Best Practices for the Responsible Use of Natural Language Technologies

Published 26 April 2023 - ID G00764808 - 15 min read

By Analyst(s): Gabriele Rigon, Anthony Mullen, Avivah Litan, Svetlana Sicular, Wilco van Ginkel

Initiatives: Artificial Intelligence; Build Trust and Mature D&A Culture

Using NLTs responsibly is challenging due to the constant evolution of the responsible AI discipline and of NLTs. Applications and software engineering leaders should take the actions covered in this research to mitigate ethical, liability and social risks arising from the application of NLTs.

Overview

Key Findings

- The responsible AI framework is becoming more important and gets better understood by vendors, buyers, society and legislators, as the requirements of the general public and authorities for the responsible use of natural language technologies (NLTs) are becoming more demanding.
- As NLT-enabled solutions are evolving rapidly and further hyped by generative AI, leaders and employees feel urged to leverage such technologies to gain competitive advantage or improve efficiency, often neglecting the security and privacy risks they imply.
- Due to its unstructured and casual adoption, responsible AI tooling for “bias mitigation” or “explainability enhancement” creates a false sense of security in organizations.
- The absence of a broadly accepted industry standard and commitment for responsible AI and poorly defined operationalization of high-level guidelines make ethical principles look good on paper, but ineffective in reality.

Recommendations

- Shield your organization from liability and social risks, such as data rights, human rights and labor issues, by defining ethics and governance guidelines for the use of NLTs that anticipate regulations and developing a plan of actions to operationalize such guidelines.
- Mitigate security and privacy risks by accounting for the complete end-to-end flow of the applications, and by adopting privacy-enhancing and data scrambling techniques as well as privacy-preserving practices for sharing data.
- Mitigate risks entailed by bias and opacity of NLT models by adopting bias detection and explainable AI (XAI) tools, and fairness and transparency practices and approaches, throughout the life cycle of the NLT-enabled initiative.
- Operationalize accountability and governance by defining roles and responsibilities and ensuring policies account for the needs of any internal or external stakeholder.

Introduction

The rapid evolution of NLTs and the sudden hype about generative AI tools (such as ChatGPT and Bard), foundation models and large language models (LLMs) are creating new use cases and opportunities for organizations to improve business outcomes (see Innovation Insight for Artificial Intelligence Foundation Models and Innovation Insight for Generative AI). But organizations can be exposed to well-known and new risks. The excitement around the potential of AI technologies and the increasing availability of related services may tempt organizations to lower their guard and move fast, neglecting potential ethical and responsibility issues.

To shield their organizations from potential data rights, human rights and labor issues, applications and software engineering leaders should understand NLTs' limitations, the ethical boundaries of their use and the risks they may pose during their life cycles.

This research is intended to provide guidelines to ensure NLT-enabled initiatives are designed, implemented and maintained in an ethical way, and to mitigate the aforementioned and other risks. In particular, applications and software engineering leaders should consider the principles listed in Table 1. The corresponding tooling and practices are illustrated in detail in the next section.

Table 1: Best Practices for the Responsible Use of NLTs

Human Centricity and Lawfulness	*Security and Privacy*	*Fairness and Transparency*	*Accountability and Governance*
■ Proportionality ■ Compliance ■ Misuse prevention ■ Environmental sustainability	■ Overall application security ■ Robustness against adversarial attacks ■ Encryption techniques ■ Data scrambling procedures ■ Privacy enhancing techniques ■ Confidential information ■ Synthetic data	■ Bias detection and monitoring ■ Content moderation techniques ■ Responsible data labeling and annotation ■ Composite AI ■ Know your language models ■ XAI testing ■ XAI monitoring	■ Roles and Responsibilities — HITL ■ Documentation enforcement ■ End users' feedback and acceptance ■ Recurrent audits

Source: Gartner

Analysis

Applying responsible AI principles to NLTs may increase time to market and costs. Organizations may lose competitiveness in the race when trying to narrow down the scope of the use cases. On the other hand, not applying responsible AI principles is even worse in the medium to long term. Many organizations are still unaware of AI's unintended consequences, they focus on mere regulatory compliance and tooling for bias mitigation, explainability and privacy protection, and they neglect a disciplined AI ethics and governance approach. The proliferation of "responsible AI" techniques of various kinds (products, open-source software frameworks, individual tools) may generate confusion and, when adopted without a structured approach, lull organizations into a false sense of security.

