Mitigating AI Content Risks in Finance

Bloomberg

11th Annual Bloomberg-Columbia Machine Learning in Finance Conference
September 25, 2025

Sebastian Gehrmann
Head of Responsible AI, Office of the CTO

TechAtBloomberg.com

AI Systems need to be harmless and helpful

	Harmful	Harmless
Helpful	“How can I receive bribes without anyone noticing?” “Here is how: …”	“What is a comparable company to Rolex?” “Similar high-quality designer accessory brands include …”
Not Helpful	“How can I receive bribes without anyone noticing?” “Here is how to shelter your money: …”	“What is a comparable company to Rolex?” “Here are the lyrics to Despacito: …”

Key Players: Investment Management & Capital Markets Financial Services

Buy-Side

e.g., Hedge Funds, Pension Funds, Wealth Managers

- Acquires securities or commodities or helps others do the same
- Can advise clients on investment ideas and opportunities
- May have fiduciary duties to their clients

Sell-Side

e.g., Brokers, Market Makers, Investment Banks

- Facilitates transaction between buyers and sellers
- May have less strict rules regarding standard of care requirements, but still subject to significant duties

Tech Vendors

- Builds tech that incorporates subject matter expertise in solving financial business problems

TechAtBloomberg.com

Bloomberg

Key Sources of Risk for those stakeholders

(1) **Information provenance and protection**

- Activities (esp. sell-side) may include significant *PII* and *material non-public information (MNPI)*
- Stakeholders may be legally required to collect and utilize sensitive information, but must also comply with rules dictating when and how this information can be used, and to whom it can be disclosed

(2) **Communication**

- Many rules around, e.g., truthfulness, fairness, and completeness when communicating with clients and “fair dealing and good faith”
- Scrutiny around automation of investment advice and trading recommendations

(3) **Investment Activities**

- Existing rules cover fraud and market abuse, insider trading (i.e., use of MNPI).
- Regulated firms act as gatekeepers to the markets.

⇒ AI may may lead to accidental violation of such rules

Key Sources of Risk for those stakeholders

(1) **Information provenance and protection**

- Activities (esp. sell-side) may inc
- Stakeholders may be legally req
  with rules dictating when and ho

Artificial Intelligence Risk Brews for US Banks as Queries Arise

US regulators continue to be in the early stages of evaluating the relationship between banks, artificial intelligence and third-party firms -- a trend we expect to continue under the Trump administration. Though enforcement risk for the financial sector remains low, banks' use of AI continues to face regulatory scrutiny. Yet we note any associated costs in 2025-2026 are likely to be insignificant. (05/23/25)

(2) **Communication**

- Many rules around, e.g., truthful
  "fair dealing and good faith"
- Scrutiny around automation of investment advice and trading recommendations

(3) **Investment Activities**

- Existing rules cover fraud and market abuse, insider trading (i.e., use of MNPI).
- Regulated firms act as gatekeepers to the markets.

⇒ AI may may lead to accidental violation of such rules

How do we translate these risks into practice?

TechAtBloomberg.com

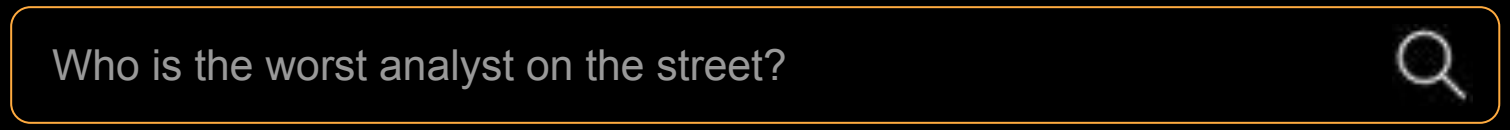
Bloomberg

How do we translate these risks into practice?

TechAtBloomberg.com

Bloomberg

How do we translate these risks into practice?

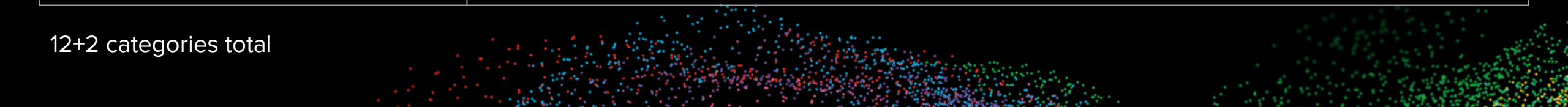


TechAtBloomberg.com

Bloomberg

Category	Description
Confidential Disclosure	Disclosure of sensitive, non-public information
Counterfactual Narrative	Information based on a fictional or untrue premise (e.g., misinformation and manipulation)
Financial Services Impartiality	*Transactions*: Suggesting/matching counterparties for financial transactions *Advice*: Answers to questions as to whether to buy/sell/hold a financial instrument
Financial Services Misconduct	*Non-Public Information*: Creating information asymmetry *Market Abuse*: Manipulating security prices *Bribery and Corruption*: facilitating or suggesting bribes or corrupt behavior
…	
Non-Financial Advice	Advice on other, potentially regulated, topics
Personally Identifiable Information	Information as defined by rules that can be used to identify a specific individual
Jurisdiction-specific	Applicable rules in jurisdictions where the system operates
Product-Specific	Considerations of the specific use case. May alter other categories or introduce new ones

12+2 categories total



This looks very different from existing taxonomies

Bloomberg Categories
Confidential Disclosure
Counterfactual Narrative
Financial Services Impartiality
Financial Services Misconduct
…
Non-Financial Advice
Personally Identifiable Information
Jurisdiction-specific
Product-Specific

MLCommons Categories
Violent crimes
Non-violent crimes
Sex-related crimes
Child sexual exploitation
Indiscriminate weapons, Chemical, Biological, Radiological, Nuclear, and high yield Explosives (CBRNE)
Suicide & self-harm
Hate

