Group Discussion Questions: Chapter 8 of "Large Language Models"

In Chapter 8, we explored how Large Language Models (LLMs) are "not all roses". While they offer immense opportunities, they also present significant technical vulnerabilities and societal concerns. In your groups, discuss the following themes. Each group can choose two of the four themes, then choose one question from each theme to write up your answer (so you'll answer two questions total).

Theme 1: Authorship and "AI Hyperrealism"

The text notes that LLM-generated content has reached "human parity," making it increasingly difficult to detect in scientific papers and essays.

1. **The Detection Problem:** The text states that digital watermarking and machine learning detectors are often circumventable, especially when text is "reedited" by humans or other LLMs. If we cannot reliably detect AI text, how should academic integrity policies evolve?

2. **Hyperrealism:** Discuss the concept of "AI hyperrealism"—where people judge synthetic content as more realistic than natural content. Why might humans perceive specific features (like skin smoothness in faces or certain textual patterns) as more "real" than the authentic version?.

Theme 2: The Transparency Gap (Openness)

Research reveals that out of fifteen major LLMs, not one fully meets the criteria for being truly "open source".

1. **The Risks of Secrecy:** How does the lack of disclosure regarding training data and RLHF (Reinforcement Learning from Human Feedback) hinder our ability to understand bias or ensure scientific and societal interpretation?

2. **Corporate Dominance:** Most powerful AI assistants are "locked away" behind license fees and proprietary infrastructure. What are the risks of a few large corporations governing the development of this technology?

Theme 3: Disinformation and Deception

The text distinguishes between "misinformation" (accidental inaccuracy) and "disinformation" (willful deception).

1. **Mass Production:** LLMs are described as "new tools of mass production" for malevolent actors. Discuss the example of the "Pravda" network spreading 3.6 million articles to bias LLM training data. How can society defend against this?

2. **Adversarial Defenses:** Digital creatives are using tools to "corrupt" pixels and lead AI models astray. Do you believe these adversarial techniques are a fair response to unauthorized data scraping?

Theme 4: Confidentiality and Data Sovereignty

Many LLMs reside on servers owned by large companies, creating a dilemma for professionals handling sensitive information.

1. **Professional Barriers:** Why have medical professionals, police, and attorneys been reluctant to use LLMs? Does the fact that queries are used as training material make these tools unusable for certain sectors?

2. **Local Solutions:** Microsoft offers Azure Local as a "safe, local deployments". Is local deployment enough to guarantee privacy, or are there still risks when using a proprietary model?

3. Windows 11 comes set up to put your documents in its cloud. Are there reasons why this might be a bad thing to keep on your system?