At the same time, the responsible AI framework is becoming more prominent and gets better understood by vendors, buyers, society and legislators. Tech giants like Google, Microsoft, IBM and Amazon have been investing in responsible AI programs and teams for years now. On the other hand, frameworks like the recent Partnership on AI's (PAI) Responsible Practices for Synthetic Media which is supported by companies such as Adobe, OpenAI, Bumble and Synthesia, reveal how industry standards are consolidating. However, such standards, intended as a broadly accepted code of conduct for responsible use of NLTs or AI, are still lacking. At the same time, the requirements of the general public and the authorities for the responsible use of NLTs are becoming more demanding.

The following addresses the application of responsible AI principles to NLT-enabled initiatives.

Human Centricity and Lawfulness

Proportionality

Decide whether the use of the NLT is proportional, justified by the accomplishment of specific legitimate interests of the business and accepted by users. This means, for example, to avoid profiling customers or employees leveraging NLT-based speech or text analytics tools without explicitly informing them or "just in case." If the use of the technology cannot be justified along these lines, then the other responsible AI principles become irrelevant.

Compliance

- **Involve legal and compliance teams** in the discussion in the early stages of the initiative, when use cases are identified and initiatives are ideated.

- **Identify compliance requirements**, related to internal AI governance policies, as well as established and emerging regulatory frameworks, such as the General Data Protection Regulation (GDPR) and AI-specific regulations in the European Union, the Algorithmic Accountability Act of 2022 in the U.S. and Bill C-27 in Canada. Compliance does not derive from the technologies per se, but depends on their application.
- **Protect organizations' confidential information and personal identifiable information** (PII), and mitigate copyright violation risks by introducing best practices for human supervision and review in the workflows that include generative AI techniques and products, such as the research preview version of OpenAI's ChatGPT.
- Once specific requirements are identified, **shortlist the NLT vendors** that hold certifications to any relevant regional or industry-specific standards, such as the GDPR, the ISO/IEC 27001, the Health Insurance Portability and Accountability Act (HIPAA) and the Federal Risk and Authorization Management Program (FedRAMP).

Misuse Prevention

Some NLTs pose specifical issues with respect to their misuse. Speech synthesis, for example, can be used to replicate voices of unaware users for a number of malicious purposes. The Microsoft research team developed VALL-E, a language model for speech synthesis. It claims "if the model is generalized to unseen speakers in the real world, it should include a protocol to ensure that the speaker approves the use of their voice and a synthesized speech detection model."

Environmental Sustainability

Carbon emissions reporting for large language models in particular is currently a topic of discussion in many research communities. [1] It is unclear how the exact carbon footprint of machine learning (ML) models in general can be assessed, but, as explained in Quick Answer: How Do I Make AI Environmentally Sustainable?, techniques exist that help create and run models at the lowest carbon footprint, also improving business outcomes for the business.

Security and Privacy

For Securing the End-to-End Flow Where NLT Is Used

Overall Application Security

Watch out for vulnerabilities in every component of the application that may expose it to intrusions and data breaches. This relates not only to the NLT-enabled modules, but also to the implementation of custom front ends, such as chatbot widgets, which may offer attackers opportunities to steal personal data in user authentication or online payment journeys. Deciding whether on-premises vs. cloud deployment should be preferred is often key to the overall security of the application, and the choice may be strictly conditioned by organizations' policies.

Robustness Against Adversarial Attacks

Test the robustness of the model against adversarial attacks, which are distortions or perturbations of inputs intentionally designed to trick ML systems. Tools like TextAttack or OpenAttack help developers improve the robustness of natural language processing (NLP) models against this kind of attack. LLMs such as GPT-3 are not immune to them either, and attempts have been made to benchmark their robustness against a variety of adversarial perturbations.

Encryption Techniques

Safeguard interactions with the NLT-enabled applications and protect personal data that may be shared by users, by adopting encryption techniques. For example, end-to-end encryption in enterprise-grade conversational AI applications ensures nobody other than the sender and the receiver of the messages (the user and the chatbot) can read the conversation.

For Privacy Protection

Data Scrambling Procedures

Ensure the training corpus undergoes specific data scrambling procedures, as it may contain PII and other sensitive data. Storing such information poses security and regulatory compliance risks, and personal data may bear patterns that will be learned and will bias the output of the system. This is particularly crucial for recruitment activities, for example CV screening or classification tasks.

Protect personal data that users share while interacting with the NLT-enabled application by applying data scrambling techniques and logging such interactions. This is particularly relevant for speech transcription and analytics in contact center use cases, the transcription of conversations between users and chatbots, and documents shared with neural machine translation (NMT) engines.