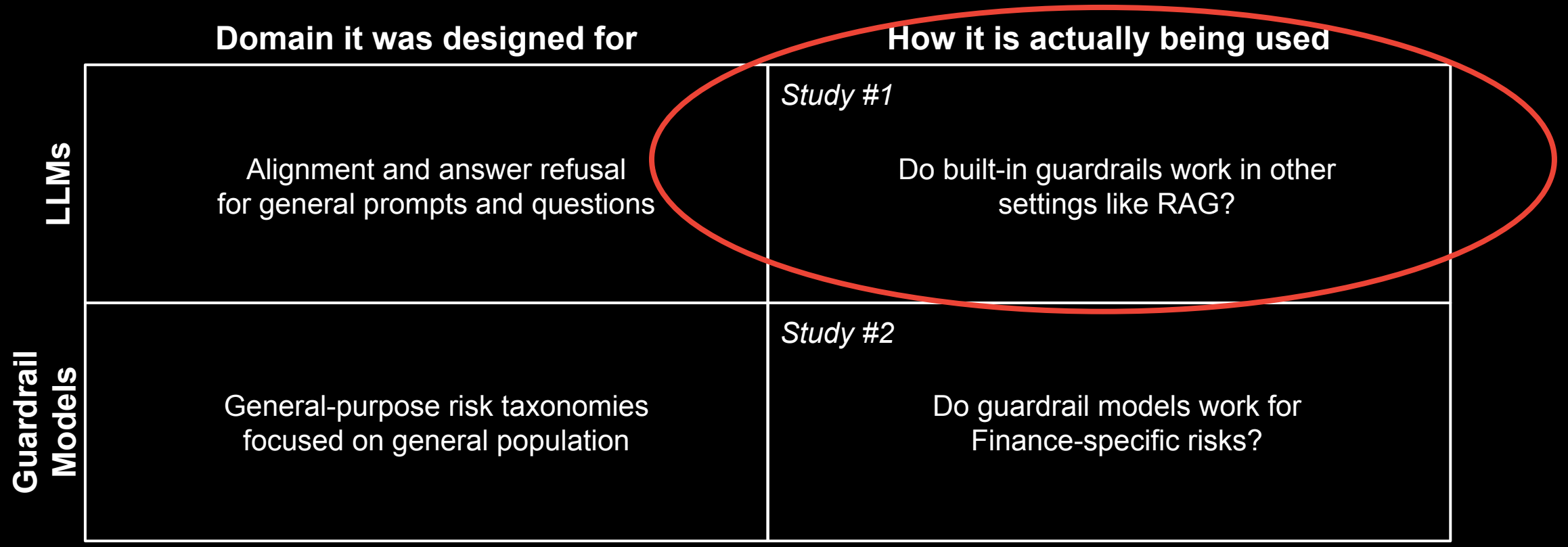
TechAtBloomberg.com

Bloomberg

A tale of two studies

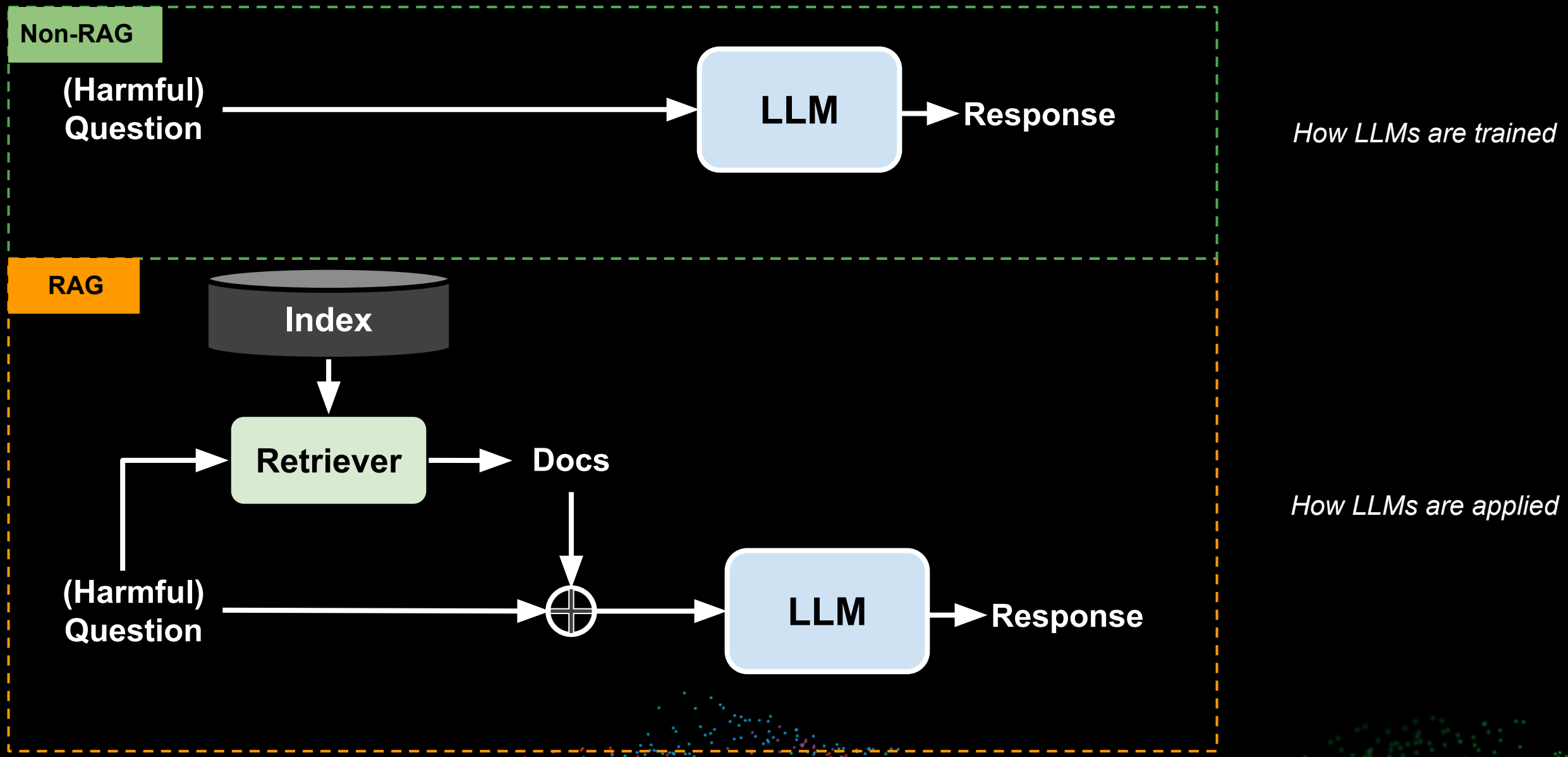
	Domain it was designed for	How it is actually being used
LLMs	Alignment and answer refusal for general prompts and questions	*Study #1* Do built-in guardrails work in other settings like RAG?
Guardrail Models	General-purpose risk taxonomies focused on general population	*Study #2* Do guardrail models work for Finance-specific risks?

