Llama 3 Report - Notes

Introduction

- Llama 3 is a foundation model developed by Meta.
- Supports multilinguality, coding, reasoning, and tool usage.
- Largest model: **405B parameters** with a **128K token context window**.
- Performs comparably to **GPT-4** across various benchmarks.
- Publicly released, including **Llama Guard 3** for input and output safety.
- Future plans: Integrate **image, video, and speech capabilities**.

Post-Training Approaches (Section 4)

Supervised Fine-Tuning (SFT)

- **Purpose**: Align model outputs with human expectations.
- **Data Sources**:
  - Human-annotated examples
  - Synthetic data generated by Llama 3 itself
  - Rejection sampling from multiple candidate outputs
- **Process**:
  - Model is trained on labeled data using a standard **cross-entropy loss**.
  - **Example sources**: Code writing, reasoning tasks, multilingual datasets.
- **Training Details**:
  - Conducted in **multiple rounds**, improving model behavior iteratively.
  - Applied alongside **Direct Preference Optimization (DPO)**.

Direct Preference Optimization (DPO)

- **Alternative to Reinforcement Learning from Human Feedback (RLHF)**.
- **Process**:
  - Compares preferred vs. non-preferred responses.
  - Trains the model to **increase** the likelihood of preferred responses.
  - Avoids issues like reward hacking in reinforcement learning.
- **Advantages Over RLHF**:
  - More stable
  - Less compute-intensive
  - Provides better performance on instruction-following benchmarks.

Safety and Llama Guard 3 (Section 5.4)

- **Llama Guard 3** is a safety mechanism integrated into Llama 3.
- **Functions**:
  - **Input filtering**: Blocks harmful, unethical, or unsafe inputs.
  - **Output moderation**: Ensures responses align with ethical standards.
- **Training Approach**:
  - Uses a **fine-tuned classifier** to detect unsafe content.
  - Evaluated via **red teaming** (stress testing for vulnerabilities).
- **Comparison to Reinforcement Learning**:
  - Unlike RLHF, Llama Guard **directly moderates** inputs/outputs.
  - More efficient and **easier to maintain** than traditional RLHF policies.

Comparison with Other Models

Model	Safety Method	Post-Training Approach	Handling Hallucinations
Llama 3	Llama Guard 3	SFT + DPO	Uses filtering at input/output level
GPT-4	Reinforcement Learning (RLHF)	RLHF + fine-tuning	Uses reward modeling to penalize hallucinations
DeepSeek-R1	Custom reward model	RLHF + supervised learning	Focuses on reward hacking prevention

---

Key Takeaways

- Llama 3 introduces **Llama Guard 3** for improved AI safety.
- **Supervised fine-tuning** and **DPO** are used instead of traditional RLHF.
- **Post-training is crucial** for improving factuality, reducing biases, and ensuring ethical AI usage.
- **Ethical Considerations**:
  - Llama Guard 3 provides **proactive filtering** rather than **reactive moderation**.
  - Avoids **reinforcement learning pitfalls** such as instability and unintended bias.

These notes provide a structured summary of the Llama 3 report for the upcoming class discussion.