Mitigating Risks in AI-Generated Content with Textmetrics Governance

written in collaboration with Extanto Technology

Authors:

Marcel Leeman, CEO, Textmetrics

Ben Berry, CFO, Extanto Technology

EXECUTIVE SUMMARY

Introduction

The proliferation of AI-generated content has transformed content creation. However, the rising challenge of apparently plausible yet inaccurate or fabricated information, together with possible bias and the potential coupled with the potential copyright issues related to GPT's training data, demands a robust governance approach. This white paper advocates for the adoption of a comprehensive ruleset, with Textmetrics positioned as the governance tool of choice. It addresses accuracy concerns, the risks associated with copyright issues, and prohibited use of AI.

Thesis Statement

For business to adopt AI, it must trust AI. To ensure ethical and reliable content creation, a ruleset is imperative for governing GPT-generated content. Textmetrics is a powerful governance tool, addressing bias and accuracy concerns and mitigating the risks associated with potential copyright issues in GPT's training data.

BACKGROUND

Brief History of GPT

Generative Pre-trained Transformer (GPT) has revolutionized content creation across diverse industries by offering powerful capabilities in generating text and assisting with various writing tasks. However, concerns have arisen regarding the accuracy and ethical implications of the content it produces. This has led to a closer examination of GPT output to ensure that the content aligns with ethical standards, is accurate, and does not propagate misinformation or bias.

Well publicized instances of GPT "hallucinations" and bias raise questions about the reliability and trustworthiness of AI-generated text. The model's ability to generate vast amounts of content quickly and efficiently has highlighted the need for careful scrutiny to verify the accuracy and authenticity of the information produced. As AI technologies such as GPT continue to evolve, there is a growing awareness of the potential ethical implications

surrounding AI-generated content. Issues such as bias in language models, lack of transparency in content creation processes, and potential copyright infringements have underscored the importance of ethical oversight and responsible use of AI in content generation. Examining these ethical implications, businesses can gain confidence that their content upholds integrity, fairness, and compliance with ethical standards.

Ethical and Copyright Concerns

Instances of GPT generating misinformation and potential copyright infringements have eroded trust and raise significant ethical and copyright concerns for businesses that utilize AI-generated content. The ethical concerns stem from the implications of disseminating inaccurate or misleading information to the public, which can erode trust, misinform individuals, and harm reputations. The use of AI-generated content that may infringe on copyright-protected material raises legal and ethical questions regarding intellectual property rights and fair use.To address these ethical and copyright concerns effectively, businesses must establish a robust ruleset governing the creation, usage, and dissemination of AI-generated content. This ruleset should outline guidelines for respecting copyright laws, ensuring accuracy and transparency in content creation, acknowledging the contributions of both humans and AI models, and upholding ethical standards in communication. By implementing a comprehensive ruleset, organizations can mitigate the risks associated with ethical lapses, copyright violations, and misinformation while promoting responsible and compliant use of AI-generated content.

THE PROBLEM STATEMENT

There are six risks inherent to the use of AI in a business setting:

Risk 1 – Users rely on the veracity of answers generated by the AI and do not feel the need to check the answers independently. Because AI can be prone to fabricating responses, **quality and reputation may suffer if this information is utilized without first being vetted and checked for accuracy.**

Risk 2 – For data to be usable, it must be structured, accessible and accurate. Therefore, there is a need to **standardize** Processes, Data utilization, Technology and Quality Control. In order to make the most of AI, one must also use tools that support data privacy and confidentiality.

Risk 3 – Implicit bias present in the model and output will not be addressed by generative AI

technology. This bias is present due to programming choices and the data used in training. In order to minimize the use of biased material, the output should be reviewed by technology that can identify the bias, **determine the quality of the output and highlight risks.**

Risk 4 - The underlying model of the AI changes over time altering or degrading the quality of the output. The LLM may demonstrate a bias in the future that was not present at the time the LLM was originally adopted due to updates to the data upon which the LLM is trained. **There is the risk that the quality of output changes.**

Risk 5 – Much of the data with which the LLM is trained is proprietary intellectual property (IP), protected by copyright and may have been used without permission. Therefore, the LLM was trained based upon data which is not the property of the owner of the LLM. **The training data used should be reviewed and copyrighted material deleted** before being processed by the LLM, unless permission to use that material is granted by its owner.

