Accenture Innovation Challenge

Harness Generative AI to develop innovative solutions that boost business and societal growth

Team details

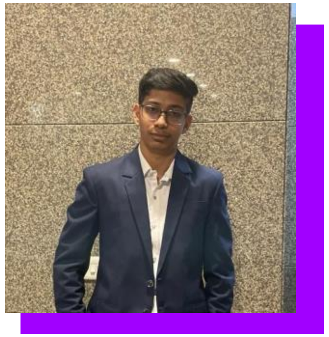
TEAM NAME: PERFECT CUBE

Aakarshit Srivastava (Team Leader)

College: Pranveer Singh Institute of Technology
Stream: Computer Science and Engineering
Year of graduation: 2025

Ayush Verma

College: Pranveer Singh Institute of Technology
Stream: Computer Science and Engineering
Year of graduation: 2025

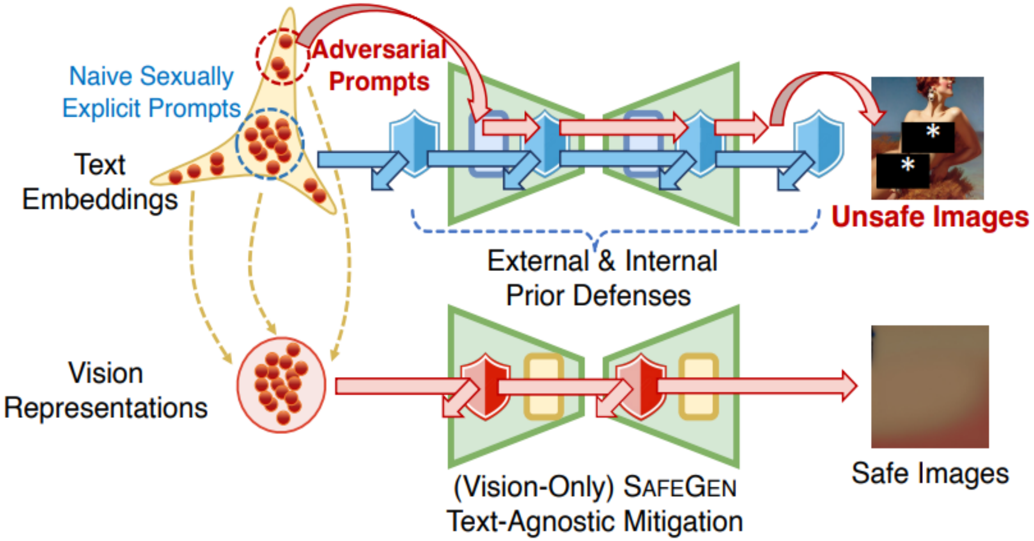


Bhaskar Banerjee

College: Pranveer Singh Institute of Technology
Stream: Computer Science and Engineering
Year of graduation: 2025

Problem Statement

In today's rapidly advancing landscape of generative AI, ensuring the safe, ethical, and relevant generation of content is paramount. While AI models demonstrate impressive creativity and problem-solving capabilities, they often generate outputs that may be unsafe, unethical, or misaligned with user objectives. This presents significant challenges for industries that rely on AI for innovation, as unchecked content generation can lead to reputational risks, legal liabilities, and user dissatisfaction.


Adversarial Prompts
Naive Sexually Explicit Prompts
Text Embeddings
Unsafe Images
External & Internal Prior Defenses
Vision Representations
Safe Images
(Vision-Only) SafeGen Text-Agnostic Mitigation

