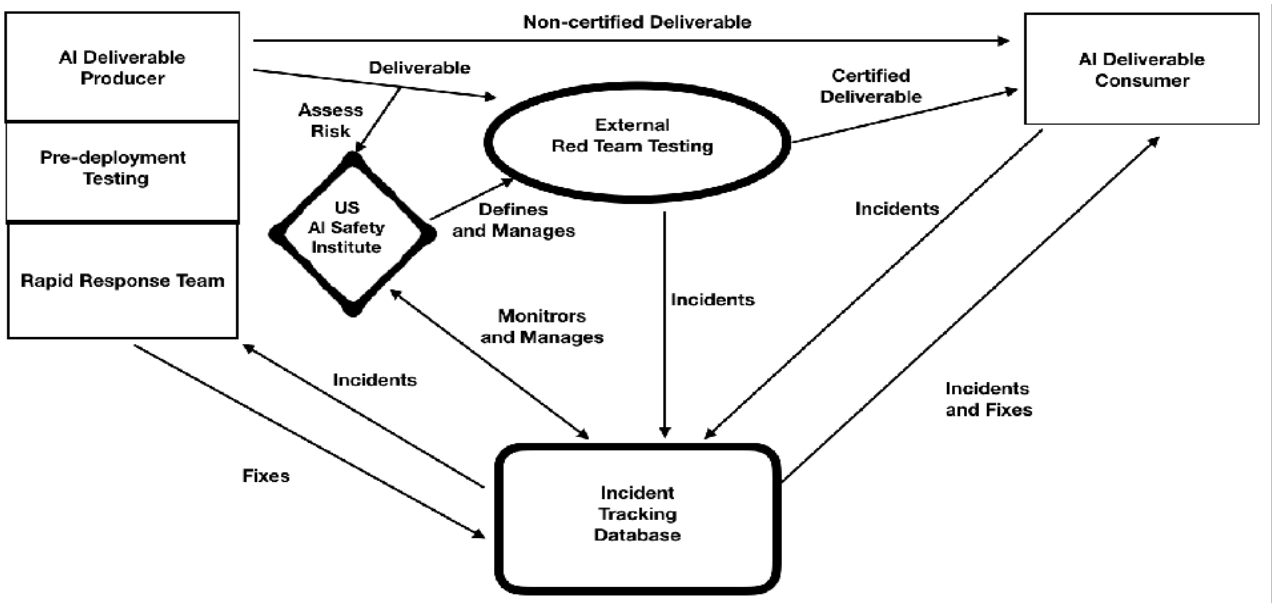
Response to NIST's RFI to "Support Safe, Secure and Trustworthy Development and Use of AI": (Role of US AI Safety Institute)

by Bob Marcus (robert.marcus@gmail.com)

The diagram below illustrates a possible role for the USAISI in enabling trustworthy AI.
See Table of Contents on page 2 for a detailed discussion outline.



The AI Safety Institute should encourage compliance with NIST's AI Risk Management Framework by AI Deliverable Producers. However this will not be sufficient to enable Trustworthy AI due to the complexity and diverse use cases of many Generative AI deliverables. Deliverables can be many outputs in the Generative AI Delivery Process including data sources, foundation models, fine tuned packages, deployed applications, and application output. .

There are two other components that will be essential for increasing reliability in AI applications. These are **External Red Team Testing** for risky deliverables and an **Incident Tracking Database** for problems discovered in testing and use of AI software. Both of these will be necessary for the AI Deliverable Consumers to have confidence that the deliverables have been thoroughly tested and that problems are being addressed and fixed. The US AI Safety Institute can play a role in initiating, monitoring, and managing these components.

