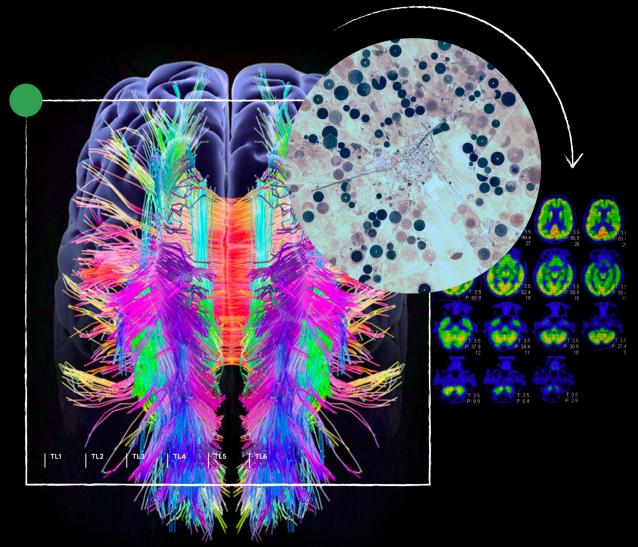
Scaling Responsible AI for Generative Models

Kathy Meier-Hellstern
Principal Engineer
Responsible Engineering
Google DeepMind

Presentation to IEEE ETR
May 22, 2024

We are at a Transformational Moment!



Why has the industry been transformed?

Scaling up model size and training data has unlocked powerful capabilities, allowing models to:

1

Breakthrough performance

in reasoning, math, science, and language-related tasks

2

Creative Potential

Generate text, code, audio, images, videos, etc....which can have a big impact on unlocking creative potential