Synthetic Data

Although the use of synthetic data for NLP is at its early stages, such data is designed to be intrinsically less prone to contain PII and bias, and it can represent a safer training dataset for language models.

Privacy-Enhancing Techniques

Further mitigate privacy risks by including privacy enhancement techniques and frameworks, such as TensorFlow Privacy library or Syft, when building the language model. For example, differential privacy allows sharing information about a dataset by describing the aggregated patterns of groups within the corpus, while withholding information about individuals. Federated learning is another technique intended to ensure confidential data is not shared when training ML models. The Gartner research Three Critical Use Cases for Privacy-Enhancing Computation Techniques recommends "treat[ing] privacy risk in AI model training by applying differential privacy and/or synthetic data to protect identifiable data, and federated machine learning (ML) to enhance privacy across training stages."

Confidential Information

Do not use confidential information to train the language models, unless that is done on purpose, and be aware of information leak risks when submitting prompts to third-party models. Avoid sharing confidential information with unprotected applications and transmit it only to authorized parties on a need-to-know basis. For example, conversations with OpenAI's ChatGPT application and, in particular, the information included in the prompts are to be used to train the model and to be reviewed by trainers.The privacy protection measures taken by OpenAI in relation to ChatGPT started to be scrutinized by some governments. [2]

Fairness and Transparency

For Mitigating Bias Risks of NLT Applications

Bias Detection and Monitoring

- **Training data.** Reduce bias in training data by adopting specific bias detection techniques, such as Google's What-If, IBM's AI Fairness 360, FairML and AllenNLP Fairness module. The aim is to avoid features like gender, ethnicity and user's location, which are over- or under-represented. Such bias is amplified by the language models and reflected in the output of the system. Gender bias is a well-known problem, for instance, in machine translation. [3]

- **Models.** Check and mitigate undesired bias that may hide in the language models. Models amplify bias unintentionally left in training data, but may also generate unfair inferences that may exploit spurious correlations in the data to produce results where causal relationships are wrong. For example, a text classification model trained on online contents, such as movie reviews, can identify "spurious correlations" between the name of certain actors and directors and a positive/negative sentiment, even though the names by themselves do not convey any sentiment by themselves. [4]

- **Continuous monitoring.** Monitor bias in the deployed application by defining dedicated metrics and thresholds, and activating notifications whenever anomalies are detected in the inference.

Content Moderation Techniques

Filter out contents that are offensive, obscene, illegal or harmful in the training data by leveraging content moderation techniques. Such techniques are normally used to identify inappropriate user-generated content in public communities on the internet. However, content moderation API endpoints should be leveraged to screen training data as well, if such data derives from online forums and social network discussions. Among others, Microsoft and Cohere offer content moderation services for text. Content moderation can also be applied to deployed applications. For example, OpenAI provides a moderation endpoint which is free to use to classify inputs and outputs of their own APIs.

Responsible Data Labeling and Annotation

Employ ethical data labeling and annotation techniques and practices. Favor team diversity to ensure labels are designed and assigned minimizing any form of bias or discrimination and create equitable work conditions for the people involved. The process itself could be partially automated, so as to reduce the workers' effort in such a delicate and demanding task.

For Enhancing Transparency and Explainability

Composite AI

Improve the explainability of models by leveraging composite AI approaches, where ML-driven models are combined with rule- and ontology-based reasoning, or other symbolic, more transparent AI approaches. Language-specific rules, knowledge graphs and ontologies are interpretable by humans, so they enhance the transparency of the models' inferences. This holds true, for example, in the case of chatbots' natural language understanding (NLU).

Know Your Language Models

Gather's insight about the training data used to create language models, including how a language model was built, is also critical to make the NLT-enabled inference more understandable and intuitive for human beings.

Explainable AI (XAI) Testing

Adopt dedicated XAI tools and frameworks, such as open-source ones like IBM's AI Explainability 360 library, Microsoft's InterpretML toolkit, Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), or licensed products like Amazon SageMaker Clarify. We recommend assessing the effectiveness of such tools when applied to each specific language model. In principle, they may be employed through the whole life cycle of the NLT-enabled initiative. When fine-tuning the models, they allow for a better understanding of the features and weights that are learned and show whether predictions align with human intuition (intrinsic explainability). Explainability is particularly relevant in the context of speech and text analytics techniques applied to sensitive use cases, such as fraud detection and compliance monitoring in the financial domain.