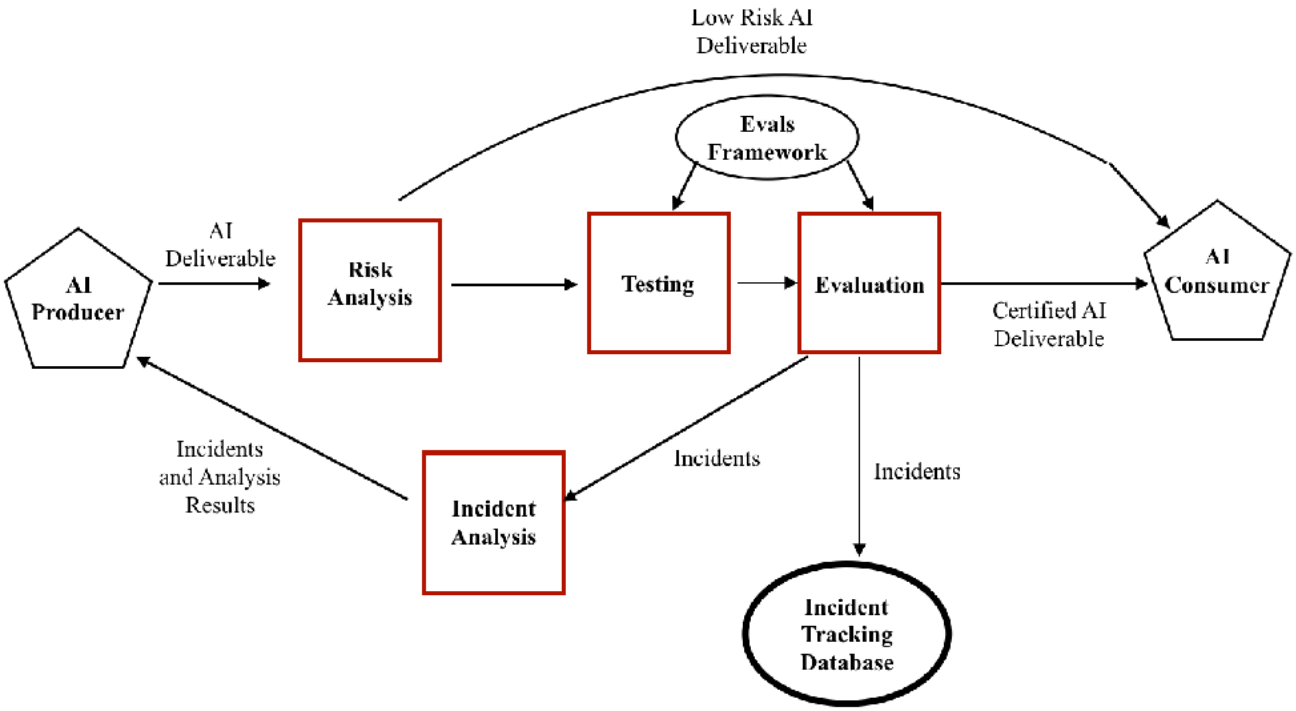
The risk associated with AI deliverables should be evaluated by a risk evaluation team. High risk deliverables should be subjected to External Red Team Testing to uncover possible problems (i.e. incidents). The extent of testing should depend on the risk associated with the deliverable. Deliverables that pass testing can be certified to increase Consumer confidence. Low risk applications can bypass the External Red Team Testing. Incidents discovered by the External Red Team or the AI Consumer should be added to the Incident Tracking Database and reported to the AI Producer. Incident fixes when available should be added to the Incident Tracking Database.

Table of Contents

1. External Red Team Testing 2
   1.1 Risk Analysis 3
   1.2 Pre-Defined Evals and Testing Frameworks 4
   1.3 Lower Risk: Generic Applications and Use Cases 5
   1.4 Higher Risk: Domain-specific Applications and Use Cases 8
   1.5 Highest Risk: Applications that Change Environment 9
   1.6 Incident Analysis and Fixes 10
2. Incident Tracking Database 11
   2.1 Incident Tracking Databases 12
   2.2 Generative AI Delivery Process 13
   2.2 LLM Interfaces to Database 13
   2.3 Notifications 14

My comments are in blue.

1. External Red Team Testing



The boxes marked in red in the diagram above are steps in External Red Team Testing. The External Red Teams could generate prompts, evaluate responses, certify, report incidences, and suggest fixes. The individual steps will be discussed below in more detail. The text in italics and quotation marks are from the linked Web site

Generative AI Trust and Governance from Singapore's AI Verify Foundation
https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

"This discussion paper proposes ideas for senior leaders in government and businesses on building an ecosystem for the trusted and responsible adoption of generative AI.. The practical pathways for governance in this paper seek to advance the global discourse and foster greater collaboration to ensure generative AI is used in a safe and responsible manner, and that the most critical outcome — trust — is sustained"

1.1 Risk Analysis

Open AI Preparedness Framework
https://cdn.openai.com/openai-preparedness-framework-beta.pdf

"We believe the scientific study of catastrophic risks from AI has fallen far short of where we need to be.To help address this gap, we are introducing our Preparedness Framework, a living document describing OpenAI's processes to track, evaluate, forecast, and protect against catastrophic risks posed by increasingly powerful models."

