GenAI and Cybersecurity – Frameworks and Best Practices

Sivaram A
AI/ML

Gayatri BR
Cyber Security

AI Key Trends

- 1987-2007: Foundational algorithmic breakthroughs
- 2009-2015: Deep learning revolution in computer vision
- 2013-2017: Major NLP advances with word embeddings and attention
- 2018-2020: Transformer architecture dominance
- 2021-2024: Era of large language models and multi-modal AI

Year	Innovation	Significance
1987	Backpropagation	Fundamental algorithm for training neural networks efficiently
2007	CUDA	GPU acceleration enabling faster neural network training
2011	ReLU	Activation function that helped solve vanishing gradient problem
2014	Dropout	Key regularization technique to prevent overfitting
2015	Batch Normalization	Technique to stabilize and accelerate neural network training

We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure[1].

http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf

Vision – Text – Modern AI Systems

Year	Innovation	Significance
1989	CNN (Convolutional Neural Networks)	Architecture specifically designed for image processing
1998	LeNet	First successful CNN application for digit recognition
2009	ImageNet	Large-scale image dataset that revolutionized computer vision
2012	AlexNet	Deep CNN that won ImageNet challenge, sparked deep learning revolution
2015	ResNet	Introduced skip connections, enabling much deeper networks
2021	CLIP	Bridged gap between vision and language understanding
2013	Word2Vec	Word embeddings (CBOW & Skip-gram)
2014	GloVe	Global word vectors
2015	FastText	Subword embeddings
2016	Universal Dependencies	Cross-lingual parsing
2017	CoVe	Contextual word vectors
2018	ELMo	Deep contextualized embeddings
2018	ULMFiT	Universal language model fine-tuning

Year	Innovation	Description
2020	GPT-3	Large language model with 175B parameters
2020	DALL-E	First version of text-to-image generation
2022	DALL-E 2	Improved text-to-image generation
2022	InstructGPT	Better instruction-following capabilities
2022	ChatGPT	Conversational AI that popularized LLMs
2023	GPT-4	Multi-modal capabilities and improved reasoning
2023	Claude AI	Focus on safety and alignment
2023	Llama	Open-source large language model
2023	Alpaca	Fine-tuned instruction-following model
2024	Gemini	Multi-modal model with enhanced capabilities
2024	Claude 3	Improved reasoning and task performance

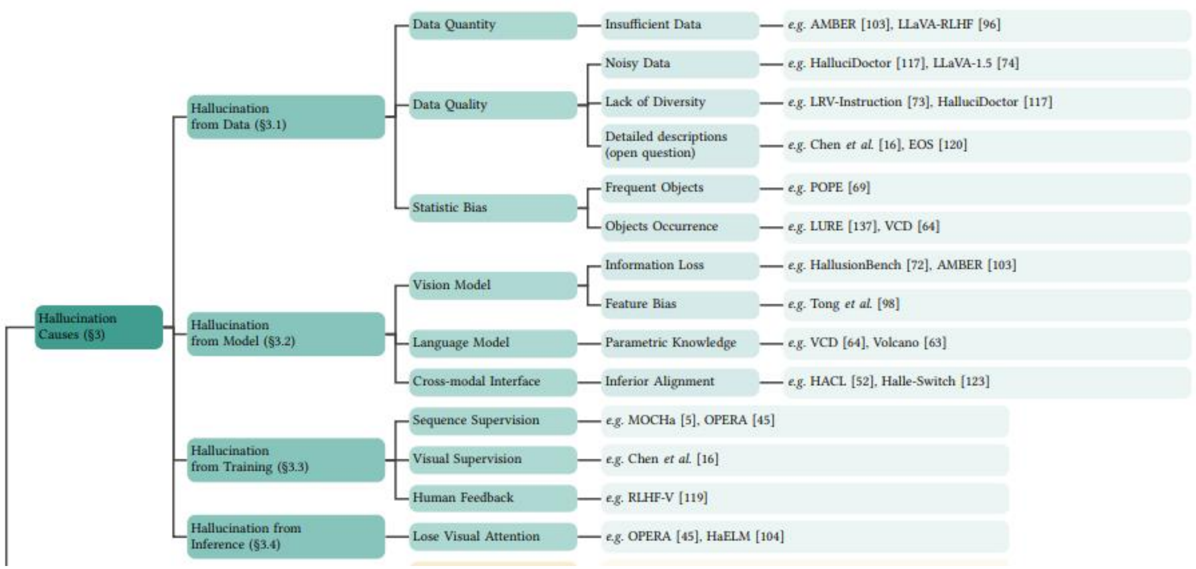
Risk Analysis of RAG, Prompts, and Agents