Risk 6 – Give users a defined set of tools instead of the whole toolbox. By giving the users **defined usage** of an LLM (e.g.: predefined prompts), the risk of misuse is lowered, while at the same time the usability of the system increases dramatically. **Defined usage will lower the possible risks while at the same time increasing the usability and adoption of AI for specific business purposes.**

(Adapted from article "Gartner Survey Shows Generative AI Has Become an Emerging Risk for Enterprises")

Textmetrics focuses on five of these risks:

Apparent Plausibility vs. Accuracy

GPT's proficiency in producing content that appears plausible yet may be inaccurate or fabricated poses substantial risks for businesses relying on AI-generated content for critical functions such as customer communication, marketing, and data analysis.

Copyright Issues in GPT's Training Data

The potential ingestion of copyrighted content without permission in GPT's training data introduces legal risks for companies using AI-generated content. The lack of transparency in the training process raises concerns about the origin and permissions associated with the data used.

Guarding Against the Use of AI in a Prohibited Setting

The use of AI as a generative tool has been banned in many business and academic settings: many businesses have prohibited the use us of AI due to the risk of divulging proprietary data or information to the AI which may then expose it inadvertently to the public. Academia has restricted its use because of its generative capabilities and noted inaccuracies.

Improving the Quality of Output

Textmetrics reviews the created content for readability, credibility, implicit bias, DEI, style, sentiment and gender target, helping content creators craft persuasive content. Textmetrics is a rules-based software, reviewing content and offering suggestions to content creators through the lenses with which it has been trained. These rules allow users to "future proof" content against changes to an LLM that may cause it to exhibit biases down the road that were not present at the time of adoption.

THE NEED FOR A RULESET

Ensuring Accuracy and Reliability

A ruleset is crucial to mitigate the risks associated with inaccurate or misleading information. Guidelines focused on accuracy and reliability act as a safeguard against unintentional dissemination of false or fabricated content:

Ruleset Implementation: Establishing a ruleset for GPT is essential to provide clear guidelines and standards for content generation, to ensure that the output is accurate, reliable, and aligns with the organization's style guide and ethical considerations. This ruleset acts as a framework to regulate the behavior of AI models and mitigate its inherent risks.

Mitigating Risks: By focusing on accuracy and reliability in GPT guidelines, organizations can reduce the likelihood of AI-generated content containing inaccuracies, biases, or fabricated information. This proactive approach helps prevent the dissemination of false narratives that could harm trust, credibility and organizational reputation.

Safeguarding Against False Content: Guidelines that emphasize accuracy and reliability serve as a protective measure against the inadvertent spread of false or misleading content by AI systems, acting as a filter to ensure that the informati generative abilities on generated is factually correct and aligned with ethical standards.

Promoting Ethical Content Generation: Implementing guidelines focused on accuracy and reliability safeguards against misinformation while promoting ethical content generation practices. By adhering to these guidelines, organizations can uphold integrity, transparency, and trustworthiness in their use of AI technologies for content creation.

Addressing Copyright Risks

It is essential to establish guidelines for the use of copyrighted material in AI training data. A ruleset can provide clarity on permissions, reducing the legal risks associated with potential copyright infringement.

INTRODUCING TEXTMETRICS AS THE GOVERNANCE TOOL

Overview of Textmetrics

Textmetrics is a comprehensive AI ruleset governance tool, addressing challenges in accuracy, credibility, bias and copyright issues. Its advanced algorithms and linguistic insights provide a robust solution to enhance content quality and reliability.

Benefits of Textmetrics

Accuracy and Fact-Checking

Textmetrics employs state-of-the-art algorithms to verify content accuracy and credibility, thereby reducing the risk of misinformation. Textmetrics' algorithms verify the quality and reliability of content, enhancing its trustworthiness and minimizing the dissemination of inaccurate or misleading information. The algorithms analyze text, identify potential errors or inconsistencies, and provide valuable insights to improve the overall accuracy, credibility and quality of the content generated using Textmetrics.

