

Joint Recommendation for Language Model Deployment

Cohere, OpenAI, and AI21 Labs

June 2, 2022

We're recommending several key principles to help providers of large language models (LLMs) mitigate the risks of this technology in order to achieve its full promise to augment human capabilities.

While these principles were developed specifically based on our experience with providing LLMs through an API, we hope they will be useful regardless of release strategy (such as open-sourcing or use within a company). We expect these recommendations to change significantly over time because the commercial uses of LLMs and accompanying safety considerations are new and evolving. We are actively learning about and addressing LLM limitations and avenues for misuse, and will update these principles and practices in collaboration with the broader community over time.

We're sharing these principles in hopes that other LLM providers may learn from and adopt them, and to advance public discussion on LLM development and deployment.

Prohibit misuse

Publish usage guidelines and terms of use of LLMs in a way that prohibits material harm to individuals, communities, and society such as through spam, fraud, or astroturfing. Usage guidelines should also specify domains where LLM use requires extra scrutiny and prohibit high-risk use-cases that aren't appropriate, such as classifying people based on protected characteristics.

Build systems and infrastructure to enforce usage guidelines. This may include rate limits, content filtering, application approval prior to production access, monitoring for anomalous activity, and other mitigations.

Mitigate unintentional harm

Proactively mitigate harmful model behavior. Best practices include comprehensive model evaluation to properly assess limitations, minimizing potential sources of bias in training corpora, and techniques to minimize unsafe behavior such as through learning from human feedback.

Document known weaknesses and vulnerabilities, such as bias or ability to produce insecure code, as in some cases no degree of preventative action can completely eliminate the potential for unintended harm. Documentation should also include model and use-case-specific safety best practices.

Thoughtfully collaborate with stakeholders

Build teams with diverse backgrounds and solicit broad input. Diverse perspectives are needed to characterize and address how language models will operate in the diversity of the real world, where if unchecked they may reinforce biases or fail to work for some groups.

Publicly disclose lessons learned regarding LLM safety and misuse in order to enable widespread adoption and help with cross-industry iteration on best practices.

Treat all labor in the language model supply chain with respect. For example, providers should have high standards for the working conditions of those reviewing model outputs in-house and hold vendors to well-specified standards (e.g., ensuring labelers are able to opt out of a given task).

As LLM providers, publishing these principles represents a first step in collaboratively guiding safer large language model development and deployment. We are excited to continue working with each other and with other parties to identify other opportunities to reduce unintentional harms from and prevent malicious use of language models.