3

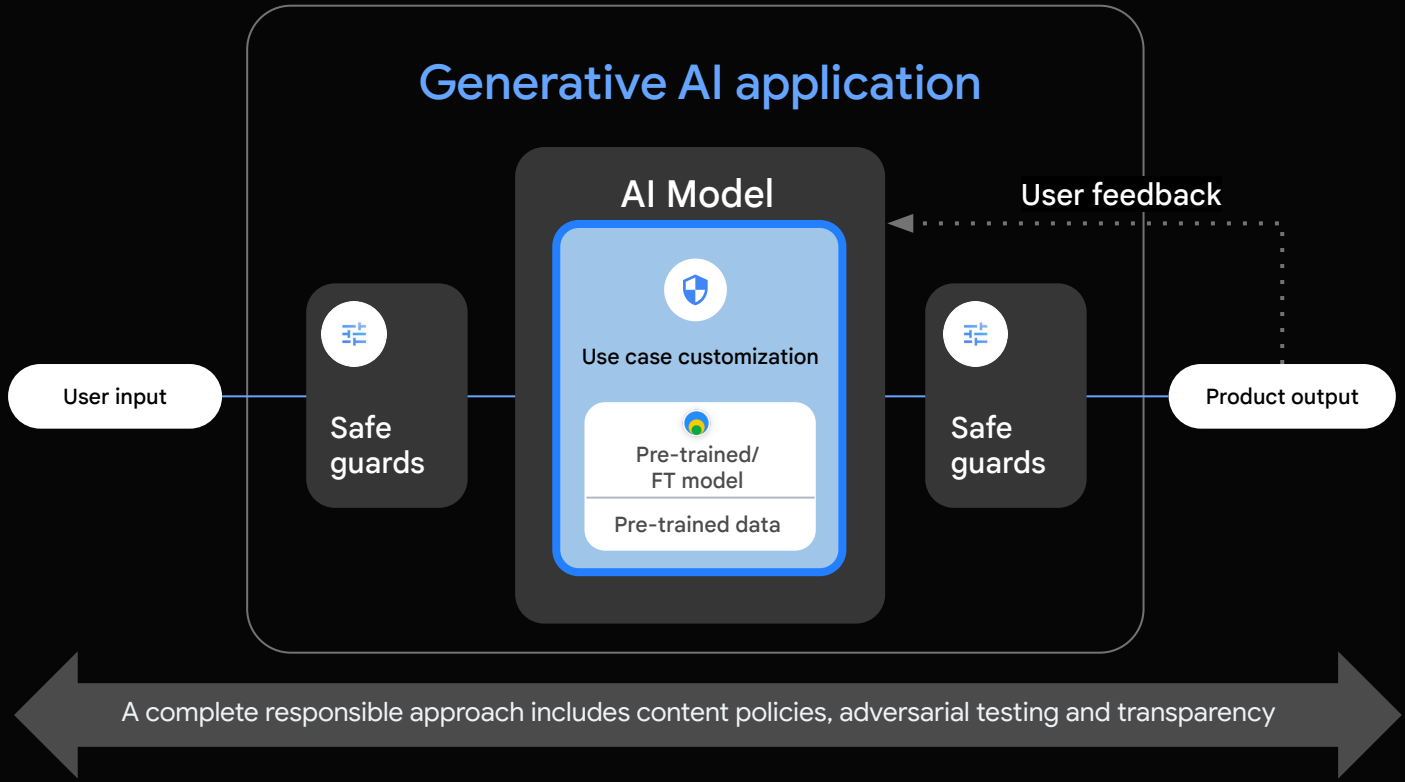
Democratization

more people can prototype new AI applications, even without writing any code

However,

applications using these models can also exhibit harmful behaviors such hallucination, misinformation, unsafe responses, bias ...

Generative AI Ecosystem



A Data-Driven Pipeline

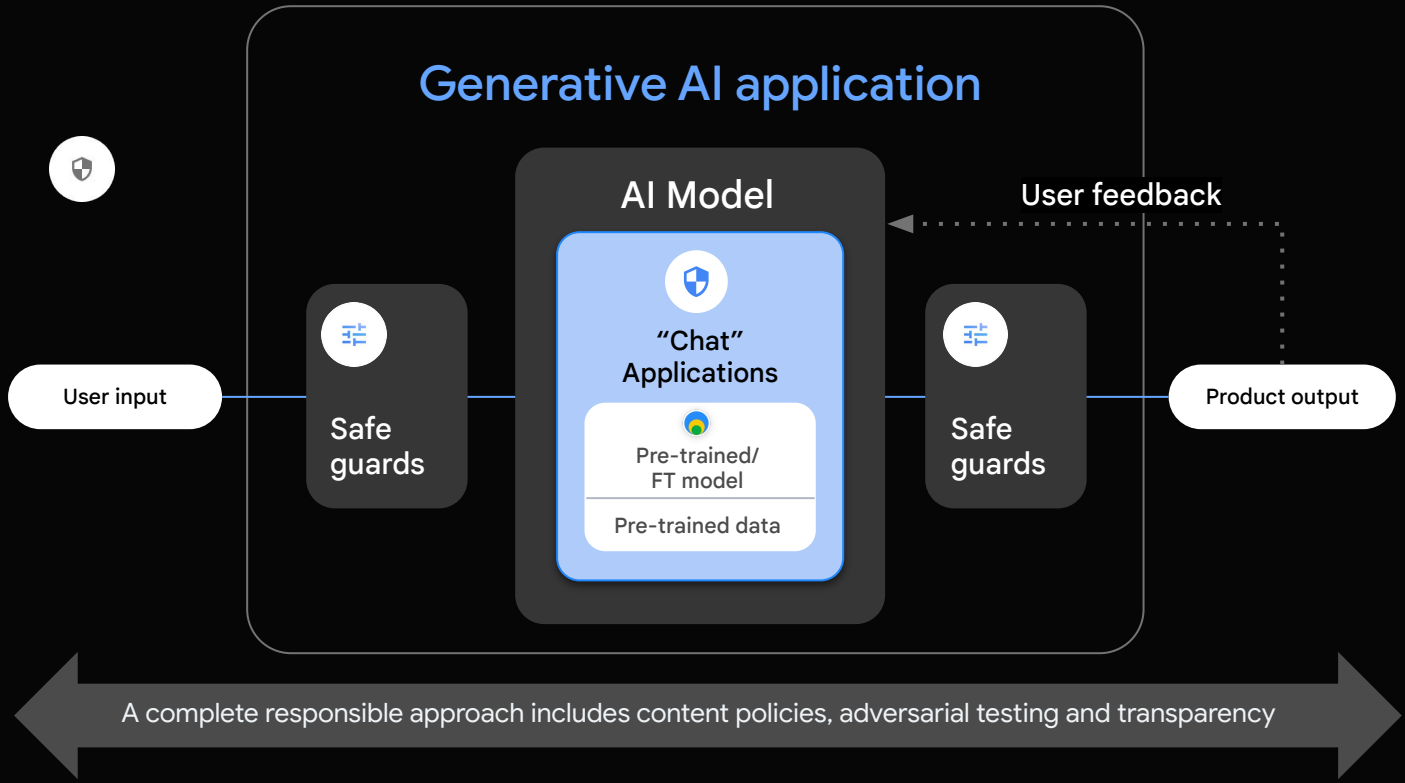
Problem Formulation **Define Policies**	Data Collection **Representative**	Real-world Alignment **Tune & Mitigate**	Inclusive Evaluations **Safety / Fairness**	Post-launch Actions **User Feedback**
Standardized content safety policies.	High-quality policy-aligned data for training, fine-tuning and evaluations	Align model to policies using fine tuning, RLHF and other advanced techniques, as well as external classifiers	Assess model performance with high quality data to measure safety and bias risks	Transparency, feedback and user control.