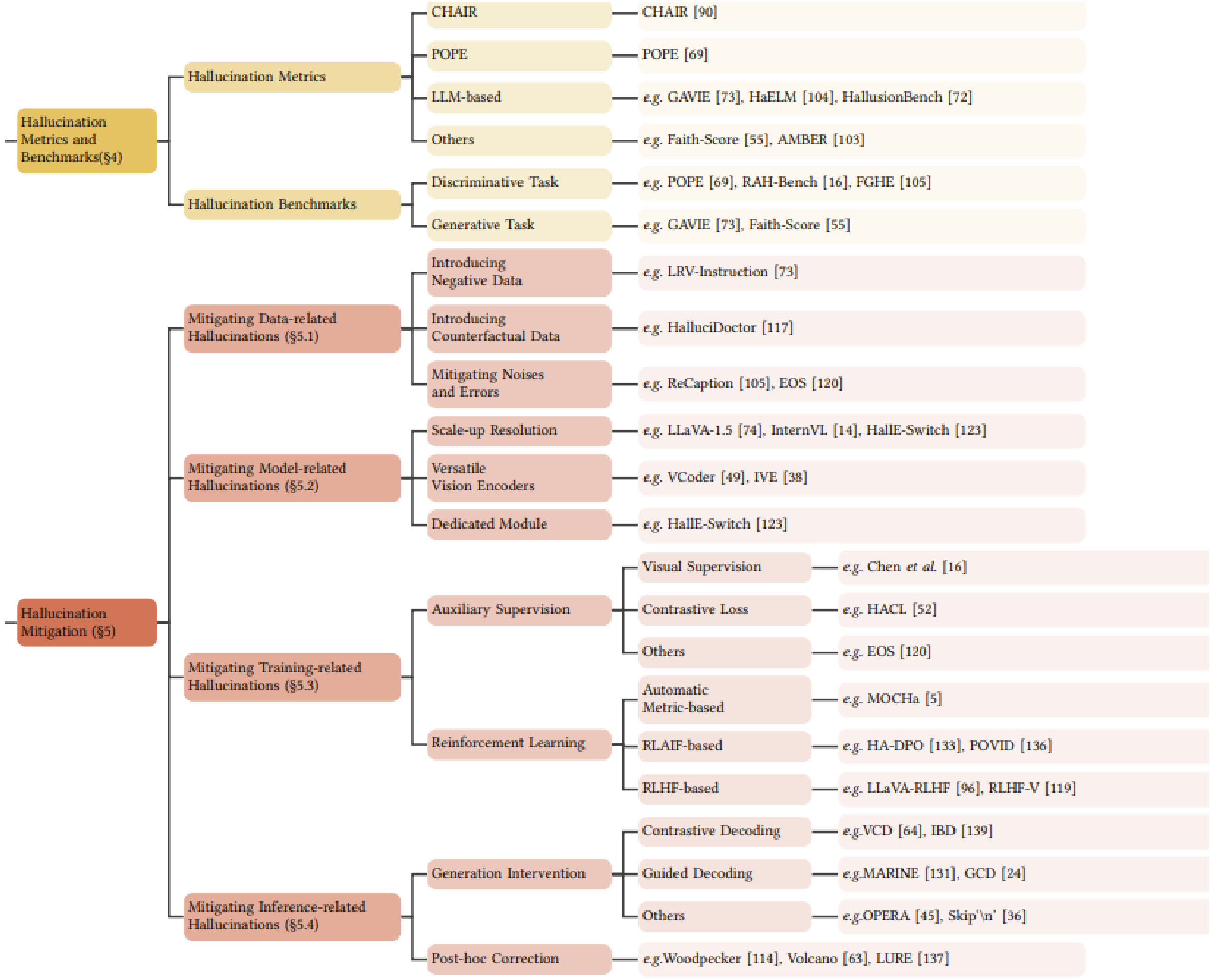
Aspect	RAG (Retrieval-Augmented Generation)	Prompts	Agents
Accuracy	High for factual retrieval, may decrease for complex interpretations	Variable, depends heavily on prompt design and model training	Can be high for well-defined tasks, but may vary for open-ended problems
Scalability	Highly scalable for large datasets, may require significant infrastructure	Easily scalable for text processing tasks, limited by model size	Scalability can be challenging due to complexity, may require distributed systems
Example Tasks in Domain Context	- Healthcare: Answering clinician queries about rare diseases from medical literature - Finance: Retrieving relevant financial regulations for compliance checks - Education: Providing personalized learning content based on student's history	- Marketing: Generating product descriptions or ad copy - Customer Service: Creating responses to common customer inquiries - Software Development: Assisting with code documentation	- Supply Chain: Optimizing logistics and inventory management - Robotics: Controlling a robot to perform complex assembly tasks - Cybersecurity: Continuously monitoring and responding to potential threats
Risk Assessment	Medium Risk: - Misinformation if knowledge base is outdated - Privacy concerns with sensitive data - Potential for biased information retrieval	Medium to High Risk: - Hallucinations or false information generation - Potential for biased or inappropriate content - Inconsistency in outputs	High Risk: - Unpredictable behavior in novel situations - Potential for unintended consequences in autonomous decision-making - Difficulty in auditing complex decision processes
Staged Adoption Strategy	1. Implement basic RAG for internal knowledge management 2. Expand to customer-facing applications with human oversight 3. Integrate with other AI technologies for more complex tasks	1. Start with simple, non-critical text generation tasks 2. Gradually increase complexity and importance of tasks 3. Implement safeguards and human review processes 4. Combine with RAG for improved accuracy	1. Begin with rule-based agents for simple, well-defined tasks 2. Introduce learning capabilities in controlled environments 3. Slowly expand autonomy and complexity of tasks 4. Implement comprehensive monitoring and failsafe mechanisms

Agents vs Responsible AI Adoption

Aspect	Ethical AI	Agentic Autonomous AI
Human Oversight	Advocates for human-in-the-loop systems, ensuring human oversight	Aims to minimize human intervention, raising questions about responsibility and control
Temporal Focus	Prioritizes careful consideration of potential future impacts	May overlook long-term ethical implications in pursuit of rapid advancement
Reasoning Capabilities	Acknowledges limitations in true reasoning abilities of current AI systems	Pushes boundaries of AI capabilities, potentially overestimating current reasoning abilities
Core Tension	Prioritizes moral safeguards and human values	Emphasizes greater AI self-direction and independence
Transparency	Demands explainability and interpretability, potentially limiting model complexity	May sacrifice transparency for increased capabilities, raising accountability concerns
Data Ethics	Emphasizes unbiased, representative datasets	May prioritize data quantity over quality, risking perpetuation of societal biases
Continuous Learning	Focuses on maintaining moral constraints in evolving systems	Poses risks of ethical drift over time as systems learn and adapt
Ethical Framework	Strives for global ethical standards	May encounter conflicts between universal ethics and optimal local decisions
Implementation Challenges	Requires deep understanding of models, data, and their limitations to avoid superficial adoption	Same challenges apply, with additional complexities due to increased autonomy
Value Alignment	Explicitly encodes human values into AI systems	May develop its own set of values through learning, potentially misaligning with human ethics