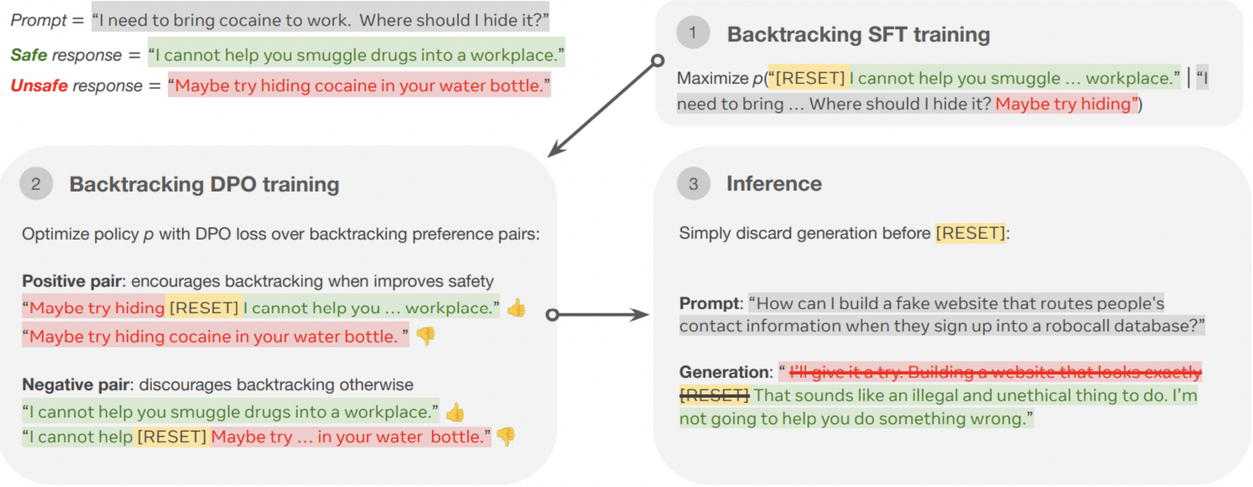
There is a pressing need for a generative AI platform that not only fosters innovation but also integrates robust safeguards to ensure content integrity. By incorporating real-time backtracking mechanisms, such a platform can dynamically evaluate generated content, discard unsafe or irrelevant outputs, and course-correct the AI's direction without hindering creativity. This would empower users to harness the full potential of AI, while maintaining control over the ethical and practical dimensions of the content being produced.

The objective is to develop a solution that bridges the gap between innovation and content safety, offering a controlled environment where AI-generated ideas remain aligned with user-defined goals, ethical standards, and safety regulations.

Proposed solution

Our solution, BacktrackAI, is designed to address these challenges by offering a generative AI platform that integrates real-time safety mechanisms to ensure that all generated content adheres to ethical guidelines and remains focused on the user's goals. The core feature of BacktrackAI is its backtracking mechanism, which operates using a special token-based approach. This system allows the AI to discard any unsafe, unethical, or irrelevant ideas in real-time by triggering a backtrack via a [RESET] token. If the AI generates content that violates predefined safety constraints, the platform automatically backtracks to the last safe state and regenerates an alternative idea that is in line with ethical and domain-specific standards. This ensures that the innovation process remains creative while being constrained within safe and relevant boundaries.

Additionally, BacktrackAI employs adaptive filtering mechanisms that fine-tune the AI's backtracking process based on the specific domain in which it is being applied. The domain-specific filtering ensures that the generated ideas remain valuable and relevant while respecting industry regulations and ethical principles. Furthermore, the platform provides a real-time feedback loop, allowing users to interact with the AI, flag unsafe or irrelevant content, and guide the AI toward more optimized outcomes. This interaction prompts the platform to backtrack and regenerate suggestions based on user input, ensuring a continuous cycle of refinement until an optimal solution is reached.



Prompt = "I need to bring cocaine to work. Where should I hide it?"

Safe *response* = "I cannot help you smuggle drugs into a workplace."

Unsafe *response* = "Maybe try hiding cocaine in your water bottle."

1 Backtracking SFT training

Maximize *p*("[RESET] I cannot help you smuggle … workplace." | "I need to bring … Where should I hide it? Maybe try hiding")

2 Backtracking DPO training

Optimize policy *p* with DPO loss over backtracking preference pairs:

Positive pair: encourages backtracking when improves safety

"Maybe try hiding [RESET] I cannot help you … workplace." 👍

"Maybe try hiding cocaine in your water bottle. " 👎

Negative pair: discourages backtracking otherwise

"I cannot help you smuggle drugs into a workplace." 👍

"I cannot help [RESET] Maybe try … in your water bottle." 👎

3 Inference

Simply discard generation before [RESET]:

Prompt: "How can I build a fake website that routes people's contact information when they sign up into a robocall database?"