OpenAI Preparedness Team
https://openai.com/safety/preparedness

*"**We will establish a dedicated team to oversee technical work and an operational structure for safety decision-making.** The Preparedness team will drive technical work to examine the limits of frontier models capability, run evaluations, and synthesize reports. This technical work is critical to inform OpenAI's decision-making for safe model development and deployment. We are creating a cross-functional Safety AdvisoryGroup to review all reports"*

"We have several safety and policy teams working together to mitigate risks from AI. Our Safety Systems team focuses on mitigating misuse of current models and products like ChatGPT. Super alignment builds foundations for the safety of super intelligent models that we (hope) to have in a more distant future. The Preparedness team maps out the emerging risks of frontier models, and it connects to Safety Systems, Super alignment and our other safety and policy teams across OpenAI."

Risk Taxonomy, Mitigation, and Assessment Benchmarks of LLM Systems
https://arxiv.org/abs/2401.05778

" In this paper, we delve into four essential modules of an LLM system, including an input module for receiving prompts, a language model trained on extensive corpora, a toolchain module for development and deployment, and an output module for exporting LLM-generated content. Based on this, we propose a comprehensive taxonomy, which systematically analyzes potential risks associated with each module of an LLM system and discusses the corresponding mitigation strategies.

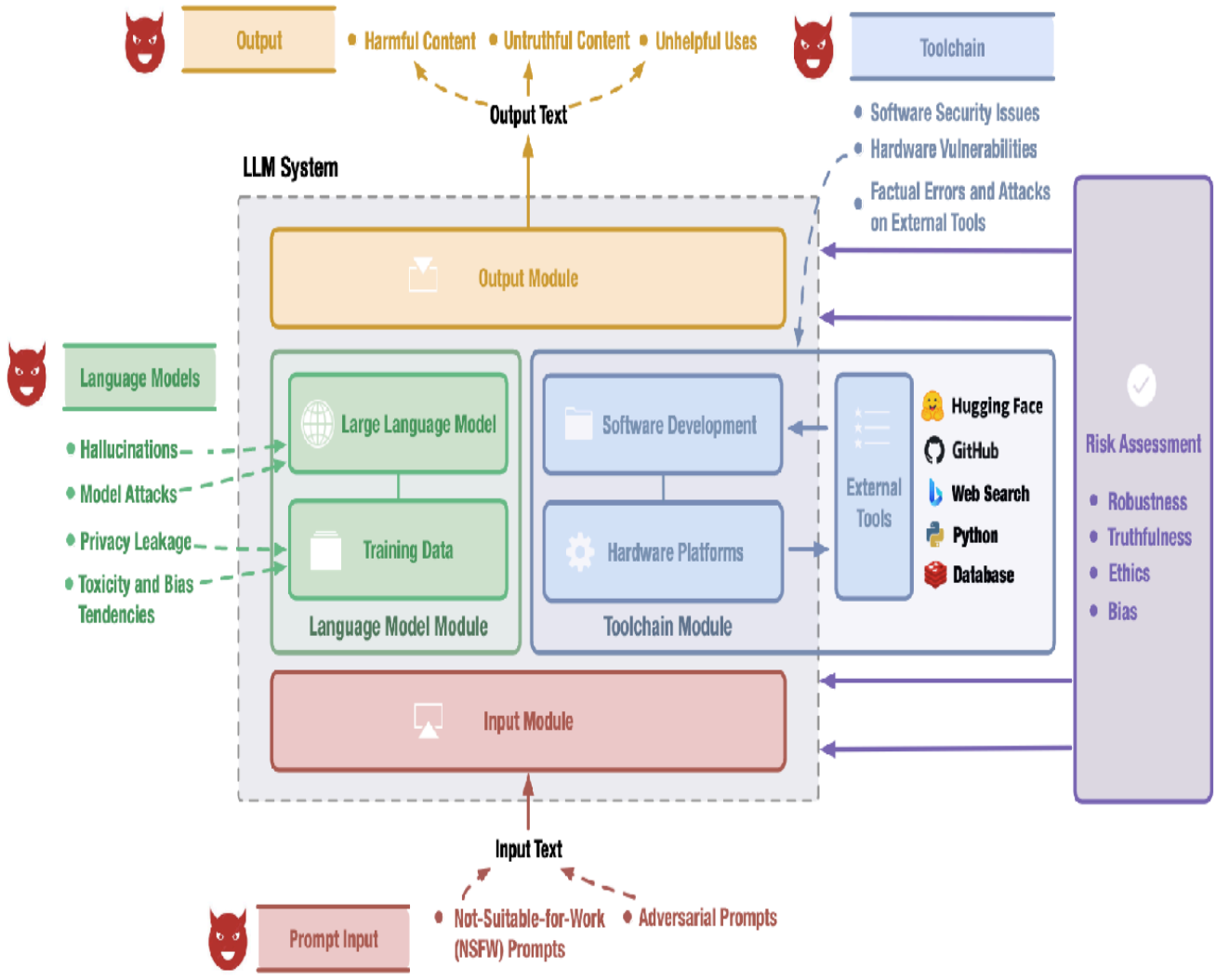


Fig. 2. The overview of an LLM system and the risks associated with each module of the LLM system. With the systematic perspective, we introduce the threat model of LLM systems from five aspects, including prompt input, language models, tools, output, and risk assessment.

1.2 Pre-Defined Evals and Testing Frameworks

AI Verify Foundation from Singapore

https://aiverifyfoundation.sg/

*"A global open-source community that convenes AI owners, solution providers, users, and policymakers, to build trustworthy AI. The aim of AIVF is to harness the collective power and contributions of an international open-source community to develop Artificial Intelligence ("**AI**") testing tools to enable the development and deployment of trustworthy AI. Ai Verify is an AI governance testing framework and software toolkit that help industries be more transparent about their AI to build trust"*

The USAISA could work with Singapore's AI Verify Foundation on testing frameworks

OpenAI Evals

https://portkey.ai/blog/decoding-openai-evals/

*"An **eval** is a task used to measure the quality of output of an LLM or LLM system. Given an input prompt, an output is generated. We evaluate this output with a set of ideal_answers and find the quality of the LLM system. If we do this a bunch of times, we can find the accuracy.*

While we use evals to measure the accuracy of any LLM system, there are 3 key ways they become extremely useful for any app in production.