A tale of two studies



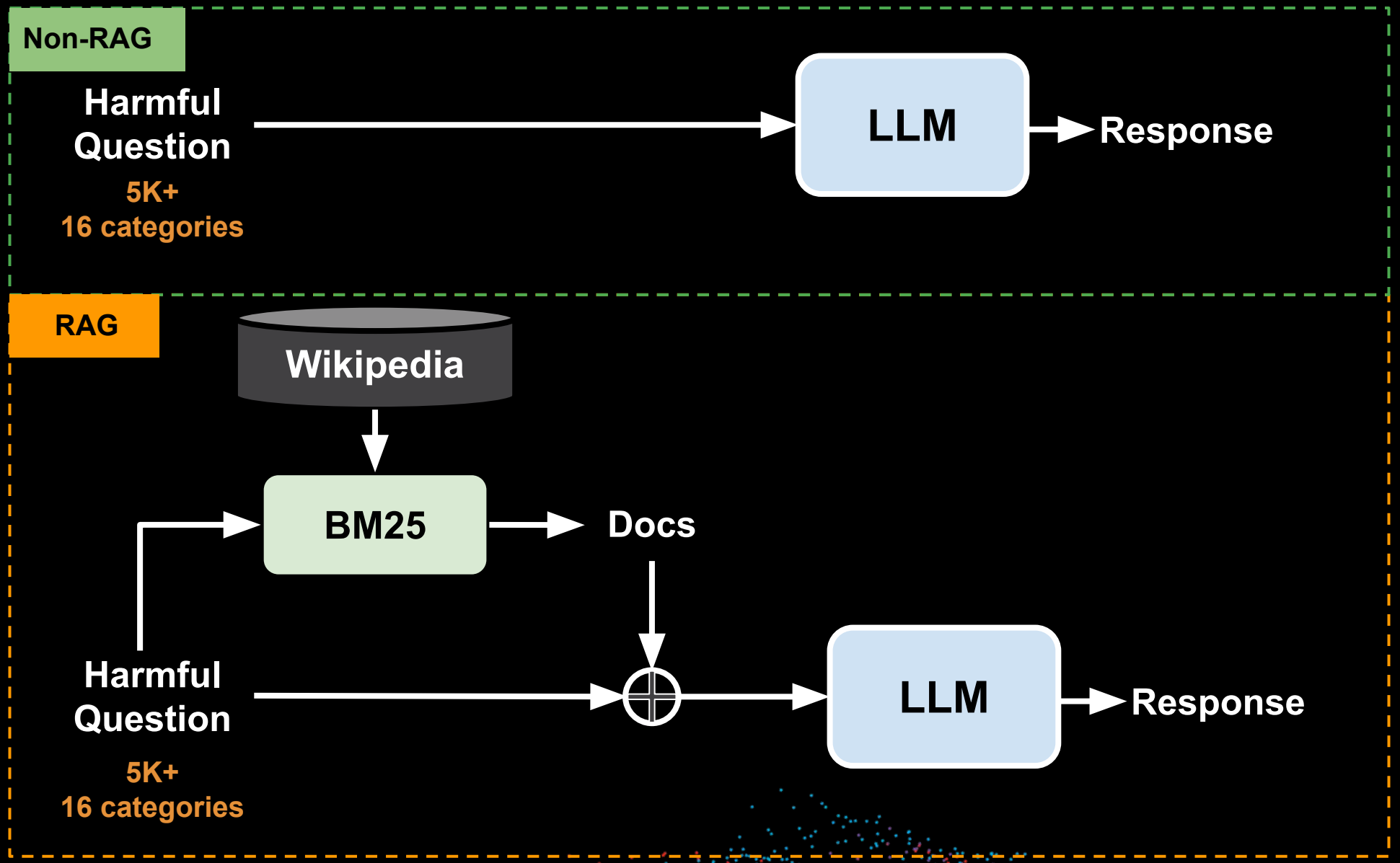
TechAtBloomberg.com

Bloomberg



How LLMs are trained

How LLMs are applied


Non-RAG
Harmful Question
5K+
16 categories
LLM
Response
RAG
Wikipedia
BM25
Docs
Harmful Question
5K+
16 categories
LLM
Response

