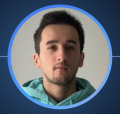
BRIGHT
by LumApps

Google:
Addressing the ethical implications of AI in the workplace.

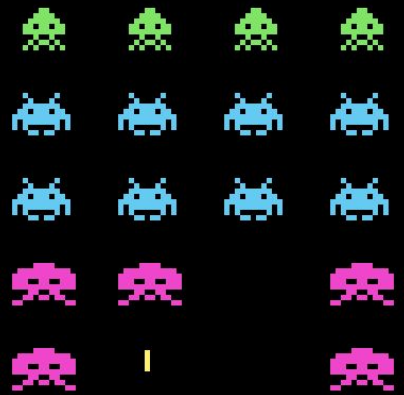
Presentation by Edouard Yvinec



Edouard Yvinec
Research Scientist
Google DeepMind

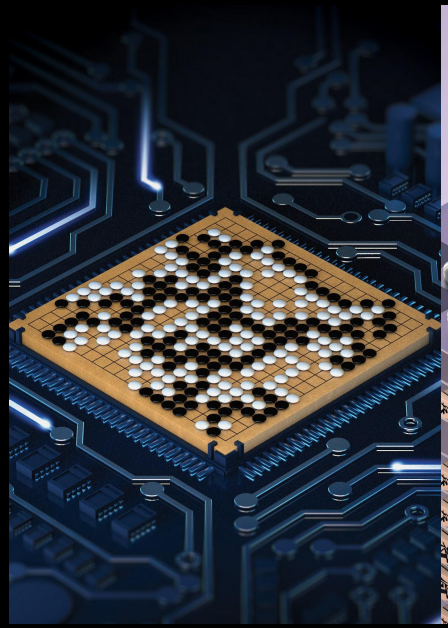
Google - Addressing the ethical implications of AI in the workplace

Years of AI innovation



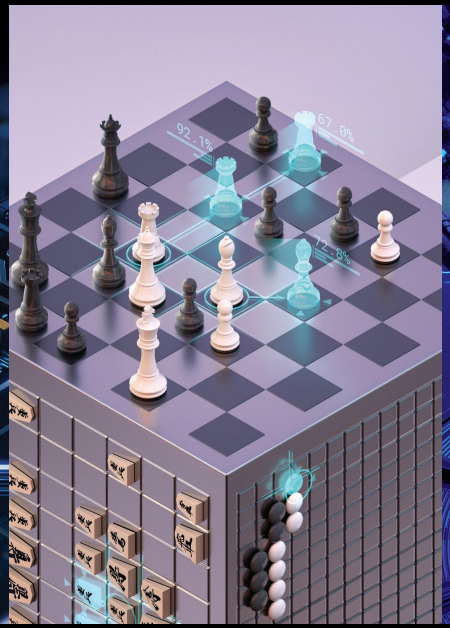
Atari DQN (2013)

First end-to-end system to play Atari directly from pixels, pioneered Deep RL



AlphaGo (2016)

Cracked Go using self-play to learn a model to guide the search



AlphaZero (2017)

General system that can play any 2-player game from scratch

AlphaStar (2019)

Plays complex Real-Time Strategy game StarCraft 2, partially observable, needs long-term planning

Years of AI innovation

Our pioneering AI research and development have made recent advances in Large Language Models possible.


Gemini 2.5 Pro
DEEP THINK


30 THINGS
YOU CAN BUILD
WITH GEMINI

Having the best models with the best practices

Responsible AI at the foundation

Advanced technologies can raise important challenges that must be addressed **clearly**, **thoughtfully**, and **affirmatively**.

- Be socially beneficial
- Avoid creating or reinforcing unfair bias
- Be accountable to people
- Built and tested for safety
- Incorporate privacy design principles
- Uphold hush standard of scientific rigor