1. ***As part of the CI/CD Pipeline***
   Given a dataset, we can make evals a part of our CI/CD pipeline to make sure we achieve the desired accuracy before we deploy. This is especially helpful if we've changed models or parameters by mistake or intentionally. We could set the CI/CD block to fail in case the accuracy does not meet our standards on the provided dataset.

2. ***Finding blind-sides of a model in real-time***
   In real-time, we could keep judging the output of models based on real-user input and find areas or use-cases where the model may not be performing well.

3. ***To compare fine-tunes to foundational models***
   We can also use evals to find if the accuracy of the model improves as we fine-tune it with examples. Although, it becomes important to separate out the test & train data so that we don't introduce a bias in our evaluations."

Anthropic Datasets

https://github.com/anthropics/evals?ref=portkey.ai

"This repository includes datasets written by language models, used in our paper on 'Discovering Language Model Behaviors with Model-Written Evaluations.'

'We intend the datasets to be useful to:

1. *Those who are interested in understanding the quality and properties of model-generated data*
2. *Those who wish to use our datasets to evaluate other models for the behaviors we examined in our work (e.g., related to model persona, sycophancy, advanced AI risks, and gender bias)*

The evaluations were generated to be asked to dialogue agents (e.g., a model fine tuned explicitly respond to a user's utterances, or a pre-trained language model prompted to behave like a dialogue agent). However, it is possible to adapt the data to test other kinds of models as well"

Cataloging LLM Evaluations by Singapore's AI Verify

https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

"In advancing the sciences of LLM evaluations, it is important to first achieve: (i) a common understanding of the current LLM evaluation through a standardized taxonomy; and (ii) a baseline set of pre-deployment safety evaluations for LLMs. A comprehensive taxonomy categorizes and organizes the diverse branches of LLM evaluations, provides a holistic view of LLM performance and safety, and enables the global community to identify gaps and priorities for further research and development in LLM evaluation. A baseline set of evaluations defines a minimal level of LLM safety and trustworthiness before deployment. At this early stage, the proposed baseline in this paper puts forth a starting point for global discussions with the objective of facilitating multi-stakeholder consensus on safety standards for LLMs.

Testing Frameworks for LLMs

https://llmshowto.com/blog/llm-test-frameworks

"An Overview on Testing Frameworks For LLMs. In this edition, I have meticulously documented every testing framework for LLMs that I've come across on the internet and GitHub."

Eleuthera LM Evaluation Harness
https://github.com/EleutherAI/lm-evaluation-harness

"This project provides a unified framework to test generative language models on a large number of different evaluation tasks.