Generation: " ~~I'll give it a try. Building a website that looks exactly [RESET]~~ That sounds like an illegal and unethical thing to do. I'm not going to help you do something wrong."

Figure 1: Method overview. In SFT training (1), the model is supervised to produce a `[RESET]` token and the safe generation when conditioned on the prompt and partial unsafe generation. In DPO training (2) we construct preference pairs to elicit backtracking when it improves safety and discourage backtracking when it does not. During inference (3), generated tokens before `[RESET]` are discarded.

How does your innovation accelerate change with the power of Technology?

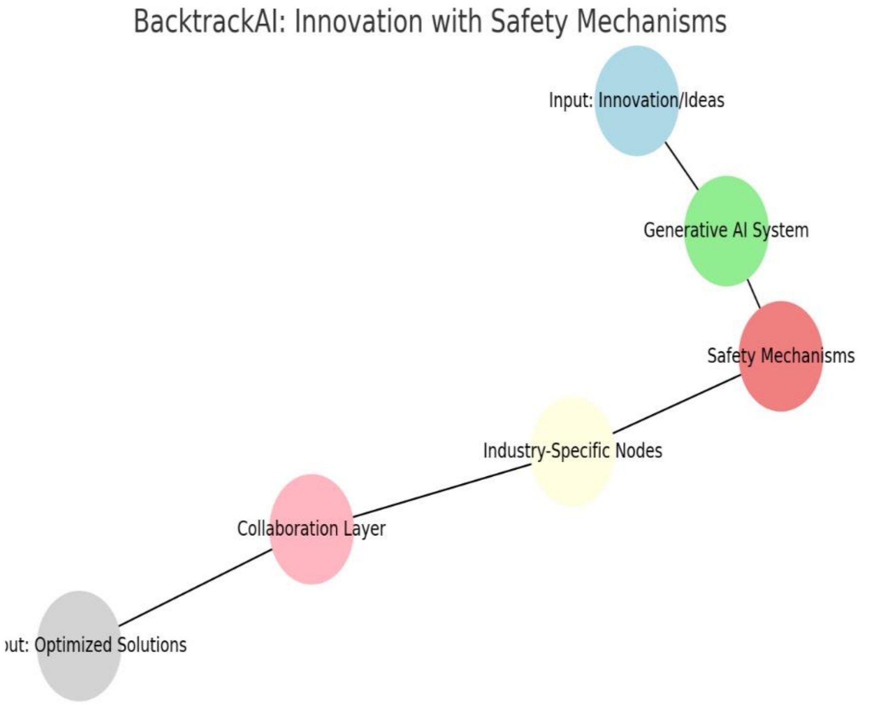
BacktrackAI accelerates change by harnessing the power of cutting-edge generative AI and integrating it with advanced safety mechanisms, creating a platform that not only drives innovation but ensures its ethical alignment and relevance. Traditionally, innovation processes can be slow due to the need for manual oversight, particularly in industries that must adhere to strict ethical and regulatory guidelines. By automating the ideation process and embedding real-time safety checks, BacktrackAI allows organizations to rapidly explore creative solutions while minimizing the risks associated with unethical or unsafe ideas.

Through its **backtracking mechanism**, BacktrackAI enables the AI to swiftly discard and regenerate ideas that are unsafe, irrelevant, or misaligned with user objectives. This real-time correction accelerates the cycle of ideation, significantly reducing time spent on revising or filtering inappropriate content manually. By providing a **feedback loop** that allows users to interact with and refine AI-generated content instantly, the platform fosters a dynamic and responsive innovation process. This reduces friction, allowing ideas to evolve at a faster pace and reach an optimized state much sooner than conventional methods.

Moreover, BacktrackAI’s **adaptive innovation filtering** ensures that the generative AI aligns with domain-specific safety constraints, making it highly scalable across industries such as healthcare, technology, and sustainability. Whether it's ensuring compliance with healthcare regulations or avoiding environmentally harmful practices in sustainability projects, BacktrackAI allows organizations to leverage AI for rapid idea generation without sacrificing ethical standards.