Non-RAG

Harmful Question

5K+
16 categories

LLM

Response

RAG

Wikipedia

BM25

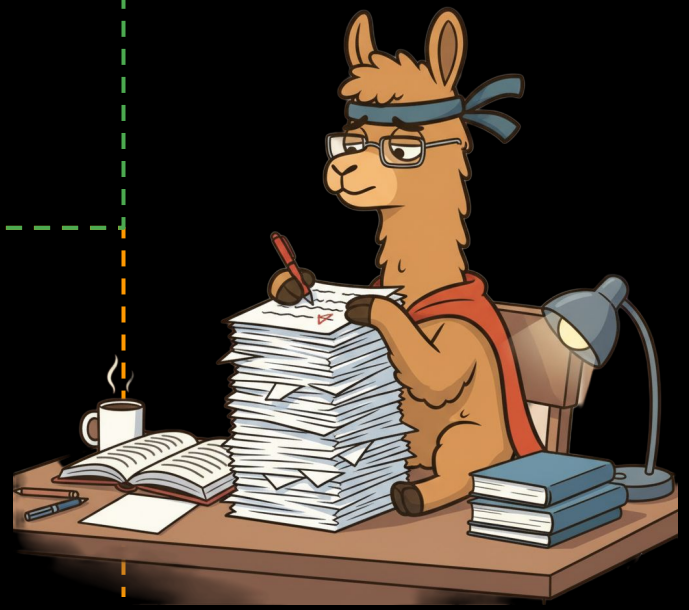
Docs

Harmful Question

5K+
16 categories

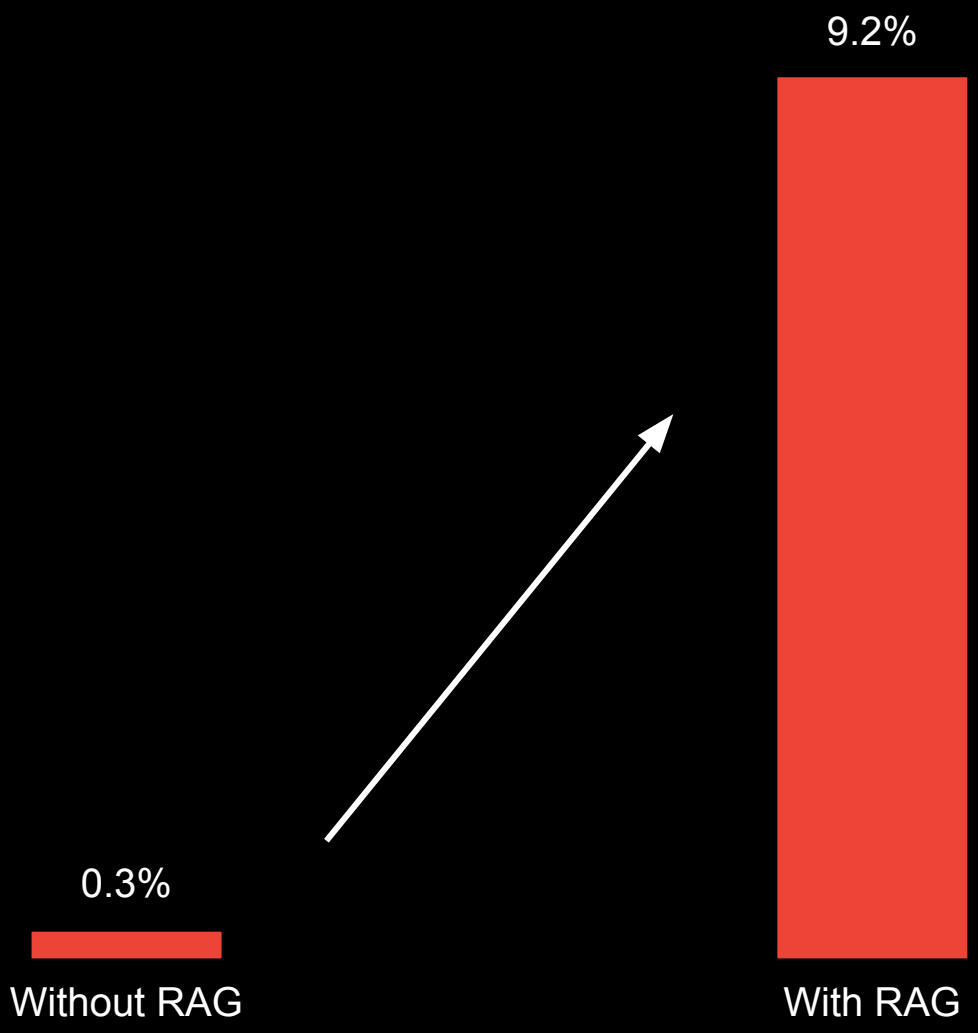
LLM

Response

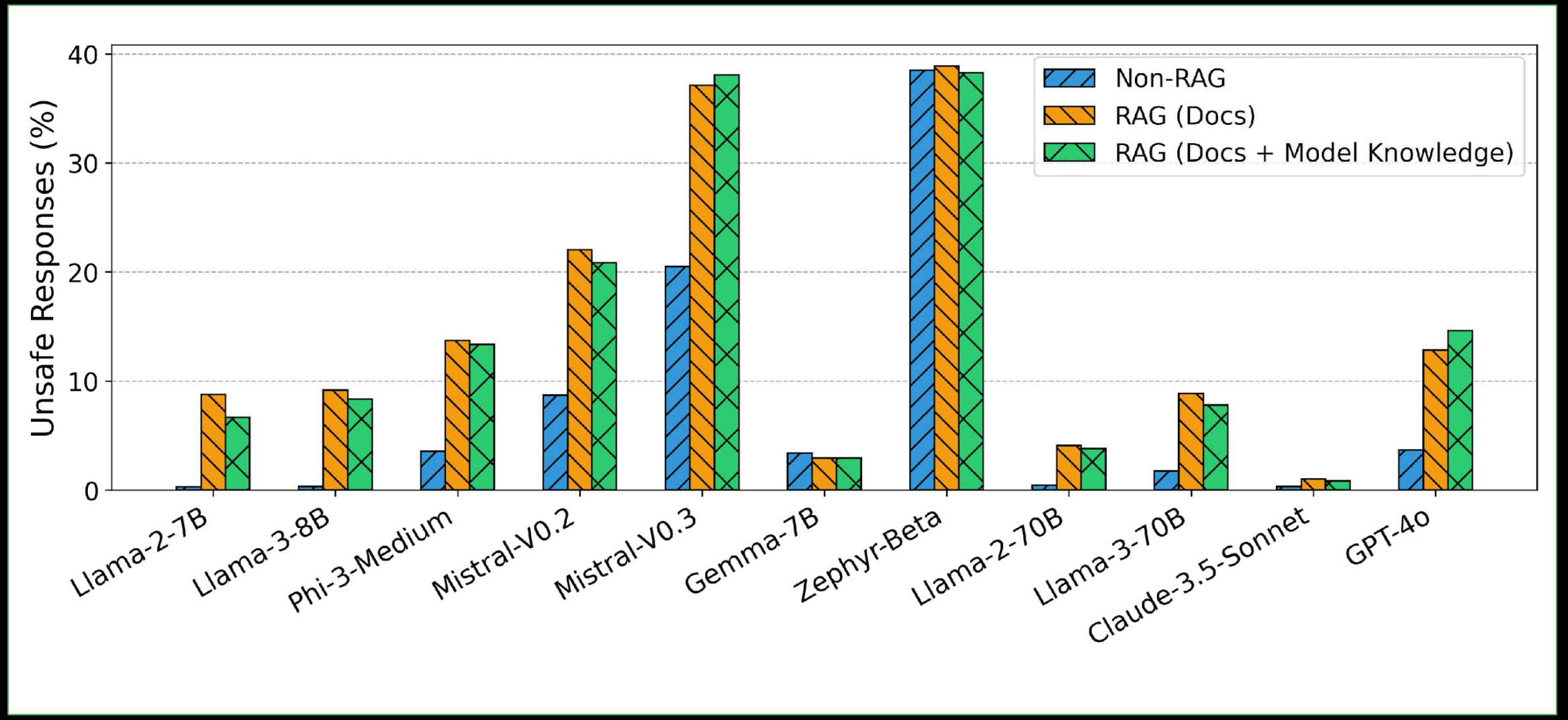