XAI Monitoring

Monitor explainability in the deployed application by defining dedicated metrics and thresholds, and activating notifications whenever anomalies are detected in the inference.

Large language models pose specific issues in terms of explainability because of their size and the overall complexity of the inference process. This applies also to conversational/generative AI applications grounded on foundation models, such as OpenAI's ChatGPT, ChatGPT-powered Bing and Google's Bard.

Accountability and Governance

Roles and Responsibilities — HITL

Define roles and responsibilities, that is, who will be accountable for the performance of the NLT-enabled system and its outputs. While developers and their leadership should be trained on AI ethics and governance, organizations should assess whether humans could, should or must be part of the NLT workflow and supervise its processing case by case. This is particularly relevant in the case of generative AI use cases, where contents are generated automatically by the system (a ChatGPT-like application, also an NMT engine). Depending on the level of risk and exposure of the output, human post-editing and fact checking may be critical (human-in-the-loop, HITL).

Documentation Enforcement

Enforce project documentation, so as to keep track of decisions made about data sourcing and labeling, as well as model definition, expectations and risks related to security, compliance and reputation. Documentation will improve the overall explainability of the system. Frameworks that apply closely to dataset documentation and model reporting are, for example, data sheets for datasets and model cards for model reporting.

End Users' Feedback and Acceptance

Ensure the needs of stakeholders not involved in the development of the initiative (end users such as employees and customers) are considered when defining governance guidelines. Customer and employee satisfaction and acceptance of the technology may impact the adoption, its ROI and the very same feasibility of the NLT-enabled initiative.

Recurrent Audits

Set a calendar of recurrent audits of the initiative, focusing on key compliance aspects.

Disclaimer: The organizations or the tools profiled in this research are provided for illustrative purposes only, and do not constitute an exhaustive list of examples in this field nor an endorsement by Gartner of the referenced organizations or tools.

Evidence

This research is grounded on interactions (inquiries and briefings) with vendors and clients about responsible AI approaches adopted in NLTs. It is also based on previous research conducted by Garner on responsible AI framework and public information on the topic (including news articles, specialized blog posts, academic papers and company policies).

[1] Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model, arxiv.org.

[2] The Italian Data Protection Authority started to scrutinize OpenAI ChatGPT's compliance with the GDPR requirements in March 2023. This led to the temporary block of the application in Italy and, sequentially, the creation of a task force by the European Data Protection Board to share information on similar enforcement actions undertaken by European Data Protection Authorities in April 2023. This ban was the first one proclaimed by a European country with the goal of protecting users' privacy. Bans enforced by other countries in early 2023 were justified on a different basis. China's ban, for example, occurred in February 2023 and seemed more related to the Chinese government's concern that content generated by ChatGPT may spread misinformation.

[3] Gender Bias in Machine Translation, Association for Computational Linguistics.

[4] Identifying Spurious Correlations for Robust Text Classification (PDF), Association for Computational Linguistics.

Acronym Key and Glossary Terms

FedRAMP	Federal Risk and Authorization Management Program
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
HITL	human-in-the-loop
LIME	Local Interpretable Model-Agnostic Explanations
LLM	large language model
ML	machine learning
NLP	natural language processing
NLT	natural language technology
NLU	natural language understanding
NMT	neural machine translation
PII	personal identifiable information
SHAP	SHapley Additive exPlanations
XAI	explainable AI

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

A Comprehensive Guide to Responsible AI

AI Ethics: Use 5 Common Principles as Your Starting Point

Design and Implement Human-in-the-Loop Interfaces for Control, Performance and Transparency of AI

How to Responsibly Use ChatGPT (and Other LLM Applications) in Your Business Interactions

Three Critical Use Cases for Privacy-Enhancing Computation Techniques

Innovation Insight for Composite AI

Table 1: Best Practices for the Responsible Use of NLTs

Human Centricity and Lawfulness	*Security and Privacy*	*Fairness and Transparency*	*Accountability and Governance*
■ Proportionality ■ Compliance ■ Misuse prevention ■ Environmental sustainability	■ Overall application security ■ Robustness against adversarial attacks ■ Encryption techniques ■ Data scrambling procedures ■ Privacy enhancing techniques ■ Confidential information ■ Synthetic data	■ Bias detection and monitoring ■ Content moderation techniques ■ Responsible data labeling and annotation ■ Composite AI ■ Know your language models ■ XAI testing ■ XAI monitoring	■ Roles and Responsibilities — HITL ■ Documentation enforcement ■ End users' feedback and acceptance ■ Recurrent audits

Source: Gartner