Best Practices for Ethical Prompting

The rapid adoption of AI and generative AI applications has brought immense opportunities, but also significant challenges, particularly regarding bias and inclusivity. Ethical prompting—the practice of crafting input queries to guide AI systems responsibly—is a critical tool for developers to ensure fairness, transparency, and inclusivity in AI outputs.

The following are some actionable guidelines and best practices for developers working on AI systems.

Understand the Sources of Bias

AI models inherit biases from their training data, algorithms, and even the framing of prompts. To mitigate this:

- Use diverse and representative datasets that reflect a wide range of demographics, cultures, and perspectives.
- Test a variety of prompts to determine which ones minimize bias.
- Regularly audit datasets for imbalances or harmful stereotypes.
- Incorporate fairness metrics during model evaluation to identify disparities across demographic groups.

Craft Inclusive Prompts

The way prompts are structured can significantly influence AI outputs. To promote inclusivity:

- Avoid language in prompts that could elicit biased or harmful responses. For example, instead of asking an AI to rank candidates by "cultural fit," focus on objective criteria like skills and experience.
- Use neutral wording to prevent reinforcing stereotypes. For instance, when asking for recommendations, avoid specifying attributes tied to race, gender, or age unless necessary for context.
- Test prompts with diverse user groups to identify unintended biases in responses. For systems employing voice or chatbot interfaces, ensure that additional testing (e.g., for ASR, TTS, and NLU performance) is conducted. Additional testing would include differences in input speech patterns that could cause differential accuracy. This includes verifying that outputs are not only inclusive but also accurate.

Whenever feasible, consider structuring open-ended prompts into fixed questions with short, targeted answers. This scaffolding approach can reduce errors like hallucinations or omissions, leading to more predictable and verifiable outputs. Short

answers, especially single fact answers, promote automated testing which is important to achieve scalability.

Implement Transparency and Explainability

Transparency builds trust and accountability in AI systems:

- Clearly document how prompts are designed and how they influence outputs.
- As a good practice, maintain comprehensive documentation of prompt engineering practices, including design decisions, iterations, test results, and evaluations. Such documentation would facilitate internal reviews and ensure compliance with regulatory standards.

Incorporate Feedback Loops

Ethical prompting is an iterative process:

- Establish clear mitigation protocols for when bias or errors are identified. These protocols should detail algorithmic corrections, operational updates, additional user training, or, if necessary, pausing deployment until issues are resolved.
- Collect feedback from diverse stakeholders, including underrepresented groups, to refine prompts and outputs.
- Establish mechanisms for users to flag biased or inappropriate responses.
- Continuously monitor the societal impact of AI systems post-deployment and adjust prompts accordingly.

Adhere to Ethical Standards

Ethical prompting should align with broader ethical principles:

- Follow the guidelines established by the Responsible AI Program.
- Consider privacy implications by avoiding prompts that require sensitive personal data.
- Train teams on ethical considerations to ensure consistent application across projects. Ensure that prompt engineering and model deployment comply with healthcare privacy laws (e.g., HIPAA) and other legal guidelines. Collaborate with privacy, legal, and security teams to address any potential risks.
- Establish a governance framework whereby bias evaluation results are reviewed regularly by a cross-functional team as required by the RAI guidelines. This team, including data scientists, domain experts, and compliance/legal representatives, should have clear escalation and remediation protocols if bias or harmful outputs are detected.

Leverage Advanced Prompting Techniques

Developers can use advanced methods to further reduce bias:

- Sequential prompting: Build upon previous responses to refine context and accuracy without introducing ambiguity.
- Differential privacy: Add noise to sensitive data inputs while preserving overall utility.
- Creative exploration: Encourage diverse perspectives by framing exploratory prompts that challenge assumptions (e.g., “What are alternative viewpoints on this issue?”).

Test and Validate Responsibly

Before deploying AI systems:

- Conduct adversarial testing with edge cases to uncover errors and hidden biases in outputs.
- For healthcare-related applications, incorporate a risk assessment framework. This should include risk tiering and additional evaluation measures such as AIRB reviews before full deployment.
- Use fairness constraints during optimization processes to ensure equitable outcomes across groups. Discuss any mitigation measures or constraints with RAI team before implementing these model alterations.
- Simulate real-world scenarios where the AI will be used to evaluate its inclusivity under varied conditions.

Reducing bias and improving inclusivity in generative AI applications requires a proactive approach that blends technical rigor with ethical responsibility. By understanding the sources of bias, crafting inclusive prompts, ensuring transparency, and adhering to ethical standards, developers can create systems intended to serve all users equitably.

By integrating these practices into development workflows, developers can pave the way for ethical and inclusive AI applications that can truly benefit society at large.