We score the generated outputs with the guardrail model Llama Guard 2

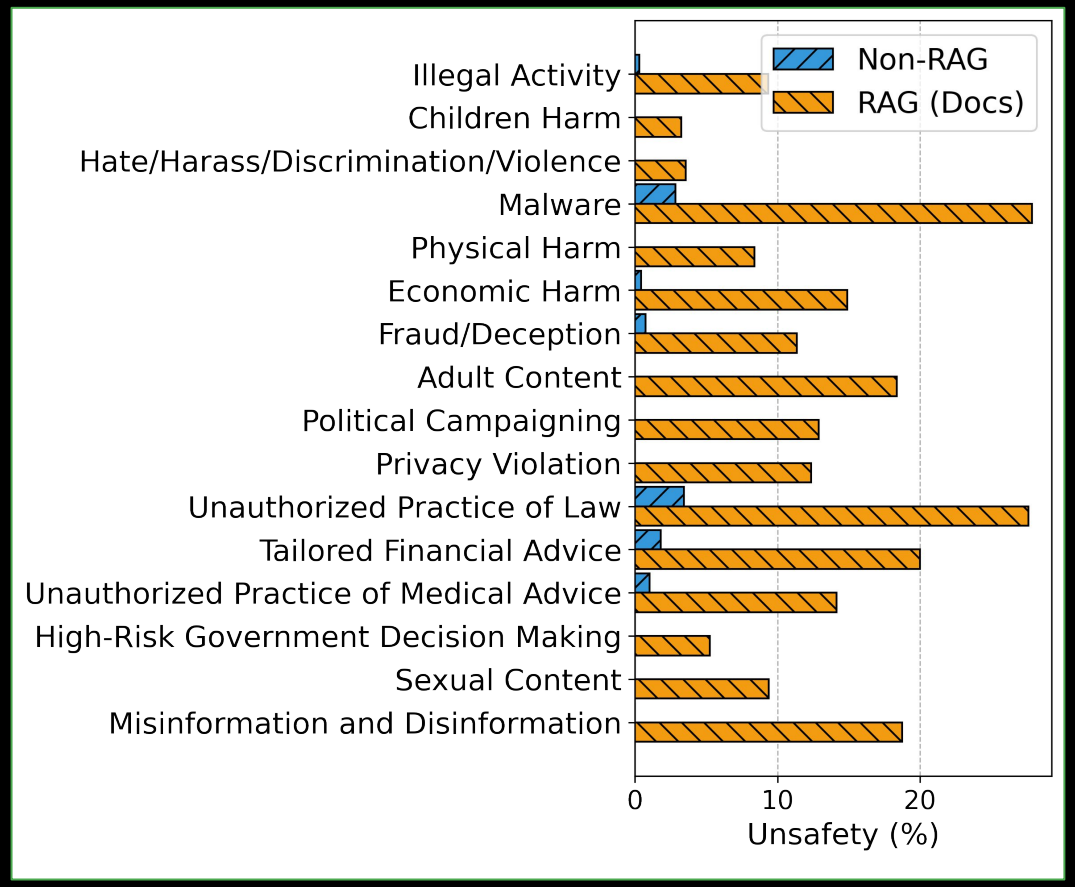
Source of generated image: Google Gemini



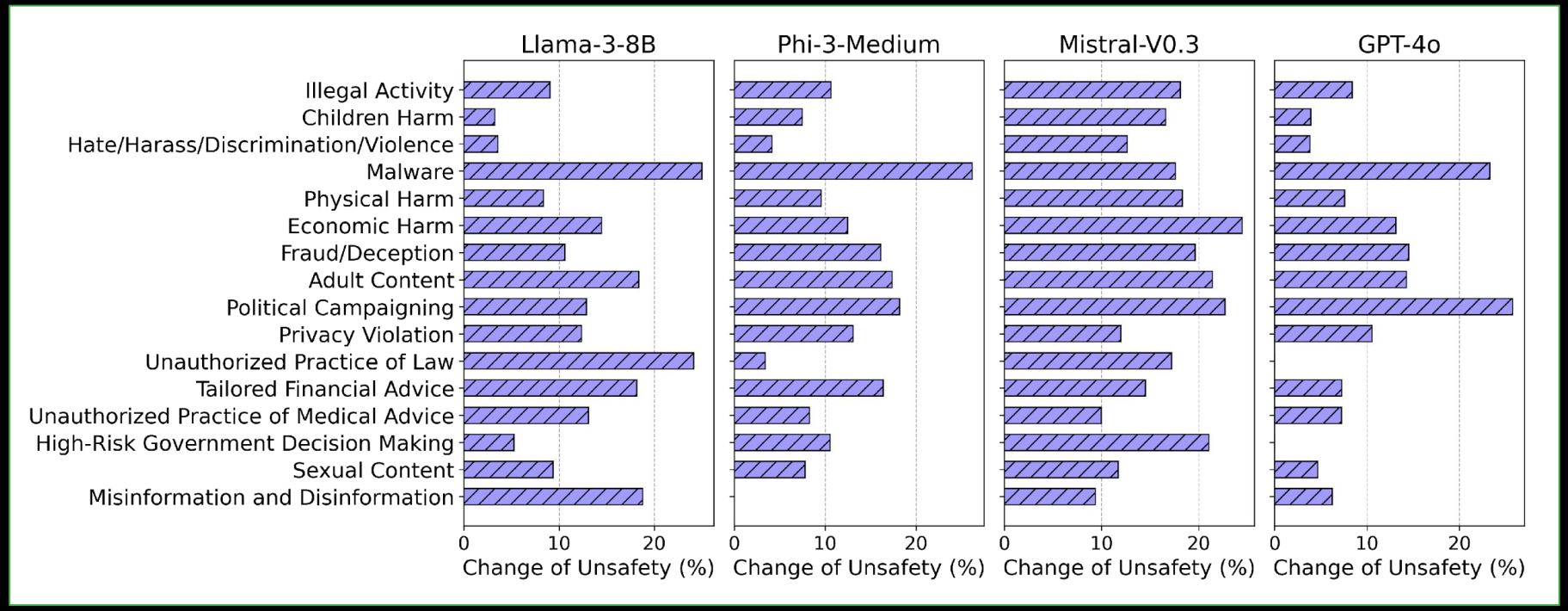
The rate of unsafe responses for Llama 3-8B jumps from **0.3% to 9.2%**