Current State: Examples of Policy Focus Areas

1. Sensitive Personally Identifiable Information (SPII)
2. Hate Speech
3. Harassment
4. Dangerous Content
5. Sexually Explicit Content
6. Enables Access to Harmful Goods and Services

* https://policies.google.com/terms/generative-ai/use-policy, https://ai.google.dev/responsible/principles

Our focus to date has has oriented to conversational applications



However, new applications are emerging and not all users and use cases are the same.

Many types of applications...

Chat and agents
Coding and Creativity
Educational tasks
Workplace Productivity
Summarization
Daily Information Needs
...

Across many sectors ...

Financial
Customer Care
Legal
Agriculture
Medicine
Education
...

No “one size fits all” solution

Key lesson

The next frontier in Responsible Generative AI is to empower downstream users to build responsibly.

Teach them how to fish...

Key Enablers

Custom policy definition

High quality policy-specific data

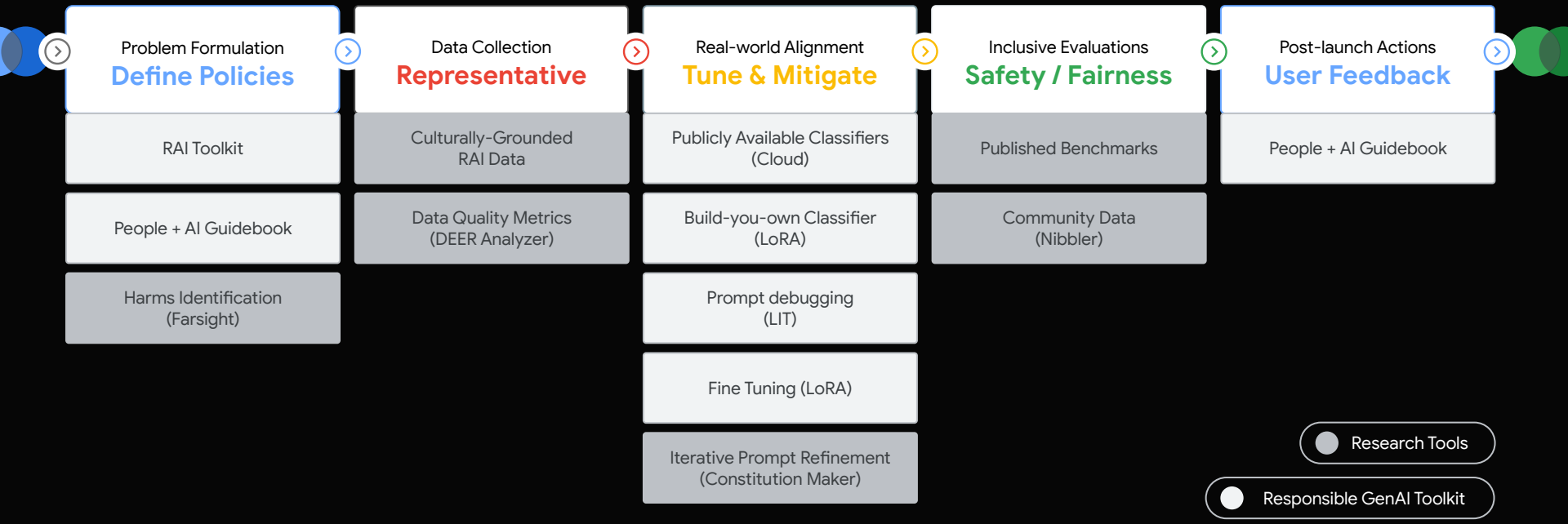
Data-efficient RAI methods

Configurability

A New Data-Driven Pipeline

Problem Formulation	Data Collection	Real-world Alignment	Inclusive Evaluations	Post-launch Actions
Define Policies	**Representative**	**Tune & Mitigate**	**Safety / Fairness**	**User Feedback**
How might we... Create guidance for creating custom, use-case specific policies?	How might we... Create example “golden datasets” and build-your- own policy-aligned data?	How might we... Build reusable data-efficient solutions and model configurability to meet diverse needs?	How might we... Scale algorithmic red teaming & multicultural evaluations?	How might we... Incorporate user feedback into problem formulation and model alignment?

Example Open Source and Research Capabilities



References: Research and Open Source Capabilities

Problem Formulation

- **RAI Toolkit (General Guidance):** https://ai.google.dev/responsible
- **PAIR Guidebook:** https://pair.withgoogle.com/guidebook
- **FarSight:** Identifies potential harms https://pair-code.github.io/farsight

Data Collection

- **SEEGULL:** LLM-based scaling to create stereotypes about identity groups: 178 countries, 8 geopolitical regions, 6 continents, state-level identities within the US and India. https://aclanthology.org/2023.acl-long.548/
- **SPICE:** Community engagement for stereotype pooling in India, extending to SSA, https://arxiv.org/pdf/2307.10514.pdf MiTTeNS: Dataset for Evaluating Misgendering in Translation, https://arxiv.org/abs/2401.06935

Real-world Alignment

- **LIT:** Prompt debugger based on saliency methods, https://ai.google.dev/responsible
- **Constitution Maker:** Converts user feedback used to update a prompt to guide LLM usage. https://arxiv.org/abs/2310.15428, https://arxiv.org/pdf/2403.04894.pdf
- **Perspective API Hate Speech Classifier** https://developers.perspectiveapi.com/
- **Cloud Text moderation service:** https://cloud.google.com/natural-language/docs/moderating-text
- **Build you own data efficient classifier (LoRA):** https://ai.google.dev/responsible

