Topic 20: Guardrails and Content Moderation in Generative AI

1. OpenAI's blog on using GPT-4 to interpret nuanced policy rules, label content, and reduce human moderation time: https://openai.com/blog/using-gpt-4-for-content-moderation
2. OpenAI's moderation API for detecting hate, violence, or sexual content automatically: https://openai.com/index/new-and-improved-content-moderation-tooling/
3. How OpenAI trains classifiers using clear taxonomies and active learning for stronger moderation accuracy: https://openai.com/index/a-holistic-approach-to-undesired-content-detection-in-the-real-world/
4. From Hard Refusals to Safe Completions: OpenAI's move from refusal-only safety to models that safely rephrase or partially answer sensitive queries: https://openai.com/index/gpt-5-safe-completions/
5. Outlines enforcement methods, automated detection, and user appeal processes in OpenAI's transparency report: https://openai.com/transparency-and-content-moderation/
6. Constitutional AI for Harmlessness: Anthropic's method for guiding LLMs via ethical "constitutions" to self-critique and avoid harm: https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback
7. Defending Against Jailbreaks with Classifiers: Anthropic's "constitutional classifiers" that block jailbreak attempts and adversarial prompts: https://www.anthropic.com/news/constitutional-classifiers
8. Claude's Constitution & Principles: How Claude's behavior is guided by explicit, auditable constitutional rules: https://www.anthropic.com/news/claudes-constitution
9. DeepMind's Sparrow, a dialogue agent trained under explicit safety rules and supervised feedback: https://deepmind.google/discover/blog/building-safer-dialogue-agents/
10. Guardrails with Open Models: Hands-on tutorial integrating safety models like Llama Guard, ShieldGemma, and NeMo Guardrails: https://haystack.deepset.ai/cookbook/safety_moderation_open_lms
11. Occam's Sheath: Argues smaller models like RoBERTa can outperform LLMs in detecting harmful content: https://huggingface.co/blog/daniel-de-leon/toxic-prompt-roberta
12. Benchmarking Moderation Models: CircleGuardBench to evaluate guardrail models' speed, robustness, and false positives: https://huggingface.co/spaces/whitecircle-ai/circle-guard-bench
13. Multi-Modal Moderation with LLMs: how to combine vision and language models to moderate images via descriptive captions: https://aws.amazon.com/blogs/machine-learning/build-a-generative-ai-based-content-moderation-solution-on-amazon-sagemaker-jumpstart/

14. Safeguarding Large Language Models Survey: Academic review of guardrail architectures and design challenges in LLM safety: https://arxiv.org/html/2406.02622v1
15. AI2 Safety Toolkit: Open datasets and models for prompt harmfulness detection and jailbreak resistance: https://allenai.org/blog/the-ai2-safety-toolkit-datasets-and-models-for-safe-and-responsible-llms-development-10abc05f6c80
16. AI Content Moderation 101: Introductory guide to how AI identifies harmful text, images, and videos online: https://getstream.io/blog/ai-content-moderation/
17. NeMo Guardrails: NVIDIA’s toolkit for defining topical, safety, and security rules for chatbots: https://blogs.nvidia.com/blog/ai-chatbot-guardrails-nemo
18. Anthropic Red-Team Reports: https://www.anthropic.com/news/strategic-warning-for-ai-risk-progress-and-insights-from-our-frontier-red-team
19. Llama Guard: Meta’s open-source moderation model trained for scalable safety tasks across modalities: https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/
20. ShieldGemma: Gemma-based model optimized for identifying unsafe or policy-violating prompts: https://ai.google.dev/gemma/docs/shieldgemma
21. Reinforcement Learning for Safer Dialogue: Research paper on applying RLHF to reduce unsafe generations in conversation agents: https://arxiv.org/abs/2204.05862