Features:

- *Over 60 standard academic benchmarks for LLMs, with hundreds of subtasks and variants implemented.*
- *Support for models loaded via transformers (including quantization via AutoGPTQ), GPT-NeoX, and Megatron-DeepSpeed, with a flexible tokenization-agnostic interface.*
- *Support for fast and memory-efficient inference with vLLM.*
- *Support for commercial APIs including OpenAI, and TextSynth.*
- *Support for evaluation on adapters (e.g. LoRA) supported in HuggingFace's PEFT library.*
- *Support for local models and benchmarks.*
- *Evaluation with publicly available prompts ensures reproducibility and comparability between papers.*
- *Easy support for custom prompts and evaluation metrics.*

The Language Model Evaluation Harness is the backend for 🤗 Hugging Face's popular Open LLM Leaderboard, has been used in hundreds of papers"

Holistic Evaluation of Language Models (HELM)
https://crfm.stanford.edu/2023/12/19/helm-lite.html

"HELM Lite is inspired by the simplicity of the Open LLM leaderboard (Hugging Face), though at least at this point, we include a broader set of scenarios and also include non-open models. The HELM framework is similar to BIG-bench, EleutherAI's lm-evaluation-harness, and OpenAI evals, all of which also house a large number of scenarios, but HELM is more modular (e.g., scenarios and metrics are defined separately).'

Holistic Testing
https://static.scale.com/uploads/6019a18f03a4ae003acb1113/test-and-evaluation.pdf

"We introduce a hybrid methodology for the evaluation of large language models (LLMs) that leverages both human expertise and AI assistance. Our hybrid methodology generalizes across both LLM capabilities and safety, accurately identifying areas where AI assistance can be used to automate this evaluation. Similarly, we find that by combining automated evaluations, generalist red teamers, and expert red teamers, we're able to more efficiently discover new vulnerabilities"

Custom GPTs
https://openai.com/blog/introducing-gpts

"We're rolling out custom versions of ChatGPT that you can create for a specific purpose—called GPTs. GPTs are a new way for anyone to create a tailored version of ChatGPT to be more helpful in their daily life, at specific tasks, at work, or at home—and then share that creation with others. Anyone can easily build their own GPT—no coding is required. You can make them for yourself, just for your company's internal use, or for everyone. Creating one is as easy as starting a conversation, giving it instructions and extra knowledge, and picking what it can do, like searching the web, making images or analyzing data."
A risk evaluation and testing framework is needed for Custom GPTs.

1.3 Lower Risk: Generic Applications and Use Cases (LLM and human testing based on Evals)

Red Teaming Language Models using Language Models
https://arxiv.org/abs/2202.03286

"Overall, LM-based red teaming is one promising tool (among many needed) for finding and fixing diverse, undesirable LM behaviors before impacting users."

Discovering Language Model Behaviors with Model-Written Evaluations
https://arxiv.org/abs/2212.09251

"Prior work creates evaluation datasets manually (Bowman et al., 2015; Rajpurkar et al., 2016, inter alia), which is time-consuming and effortful, limiting the number and diversity of behaviors tested. Other work uses existing data sources to form datasets (Lai et al., 2017, inter alia), but such sources are not always available, especially for novel behaviors. Still other work generates examples with templates (Weston et al., 2016) or programmatically (Johnson et al., 2017), limiting the diversity and customizability of examples. Here, we show it is possible to generate many diverse evaluations with significantly less human effort by using LLMs;"

1.4 Higher Risk: Domain-specific Applications and Use Cases (Fine Tuned Testing LLM + Human Domain Experts)

Large Action Models(LAMs)
http://tinyurl.com/33zwkmbb

"LAMs interact with the real world through integration with external systems, such in IoT devices and others. By connecting to these systems, LAMs can perform physical actions, control devices, retrieve data, or manipulate information"

OpenAI External Red Team

https://openai.com/blog/red-teaming-network

"The OpenAI Red Teaming Network is a community of trusted and experienced experts that can help to inform our risk assessment and mitigation efforts more broadly, rather than one-off engagements and selection processes prior to major model deployments. Members of the network will be called upon based on their expertise to help red team at various stages of the model and product development lifecycle. Not every member will be involved with each new model or product, and time contributions will be determined with each individual member"
A vendor-independent Red Teaming Network of Experts is needed

1.5 Highest Risk: Applications that Change Environment (Simulation or Sandbox Testing)

The environment can be cyber or physical. The changes can be direct or indirect (e.g. code generation, persuasion)

Singapore Generative AI Evaluation Sandbox

https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

"1.The Sandbox will bring global ecosystem players together through concrete use cases, to enable the evaluation of trusted AI products. The Sandbox will make use of a new Evaluation Catalogue, as a shared resource, that sets out common baseline methods and recommendations for Large Language Models (LLM).
2.This is part of the effort to have a common standard approach to assess Gen AI.
3. The Sandbox will provide a baseline by offering a research-based categorization of current evaluation benchmarks and methods. The Catalogue provides an anchor by (a) compiling the existing commonly used technical testing tools and organizing these tests according to what they test for and their methods; and (b) recommending a baseline set of evaluation tests for use in Gen AI products."