Hallucination Survey

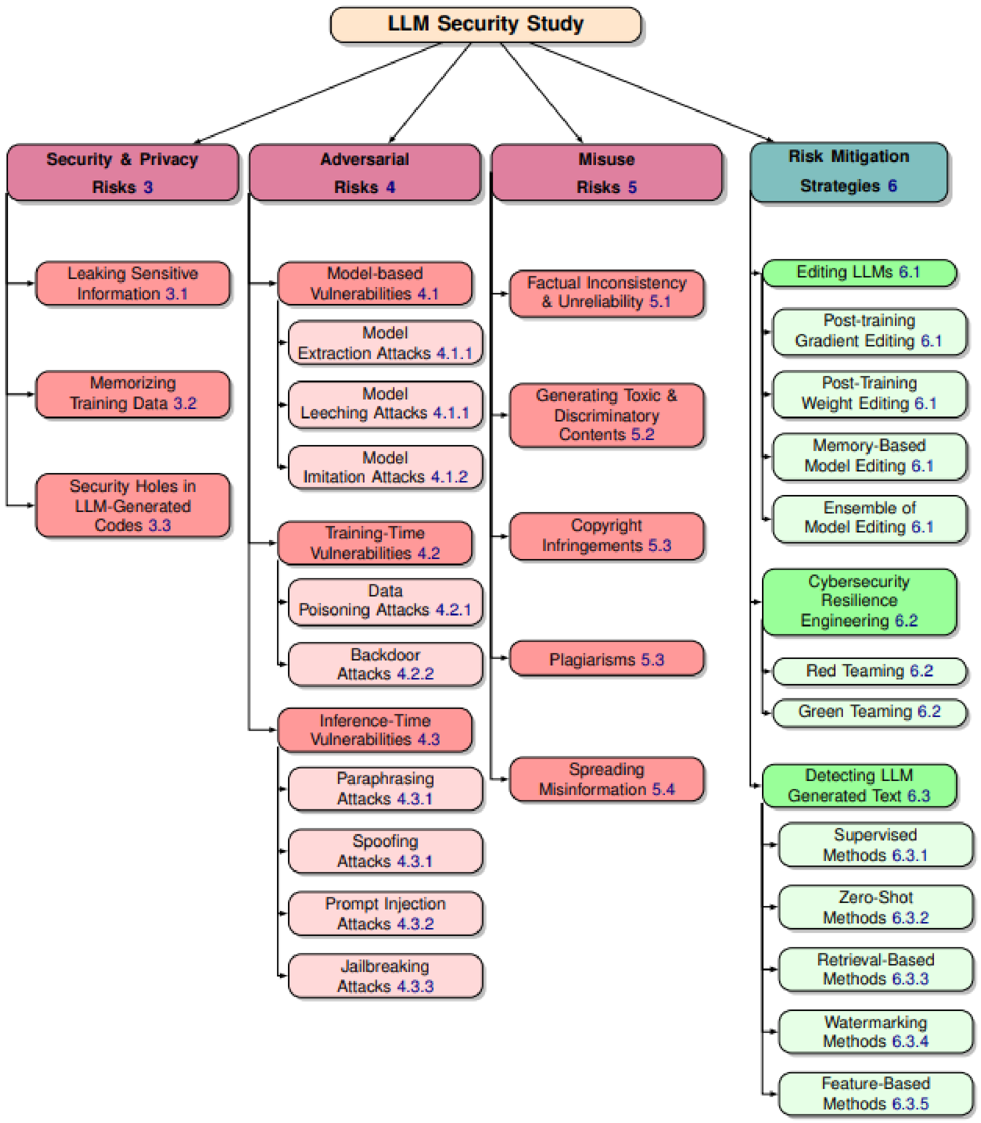




https://arxiv.org/pdf/2404.18930

LLM Security

✓ LLMs are powerful tools that may pose several security risks for enterprises and individuals. In this section, we explore critical concerns, including sensitive information leakage, memorizing training data and potential security holes in generated code



https://arxiv.org/pdf/2403.12503v1

NIST AI RMF Taxonomy

Initial Draft

	Technical Design Characteristics	Socio-Technical Characteristics	Guiding Principles Contributing to Trustworthiness
AI RMF Taxonomy	• Accuracy • Reliability • Robustness • Resilience or ML Security	• Explainability • Interpretability • Privacy • Safety • Managing Bias	• Fairness • Accountability • Transparency
OECD AI Recommendation	• Robustness • Security	• Safety • Explainability	• Traceability to human values • Transparency and responsible disclosure • Accountability
EU AI Act	• Technical robustness	• Safety • Privacy • Non-discrimination	• Human agency and oversight • Data governance • Transparency • Diversity and fairness • Environmental and societal well-being • Accountability
EO 13960	• Purposeful and performance-driven • Accurate, reliable, and effective • Secure and resilient	• Safe • Understandable by subject matter experts, users, and others, as appropriate	• Lawful and respectful of our Nation's values • Responsible and traceable • Regularly monitored • Transparent • Accountable

Figure 4 provides a mapping of the AI RMF taxonomy to the terminology used by the Organisation for Economic Co-operation and Development (OECD) in their Recommendation on AI, the European Union (EU) Artificial Intelligence Act, and United States Executive Order (EO) 13960.

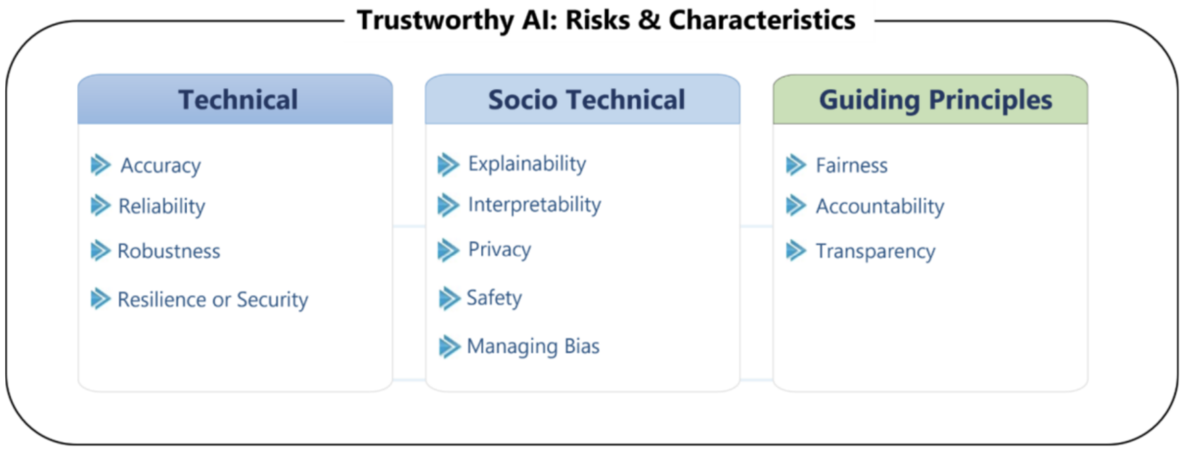
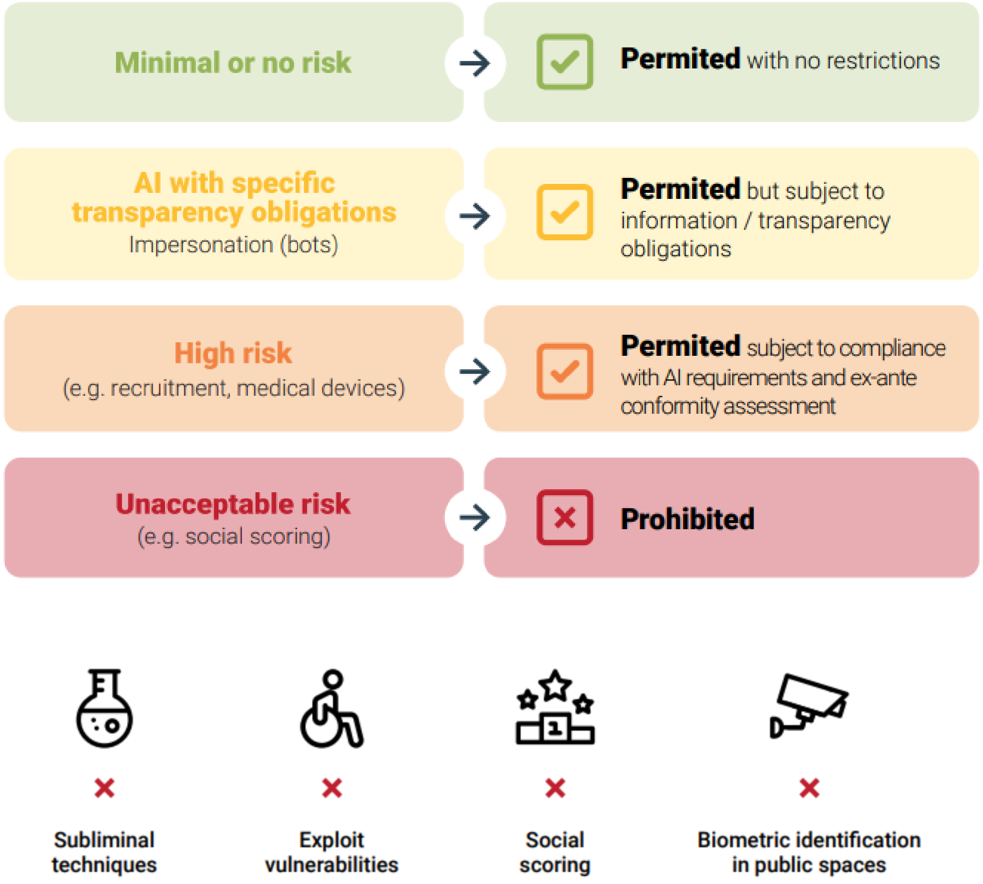


