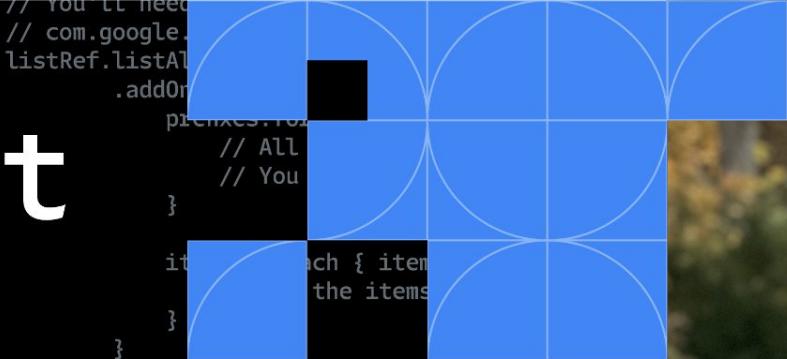


# devfest



## Wild to Wise - Taming GenAI in a Responsible Ecosystem

 Google Developer Groups  
DC Startup Week, 2023





[pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election](https://pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election)

News Finances Startup Stemsible Metashore Metashore Federal Big Data

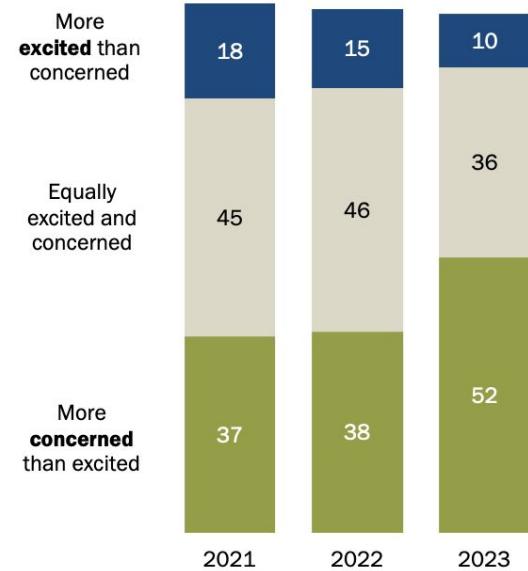
# AI-generated disinformation poses threat of misleading voters in 2024 election



Politics May 14, 2023 7:52 PM EDT

## Concern about artificial intelligence in daily life far outweighs excitement

% of U.S. adults who say the increased use of artificial intelligence in daily life makes them feel ...



Note: Respondents who did not give an answer are not shown.  
Source: Survey conducted July 31-Aug. 6, 2023.

PEW RESEARCH CENTER

The r



Duke  
UNIVERSITY

# My Journey

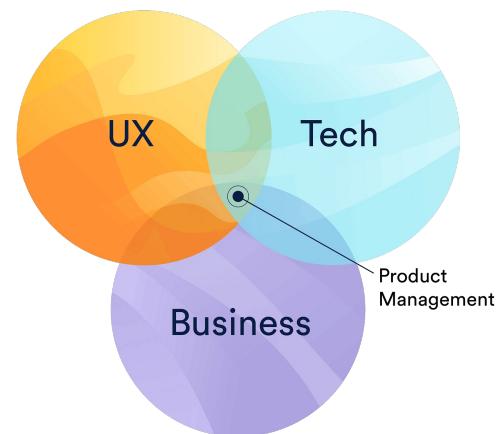
Commercial, Non-Profit, & Federal



Startups & Product Consulting



Interests & Expertise

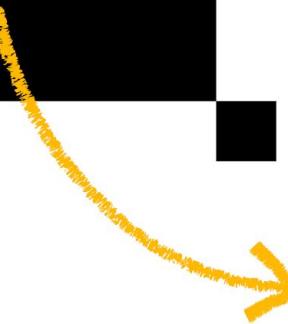
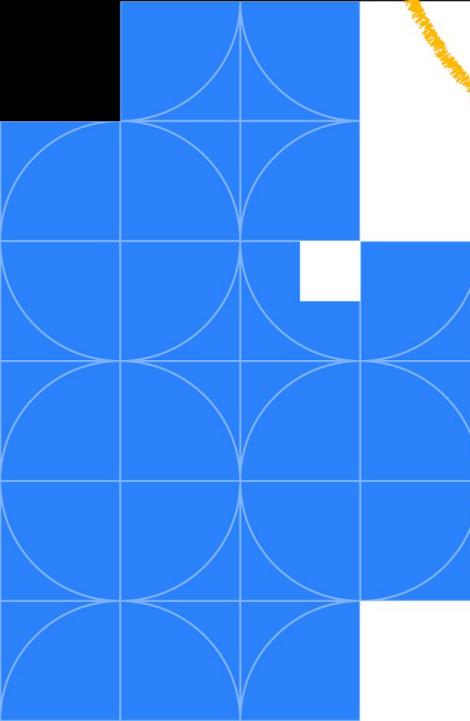