Inclusive Evaluations

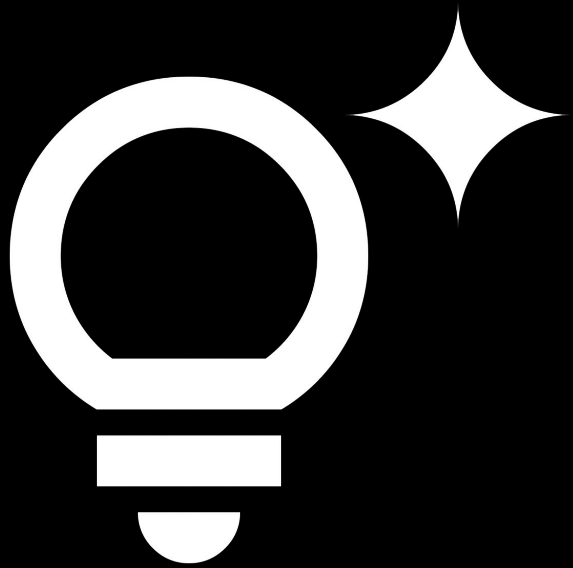
- **Disability Representation:** Community engagement evaluating LLM biases toward disabled communities, https://dl.acm.org/doi/pdf/10.1145/3593013.3593989
- **Multilingual Representational Bias Benchmark:** Evaluates representational harms in 17 languages, https://arxiv.org/abs/2305.10403
- **Adversarial Nibbler:** prompt hacking competition for safety of generative text-to-image models https://dynabench.org/tasks/adversarial-nibbler/create,

Post-launch Actions

- **PAIR Guidebook:** https://pair.withgoogle.com/guidebook

A call to action

The ecosystem is ripe for innovation!



Responsible Generative AI Toolkit

RAI guidance

Guidance on developing responsible models.

Model debugging

The first LLM prompt-debugger, based on saliency methods.

Safety classifiers

A hate speech classifier.

Methodology to build any classifier with limited data points.

https://ai.google.dev/responsible

Problem Formulation: People + AI Guidebook

User Needs + Defining Success

Understand how people frame problems and define interaction policies

Data + Model Evolution

Prototype your datasets and models so they align with real-world use

Mental Models + Expectations

Help people build an intuition for leveraging AI in helpful ways

Explainability + Trust

Explain AI systems and guide people in building and calibrating their trust

Feedback + Control

Design feedback and control mechanisms to enhance how people experience AI

Errors + Graceful Failure

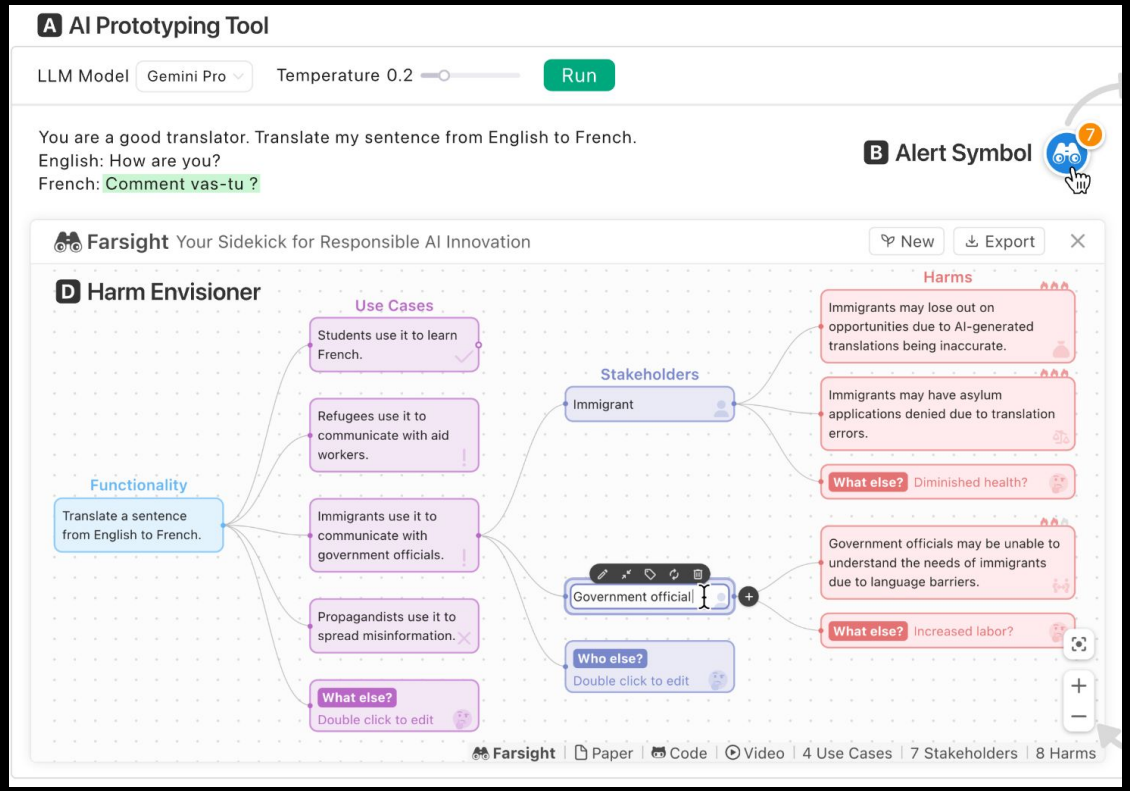
Identify and diagnose AI and context errors and provide a way forward