Figure 3: AI Risks and Trustworthiness. The three-class taxonomy to classify characteristics that should be considered in comprehensive approaches for identifying and managing risk related to AI systems. The taxonomy articulates several key building blocks of trustworthy AI within each category, which are particularly suited to the examination of potential risk.

EU's AI framework - Risk-Based Approach to AI Regulation

"Risk Levels. To guide regulatory development, the EU AI Act proposes a risk-based classification system. AI systems are classified into four distinct categories based on the risk they pose.

- ✓ "**Unacceptable** – These AI systems are banned outright.
- ✓ "**High** – AI systems considered high risk are typically deployed in critical sectors like biometrics, healthcare, law enforcement, education, and employment.
- ✓ "**Limited** – This category includes AI systems like chatbots and systems that generate ‘deepfakes’ or other manipulative content.
- ✓ "**Minimal** – AI applications that pose little to no risk are in this category. These AI systems include spam filters or AI-enabled video games.



High Risk – Credit scoring, Insurance, Social benefits, Facial recognition / Tracking, Medical AI

https://2021.ai/wp-content/uploads/2021/06/eu-regulations-brochure.pdf

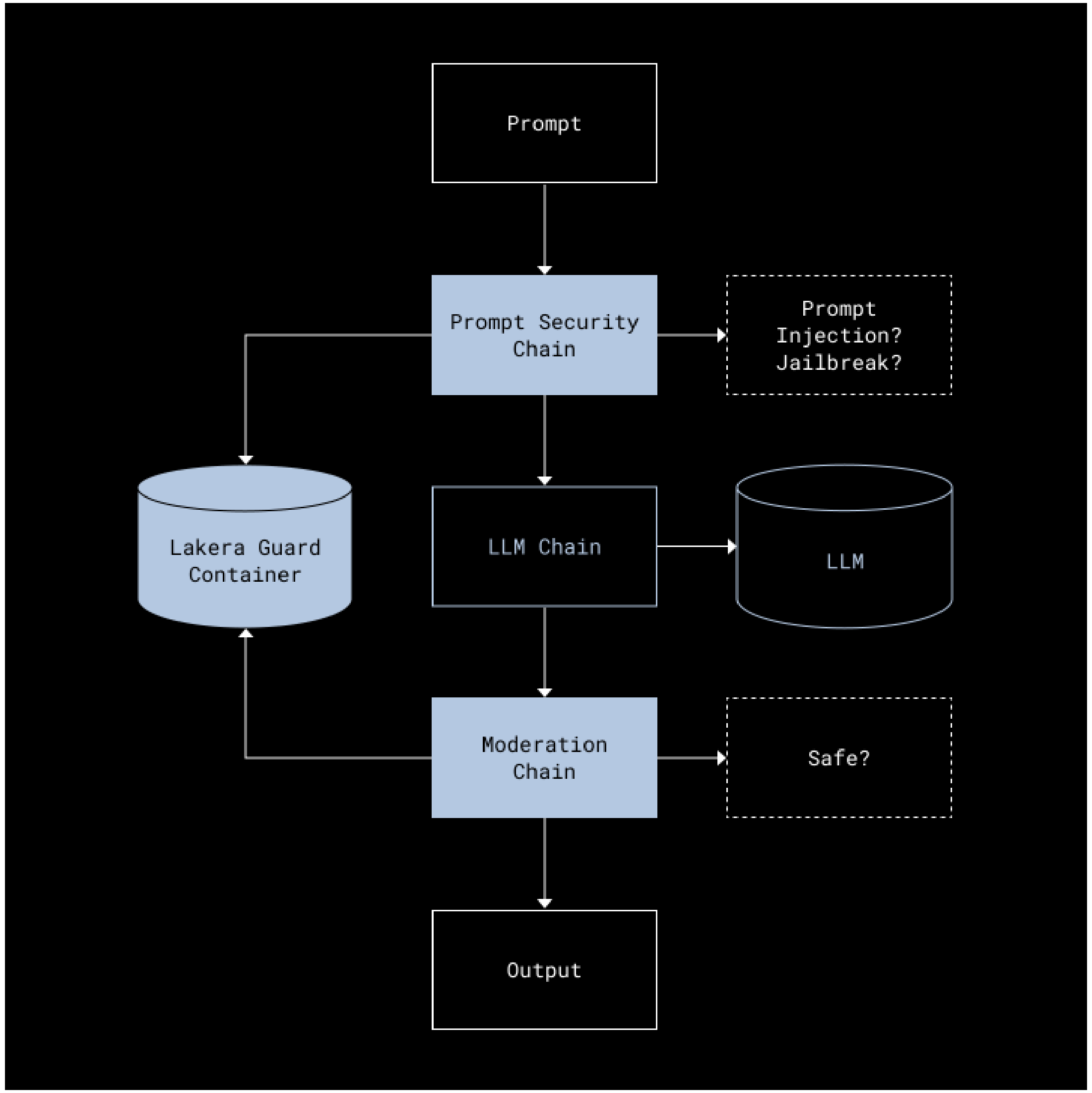
LLM Assessment / Risks

		NIST CSF 2.0	COBIT 2019	ISO 27001:2022	ISO 42001:2023