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.blue[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.blue[500],
```

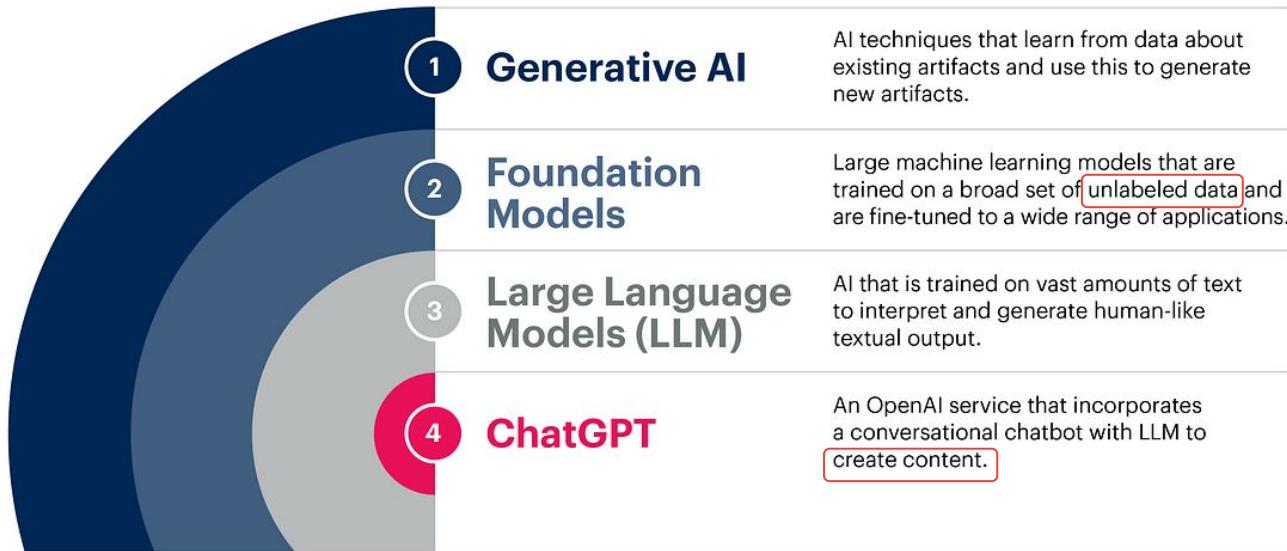
```
Text('23'),
```



Google Developer Groups

# Generative AI Landscape & Basics

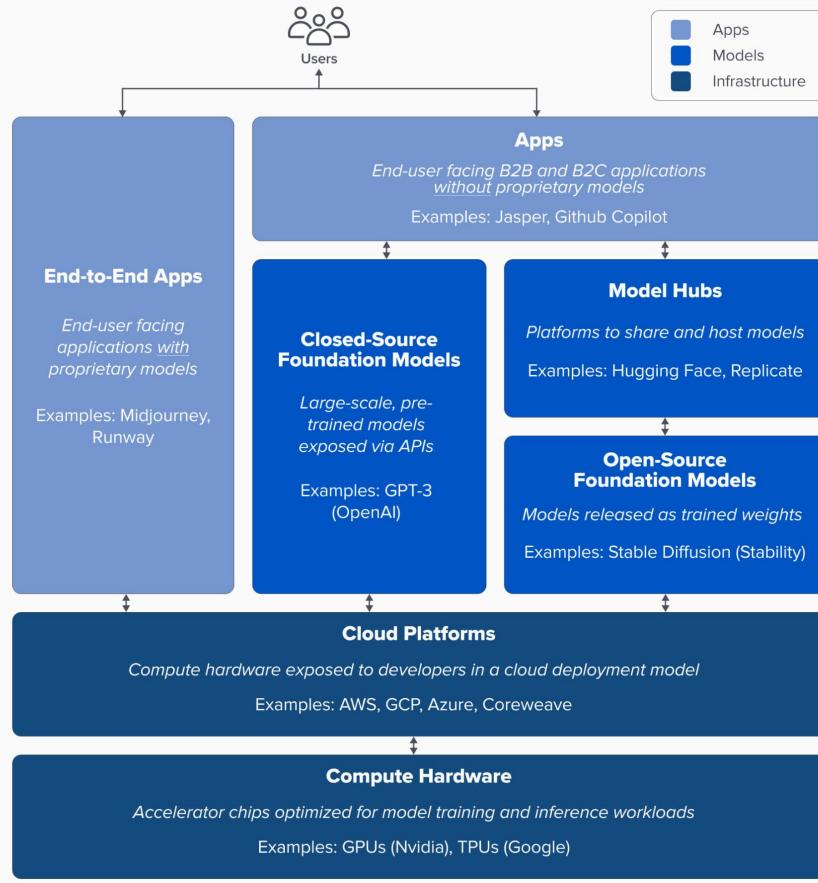
# What Is Generative AI?



Source: Gartner  
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. 2421958

Gartner

## Preliminary generative AI tech stack



# There are lots of opportunities GenAI is unlocking



## Text generation

Create new pieces of original content, such as short stories, essays, social media posts, and webpage copy.



## Chatbots

Build conversational interfaces such as chatbots and virtual assistants to enhance the user experience for your customers.



## Search

Search, find, and synthesize information to answer questions from a large corpus of data.



## Text summarization

Get a summary of textual content, such as articles, blog posts, books, and documents, to get the gist without having to read the full content.



## Image generation

Create realistic and artistic images of various subjects, environments, and scenes from language prompts.



## Personalization

Help customers find what they're looking for with more relevant and contextual product recommendations than word matching.

## Top 50 GenAI Web Products, By Monthly Visits

1.  ChatGPT	11.  YOU <sup>1</sup>	21.  NightCafe	31.  GPTGO.ai	41.  Fliki