Copyright Compliance

Textmetrics facilitates compliance with copyright regulations by ensuring transparency and adherence to permissions for content used in AI training data. By incorporating features that promote transparency and respect for intellectual property rights, Textmetrics helps organizations navigate the complexities of copyright laws when utilizing AI-generated content.

Uncovering Prohibited Use of AI

According to the results of a 2023 study performed by cybersecurity firm ExtraHop in which 1200 CTOs were surveyed, while having a corporate policy in place that bans the use of AI in the workplace is good business, it is nonetheless insufficient to fully prevent its use. In order to determine if AI has been used in a restricted setting, it is necessary to evaluate the quality and complexity of the content; analyze the language patterns for inconsistencies in tone, vocabulary or sentence structure and examine metadata and source information for signs of AI tools or platforms.

Real-time Monitoring and Compliance

Continuous monitoring in real-time through the integration of Textmetrics' editing functionality enables businesses to stay compliant with evolving regulations, effectively mitigating risks associated with accuracy, implicit bias and copyright concerns. By embedding Textmetrics into their content creation workflows, organizations can proactively monitor and enhance the accuracy, quality, and compliance of their content in real-time. This dynamic editing functionality ensures that content aligns with the latest regulatory requirements and enables businesses to address potential inaccuracies, inconsistencies, or copyright issues promptly.

Textmetrics' real-time monitoring capabilities allow businesses to detect and rectify any deviations from regulatory standards or copyright guidelines as they occur, minimizing the likelihood of compliance breaches or legal infringements. A proactive approach to content editing and compliance management will help organizations maintain a high level of accuracy, reliability, and ethical standards in their content creation processes.

IMPLEMENTATION STRATEGY

Integrating Textmetrics into AI Systems

Textmetrics easily integrates into existing AI infrastructure, ensuring compatibility with diverse content generation workflows. Its user-friendly interface facilitates a smooth incorporation into existing processes.

Because Textmetrics is a mature, rules-based writing tool, it is not subject to the inconsistencies found in the content created by AI. In order to improve the quality of AI’s output, we propose an integration model that is commonly used to implement user interfaces, data and controlling logic. This integration model separates the business logic and

presentation layer from one another, giving reliable structure to the business logic results and predictability to how those results are presented to the UI and, by extension, to its end users. This "separation of concerns" is illustrative of the relationship between Model, View and Controller.

This separation isolates the application's concerns into three distinct layers. The GPT layer, as the Model, is responsible for the application's logic, storing and retrieving data from back-end data stores. It will likely include mechanisms for validating data and carrying out other data-related tasks. The UI serves as the View and is necessary for the user to interact with the application. It displays the data and enables users to interact with it. The Textmetrics layer becomes the Controller, containing the business rules necessary to facilitate communications with clear, consistent content that can be relied upon to convey the organization’s high quality content.

GPT is a Large Language Model (LLM), creating content with its extensive understanding of language patterns and knowledge. This LLM acts as the foundational element, processing and generating human-like text responses. However, GPT operates with certain limitations. It can exhibit a lack of genuine understanding, has a sensitivity to input phrasing, can be prone to bias (because of the implicit bias sometimes present in source data) and contextual dependence as well as exhibit difficulty handling complex tonal shifts.On the user-facing side, the UI provides an intuitive interface for clients to engage with the language model. It ensures a user-friendly experience and facilitates meaningful interactions. Textmetrics, as the Controller, interprets user inputs, making use of Natural Language Processing (NLP) to understand context and intent, and guides the LLM to generate coherent and contextually relevant responses. Textmetrics acts as the orchestrator, harmonizing user interactions with the vast language capabilities of GPT.

Training and Adoption

Effective training programs empower teams to utilize Textmetrics optimally. Comprehensive resources and support facilitate widespread adoption within the organization, ensuring maximum benefits for business and user.