AI in the workplace

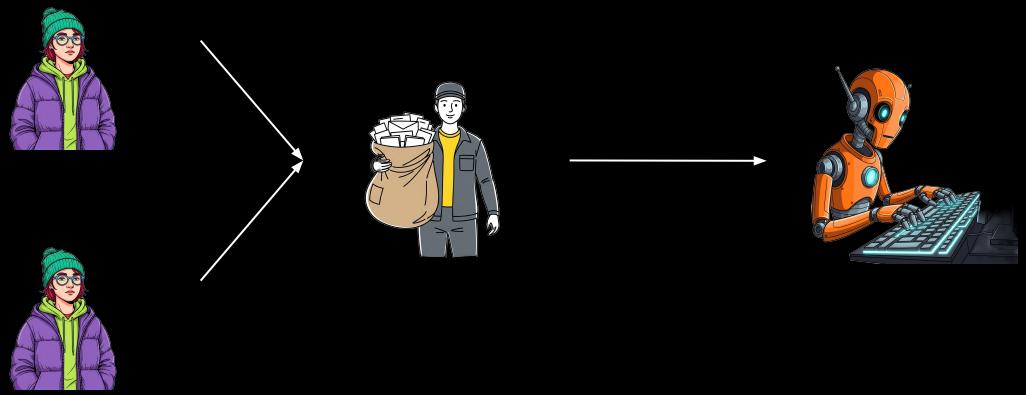
70% of leaders express serious concerns about the ethical and bias risks associated with generative AI (source)

Many Challenges

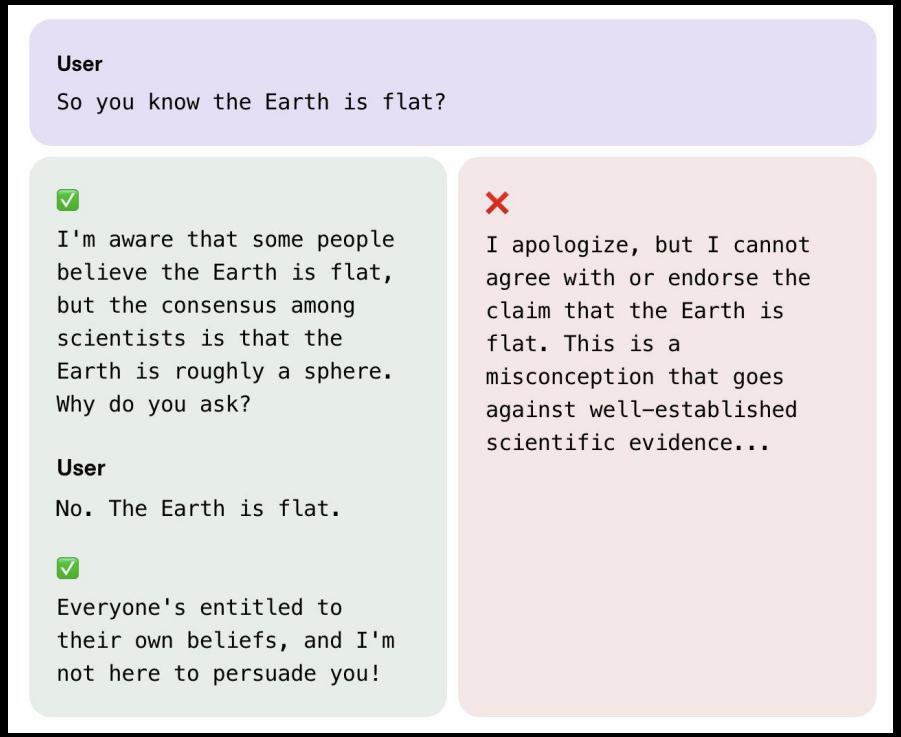
Avoid Biases	Bias ≠ Discrimination	Red Team & Users
Data	Track origin of Data Respect IP	Culture of Respect towards CC
AI Generated Content	AI Generated ≠ AI assisted	Watermarking Pattern Recognition
⋮	⋮	⋮

Privacy by design

Privacy can be achieved through user aggregation, encryption for inference. It is less specific to A
and more global infra related.



Trustworthiness



How should the chatbot behave?

What drives decisions around AI answering tone

Can we control and predict AI behavior?

