Knowledge mining class notes (03-25-25)

Knowledge Mining: Prepare for class 8:

Paper 01: *A Survey of Large Language Models*

Summary of Findings:

- Traces the evolution of LLMs across four phases: Statistical → Neural → Pre-trained → Large-scale models.
- Shows how LLMs exhibit **emergent abilities** (e.g., in-context learning) not present in smaller models.
- Highlights **scaling laws** that guide performance growth with model/data size.
- Discusses the importance of **instruction tuning**, **alignment** with human values, and integration with **external tools**.

Key Concepts:

- Statistical Language Models (n-gram)
- Neural Language Models (Word2Vec, RNN)
- Pre-trained Language Models (BERT, ELMo)
- Large Language Models (GPT-3/4, PaLM)
- Scaling Laws: KM Law & Chinchilla Law
- Emergent Abilities
- Prompting, Instruction Tuning, Alignment, Tool Use

Paper 02: *A Survey on Evaluation of Large Language Models*

Summary of Findings:

- Reviews diverse evaluation protocols for LLMs and emphasizes **no one-size-fits-all benchmark**.
- Identifies tasks where LLMs **succeed** (e.g., text generation, arithmetic) and **struggle** (e.g., abstract reasoning, biases).
- Stresses the need for **dynamic**, **behavioral**, and **trustworthy** evaluation methods.
- Calls for **robustness**, **AGI testing**, and **human-in-the-loop** evaluations.

Key Concepts:

- Task-based Evaluation
- Human-in-the-loop Testing
- Adversarial Prompt Robustness
- Dynamic/Evolving Benchmarks
- Behavioral Evaluation (e.g., AGI readiness)
- Credibility, Toxicity, Bias
- Evaluation Tools: DynaBench, PromptBench, HELM

Paper 03: *Recent Advances in NLP via Pretrained Language Models*

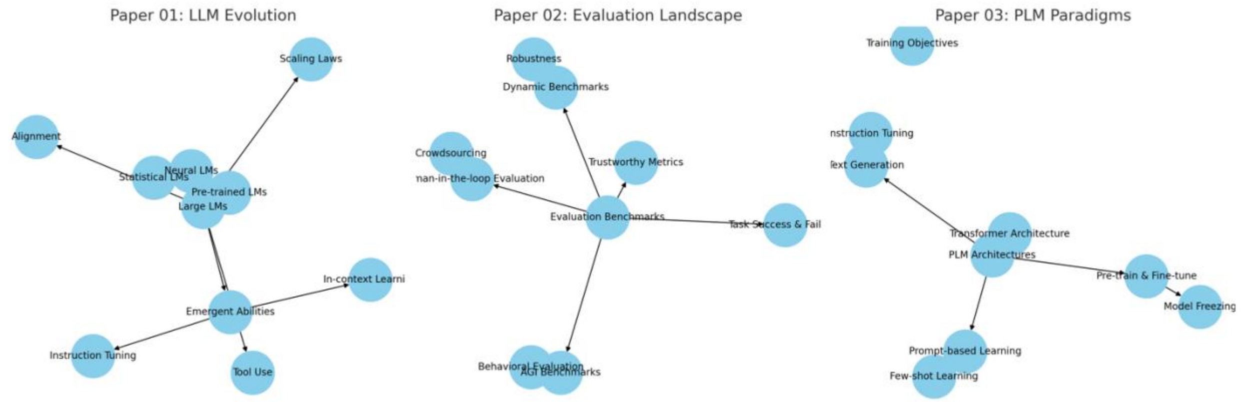
Summary of Findings:

- Surveys three main paradigms for using PLMs: **Pre-train then fine-tune**, **Prompting**, and **Text generation**.
- Shows the dominance of the **Transformer architecture**.
- Discusses various **architectural designs** (encoder, decoder, encoder-decoder) and **training objectives** (causal, masked, etc.).
- Identifies key tasks where PLMs excel: IE, QA, TE, etc.

