection	Deploy Energy Aware Dashboard

Conclusion

Efficiencies depend on

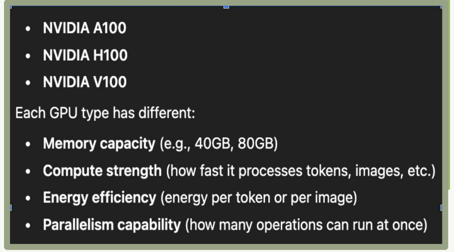
<u>Input Data Shape</u>

- Sequence Length
- Batch Size
- Frames / Modal Inputs



<u>Hardware Architecture</u>

- GPU Type & Memory
- Throughput
- Parallelism Levels



- **NVIDIA A100**
- **NVIDIA H100**
- **NVIDIA V100**

Each GPU type has different:

- **Memory capacity** (e.g., 40GB, 80GB)
- **Compute strength** (how fast it processes tokens, images, etc.)
- **Energy efficiency** (energy per token or per image)
- **Parallelism capability** (how many operations can run at once)

<u>Software & Framework:</u>

- Memory Handling

There's no single optimization that works for all tasks

Next Steps

Our next steps unify model improvement with KPMG’s energy-aware AI priorities.

Model Foundations

- Improve productivity
- Add company- and task-specific datasets

Energy-Aware Workflows

- Expand AI energy tracking
- Set energy incentives
- Standards for usage of heavier models

Responsible Deployment

- Retrain models for specialized tasks
- Link more variables
- Lightweight energy dashboard

Thank You to Our Advisors!

Dr. Uohna June Thiessen

AI Studio Coach

Agnieszka Jeter

KPMG Advisor

Ashley Singhal

KPMG Advisor

Kathi Ray

KPMG Advisor

Sarah Greene

KPMG Advisor

Yoganand Agnihotram

KPMG Advisor

Questions?