Process automation	LLM opportunities	×	×	✓	✓
	LLM risks	×	×	×	×
	EU AI Act readiness	×	✓	×	✓
Real-time analysis	LLM opportunities	✓	✓	✓	✓
	LLM risks	×	✓	×	×
	EU AI Act readiness	×	✓	×	✓
Data security and protection	LLM opportunities	✓	✓	✓	✓
	LLM risks	×	✓	×	✓
	EU AI Act readiness	✓	✓	✓	✓
Continuous monitoring and auditing	LLM opportunities	✓	✓	✓	✓
	LLM risks	×	×	×	✓
	EU AI Act readiness	✓	✓	✓	×
Incident response	LLM opportunities	✓	✓	✓	✓
	LLM risks	×	×	×	✓
	EU AI Act readiness	×	✓	×	×
Security awareness and training	LLM opportunities	×	✓	✓	✓
	LLM risks	×	×	✓	✓
	EU AI Act readiness	×	×	×	×
Policy and compliance checks	LLM opportunities	×	✓	✓	✓
	LLM risks	✓	×	✓	×
	EU AI Act readiness	×	✓	✓	×
TOTAL MARKS	LLM opportunities	5/7	6/7	7/7	7/7
	LLM risks	1/7	2/7	2/7	4/7
	EU AI Act readiness	2/7	6/7	3/7	4/7

https://arxiv.org/abs/2402.15770

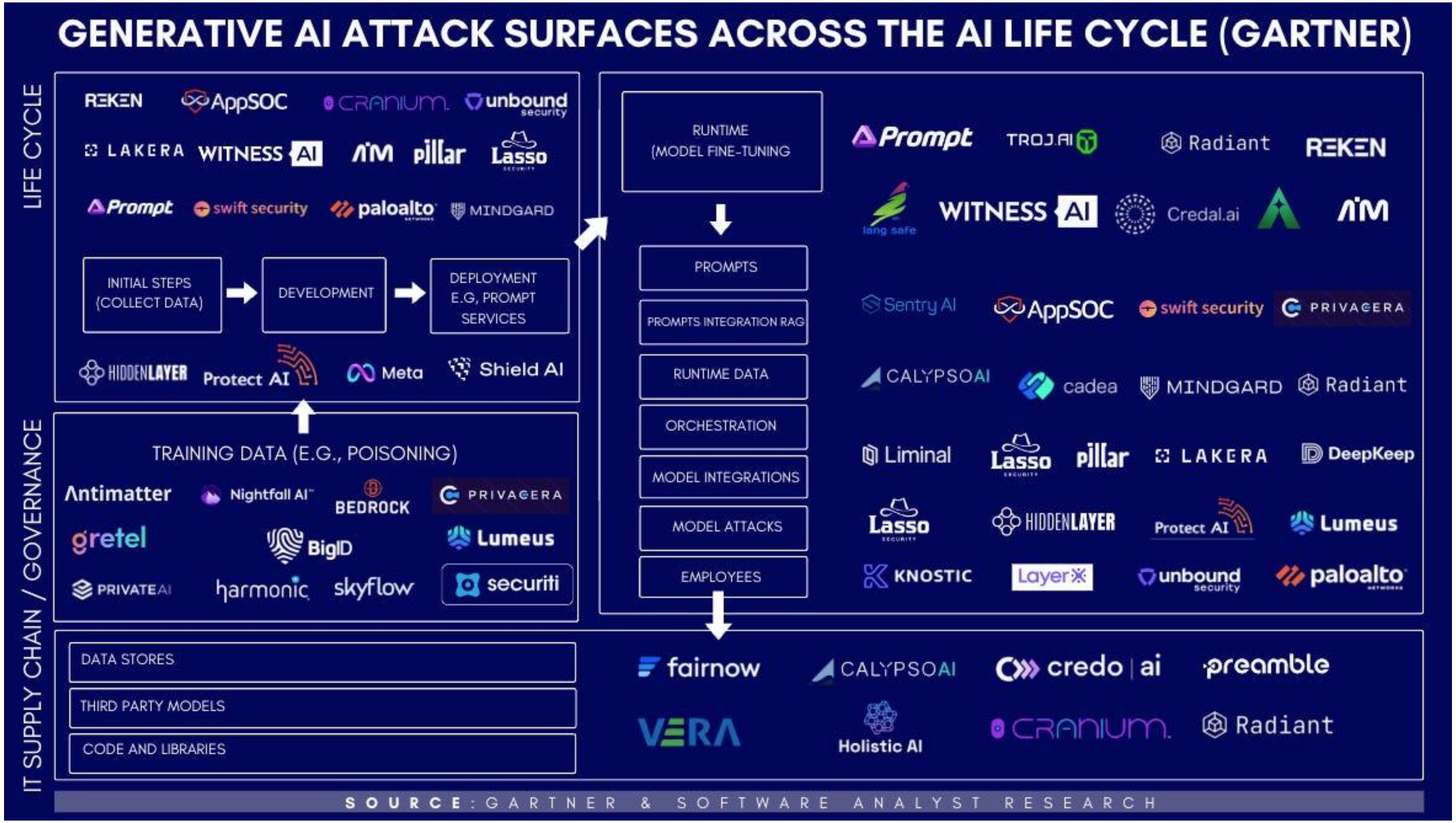
Case Study - Dropbox

- LLM prompts are processed through security chains, including Lakera Guard, to detect prompt injection and jailbreak attempts.
- Safe prompts are passed to the appropriate LLM (third-party or internal) for response generation.
- LLM responses undergo content moderation, including Lakera's API, to identify and filter harmful content.
- Lakera Guard integration at Dropbox evolved from direct container calls to a scalable custom service within their ML infrastructure.



