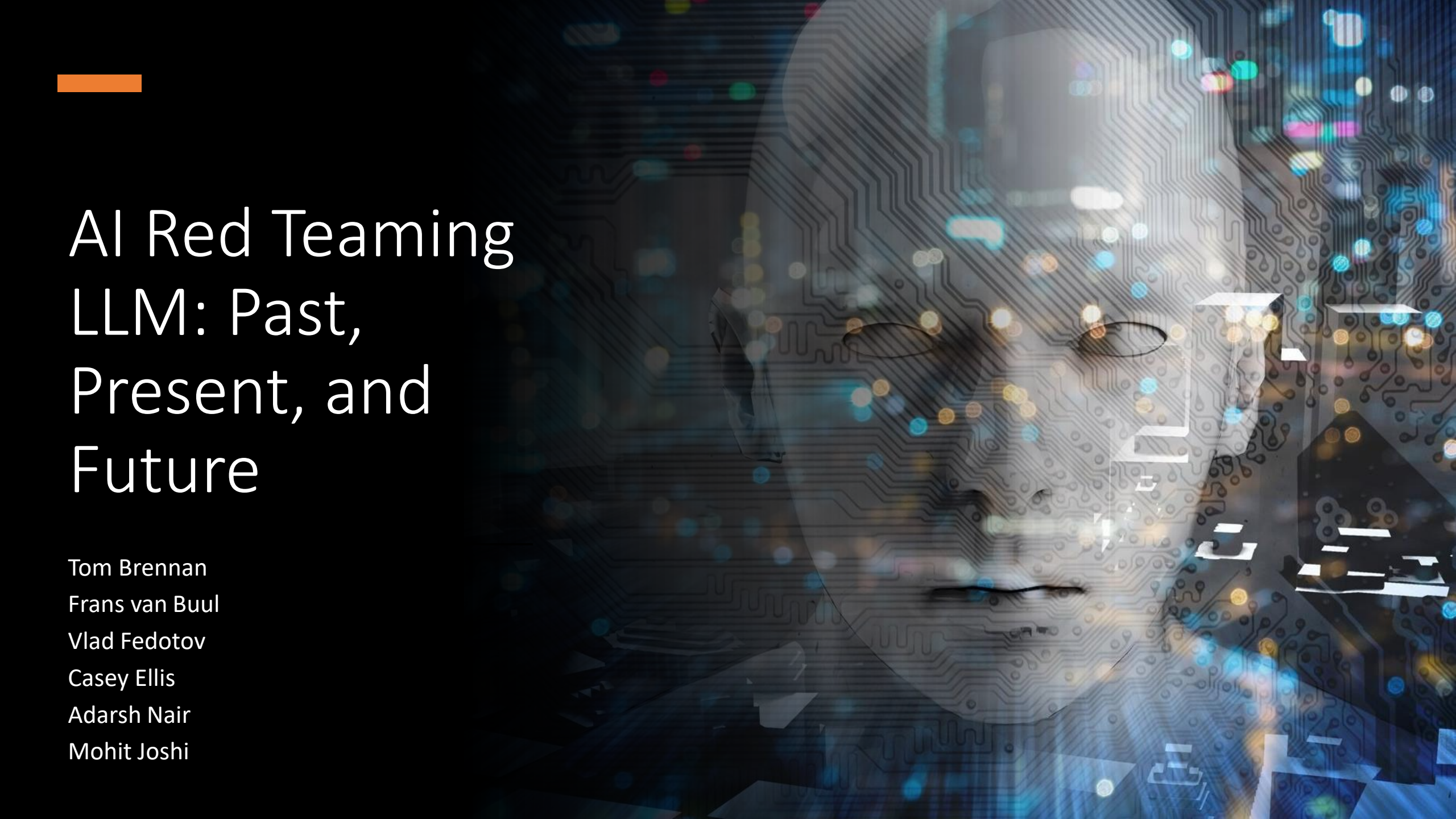
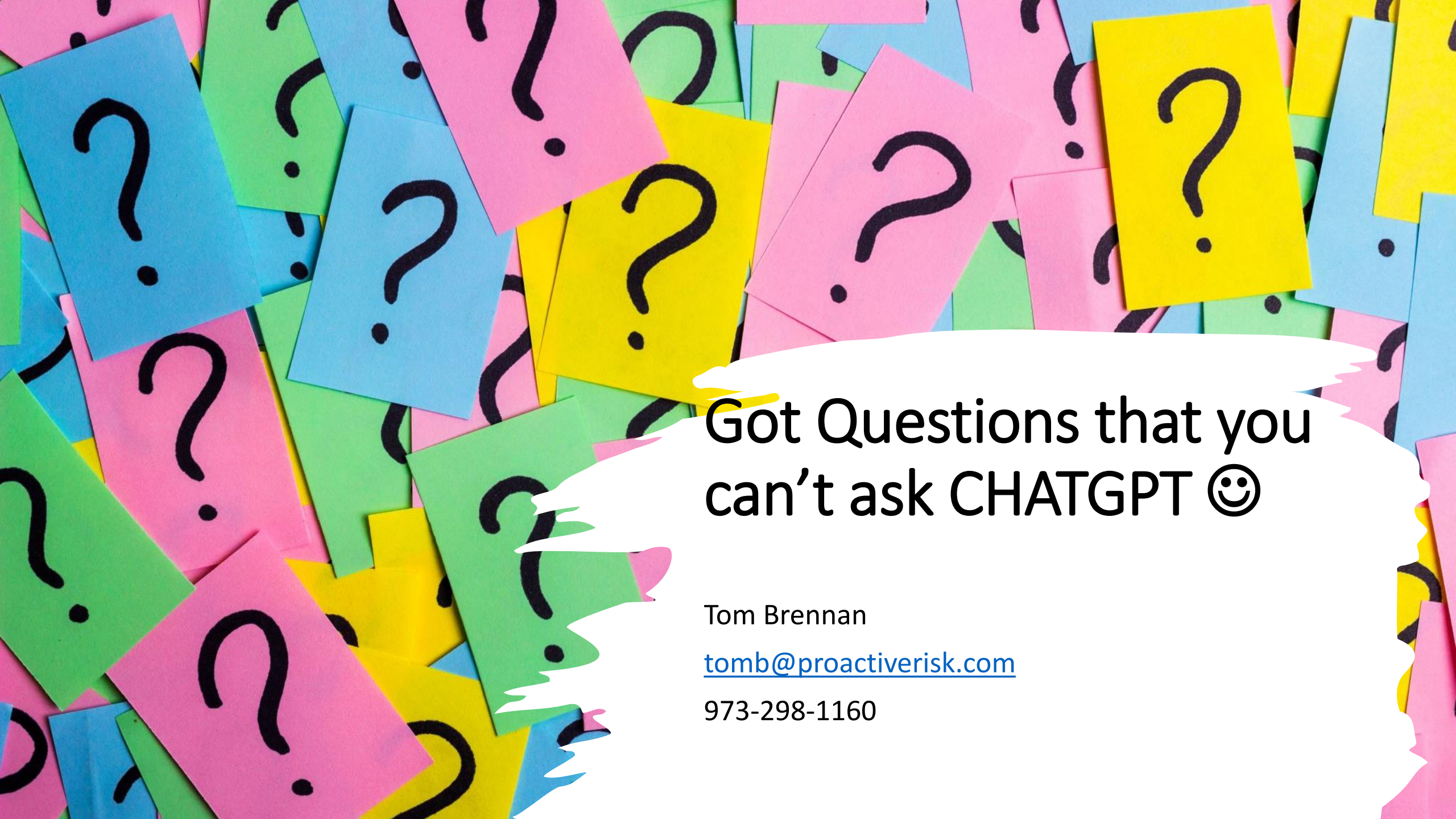
AI Red Teaming LLM: Past, Present, and Future