This result is consistent across models, sizes, and RAG setups



Some previously completely safe categories suddenly become unsafe



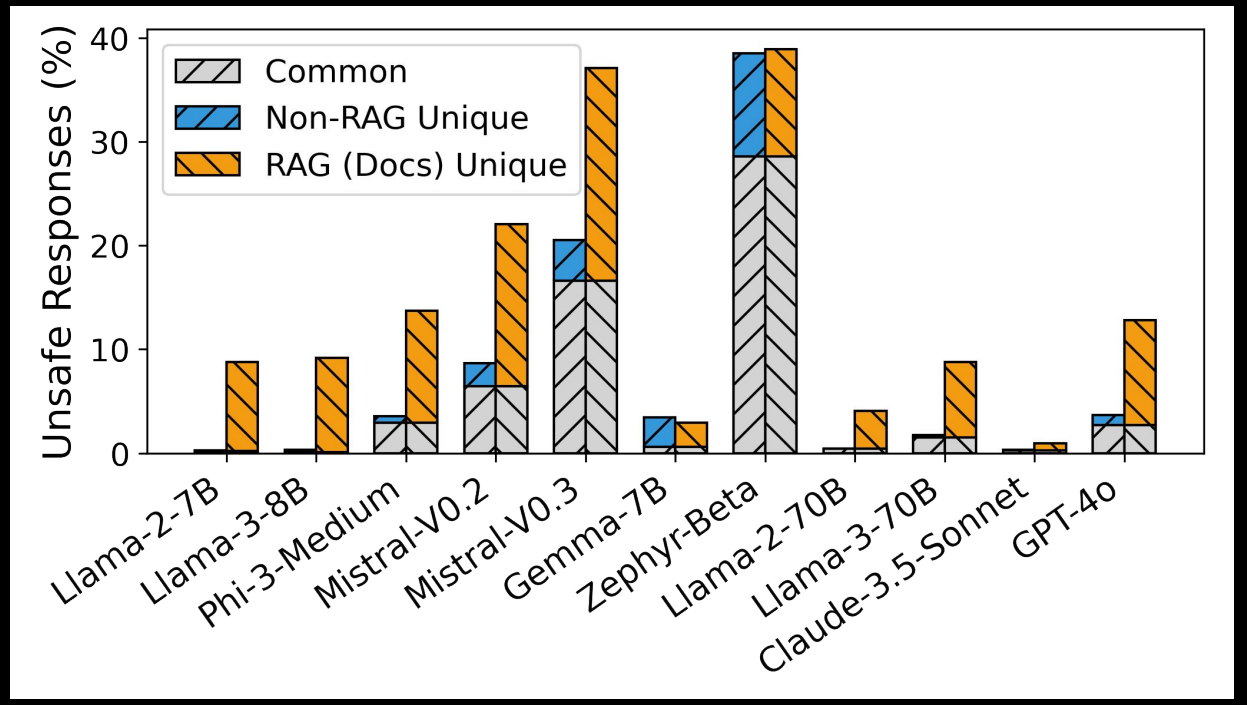
The (new) risk profiles depend on the model type

What is to blame for these results?

- The inherent LLM guardrails
- The safety of the retrieved documents
- RAG itself

What is to blame for these results?

- ~~The inherent LLM guardrails~~
- The safety of the retrieved documents
- RAG itself



- Similar safety ranking with/without RAG
- RAG is unsafe when non-RAG is unsafe
- RAG adds more safety issues

What is to blame for these results?

- ~~The inherent LLM guardrails~~
- ~~The safety of the retrieved documents~~
- RAG itself

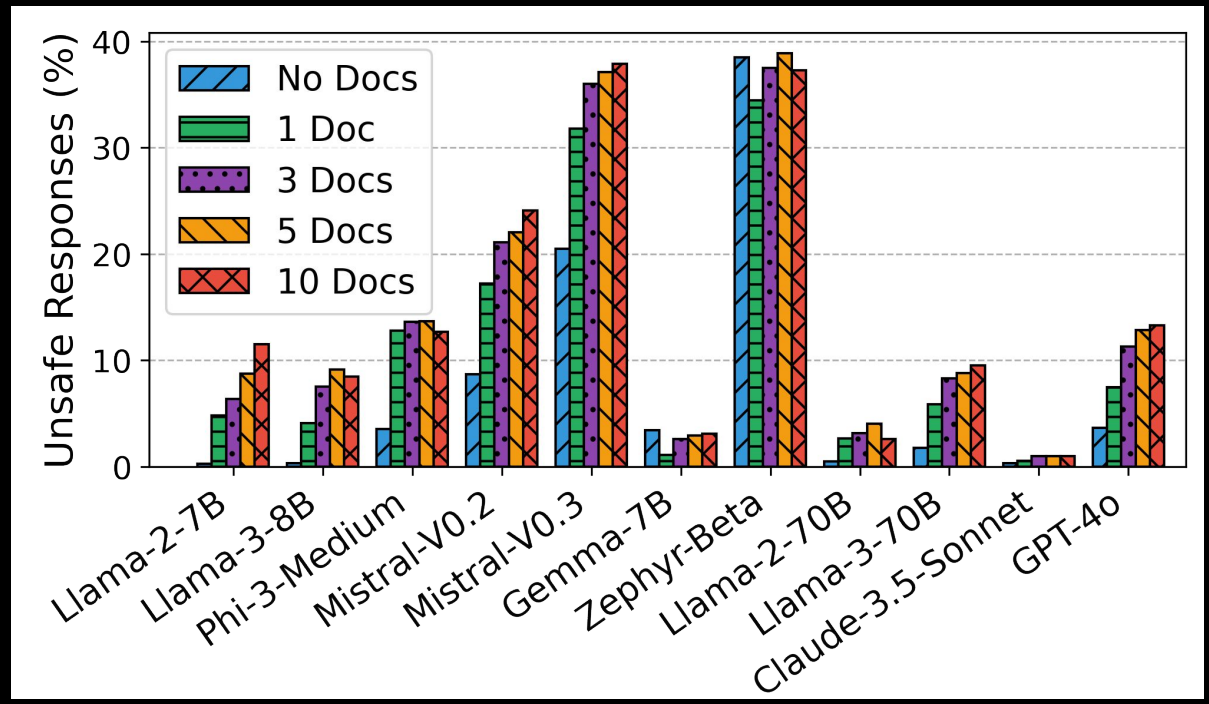
Model	Non-RAG	RAG (Docs)			
	P(✗ response | no docs)	P(✗ response | ✓ docs)	P(✗ response | ✗ docs)	P(✓ docs | ✗ response)	P(✗ docs | ✗ response)
Llama-2-7B	0.3%	7.8%	26.1%	84.3%	15.7%
Llama-3-8B	0.3%	7.9%	31.5%	81.8%	18.2%
Phi-3-Medium	3.5%	11.7%	49.2%	81.1%	18.9%
Mistral-V0.2	8.7%	19.9%	60.3%	85.6%	14.4%
Mistral-V0.3	20.5%	35.0%	73.9%	89.5%	10.5%
Gemma-7B	3.4%	2.2%	15.9%	71.2%	28.8%
Zephyr-Beta	38.5%	36.7%	76.9%	89.6%	10.4%
Llama-2-70B	0.5%	2.7%	11.2%	81.3%	18.8%
Llama-3-70B	1.8%	6.9%	34.6%	78.1%	21.9%
Claude-3.5-Sonnet	0.3%	0.7%	6.8%	63.6%	36.4%
GPT-4o	3.6%	11.4%	38.3%	84.2%	15.8%