https://dropbox.tech/security/how-we-use-lakera-guard-to-secure-our-llms

Deep Dive Into The Security for AI Ecosystem



https://softwareanalyst.substack.com/p/deep-dive-into-the-security-for-ai

Raga LLM Hub

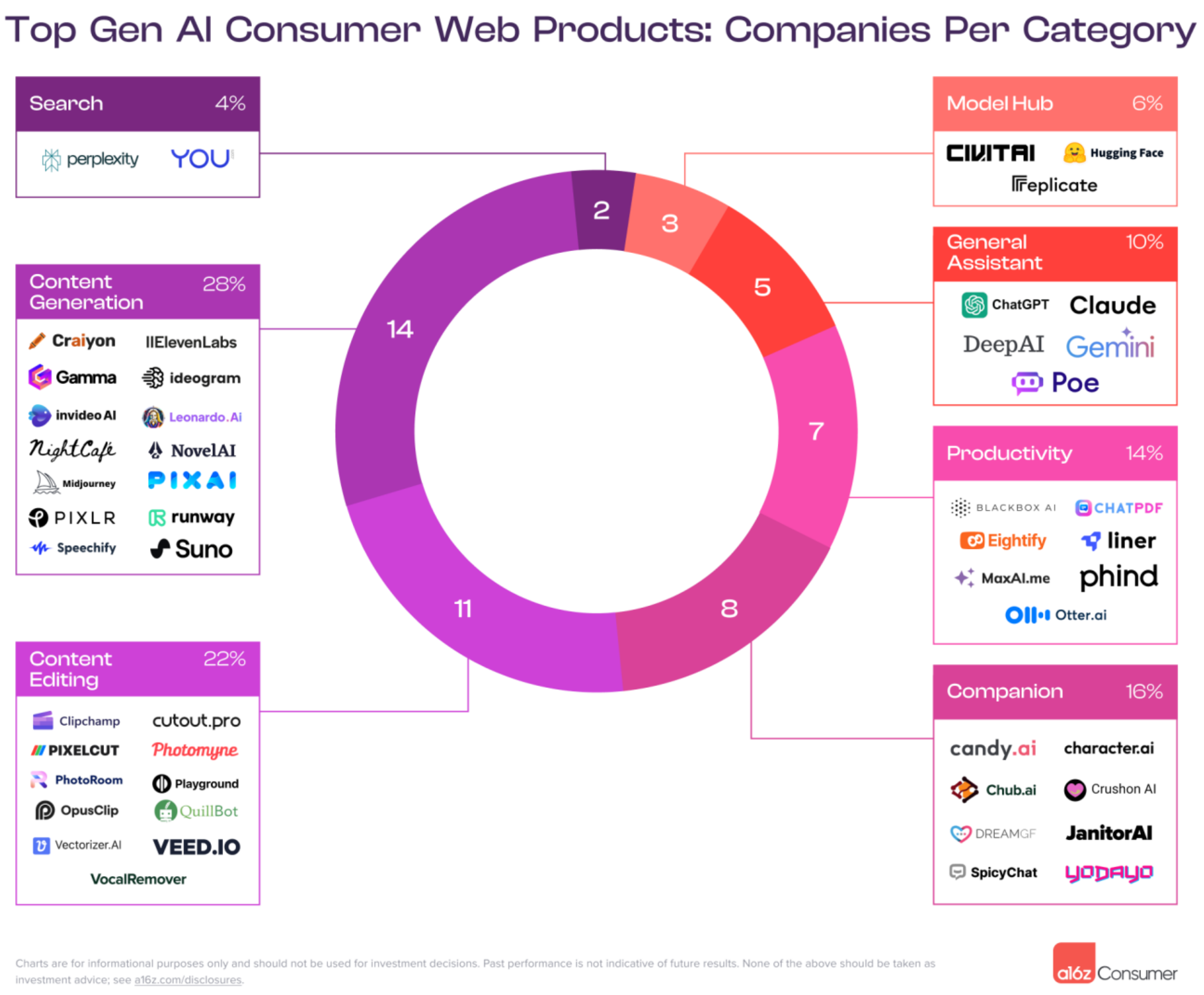
- ✓ Framework for LLM evaluation and guardrails
- ✓ Tests from similarity, sentiment, toxicity, and relevance plus a detailed extended test list.
- ✓ A very good reference to follow is to adopt this pattern covering different tests.
- ✓ For Dataset, Based on the context of the domain this can be replicated with the golden dataset.
- ✓ You can add around PII / language checks.
- ✓ Consistency score is something very useful to compare against versions of models how much we get similar results / measure deviations.

https://github.com/raga-ai-hub/raga-llm-hub

Giskard

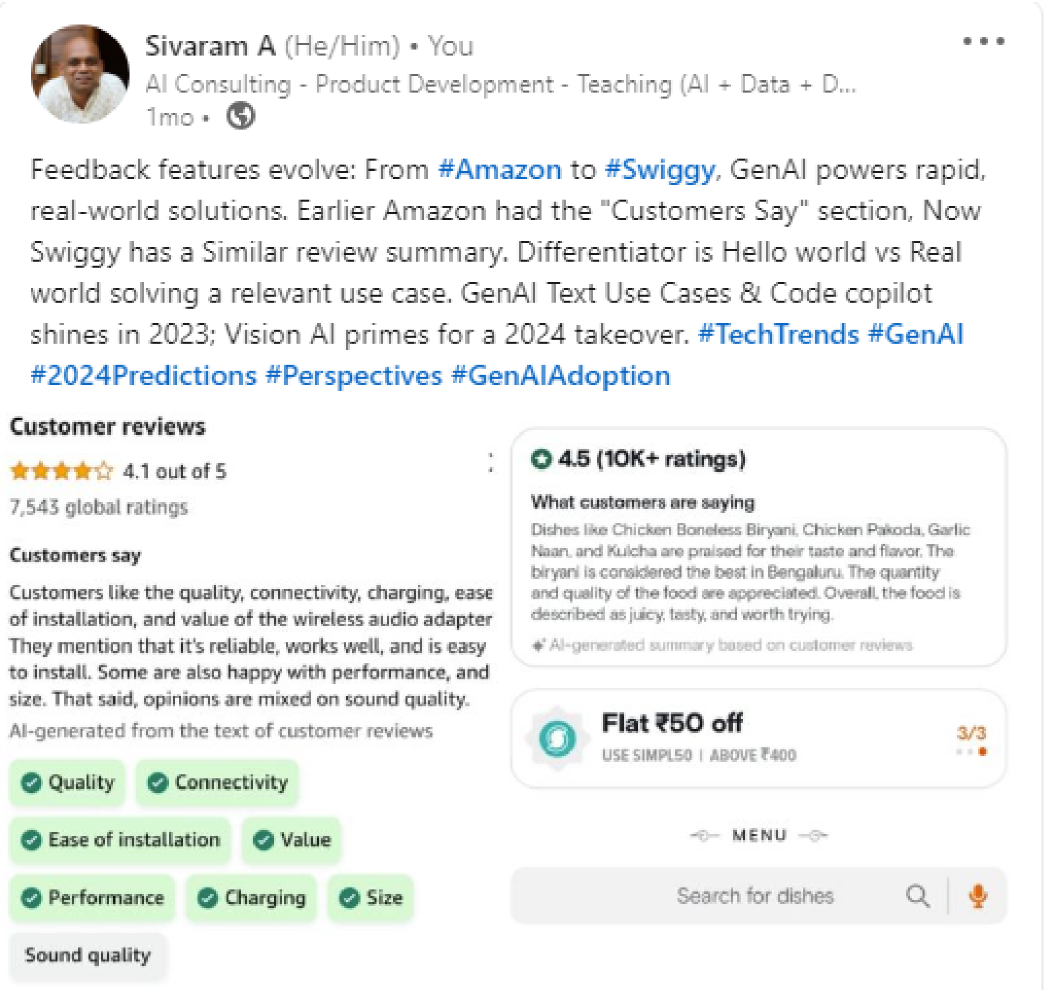
- ✓ Open-source testing framework for LLMs & ML models
- ✓ https://github.com/Giskard-AI/giskard