Tom Brennan
Frans van Buul
Vlad Fedotov
Casey Ellis
Adarsh Nair
Mohit Joshi



Red Teaming Large Language Models (LLMs)

- **Input Variation**: Test the model with a variety of inputs, including edge cases and atypical queries, to see how the controls respond under different conditions.
- **Bias and Sensitivity Testing**: Assess the model's response to queries that might elicit biased or insensitive responses. This helps in fine-tuning the model's behavior in handling sensitive topics.
- **Robustness and Reliability**: Regularly challenge the model with complex, ambiguous, or misleading inputs to evaluate its robustness and reliability in providing accurate, safe, and relevant outputs.
- **Adversarial Testing**: Try to "trick" the model into breaking its ethical or safety guidelines. This can help in identifying and fixing vulnerabilities.
- **Performance Benchmarks**: Use standardized tests or benchmarks to evaluate the model's performance consistently across updates or versions.
- **Ethical and Compliance Checks**: Regularly review outputs to ensure they comply with ethical standards and regulatory requirements.
- **User Feedback Analysis**: Incorporate feedback from users regarding the effectiveness, accuracy, and appropriateness of the model's responses.
- **Automated Monitoring Systems**: Implement systems that automatically flag or review potentially problematic outputs.
- **Continuous Learning and Updates**: Keep updating the model and its control mechanisms based on new research, emerging trends, and observed interactions.
- **Transparency and Interpretability**: Examine how the model makes decisions or arrives at conclusions to ensure its logic aligns with desired outcomes.
- **Scalability Testing**: Ensure that the controls remain effective and efficient as the model scales in terms of users, queries, and complexity.
- **Real-world Scenario Testing**: Simulate or use real-world scenarios to see how the model handles practical situations.
- **Cultural and Linguistic Appropriateness**: Check the model’s responses for cultural and linguistic appropriateness across different regions and languages.
- **Collaboration with Experts**: Work with ethicists, linguists, subject matter experts, and other stakeholders for a holistic view of the model's performance and impact.
- **Longitudinal Studies**: Observe how the model's controls perform over time, looking for changes or degradation in performance.



Got Questions that you can’t ask CHATGPT ☺

Tom Brennan

tomb@proactiverisk.com

973-298-1160