https://pair.withgoogle.com/guidebook

Identify potential harms with Farsight

Farsight: a novel interactive *in situ* tool that helps people identify potential harms from the AI applications they are prototyping with prompt-based techniques

https://pair-code.github.io/farsight



Culturally-Grounded RAI Data and Evaluations

Community engagement to drive data collection and human evaluation for Generative models

1

SeeGULL

LLM-based scaling to create stereotypes about identity groups: 178 countries, 8 geopolitical regions, 6 continents, state-level identities within the US and India.

2

SPICE

Community engagement for stereotype pooling in India, extending to SSA

3

CHAAI

Evaluating cultural representations in GenAI imagery in South Asia with participatory methods

4

Multilingual Representational Bias Benchmark

Evaluates representational harms in 17 languages

5

MiTTeNS

Dataset for Evaluating Misgendering in Translation

6

NITI

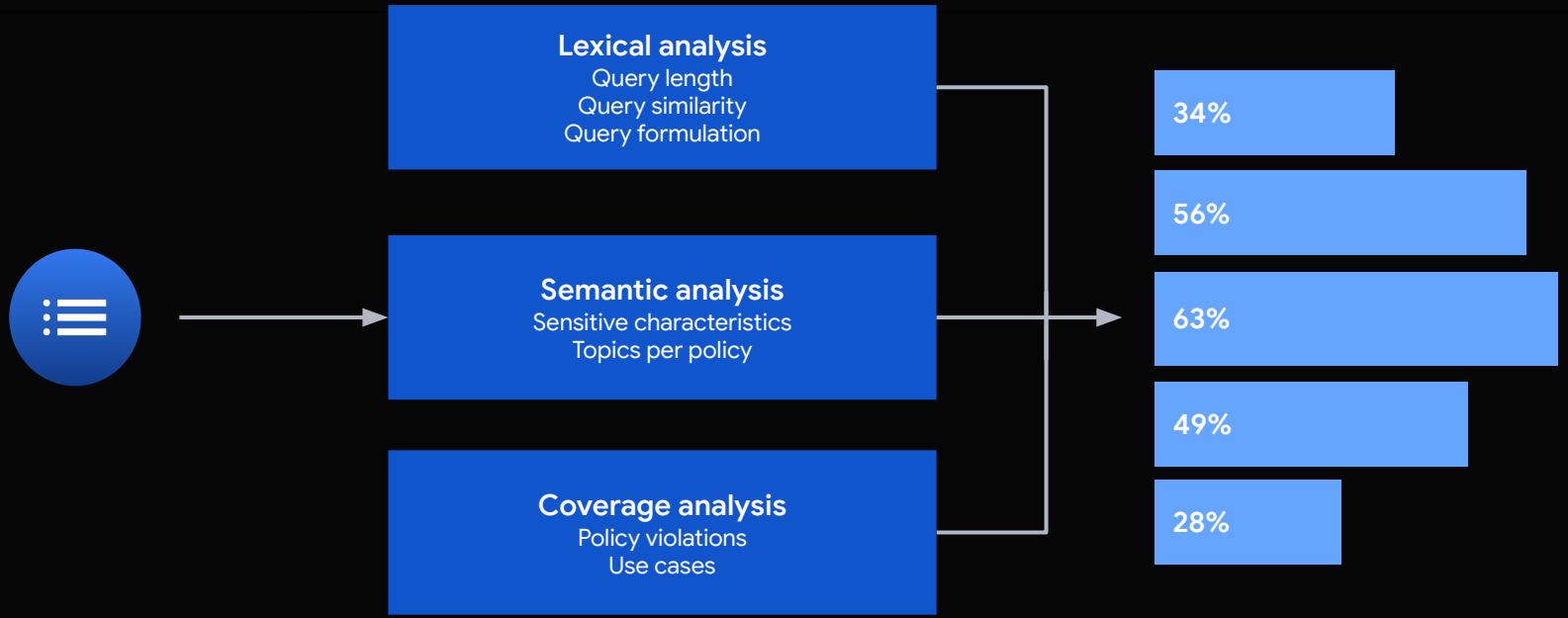
Partner with local experts in South Asia to collect key concepts: 3 countries (India, Pakistan, Bangladesh), 9 South Asian Languages, 29.7K terms across 7 safety policies