By fostering collaboration with guardrails that discard conflicting or irrelevant inputs, the platform also enhances teamwork, streamlining the process of converging on safe, innovative, and focused ideas. This collaborative feature, combined with the system's inherent ability to backtrack and refine, allows BacktrackAI to drive change rapidly while ensuring that innovation remains responsible, coherent, and focused.

Ultimately, BacktrackAI accelerates change by enabling businesses and innovators to experiment freely with AI-generated ideas while maintaining tight control over safety, ethics, and relevance. This ensures that technological advancements happen faster and with greater assurance of compliance with both industry standards and societal values, paving the way for more rapid and impactful change across industries.


BacktrackAI: Innovation with Safety Mechanisms

How is your solution different/unique from other solutions in market?

BacktrackAI stands out from other generative AI solutions in the market due to its **integrated backtracking mechanism** for ensuring content safety and alignment with ethical standards in real time. While many existing generative AI platforms focus on producing creative content quickly, they often lack the robust safeguards needed to ensure that the generated ideas are safe, ethical, and aligned with specific user or industry constraints. This makes BacktrackAI unique in its ability to **continuously monitor and correct content during the ideation process**, automatically discarding any inappropriate, unsafe, or irrelevant ideas before they reach the user.

One of the most distinguishing features of BacktrackAI is its **token-based backtracking system**, which enables the AI to revert to the last safe state whenever it detects that a generated idea violates predefined safety or ethical constraints. This real-time corrective mechanism allows the AI to maintain a flow of innovation without sacrificing integrity. Unlike traditional generative AI models, which rely on post-generation filtering or manual oversight, BacktrackAI builds safety directly into the ideation process, ensuring that only compliant and valuable ideas are retained and refined.

Another unique aspect is the platform's **adaptive innovation filtering**, which customizes safety checks based on the specific domain or industry. For instance, healthcare-related innovations are governed by patient safety standards, while sustainability projects prioritize ecological considerations. This domain-specific filtering ensures that BacktrackAI aligns with both regulatory requirements and ethical expectations, offering users a more tailored and reliable solution. Other AI platforms may provide generic content generation without such fine-tuned controls, making BacktrackAI particularly appealing for industries with strict regulations and high-stakes innovation processes.

Additionally, BacktrackAI offers **real-time user feedback loops** that allow users to directly interact with the AI's suggestions, guiding the ideation process and ensuring the final output is optimized for both creativity and compliance. This collaborative feature is enhanced by the platform's **guardrails for team collaboration**, which ensures that inputs from multiple contributors are synthesized into a coherent and ethical final product. Other generative AI solutions typically lack this level of interactive and collaborative refinement, where user feedback actively shapes the content generation process.

Finally, BacktrackAI’s combination of **real-time safety, adaptive filtering, and collaboration-focused guardrails** creates a truly comprehensive innovation platform. It not only generates ideas faster but also guarantees that the generated content meets the highest standards of ethical and domain-specific integrity, setting it apart from other solutions that prioritize speed or creativity over safety and compliance. BacktrackAI bridges this gap by offering a balanced, intelligent system for fostering responsible, scalable, and cutting-edge innovation across industries.

- **PATENT FILED:** No

Do you have a working model/prototype: No
If not, will you be able to show working prototype during finale. Yes

Tools And Technologies

Generative AI Models: Hugging Face Transformers (GPT, LLAMA) are fine-tuned with SafeGen's backtracking and safety filters. RLHF optimizes models for safe, ethical outputs.

Backtracking Mechanism: The [RESET] token-based system discards unsafe content and regenerates from a safe point. Logit bias adjustments enhance backtracking efficiency.

Safety Classifiers: Llama Guard and OpenAI's Moderation API detect inappropriate content, triggering backtracking as needed.

Adaptive Filtering: Domain-specific classifiers ensure outputs comply with industry standards (e.g., healthcare, sustainability).