2. character.ai	12.  leonardo.	22.  Replicate	32.  runway	42.  pornpen.ai
3. Bard	13.  PIXLR	23.  Speechify	33.  Playground	43.  KAPWING
4.  Poe	14.  VEED.IO	24.  ElevenLabs	34.  Kaiber	44.  Gamma
5.  QuillBot	15.  tome	25.  Lexica	35.  Hotpot	45.  Looka
6.  PhotoRoom	16. AI-Novel	26. VocalRemover	36. Stable Diffusion	46.  human or not?
7.  CIVITAI	17. cutout.pro	27.  Writesonic	37.  copy.ai	47.  PIXAI
8.  Midjourney	18.  ForefrontAI	28.  CHATPDF	38.  ZeroGPT	48.  WRITER
9.  Hugging Face	19.  Clipchamp	29.  D-ID	39.  Smodin	49.  NovelAI
10.  Perplexity	20. TheB.AI	30.  Chub.ai	40.  ZMO.AI	50. DeepSwap

Note: This list was generated based on global desktop and mobile web visits with data from SimilarWeb as of June 2023. However, for companies on the list that also have a mobile app, we added an estimate of their app "traffic" (MAUs x sessions per MAU per month), with data from Sensor Tower as of June 2023, alongside web traffic to determine their spot on the list.

This list does not include app-only companies, and does not include activity happening within a Discord server.

Charts are for informational purposes only and should not be used for investment decisions. Past performance is not indicative of future results. None of the above should be taken as investment advice; see [al6z.com/disclosures](https://al6z.com/disclosures) for more.

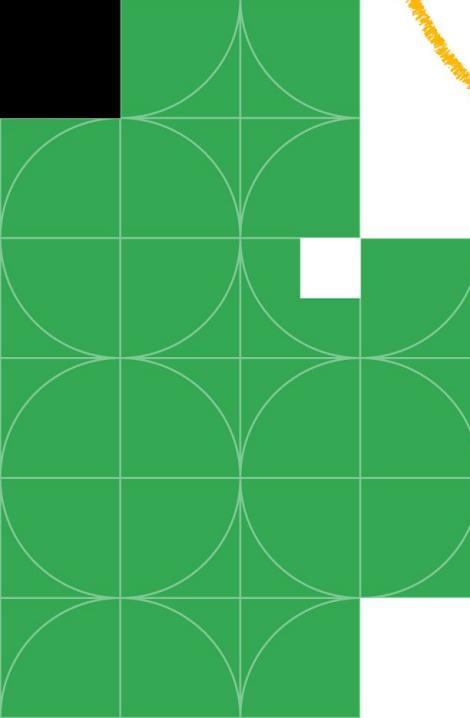


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],
```

```
Text('23'),
```