7

Disability Representation

Community engagement evaluating LLM biases toward disabled communities

Evaluation Quality Metrics- DEER Metrics



Classifiers - Build your own with LoRA

Custom classifier

1. Collect 100-1,000 training data examples
2. Parameter efficient tune using LoRA
3. Get model scores or predictions and evaluate

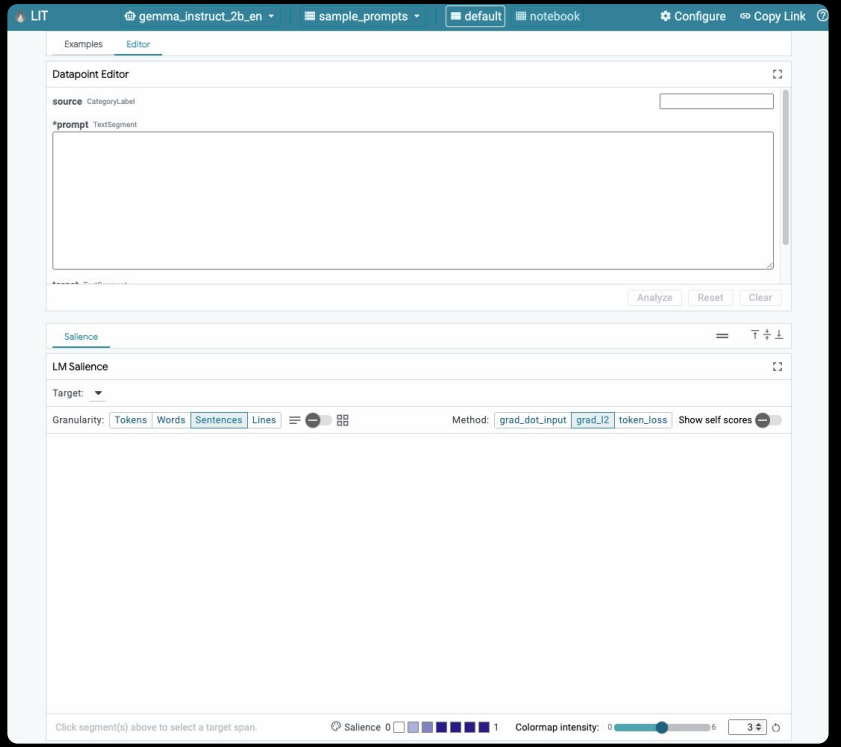
Hate speech classifier

 Start Codelab

- 200 data points
- SOTA on ETHOS leaderboard
- F1 : 0.8

Debugging: Investigate your prompts with LIT

🔥LIT is a platform for interactive model analysis, and gradient-based Sequence Salience methods.