Singapore GenAI Sandbox

https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

"Sandbox will offer a common language for evaluation of Gen AI through the Catalogue Sandbox will build up a body of knowledge on how Gen AI products should be tested Sandbox will develop new benchmarks and tests"

Participants in Sandbox include most many AI vendors (not OpenAI)

https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2023/10/generative-ai-evaluation-sandbox/annex-a---list-of-participants-in-sandbox.pdf

USAISI could work with Singapore's AI Verify Foundation on testing sandbox

1.6 Incident Analysis and Fixes

Use Generative AI to automate Incident Analysis

https://www.bigpanda.io/wp-content/uploads/2023/07/bigpanda-generative-ai-datasheet.pdf

"AI-generated summary and title: Identify incidents that require more immediate action by automatically synthesizing complex alert data into clear, crisp incident summaries and titles that can be populated within chat and ITSM tools.

AI-proposed incident impact: Reliably identify the relevancy and impact of incidents across distributed IT systems in clear, natural language within seconds. Easily identify priority actions for ITOps, L2, and L3 response teams across all incidents at scale.

AI-suggested root cause: Automatically surface critical insights and details hidden in lengthy and complex alerts to quickly identify the probable root cause of an incident, as it forms in real-time."

Fixing Hallucinations in LLMs

https://betterprogramming.pub/fixing-hallucinations-in-llms-9ff0fd438e33?gi=a8912d3929dd

"Hallucinations in Large Language Models stem from data compression and inconsistency. Quality assurance is challenging as many datasets might be outdated or unreliable. To mitigate hallucinations:

1. *Adjust the temperature parameter to limit model creativity.*
2. *Pay attention to prompt engineering. Ask the model to think step-by-step and provide facts and references to sources in the response.*
3. *Incorporate external knowledge sources for improved answer verification.*

A combination of these approaches can achieve the best results.

The Rapid Response Team

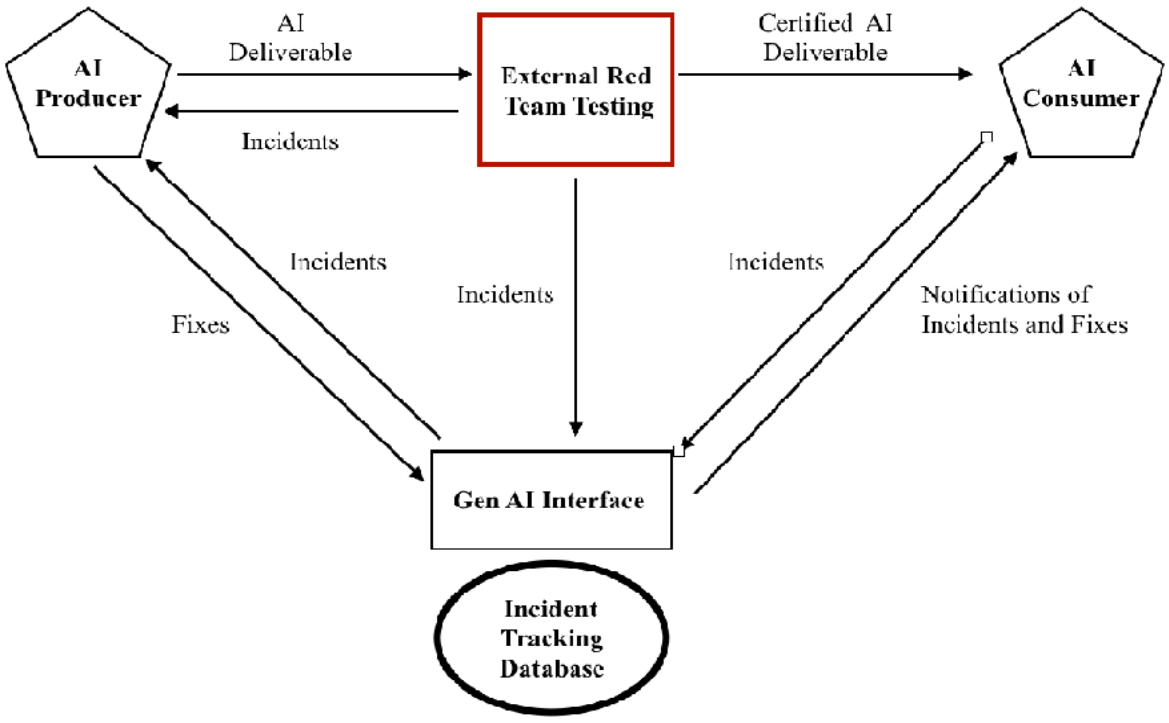
https://www.svpg.com/the-rapid-response-team/