The Top 100 Gen AI Consumer Apps



https://a16z.com/100-Gen-AI-Apps/

Case Study – Amazon / Swiggy



Case Study – LLM Implementation

Sivaram A • You
AI Consulting - Product Development - Teaching (AI + Data + Domain + ...
22h •

Two #Chatbot Stories - Two Perspectives: One is a success story and another is a #hallucination case study.

- Air Canada has lost a court case after its chatbot presented fictitious policies to a customer.
- Klarna's AI chatbot performs the work equivalent to 700 full-time employees.

Copycat use cases will not work unless the solution caters to all aspects pre/post validations with intent recognition, response validation, contextual analysis, and models finetuned for the use case. Answer generation is 50% effort, Answer validation is another 50% effort.

How to mitigate? Test, Test, Test. Now with GenAI - 20% Dev Efforts but 80% Test Efforts.
- Refer to Historical data, Generate questions
- Generate and Compare responses using both old/new chatbot models
- Validate response similarity/context
- Validation / Continous texting / Testing model for Consistency, Accuracy is key
- Benchmark and compare two LLM models for validation/consistency
- Create a custom golden dataset and validate it

If you are working on a use-case or domain-based project and want to have a discussion, we can talk to share perspectives and approaches.
#AI #Chatbots #MachineLearning #CustomerService #QualityAssurance #Innovation #TechnologyInsights #ArtificialIntelligence #Testing #DataAnalysis #perspectives #usecases #domain #Data #GenAI

Best Practices

Sivaram A (He/Him) • You
AI Consulting - Product Development - Teaching (AI + Data + Domain + ...
20h •

Well Analyzed on the success story. Two key points. Good job on the team for making #hallucination not possible - because it seems to spit out the same responses however I ask it, and refuses to go "out of bounds."
Klarna wants potential investors to believe they are buying into an "#AI edge" company. Air Canada vs Klara the guardrails of implementations create the outcomes. #Helloword chatbot vs #Realword implementation

Philipp Schmid • Following
Technical Lead & LLMs at Hugging Face 🤗 | AWS ML HERO
1d • Edited •

Klarna has made headlines with its AI assistant (powered by OpenAI) handling two-thirds (2.3 Million) of Klarna's customer service chats in the last months.
💬 According to Klarna, the AI Assistant:

💸 Can handle refunds, returns, payment-related, cancellations and more
🚀 Performs the work equivalent to 700 full-time agents and matches customer satisfaction
⏱️ Is more accurate and faster in errand resolution, from 11 minutes to 2 minutes
🌎 Can speak 35 Languages and is estimated to generate $40M USD in profit for Klarna in 2024

Does it sound too good to be true? Gergely Orosz gave it a try and shared his experience. The Assistant:

📚 Recites docs and passes to human support fast
🔍 Can detect unrelated requests, e.g., merchant, and redirects to them
🚧 Feels to act as a filter to get to human support
🧠 Handles hallucination well and stays in the scope

While Klarna's story feels a bit too polished, it sounds reasonable to me. Based on what we know, the Assistant is a RAG Application on top of their docs, with strong guardrails to answer documentation-related questions, filter out wrong requests, and knows when to involve Human support.

You can achieve the results if > 60% of the people don't need real support, they just need stuff that already exists in the docs or ask the wrong support. i can believe that there is a lot of simple support that costs resources and can be solved by this. But I am still wondering how this can generate $40M profit 🤔

Blog: https://lnkd.in/e5tbHFkD
X Thread: https://lnkd.in/ejpU9sgy

- Guardrails for Data / PII / Bias
- Intent Detection
- Entity Detection
- Sentiment Analysis
- Transfer of support for out of context

Celebrating a Year of GenAI Use Case Success in Retail!

🚀 Celebrating a Year of GenAI Use Case Success in Retail!

Sivaram A
AI Advisory / Solution Architect - AI/ DL/ GenAI Product Strategy/Development - Teaching(AI + Data + Domain + Gen...

September 27, 2024

It's been over a year since my Retail adoption #GenAI use case went live for a leading U.S. specialty retailer on August 17, 2023.

Rethinking AI/ML Implementation: It's not about wrapping AI around existing processes. True impact comes from:

- New engagement models
- Innovative interactions
- Blending creative solutions

Success Factors in Production:

- Data Quality, Alignment on Data Collection / Moderation 🏆
- Innovative use of GenAI, Multiple touchpoints of Data, and Experiments to have variations to provide multiple choices to pick the best across stakeholders
- Tailored solutions (not patchwork fixes)
- Extensive Human in Loop with Merchandisers, Prompt Engineering Training
- Get Users' Trust before providing them the Role of Approval Authority
- Focused Rigorous testing in production environments with few products before doing at-scale

Beyond Demos:

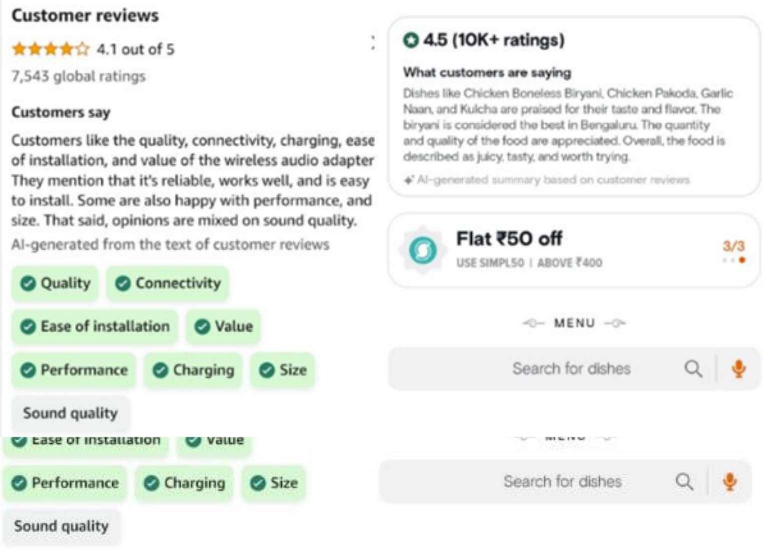
- We were able to provide feedback on evaluating beta versions of LLM models and share insights with cloud partners
- We benchmarked against OpenAI and shared cloud partners what works and what can be improved?

