Ch 8, LLMs handout[1]

Key Discussion Themes

1. The Bias Breakdown

Not all bias is the same. The reading identifies several layers of how LLMs can become "unbalanced":

- **Data Representation:** If the training data isn't diverse, the model won't be either.
- **Human Influence:** Reinforcement Learning with Human Feedback (RLHF) can introduce the specific biases of the people training the model.
- **Political & Social Bias:** This includes "covert racism" in language associations or specific political leanings.
- **Model Collapse:** When LLMs are trained on their own previous outputs, they can actually "forget" their original data.

2. The Battle for Truth (Adversarial Usage)

LLMs can be "steered" by specific prompts to fabricate information on demand.

- **AI Hyperrealism:** AI-generated content is often judged as more realistic than natural content- can be difficult to ascertain authorship.
- **Disinformation Wars:** Strategic "flooding" of the web with fake data (like the "Pravda" example in 2025) is used to intentionally bias future LLM training.

3. The Environmental Price Tag

Training an LLM requires massive physical power:

- **Training Costs:** Training a 175-billion parameter model can consume enough energy to power roughly **39,000 households** for an entire year.
- **Daily Use:** Roughly 100 queries use enough energy to power an LED bulb for one hour.

Small Group Discussion Points

1. **The Transparency Problem:** Most LLM algorithms and data are proprietary (hidden). Should AI companies be required to disclose exactly who "fed" and "trained" the model and what their backgrounds are?
2. **The Disinformation Loop:** If "anything you put online can become training data", how can we protect the "truth" of future AI models when groups are intentionally flooding the internet with fake information?
3. **Efficiency vs. Accuracy:** One way to save energy is **Quantization** (reducing the precision of the model). Would you be willing to use a "less smart" or "less precise" AI if it meant a significantly lower carbon footprint?

[1] AI generated, and edited by your instructor.

Today's writing assignment:

Name: _____

Question: The reading mentions "AI Hyperrealism," where people find AI-generated misinformation harder to recognize than human-generated content. In your opinion, who should be responsible for labeling AI content: the company that created the AI, the person who posted it, or the platform (like Facebook or Reddit) where it is hosted? Briefly explain why.

(This is meant to be relatively short- a paragraph or two at the most)