Google Developer Groups

# Generative AI User Experience & Outcomes

## Case in Point #1

R is the square root of 4%

< 1/2 >  Yes, the square root of 4% is 2%. The square root of a percentage is found by taking the square root of the decimal equivalent of the percentage. In this case, the decimal equivalent of 4% is 0.04, and the square root of 0.04 is 0.2. When expressed as a percentage, it becomes 2%.

R answers are wrong

 I apologize for the confusion. Let me clarify.

The square root of 4% is not 2%. To find the square root of a percentage, you need to take the square root of the decimal equivalent. The decimal equivalent of 4% is 0.04.

So, the square root of 0.04 is approximately 0.2, not 2%. I appreciate your patience and thank you for bringing this to my attention.

R .2 is not 2%, it is 20%

 You're absolutely correct, and I appreciate your clarification. I misspoke in my response. I apologize for any confusion.

To correct: the square root of 0.04 is 0.2, and when expressed as a percentage, it is indeed 20%, not 2%. Thank you for catching that mistake, and I appreciate your understanding.

## Case in Point #2



**Nitin Aggarwal**  (He/Him)  
Head of Cloud AI Services (GenAI) at Google  
 Talks about #strategy, #leadership, #generativeai, #machinelearning, and #artificialintelligence  
Sunnyvale, California, United States · [Contact info](#)

 Google  
 Stanford University Graduate School of Business

### Here's what else to consider

This is a space to share examples, stories, or insights that don't fit into any of the previous sections. What else would you like to add?



