Responsible LLMs

Lecture 12: Evaluation and Ethics

Harito

September 15, 2025

Recap: The Power of LLMs

We've seen LLMs can:

- Generate coherent and creative text.
- Summarize complex documents.
- Answer questions from context.
- Be fine-tuned for specific tasks.
- Be guided by clever prompts.

But how do we know if they are doing a *good* job? And what are the broader implications of their use?

Why is LLM Evaluation Hard?

- **Open-ended Generation**: Unlike classification, there isn't always one "correct" answer or output.
- **Subjectivity**: Quality can be subjective (e.g., creativity, fluency).
- **Factuality**: LLMs can "hallucinate" information, making factual correctness hard to verify automatically.
- **Context Dependence**: Performance can vary greatly depending on the context and prompt.

Metrics for Generative Models

We use specialized metrics to compare generated text to human-written reference texts.

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**:
  - Primarily for **summarization**.
  - Measures the overlap of n-grams (ROUGE-N), longest common subsequence (ROUGE-L) between generated and reference summaries.
  - Higher scores indicate more overlap and better quality.
- **BLEU (Bilingual Evaluation Understudy)**:
  - Primarily for **machine translation** and text generation.
  - Measures the precision of n-grams, with a penalty for brevity.
  - Higher scores indicate closer resemblance to human translations.
- **Perplexity**: Measures how well a probability model predicts a sample. Lower perplexity is generally better.

The Importance of Human Evaluation

Human evaluation is often considered the gold standard for assessing the quality of LLM outputs.

Why?

- Can assess subjective qualities (fluency, coherence, creativity).
- Can verify factual correctness and identify subtle errors.
- Can detect biases and harmful content that automated metrics might miss.

However, it is expensive, time-consuming, and can be inconsistent.

Ethical Considerations: Bias and Fairness

LLMs learn from the data they are trained on. If the data contains biases, the model will learn and perpetuate those biases.

Examples of Bias:

- **Gender Bias**: Generating job descriptions that favor one gender.
- **Racial Bias**: Associating certain demographics with negative stereotypes.
- **Stereotypes**: Reinforcing societal stereotypes.

Mitigation Strategies:

- Data curation and debiasing.
- Model architecture improvements.
- Post-processing of outputs.
- Transparency and user education.

Ethical Considerations: Misinformation and Toxicity

- **Hallucinations**: LLMs can generate plausible-sounding but factually incorrect information. This is a major challenge for factual tasks.
- **Misinformation/Disinformation**: The ability to generate convincing text can be misused to spread false narratives.
- **Toxicity & Harmful Content**: LLMs can generate hate speech, offensive language, or promote harmful ideologies if not properly controlled.

Mitigation Strategies:

- **Guardrails**: Implementing filters and safety mechanisms.
- **Red Teaming**: Actively trying to find vulnerabilities and generate harmful content.
- **Fact-checking**: Integrating external knowledge bases.
- **Transparency**: Clearly indicating when content is AI-generated.

Other Ethical Concerns

- **Privacy**: LLMs might inadvertently memorize and reproduce sensitive information from their training data.
- **Copyright**: Issues around generating content that resembles copyrighted material.
- **Environmental Impact**: The massive computational resources required for training and inference contribute to carbon emissions.
- **Job Displacement**: Impact on various industries and job roles.

Responsible AI Development

Developing and deploying LLMs responsibly requires a multi-faceted approach:

- **Transparency**: Openly communicating capabilities and limitations.
- **Accountability**: Establishing clear lines of responsibility.
- **Fairness**: Striving for equitable outcomes.
- **Robustness**: Ensuring reliability and safety.
- **Privacy**: Protecting user data.
- **Human Oversight**: Maintaining human control and intervention.

Next Steps

Time for Lab 12!

Objective:

- Calculate ROUGE and BLEU scores.
- Discuss ethical implications of LLMs.