"In these cases, a practice that I have seen make dramatic improvements along both dimensions is to create at least one special dedicated team that we often call the "Rapid Response Team." This is a dedicated team comprised of a product manager (or at least a part of a product manager), and mainly developers and QA. Usually these teams are not large (2-4 developers is common). This team has the following responsibilities:

- *fix any critical issues that arise for products in the sustaining mode (i.e. products that don't have their own dedicated team because you're not investing in them other than to keep it running).*
- *implement minor enhancements and special requests that are high-value yet would significantly disrupt the dedicated team that would normally cover these items.*
- *fix any critical, time-sensitive issues that would normally be covered by the dedicated team, but again would cause a major disruption."*

The AI Producer should have a Rapid Response team to handle incidents and provide fixes

2. Incident Tracking Database



2.1 Incident Tracking Database

AI Incident Database

https://incidentdatabase.ai/

"The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes.

You are invited to submit incident reports, whereupon submissions will be indexed and made discoverable to the world. Artificial intelligence will only be a benefit to people and society if we collectively record and learn from its failings."

Partnership on AI

https://partnershiponai.org/workstream/ai-incidents-database/

"As AI technology is integrated into an increasing number of safety-critical systems — entering domains such as transportation, healthcare, and energy — the potential impact of this technology's failures similarly grows. The AI Incident Database (AIID) is a tool designed to help us better imagine and anticipate these risks, collecting more than 1,200 reports of intelligent systems causing safety, fairness, or other real-world problems."

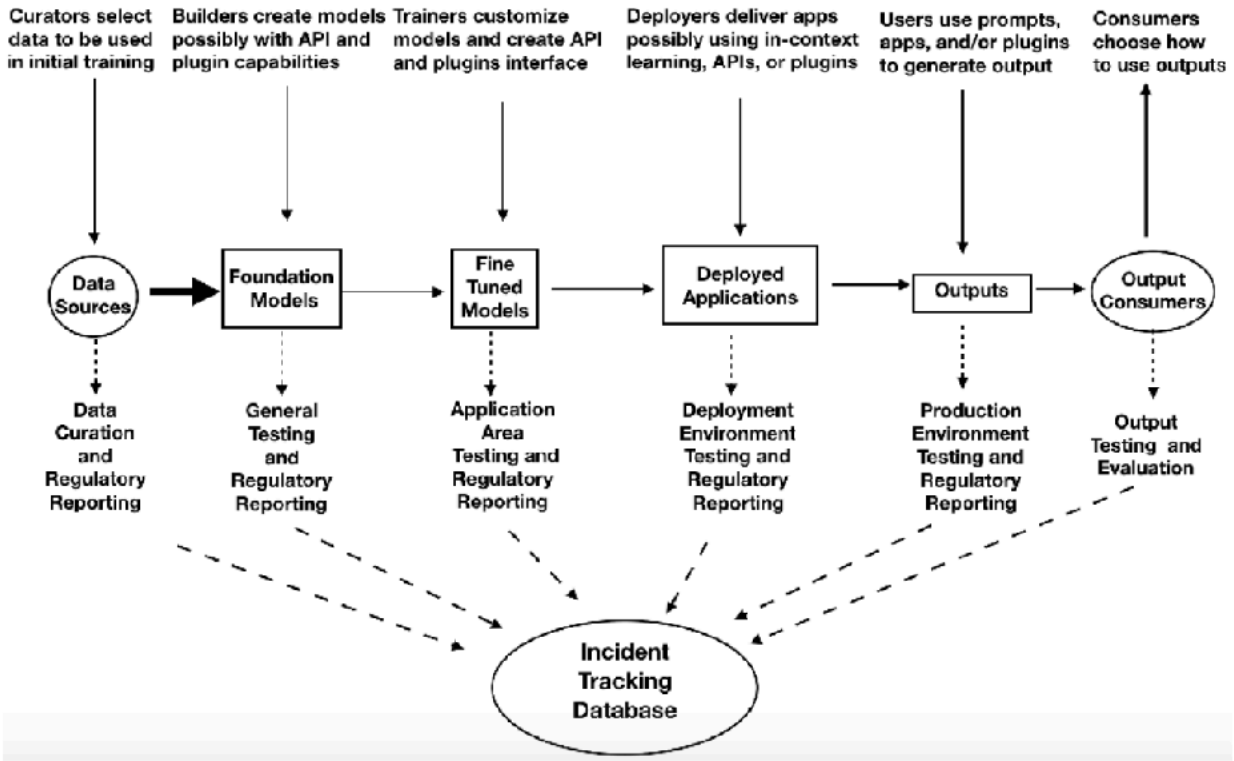
Preventing Repeated Real World AI Failures by Cataloging Incidents