Textmetrics has a standardized onboarding process consisting of 4 major phases. A phase is comprised of multiple sessions to gather all necessary information and to discuss choices that need to be made. Over the course of eight weeks, Textmetrics will execute the following project phases: Project definitions (kickoff, scoping, reporting, customizations); Configuration & Customizations (Content, SEO, IT); Go Live (Training, Monitoring, Feedback); Evaluation and Support (Feedback, Fine-tuning, Monitoring & Support).

The first phase defines the issues that Textmetrics is to address, who and how the system is to be used and the rulesets that will be integrated. This is done, in part, by determining the LLM (GPT) to be integrated, understanding its strengths, weaknesses and its approach to the risks, developing a strategy to address those risks with the tools that Textmetrics provides, as well as providing barriers around the organization's intellectual property (IP) to keep it isolated to the organization. The next phase further uncovers and implements the customizations required to get the most out of the integration, determines impact of the platform to the technical environment, and fine-tunes the algorithms. The third phase, the go-live phase, activates accounts, trains users and sets up monitoring. The final phase, that of evaluation and support, further fine-tunes operations, monitors use and addresses any issues that arise during early adoption. This final phase incorporates feedback from the user base, captured information that may have been missed or not presented in the trainings.

CASE STUDIES

GPT Generating Misinformation

According to recent articles published by the Associated Press there have been several high-profile instances of the deliberate creation and dissemination of misinformation with the help of AI targeting political figures:

- An AI-generated robocall using the apparent voice of Joe Biden just prior to New Hampshire's recent primary election discouraging democrat voters from voting in the actual primary.
- A video of Moldova’s pro-Western president throwing her support behind a political party friendly to Russia.
- Audio clips of Slovakia’s liberal party leader discussing vote rigging and raising the price of beer.
- A video of an opposition lawmaker in Bangladesh — a conservative Muslim majority nation — wearing a bikini.

GPT's Use of Copyrighted Content

There are several pending court cases related to AI companies' use of copyrighted content, particularly in the context of Generative Pre-trained Transformer (GPT) models, which are shaping the landscape of how AI may utilize copyrighted materials for training purposes. The outcome of these cases, as highlighted by the National Law Review, will shape the

boundaries and implications of AI systems accessing and processing copyrighted works.

- **Fair Use Defense:** AI companies, including OpenAI and Microsoft, are invoking the "Fair Use" doctrine as a defense in these court cases. They argue that the training of AI models with copyrighted materials falls within the realm of fair use, comporting with established precedents that recognize limited uses of copyrighted works for transformative purposes.
- **Black Box Defense:** AI companies are calling on the "black box" defense, referring to the opacity surrounding how AI tools like GPT ingest images and content, as well as the subsequent actions taken with this information. This lack of transparency raises questions about the control and understanding AI companies have over the data ingested by their systems, leading to debates about accountability and responsibility for potential copyright infringements.

The outcomes of these court cases will likely have a significant impact on the future practices of AI companies regarding the use of copyrighted materials for training AI models. The decisions made in these lawsuits will set precedents for how AI systems can access, process, and generate content derived from copyrighted sources while navigating the boundaries of fair use and intellectual property rights.

FUTURE PROSPECTS

Evolving Regulations and Standards

The collaborative efforts of industry and government are crucial to shape evolving standards for AI content governance . Predictions for future developments underscore the need for a collective approach to address emerging challenges.

"Show Your Work"

In order to overcome the lack of trust in AI and its work product, AI should demonstrate how it developed its conclusions, in other words, AI should be asked to show its work. In an opinion piece in the Government Technology blog from March 2024, author Ben Miller suggests that users must be able to trust the algorithms AI is using to perform its work, that they are running correctly and with minimal bias. According to Mr. Miller, it must be able to evidence:

- **Transparency** - make it clear when AI is being used and how it’s being used. This might include “model cards” or “fact sheets” outlining basics such as where it pulls data from and its performance metrics.

- **Explainability** - provide narratives, statistics and other tools to help users understand how an algorithm works, especially for specific outputs.
- **Auditability** - AI tool design allows users to monitor it for key indicators of success and failure such as bias and accuracy. This might involve providing data provenance and lineage, allowing one to trace an output back to the data it was based on.