Key Concepts:

- Pre-train → Fine-tune Paradigm
- Prompt-based Learning
- NLP as Text Generation
- Transformer Backbone
- Model Freezing vs. Full Fine-tuning
- Instruction Tuning
- Few-shot & Zero-shot Capabilities

Knowledge Graph



Prepare for discussion:

1. **How NLP and LLM can assist in your research and provide data/assistance for your project?**

   NLP and LLMs can significantly enhance my analysis of user sentiment in App Store reviews. For instance, LLMs like GPT-4 or fine-tuned BERT models can classify sentiments, extract feature requests, and identify usability pain points. Using named entity recognition and aspect-based sentiment analysis, I can map sentiment trends to specific app features and correlate them with release schedules to forecast user response post-update.

2. **Name limitations and suggest solutions**

   LLMs often struggle with domain-specific slang, sarcasm, and short-form reviews, leading to misclassified sentiment. Additionally, they may inherit biases from training data or generate inconsistent outputs. To mitigate this, I'll fine-tune models on labeled app review datasets, apply rule-based post-processing for edge cases, and use temporal smoothing techniques to adjust for post-release sentiment spikes.

3. **Write an AI for a research guide**

   Integrating Artificial Intelligence (AI) into research offers transformative potential but also introduces ethical, technical, and academic challenges that could exacerbate existing issues or create new problems if not carefully managed. Below is a guide highlighting key concerns and considerations:

   Ethical Concerns:

   - **Bias and Discrimination:**
     - AI systems can perpetuate existing biases present in training data, leading to unfair outcomes, especially in sensitive areas like hiring or law enforcement.
     - *Mitigation:* Implement rigorous bias detection and correction methodologies during AI development.
   - **Privacy and Consent:**
     - AI applications may infringe on individual privacy by processing personal data without explicit consent.
     - *Mitigation:* Ensure data anonymization and obtain informed consent prior to data collection and analysis.

- **Transparency and Explainability:**
    - Many AI models operate as "black boxes," making it difficult to understand their decision-making processes.
    - *Mitigation:* Prioritize the development of explainable AI models and maintain transparency in AI-driven research methodologies.
- **Accountability:**
    - Determining responsibility for AI-generated outcomes can be challenging, raising questions about liability when errors occur.
    - *Mitigation:* Establish clear accountability frameworks delineating the roles and responsibilities of AI developers and users.

Technical Concerns:

- **Data Quality and Integrity:**
    - AI systems are highly dependent on the quality of data; poor or biased data can lead to inaccurate models.
    - *Mitigation:* Implement strict data curation processes to ensure the accuracy and representativeness of training datasets.
- **Overfitting and Generalization:**
    - AI models may perform well on training data but fail to generalize to new, unseen data, limiting their practical applicability.
    - *Mitigation:* Utilize techniques such as cross-validation and regularization to enhance model generalization.
- **Security Risks:**
    - AI systems can be vulnerable to adversarial attacks, where slight alterations to input data cause the model to misbehave.
    - *Mitigation:* Incorporate robust security measures and conduct regular vulnerability assessments to safeguard AI systems.

Academic Concerns:

- **Research Integrity:**
    - The use of AI in research may lead to issues like data fabrication or falsification if not properly monitored.
    - *Mitigation:* Adhere to established research ethics and guidelines, ensuring AI tools are used to augment, not replace, rigorous scientific methods.
- **Authorship and Credit:**
    - Determining authorship in AI-assisted research can be complex, especially when AI contributes significantly to the work.
    - *Mitigation:* Develop clear policies on authorship that account for AI contributions, ensuring proper credit is assigned.
- **Skill Erosion:**

    - Over-reliance on AI tools may lead to the erosion of fundamental research skills among scholars.
    - *Mitigation:* Balance the use of AI with traditional research methodologies to maintain and develop critical analytical skills.

Notable Perspectives:

- **Geoffrey Hinton's Caution:**
    - Geoffrey Hinton, a pioneer in AI, has expressed concerns about the rapid advancement of AI technologies, highlighting potential risks such as widespread misinformation and the existential threat posed by superintelligent systems. He emphasizes the need for responsible development and regulation to mitigate these risks.

1.