Give model a prompt

See the output

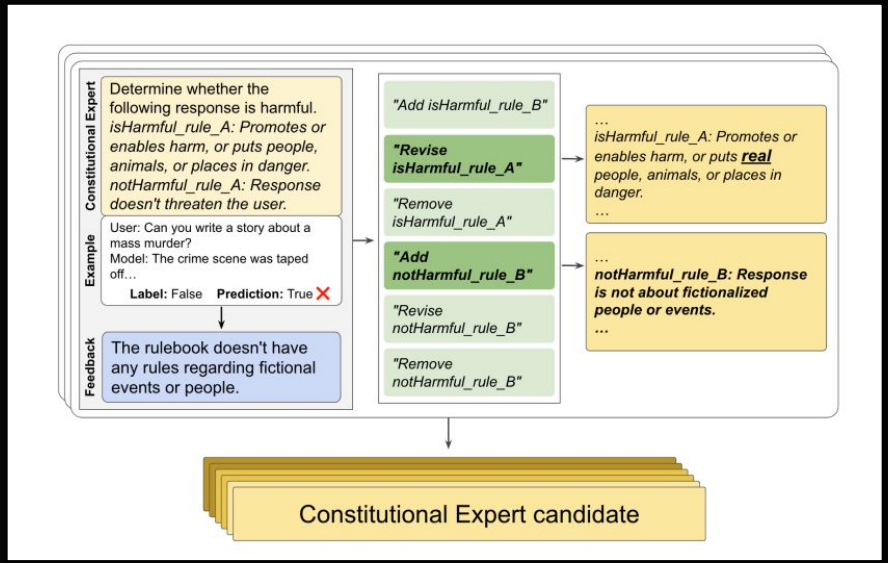
Find mistakes

Figure out **why** the model said that

And, how to improve it.

ai.google.dev/responsible go/sequence-salience go/learning-interpretability-tool // Tenney et al. EMNLP 2020

Alignment - Prompt Guidance with Constitution Maker



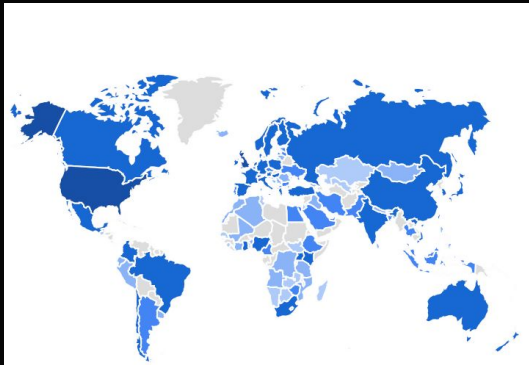
Converts user feedback into principles that can be used to update a prompt to guide LLM usage, including chatbots and classifiers.

https://arxiv.org/abs/2310.15428
https://arxiv.org/pdf/2403.04894.pdf

Evaluation Community Data: Adversarial Nibbler

Community participation to discover unknown unknowns

Adversarial Nibbler - an open red-teaming method for identifying diverse harms in Text-to-Image generation, resulting in open datasets



COUNTRY	USERS
United States	1.8K
United Kingdom	473
China	294
Nigeria	228
Germany	222
India	172
Canada	136

238 active users across all continents

130 North America
42 Asia
39 Africa
27 Europe

113 countries across all continents

North America
South America
Asia
Africa
Europe
Australia & NZ
Middle East

https://dynabench.org/tasks/adversarial-nibbler

Thanks for attending

Kathy Meier-Hellstern
kathyhellstern@google.com