The application of a strong ruleset is a firm step in this direction.

CONCLUSION

Recapitulation of the Problem

As stated above, for business to adopt AI, it must be able to trust AI. To ensure ethical and reliable content creation, a ruleset is imperative to govern GPT-generated content. Textmetrics is a powerful governance tool, addressing accuracy concerns and mitigating the risks associated with potential issues associated with GPT.

The Role of Textmetrics

AI-generated content has transformed content creation. Because of the rising challenge of superficially plausible yet inaccurate or fabricated information, together with potential copyright issues related to GPT's training data, business users will be wise to look for a robust governance approach. We believe that Textmetrics is the governance tool of choice as it easily addresses accuracy concerns and the risks associated with copyright issues.

- Textmetrics focuses on these primary risks:
- Apparent Plausibility vs. Accuracy
- Copyright Issues in GPT's Training Data
- Guarding Against the Use of AI in a Prohibited Setting
- Risks to the Quality of Output, Now and in the Future

Call to Action

As fully integrated functionality, Textmetrics offers a robust feature set to organizations; a mature content creation assistant and a safety net against the risks of AI.

Artificial Intelligence is the shape of things to come and Textmetrics molds AI into a trustworthy partner for success.

Contact us to learn more at: https://extanto.com/contact

REFERENCES

"Disinformation Researchers Raise Alarms About A.I. Chatbots"; February 8, 2023; New York Times; https://readwise.io/reader/shared/01grvprtzgpj1saf78nk17x4bk/

"'New York Times' sues ChatGPT creator OpenAI, Microsoft, for copyright infringement"; National Public Radio; December 27, 2023; Bobby Allyn; https://www.npr.org/2023/12/27/1221821750/new-york-times-sues-chatgpt-openai-microsoft-for-copyright-infringement

"Election disinformation takes a big leap with AI being used to deceive worldwide"; March 14, 2024; Associated Press; Ali Swenson and Kelvin Chan; https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd

"New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary"; January 22, 2024; Associated Press; Ali Swenson and Will Weissert; https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5

"Library Copyright Alliance Principles for Copyright and Artificial Intelligence"; Library Copyright Alliance; July 10, 2023; https://www.librarycopyrightalliance.org/wp-content/uploads/2023/06/AI-principles.pdf

"Open AI and Journalism"; **Open AI Blog; January 8, 2024; Open AI;** https://openai.com/blog/openai-and-journalism#OpenAI

"Generative AI systems Tee Up Fair Use Fight"; National Law Review, February 29, 2024, Ariba A. Ahmad , Andrew M. Gross of Foley & Lardner LLP, https://www.natlawreview.com/article/generative-ai-systems-tee-fair-use-fight

"Making AI Work for Government: It All Comes Down to Trust"; March 2024; Government Technology Opinion Blog; Ben Miller; https://www.govtech.com/opinion/making-ai-work-for-government-it-all-comes-down-to-trust

"AI's mysterious black box problem explained"; University of Michigan-Dearborn News, March 6, 2023, Lou Blouin; https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained

"The Generative AI Tipping Point"; Extrahop Networks; October 6, 2023; https://cloud-assets.extrahop.com/resources/ebooks/extrahop-generative-ai-survey-ebook.pdf

"Gartner Survey Shows Generative AI Has Become an Emerging Risk for Enterprises"; August 8, 2023, Gartner Press Office - Rob Van Der Meulen; https://www.gartner.com/en/newsroom/press-releases/2023-08-08-gartner-survey-shows-generative-ai-has-become-an-emerging-risk-for-enterprises

"LICENSING RESEARCH CONTENT VIA AGREEMENTS THAT AUTHORIZE USES OF ARTIFICIAL INTELLIGENCE"; Authors Alliance Blog; January 10, 2024; Rachael G. Samberg, Timothy Vollmer, and Samantha Teremi; https://www.authorsalliance.org/2024/01/10/licensing-research-content-via-agreements-that-authorize-uses-of-artificial-intelligence/