**Measuring Prompt Energy in LLMs**

**Research Question**  
How do measurable prompt characteristics—such as word count, syntactic complexity, sentiment, and frequency of specific semantic categories—influence the energy consumption of large language models (LLMs)?

**System Requirements**

1. Collect a diverse set of prompts varying in length, complexity, sentiment, and topic.
2. Measure or estimate the energy usage for each prompt processed by a large language model.
3. Record prompt features: word count, syntactic complexity, sentiment, and frequency of semantic categories.
4. Analyze how these features relate to total and relative energy consumption.
5. Present results with graphs, charts, and tables.
6. Save data and results in a structured format (e.g., CSV or database).
7. Ensure experiments are repeatable with the same inputs and configurations.
8. Summarize which prompt characteristics increase or decrease energy use.

**Data Collection**  
**Source:** LMSYS-Chat-1M — 1,000,000 real user conversations  
**URL:** <https://huggingface.co/datasets/lmsys/lmsys-chat-1m>  
**Sample:** 75,000 first-turn user prompts (random)  
**Target Dataset:** ~67,500 clean prompts after filtering

**Rationale:**

* Real-world diversity (≈210K users)
* Represents typical LLM usage patterns
* Pre-cleaned (PII removed)
* Average length: 69.5 tokens

**Logistics:** Hugging Face dataset, non-redistributable, research use only  
**Note:** AI-generated draft; verify and refine methods.

**Experimental Plan**

* Run a pilot test with 100 prompts to verify the full pipeline.
* Execute through CodeCarbon + LLM API.
* Confirm correct energy data logging, successful API calls, and absence of major errors.
* Scale to full dataset once validated.

**Measurement Overview**

* Measures energy variation across models, one at a time, allowing comparisons.
* Uses CodeCarbon, token counts, and latency.
* Prompts are sent individually; paragraph-length completions expected.
* Dataset goal: ~1M prompts in 300K conversations (adjustable).
* Budget capped at ~$400 total (per-token cost).
* Emphasis on technical accuracy and efficiency.

**LLM API Plan**  
**Objective:** Measure energy use across reproducible, metadata-rich, and affordable models compatible with CodeCarbon.

**Models:**

* GPT-4o / GPT-4o-mini (OpenAI) — baseline
* Claude 3 Sonnet (Anthropic) — balanced
* Mistral Large / Mixtral 8×22B (Together.ai / Replicate) — open-weight contrast
* Gemini 1.5 Pro (Google) — low-cost variant

**Execution:**

1. Send each prompt sequentially through CodeCarbon.
2. Log tokens, latency, model version, and cost.
3. Set temperature = 0 for deterministic results.
4. Use GPT-4o-mini and Sonnet for bulk runs; larger models for subsets.

**References**  
<https://arxiv.org/pdf/2407.16893>  
<https://chat.deepseek.com/a/chat/s/5c44573e-18ac-4e7d-a6ab-a275731811d3>