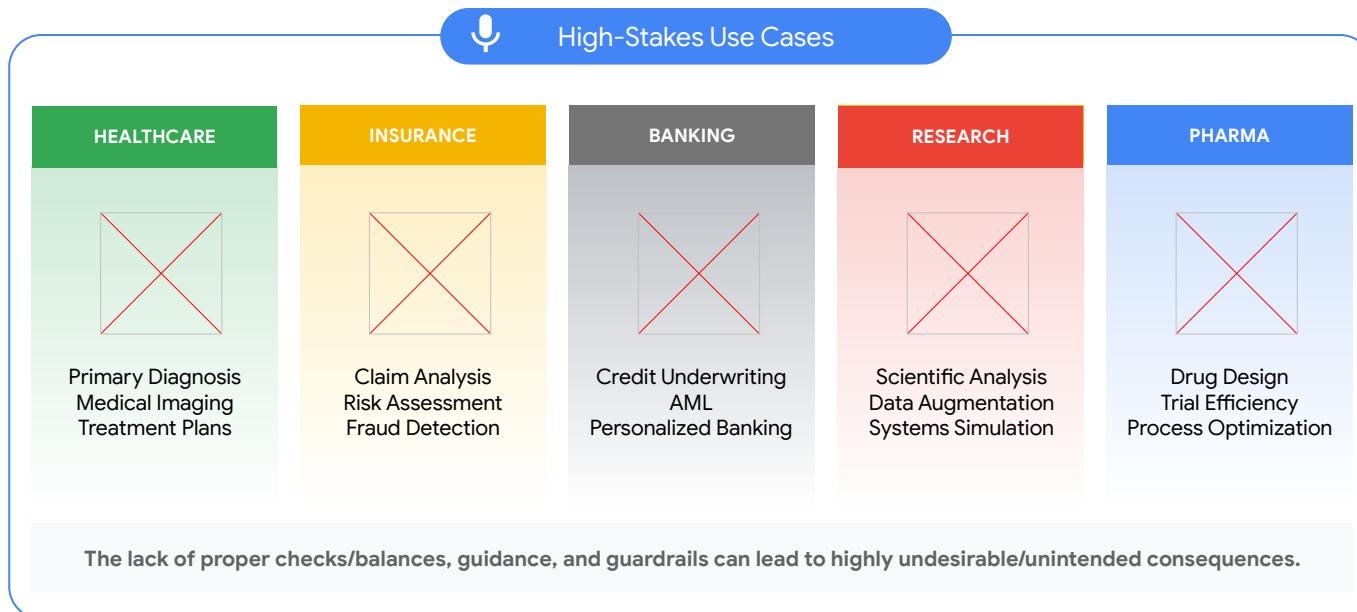
Add your perspective

**Nitin Aggarwal** Head of Cloud AI Services (GenAI) at Google

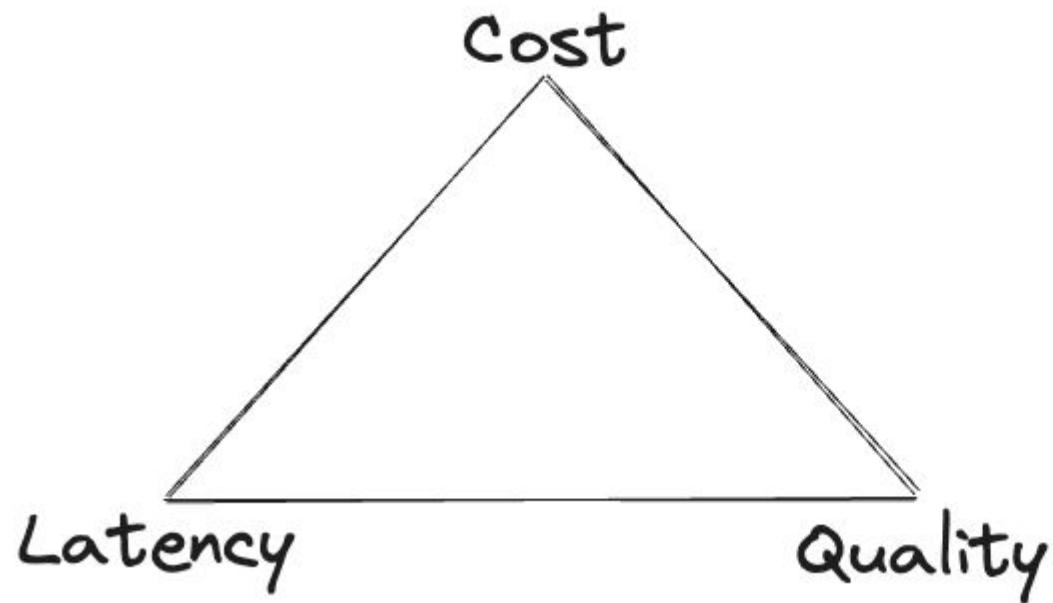
I don't have anything else to share or add to this article. I'm just experimenting if my content/comment/contribution is even being explored/evaluated or it's just a random way to get a "top voice" badge.

 Like ...

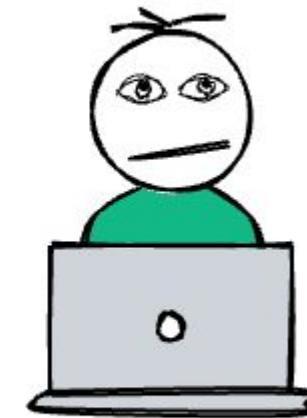
# High-stakes use cases can potentially do a lot more harm!



**BUT, we have to make trade-offs to get to prod-ready state**



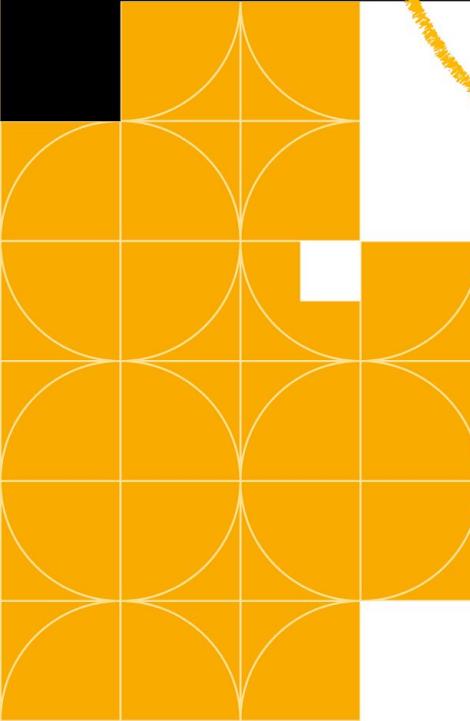
Which Embedding model?



```
text('Section Title',  
    style: TextStyle(  
        color: Colors.yellow[200],  
    ),  
)  
,
```

# devfest

```
s.star,  
r: Colors.yellow[500],  
Text('23'),
```



Google Developer Groups

# Challenges posed to Responsible Generative AI

# devfest

The usual concerns of responsible AI become more difficult for GenAI. But GenAI also gives rise to new challenges that simply don't exist for predictive models that are more narrow.