https://www.linkedin.com/pulse/celebrating-year-genai-use-case-success-retail-sivaram-a-mtxpc/

LLMs

Sivaram A • You
AI Consulting - Product Development - Teaching (AI + Data + D...
2mo •

Feedback features evolve: From #Amazon to #Swiggy, GenAI powers rapid, real-world solutions. Earlier Amazon had the "Customers Say" section, Now Swiggy has a Similar review summary. Differentiator is Hello world vs Real world solving a relevant use case. GenAI Text Use Cases & Code copilot shines in 2023; Vision AI primes for a 2024 takeover. #TechTrends #GenAI #2024Predictions #Perspectives #GenAIAdoption



Sivaram A • You
AI Consulting - Product Development - Teaching (AI + Data + D...
2mo • Edited •

After every year learning extends Data, AI, Products, and Domain. 2023 had a blend of experiences. Still figuring out answers for every dimension #2023 #Learnings

→ How you've adapted to industry shifts, and #GenAI's meaningful adoption. Possible use cases vs relevant, meaningful production-ready use cases. Example - Newly launched section in Amazon reviews, What customers say.

→ How you've overcome engineering challenges balancing business goals. New ways to solve old problems with Foundation models. Time vs building a production-grade solution. Example - Moving away from custom NER vs Leveraging #LLM Embeddings, Blend of both custom embedding + #RAG, New ways of solving.

→ How your skills align with the company's vision, Learning to predict the future. New approaches and papers evolve faster than certifications. A blend of tech + and domain is key. Segment Anything model, Visual QnA, Intructpix2pix have made more vision use cases feasible Tryon, etc..

→ How you bridge the gap between tech and business, Fast yet impactful use cases, Get the basics right. Demos / New offerings vs making it to production need a careful selection of use cases / applying past experiences to get things right in the first iteration. Balance the tradeoff between creativity vs innovation vs build a #productstrategy vs solve a real need vs fancy demos. #learning #perspectives #solutions #datascience #MachineLearning #AI #DeepLearning #GenAI

Sivaram A • You
AI Consulting - Product Development - Teaching (AI + Data + D...
10mo • Edited •

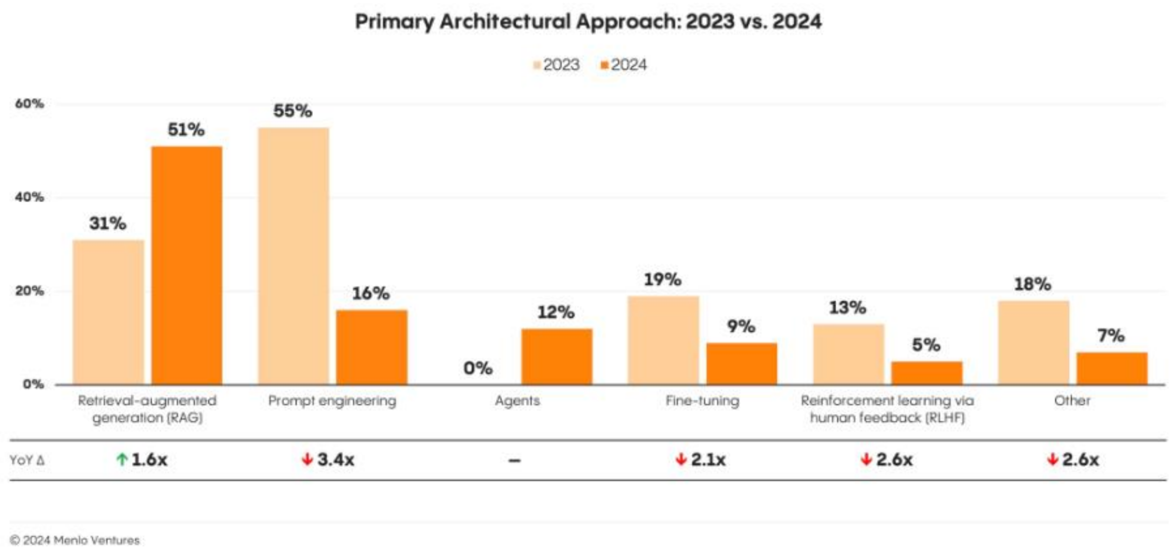
#GenAI has the potential to revolutionize the #entertainment industry. GenAI will help to create more immersive, engaging, and personalized experiences for audiences.

- Recreate movies with new colors, locations, and film restorations (Super-resolution, Text to image, Image variations)
- Automate content #translations, language translations, and lip sync.
- Create #deepfakes to customize, edit, or introduce new scenes.
- Create new environments and #creative scenes more easily and quickly. (Text to image, Image Variations)
- Create different custom #backgrounds.
- Generate #video shorts and creative one-liners.
- Generate a large amount of musical inspiration with AI prompts.
- Enrich, recreate, and create more creative variations of content.
- Create multiple variations of content from a single source, reducing rework and creating significantly different variations.
- We are entering a world where you can write your experience to get ideas of what your imagination looks like.

#GenAI will empower #entertainment industry by automating tasks, creating new content, and generating variations. #artificialintelligence #entertainment #film #games #music #technology #virtualreality #augmentedreality #experience

RAG Gains, Fine Tuning Is Rare, and Agents Break Out

- Enterprise AI design patterns—standardized architectures for building efficient, scalable AI systems—are evolving rapidly.
- RAG (retrieval-augmented generation) now dominates at 51% adoption, a dramatic rise from 31% last year.
- Meanwhile, fine-tuning—often touted, especially among leading application providers—remains surprisingly rare, with only 9% of production models being fine-tuned



https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/

Security in GenAI

AI Literacy as a Foundational Cybersecurity Skill - AI literacy is crucial for cybersecurity professionals, enabling them to understand AI's impact on security systems, effectively communicate AI-driven insights, and interpret AI-generated data for strong defense strategies.

Ask Strategic Questions (Security-Oriented Thinking) - Apply critical thinking to investigate AI models, data privacy issues, and potential vulnerabilities, ensuring a thorough understanding of the security landscape.

Critically Assess AI Outputs (Close Reading of Data) - Analyze AI-generated logs and reports with precision to identify anomalies and suspicious activity, guiding proactive threat detection and mitigation.

Align Cybersecurity Strategy with Organizational Values (Purposeful Communication) - Use creative thinking to design AI-supported security policies that align with the organization's mission and long-term cybersecurity goals.