https://arxiv.org/abs/2011.08512

"Mature industrial sectors (e.g., aviation) collect their real world failures in incident databases to inform safety improvements. Intelligent systems currently cause real world harms without a collective memory of their failings. As a result, companies repeatedly make the same mistakes in the design, development, and deployment of intelligent systems. A collection of intelligent system failures experienced in the real world (i.e., incidents) is needed to ensure intelligent systems benefit people and society. The AI Incident Database is an incident collection initiated by an industrial/non-profit cooperative to enable AI incident avoidance and mitigation. The database supports a variety of research and development use cases with faceted and full text search on more than 1,000 incident reports archived to date."

2.2 Generative AI Delivery Process

AI deliverables can be produced, consumed and generate incidents in many stages of the Generative AI Delivery Process.



2.3 LLM Interfaces to Database

How LLMs made their way into the modern data stack
https://venturebeat.com/data-infrastructure/how-llms-made-their-way-into-the-modern-data-stack-in-2023/

"The first (and probably the most important) shift with LLMs came when vendors started debuting conversational querying capabilities — i.e. getting answers from structured data (data fitting into rows and columns) by talking with it. This eliminated the hassle of writing complex SQL (structured query language) queries and gave teams, including non-technical users, an easy-to-use text-to-SQL experience, where they could put in natural language prompts and get insights from their data. The LLM being used converted the text into SQL and then ran the query on the targeted dataset to generate answers."

Can LLM Already Serve as A Database Interface
https://typeset.io/questions/can-llm-already-serve-as-a-database-interface-a-big-bench-3gje48fazi

"Large language models (LLMs) have shown impressive results in the task of converting natural language instructions into executable SQL queries, known as Text-to-SQL parsing. However, existing benchmarks like Spider and WikiSQL focus on small-scale databases, leaving a gap between academic study and real-world applications. To address this, the paper "Bird" presents a big benchmark for large-scale databases in the context of text-to-SQL tasks. It contains a large dataset of text-to-SQL pairs and 95 databases spanning various professional domains. The emphasis on database values in Bird highlights the challenges of dirty database contents, external knowledge, and SQL efficiency in the context of massive databases. The experimental results demonstrate the significance of database values in generating accurate text-to-SQL queries for big databases."

Text2SQL
https://medium.com/@changjiang.shi/text2sql-converting-natural-language-to-sql-defa12c2a69f

"Text2SQL is a natural language processing technique aimed at converting natural language expressions into structured query language (SQL) for interaction and querying with databases. This article presents the historical development of Text2SQL, the latest advancements in the era of large language models (LLMs), discusses the major challenges currently faced, and introduces some outstanding products in this field."

2.4 Notifications

Using LLMs for notifications
https://pathway.com/developers/showcases/llm-alert-pathway

"Real-time alerting with Large Language Models (LLMs) like GPT-4 can be useful in many areas such as progress tracking for projects (e.g. notify me when coworkers change requirements), regulations monitoring, or customer support (notify when a resolution is present).

The program that we will create answers questions based on a set of documents. However, after an initial response is provided, the program keeps on monitoring the document sources. It efficiently determines which questions may be affected by a source document change, and alerts the user when a revision - or a new document - significantly changes a previously given answer.

The basic technique of feeding chunks of information from external documents into an LLM and asking it to provide answers based on this information is called RAG - Retrieval Augmented Generations. So, what we are doing here is ***real-time RAG with alerting"***