# A whole new breed of challenges never seen before

Hallucinations

Factual  
Inaccuracies

Toxicity

Outdated  
Information

Fake Content

IP Theft

Data Privacy

No References

Ethical Concerns

Age  
Appropriateness

Misinformation

Plagiarism &  
Cheating

# It's like going from the minor leagues to the majors



# Generative AI takes Stereotypes and Bias from Bad to Worse



## Working Women Misrepresented Across the Board

Stable Diffusion results compared to US demographics for each occupation

Average US income in 2022

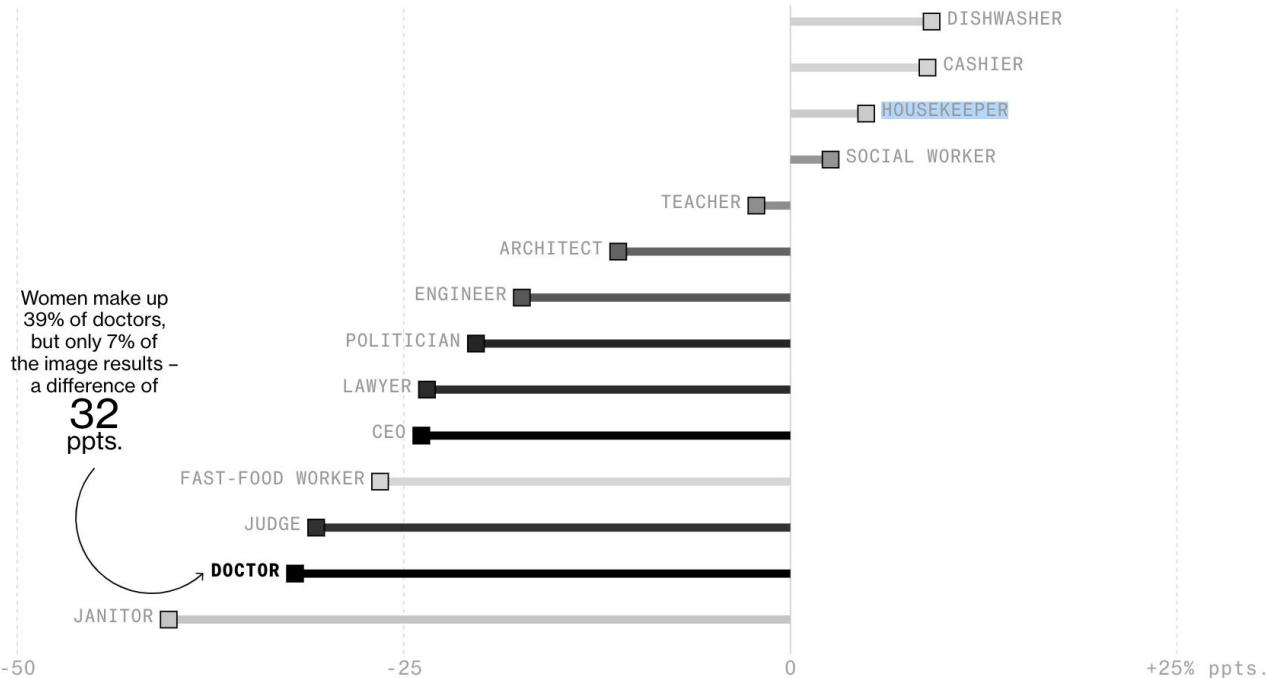
\$20K

\$242K

← GENERATED FEWER WOMEN

US

GENERATED MORE WOMEN →



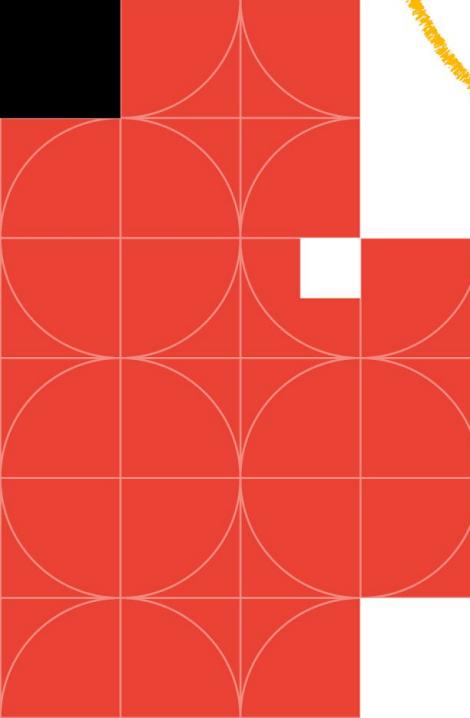
Sources: Bureau of Labor Statistics, American Medical Association, National Association of Women Judges, Federal Judicial Center, Bloomberg analysis of Stable Diffusion

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.red[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.red[500],
```

