Grammarly's Responsible AI Standards

Out of our commitment to improving communication while safeguarding users' thoughts and words, Grammarly established our guiding responsible AI standards—transparency, fairness and safety, user agency, accountability, and privacy and security. These pillars not only guide our AI developments but also reflect our dedication to building AI that is safe, fair, and reliable, ensuring ethical innovation and high standards.

- Transparency
- Fairness and safety
- User agency
- Accountability
- Privacy and security

Upholding transparency

At Grammarly, we believe users should recognize when they're engaging with AI and understand the logic behind AI-generated suggestions. We're committed to transparency about the risks, limitations, and potential biases of our AI systems, as well as our business motivations, to build trust and ensure our technology is objective. Providing documentation on how our AI operates, including its data sources and training methods, informs users about its capabilities and guides appropriate use.

We disclose how user data is processed, stored, and protected, especially when AI uses personal data to influence decisions. Grammarly details this in our technical specifications and gives all users the ability to opt out of having their content used for model training. Grammarly makes money by selling subscriptions, not through ads, aligning our incentives with providing the highest-quality experience.

Championing fairness and safety

We prioritize fairness and safety in our AI systems, ensuring they deliver unbiased outputs without causing harm. We're committed to identifying and mitigating biases to guarantee equitable performance for all users. Our sensitivity guidelines include supporting inclusive language, avoiding bias, and never suggesting offensive content.

- **Grammarly's risk assessment process:** Every feature must undergo a risk assessment process and receive approval from our Responsible AI (RAI) team before launching. Each feature receives a risk categorization according to criteria developed by our team, and launches are not approved unless the recommended mitigations are implemented.

- **Seismograph:** Grammarly's RAI team has developed a tool called Seismograph to ensure product compliance with sensitivity guidelines. Seismograph uses machine learning classifiers and dictionaries to identify and mitigate potentially sensitive language, continuously updating to address lexical changes.

Fostering user agency

We believe AI should enhance skills, respect autonomy, and empower users to make decisions. We give users the ability to evaluate AI's suitability for specific use cases, understanding that it varies depending on their organization's needs and policies. Users can accept or disregard AI suggestions and control the types of suggestions they receive in their settings. Our commitment is to empower every user to use AI to express themselves in the most effective way possible.

Building on our commitment to helping users transparently and responsibly use AI, Grammarly offers several tools to equip them with the information necessary to make appropriate judgments about the use of AI for their specific purposes.

Authorship is a Grammarly tool (currently in beta testing) that enables users to trace text origins in their writing, reflecting our commitment to responsible AI use. Grammarly's AI detector is another feature that evaluates the originality of selected text and also provides guidance on interpreting and acting on the results responsibly, giving users the opportunity to appropriately attribute sources, rewrite content, and mitigate the risk of being incorrectly accused of AI plagiarism.

Embracing accountability

Accountability is the commitment to ethical and responsible AI. We work to anticipate potential abuse, assess its frequency, and pledge to take ownership of our model's outcomes.

We perform regression testing to monitor changes in our model, reflecting our accountability to designing AI that supports fair communication. We also conduct continuous user feedback reports, and our robust Support team collects instances of adverse model outcomes. This allows us to review and iterate on incorrect or harmful information that our product might generate. We understand that our users rely on us to ensure that our suggestions are helpful, and our team works hard to ensure that we take ownership and accountability when we get it wrong. Users can report offensive and incorrect suggestions in product, and their submissions are reviewed and prioritized by our RAI team.

Preserving privacy and security

Ensuring the privacy and security of our users is of utmost importance. We maintain industry-standard **data protection**, **secure infrastructure**, and **third-party verification**. Our teams comply with legal and internal standards, upholding privacy and implementing tight controls to safeguard user data.

Users can control whether their content is used to improve our products, and we do not allow third-party AI processors to store user data or use it to train their models. Grammarly maintains a thorough vendor review process, including multistep security and privacy assessments.

We proactively prevent adversarial attacks with a red team of security experts and maintain a bug bounty program through HackerOne.

To learn more about Grammarly's approach to responsible AI, read our white paper "The Responsible AI Advantage: Guidelines for Ethical Innovation" or visit grammarly.com/responsible-AI