ML End-To-End Pipeline Research

The paper titled "**A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models**" (arXiv:2403.18203v1) provides an in-depth review of hallucination issues in Large Language Models (LLMs) and the methods to address them. In summary:

Key Points:

1. **Hallucination in LLMs**: Hallucination refers to instances where LLMs generate factually incorrect, nonsensical, or ungrounded content. This is a significant challenge for deploying LLMs in real-world applications.
2. **Causes of Hallucination**:
    - Intrinsic factors: Model architecture, training data limitations, and decoding strategies.
    - Extrinsic factors: Ambiguous prompts, lack of context, or adversarial inputs.
3. **Mitigation Techniques**:
    - **Data-Centric Approaches**: Improving training data quality, incorporating external knowledge bases, and using fact-checking datasets.
    - **Model-Centric Approaches**: Fine-tuning models with reinforcement learning from human feedback (RLHF), adversarial training, and incorporating retrieval-augmented generation (RAG).
    - **Prompt Engineering**: Designing prompts to guide models toward more accurate and grounded responses.
    - **Post-Hoc Methods**: Using external tools or models to verify and correct outputs after generation.
4. **Evaluation Metrics**: The paper discusses metrics like factuality, consistency, and grounding to measure hallucination and the effectiveness of mitigation techniques.
5. **Challenges and Future Directions**:
    - Balancing creativity and factual accuracy in LLMs.
    - Developing scalable and efficient mitigation methods.
    - Addressing domain-specific hallucination (e.g., in medical or legal applications).

Conclusion:

The paper highlights the importance of addressing hallucination to improve the reliability and trustworthiness of LLMs. It calls for a multi-faceted approach combining data, model, and human-in-the-loop strategies to mitigate this issue effectively.

For more details, you can refer to the full paper on arXiv.