```
Text('23'),
```



# A Responsible Generative AI Ecosystem

# A responsible GenAI ecosystem requires developing 5 key capabilities

01  
**Work  
Backwards**

02  
**Embrace  
Standards**

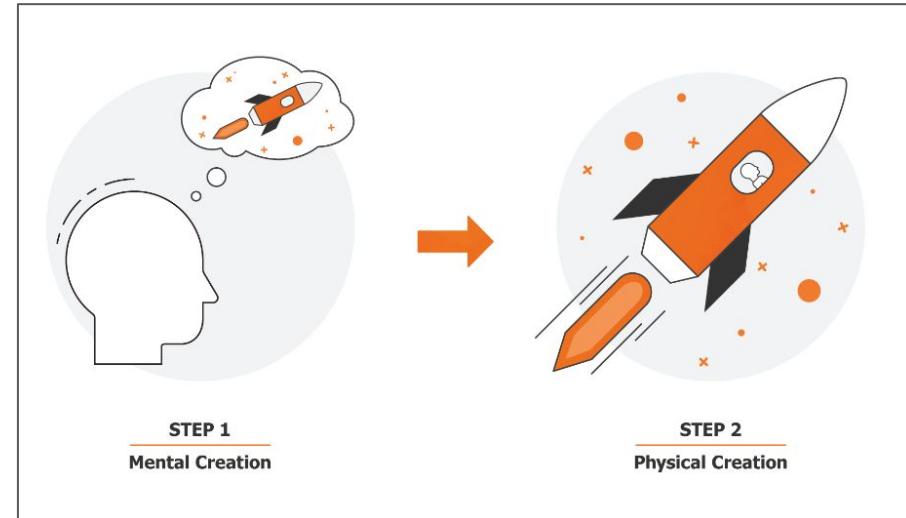
03  
**Instrument  
Observability**

04  
**Fail  
Fast**

05  
**Inspect &  
Adjust**

# Work Backwards - Envision and start from the end

- Take time understanding the problem you are looking to solve, **jobs to be done, needs to be met**, and the people you are solving for
- Imagine what may be “**good**” and “**great**” experience and outcome(s) for the end user
- Understand the **functional domain, regulations, known risks, and constraints** within the problem space
- Take time understanding the **user’s context** - Where will they be when using the application? Home? Road? How? When?
- The above will tell you which foundational model is right for you, how to **fine-tune it**, and what **guardrails** to put in place
- Don’t know what problem to solve? Start with something **you are passionate about?** IT doesn’t have to be a money-maker from day 1.
- Go slow, go small, **build a prototype**, MVP, test it



# Embrace Standards - Learn from the Experts

- There are a variety of **frameworks, principles, and standards** available from the industry, academia, and government
- Stanford's **HELM** (Holistic Evaluation of Language Models) is a great place to understand how they **benchmark** leading LLMs
- Google has proposed **3 principles for designing** responsible generative AI systems (*listed on the right*)
- Augment LLMs with independent, verified citation databases and similar sources, using approaches such as **retrieval-augmented generation** (RAG)
- **Attribute generated outputs** to particular pieces of training data, allowing users to assess the validity of those sources
- Develop a **prohibited use policy** in advance of release, based on harms identified early in the research, development and ethics review process



# Instrument Observability - To manage, you must measure

- Check out **LangKit** (from WhyLabs): An **open-source** toolkit for monitoring Large Language Models (LLMs). It allows you to extract signals from prompts & responses, ensuring safety & security.
- Go back to basics and determine the **2-3 things** that determine whether or not you are **delivering value** to your user/customer.
- **Text quality metrics**, such as readability, complexity and grade level, can provide important insights into the quality and appropriateness of generated responses.
- **Measure the similarity** of prompts and responses **against certain topics** or known examples, such as jailbreaks or controversial subjects.
- Measure **sentiment** and monitor **toxicity** can ensure that the user experience is positive and respectful. Ex: Think of a customer support/service app!
- Monitoring for **security and privacy** can ensure that the system is neither recording or serving PII data or credit card information.



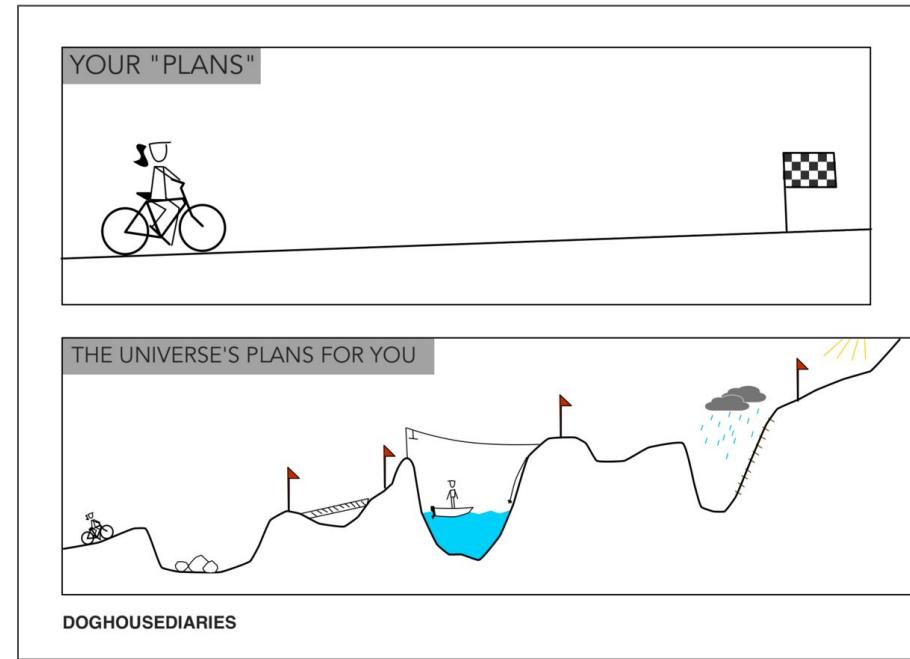
```
from langkit import llm_metrics
import whylogs as why
schema = llm_metrics.init()