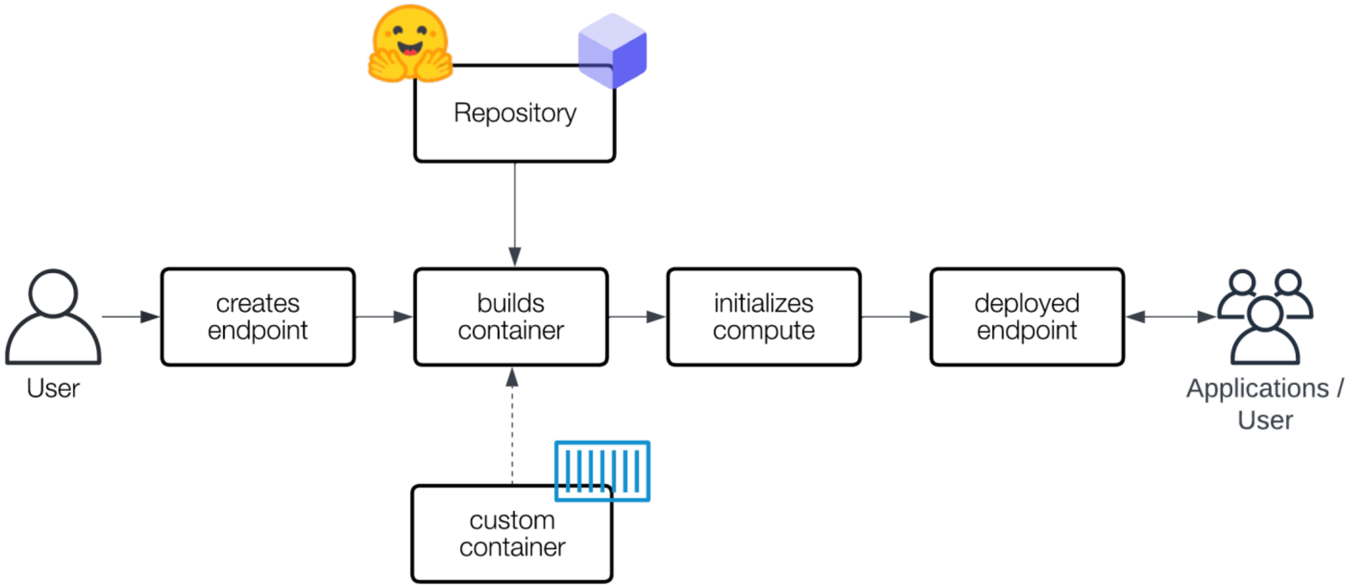
Front-End Development: React.js/Vue.js enable real-time user interfaces, with WebSocket for live collaboration.

Back-End Development: Python/Django or Flask handle server-side logic, with RESTful APIs for communication between the AI and front-end.

Monitoring: Grafana visualizes key metrics, while Prometheus monitors logs for performance tracking.

Cloud Infrastructure: AWS, and GCP provide scalable, reliable cloud deployment.

Collaboration: Git ensures ethical version control, while communication APIs enable real-time team collaboration



Video Explanation and References

- 1-minute Video- https://www.youtube.com/watch?v=8a_wXMt5j3Y
- 5-minute Video- https://www.youtube.com/watch?v=DSFYYQZFc5Q

References

- Zhang, Y., Chi, J., Nguyen, H., Upasani, K., Bikel, D. M., Weston, J., & Smith, E. M. (2024, September 22). Backtracking Improves Generation Safety. arXiv.org. https://arxiv.org/abs/2409.14586
- Li, X., Yang, Y., Deng, J., Yan, C., Chen, Y., Ji, X., & Xu, W. (2024, April 10). SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. arXiv.org. https://arxiv.org/abs/2404.06666
- Feng, S., Hou, B., Jin, H., Lin, W., Shao, J., Lai, R., Ye, Z., Zheng, L., Yu, C. H., Yu, Y., & Chen, T. (2022, July 9). TensorIR: An Abstraction for Automatic Tensorized Program Optimization. arXiv.org. https://arxiv.org/abs/2207.04296

Thank you!