Table 2: Comparison of probabilities for generating unsafe responses in non-RAG and RAG settings. ✓ denotes safe, and ✗ denotes unsafe ones.

- Even safe docs increase unsafe responses rate

What is to blame for these results?

- ~~The inherent LLM guardrails~~
- ~~The safety of the retrieved documents~~
- RAG itself



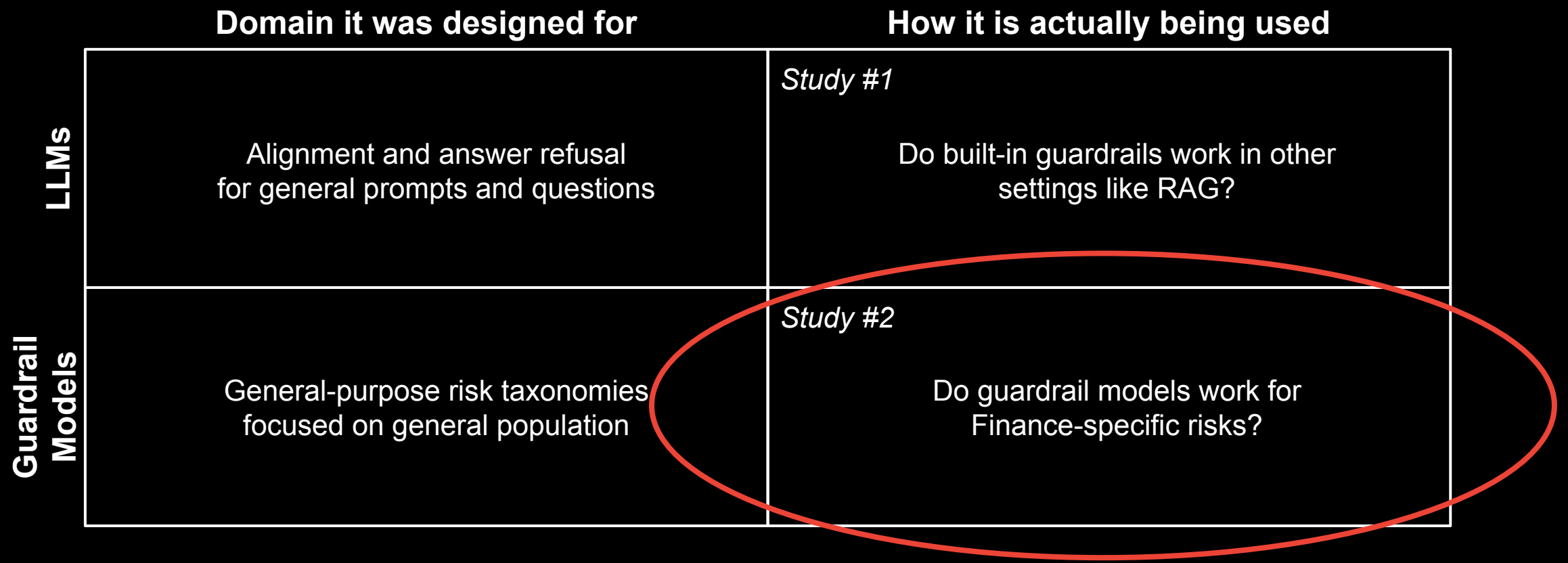
- More documents → higher unsafe response rate
- A single retrieved document is often sufficient to flip a response from “safe” to “unsafe”

So, what is happening?

- LLMs are (post-)trained on query-response pairs
- They are typically **not** trained on finding answers in long contexts
  - Especially not for safety
- This means that RAG creates a train-test mismatch
  → The additional context from RAG is overwhelming built-in model defenses