profile = why.log(prompt_and_response,
                  name="HF prompt & response",
                  schema=schema)
```

Safety Attribute	Description
Derogatory	Negative or harmful comments targeting identity and/or protected attributes.
Toxic	Content that is rude, disrespectful, or profane.
Sexual	Contains references to sexual acts or other lewd content.
Violent	Describes scenarios depicting violence against an individual or group, or general descriptions of gore.
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
Profanity	Obscene or vulgar language such as cursing.
Death, Harm & Tragedy	Human deaths, tragedies, accidents, disasters, and self-harm.
Firearms & Weapons	Content that mentions knives, guns, personal weapons, and accessories such as ammunition, holsters, etc.
Public Safety	Services and organizations that provide relief and ensure public safety.
Health	Human health, including: Health conditions, diseases, and disorders Medical therapies, medication, vaccination, and medical practices Resources for healing, including support groups.
Religion & Belief	Belief systems that deal with the possibility of supernatural laws and beings; religion, faith, belief, spiritual practice, churches, and places of worship. Includes astrology and the occult.

# Fail Fast - Break it because somebody else will

- Regular ML models for classification and prediction are pretty stable but LLMs is a whole new space, with a **bunch of unpredictability**. So try real hard breaking things up.
- Try not to get caught up with the hype ... work with **proven methodologies**, tried and tested stuff.
- **Be creative** and take on different personas, both desired and **undesired**, intended and **unintended**, so that your application gets tested in known and unknown ways.
- Stress, load, and **performance test** with different thruputs and settings to get a sense of user experience as well as system latencies and cost implications
- Put it in the hands of a small cohort of **target users** and see if they well or not they are able to use your system. Because your users will come up with ways to **interact with your model** that you never imagined.
- **Continue observing** and monitoring key model metrics



# Inspect & Adjust - Continuous improvement is key

- Implement a **user feedback loop** (ex: thumbs up/down) to get direct, real-time user satisfaction and value rating.
- Watch out and measure new and/or **different ways the users interact** with the system and prompt distribution.
- Continue to **study emerging best practices** and guidelines, and implement the ones that make sense with your context
- Scale the system at a pace that allows you to **test and verify at every stage**, and keep tech debt manageable at all times.
- Keep an eye out for **federal and state regulations**, and always stay on the right side of things
- Analyze prompt and response metrics, and **fine-tune the model safety attributes** accordingly
- **UI/UX** are extremely **important** in this extremely non-deterministic system!
- Winning strategy: **Solve for specific domains** and use cases



# Responsible GenAI Ecosystem - Summary

## Work Backwards

- Understand target user context and needs
- Envision risks, regulations, and implement guardrails
- Start small, be specific, build and test a prototype

## Instrument Observability

- Use open-source LangKit for LLM monitoring
- Identify and optimize 3 safety and value metrics that truly matter
- Monitor security and privacy to protect the user and your brand

## Embrace Standards

- Use HELM LLM benchmarking metrics
- Integrate with RAG to improve response quality
- Publish a Prohibited Use Policy before product launch

## Fail Fast

- Test model in random and unpredictable ways
- Stress and load test for system latency and cost implications
- Test model early and often with early adopters / users

## Inspect & Adapt

- Implement user feedback loop
- Watch user interaction and prompt distribution
- Fine-tune model safety attributes

# Thank you.