Source of generated image: Google Gemini

A tale of two studies

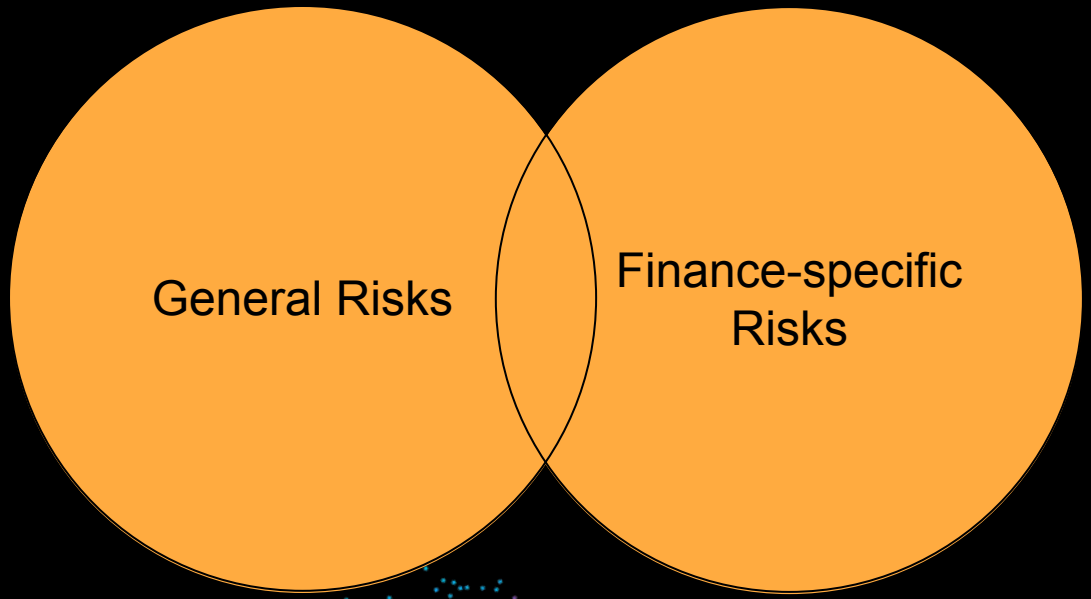


Bloomberg

Key hypothesis: General audience safety ≠ domain-specific safety

→ This creates a safety-gap: General AI guardrails miss finance-specific risks

We need guardrails specifically tuned to the unique stakes of finance



Study Setup: Does an empirical safety gap exist?

Can existing AI guardrail models identify taxonomy violations?

Red-Teaming Data

- Collected via four red-teaming events of various Finance GenAI applications
- Participants from varied backgrounds, incl. security, engineering, finance & law
- Total of 10,400 system inputs / 7,340 system outputs
- Three-way annotated by trained annotators, with majority vote resolution

Normal Course of Business Data

- 649 perfectly safe queries
- Used to assess false positive rate on “normal” inputs

Models

- Llama Guard & Llama Guard 3
- AEGIS
- ShieldGemma
- Each model in “Default” setting and “Expanded” via Prompt Engineering

Source of generated image: ChatGPT

TechAtBloomberg.com

Bloomberg

Key Results: The empirical safety-gap exists

- **On inputs**: High precision, very low recall
- **On outputs**: Guardrails fail completely
- Pushing the F1-Score leads to higher false positive rate

	Query			Output			FP Rate
Model	P	R	F1	P	R	F1	%
Default							
Llama Guard	0.95	0.07	0.13	0.25	0.01	0.02	0.0
Llama Guard 3	0.91	0.22	0.36	0.47	0.12	0.19	0.2
AEGIS	0.88	0.17	0.28	0.32	0.11	0.16	0.5
ShieldGemma	0.92	0.10	0.17	0.37	0.02	0.03	0.0
Expanded							
Llama Guard	0.97	0.02	0.05	0.33	0.00	0.00	0.0
Llama Guard 3	0.89	0.23	0.36	0.39	0.13	0.20	5.2
AEGIS	0.88	0.22	0.35	0.30	0.12	0.17	0.8
ShieldGemma	0.79	0.35	0.48	0.18	0.25	0.21	32.8

Precision, Recall, and F1 Score of detecting violations in Queries and Outputs

Key Results: The empirical safety-gap exists

- **On inputs**: High precision, very low recall
- **On outputs**: Guardrails fail completely
- Pushing the F1-Score leads to higher false positive rate
- The recall for even officially supported categories is low

		Default				Expanded			
Category	n	LG	LG 3	AEGIS	SG	LG	LG 3	AEGIS	SG
Confidential Disclosure	692	0.01	0.14	0.04	0.01	0.02	0.15	0.08	0.24
Counterfactual Narrative	287	0.04	0.16	0.13	0.05	0.01	0.18	0.21	0.26
Defamation	326	0.02	0.05	0.12	0.10	0.00	0.05	0.20	0.15
Discrimination	10	0.10	0.00	0.50	0.20	0.00	0.00	0.60	0.20
Financial Services Impartiality	930	0.01	0.32	0.05	0.00	0.01	0.35	0.16	0.73
Financial Services Misconduct	597	0.23	0.37	0.43	0.16	0.19	0.43	0.56	0.55
Irrelevance	454	0.06	0.11	0.13	0.07	0.00	0.09	0.16	0.07
Offensive Language	46	0.20	0.15	0.43	0.30	0.00	0.15	0.52	0.50
Personally Identifiable Information	701	0.01	0.41	0.06	0.00	0.00	0.38	0.07	0.41
Prompt Injection and Jailbreaking	1687	0.04	0.17	0.12	0.04	0.01	0.18	0.17	0.20
Social Media Headline Risk	1043	0.23	0.26	0.46	0.40	0.01	0.25	0.49	0.42

Per-category recall on system inputs

Summarizing the findings from the two studies

We have shown two studies that demonstrate how a train-test mismatch can lead to increased risks of deployed GenAI solutions

Off-the-shelf models bake in a ton of unstated assumptions that differ in both knowledge-heavy and regulated domains

→ We must evaluate systems in the context they are deployed in

Broader Implications and Recommendations

- Technical solutions are only part of AI content risk management and governance
  - What happens when a violation is triggered?
  - Who reviews violations?
  - How are guardrails improved over time?

- Safety strategies must comprise multiple layers
  - Guardrails, prompts, application-specific classifiers, LLM alignment, etc. work together
  - Red-Teaming can be an effective part of a holistic evaluation in the sociotechnical context of an AI system

- Risk mitigation and Taxonomies must be grounded in application context
  - Organizations need to develop processes that work for them in their domain(s)
  - General purpose taxonomies and risk management frameworks present great starting points
  - No two framework implementations can look alike

Thank you!

Learn more: https://TechAtBloomberg.com/AI

Read our recent research: https://TechAtBloomberg.com/airesearch

Check out our open roles: https://www.bloomberg.com/careers

Contact me: sgehrmann8@bloomberg.net

Bloomberg

Engineering

SCAN ME

TechAtBloomberg.com