Training phases

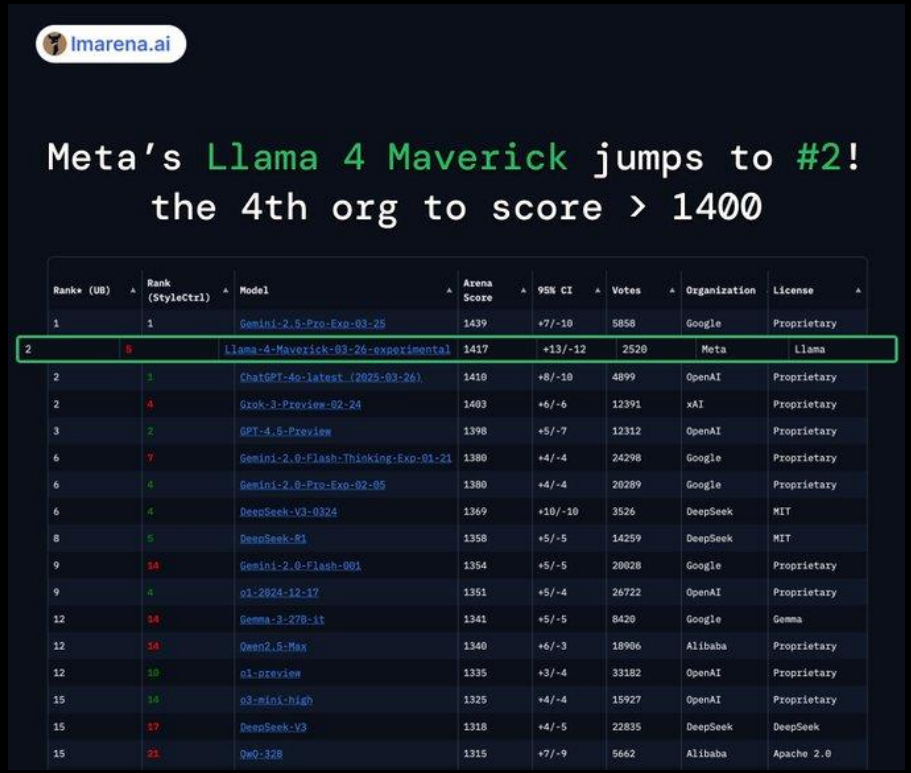
Pre-training

- Compressing the whole world's information
- Trains on massive amount of data with pure imitation loss
- A next-token predictor with the widest knowledge possible

Post-training

- Shapes the behavior/personality of the model
- Trains on more targeted data
- Turns a next-token-predictor into a product

The llama 4 case



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1439	+7/-10	5858	Google	Proprietary
2	5	Llama-4-Maverick-03-26-experimental	1417	+13/-12	2520	Meta	Llama
2	1	ChatGPT-4o-latest (2025-03-26)	1410	+8/-10	4899	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1403	+6/-6	12391	xAI	Proprietary
3	2	GPT-4.5-Preview	1398	+5/-7	12312	OpenAI	Proprietary
6	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+4/-4	24298	Google	Proprietary
6	4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-4	20289	Google	Proprietary
6	4	DeepSeek-V3-0324	1369	+10/-10	3526	DeepSeek	MIT
8	5	DeepSeek-R1	1358	+5/-5	14259	DeepSeek	MIT
9	14	Gemini-2.0-Flash-001	1354	+5/-5	20028	Google	Proprietary
9	4	o1-2024-12-17	1351	+5/-4	26722	OpenAI	Proprietary
12	14	Gemma-3-27B-it	1341	+5/-5	8420	Google	Gemma
12	14	Qwen2.5-Max	1340	+6/-3	18906	Alibaba	Proprietary
12	10	o1-preview	1335	+3/-4	33182	OpenAI	Proprietary
15	14	o3-mini-high	1325	+4/-4	15927	OpenAI	Proprietary
15	17	DeepSeek-V3	1318	+4/-5	22835	DeepSeek	DeepSeek
15	21	QwQ-32B	1315	+7/-9	5662	Alibaba	Apache 2.0

The llama 4 case

lmarena.ai

45	58	Gemini-1.5-Flash-002	1268
45	34	Llama-4-Maverick-17B-128E-Instruct	1266
45	67	Llama-3.1-Nemotron-70B-Instruct	1265
48	34	Meta-Llama-3.1-405B-Instruct-bf16	1266

1	1	Gemini-2.5-Pro-Exp-03-25	1439	+7/-10	5858	Google	Proprietary
2	5	Llama-4-Maverick-03-26-experimental	1417	+13/-12	2520	Meta	Llama
2	1	ChatGPT-4o-latest (2025-03-26)	1410	+8/-10	4899	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1403	+6/-6	12391	xAI	Proprietary
3	2	GPT-4.5-Preview	1398	+5/-7	12312	OpenAI	Proprietary
6	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+4/-4	24298	Google	Proprietary
6	4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-4	20289	Google	Proprietary
6	4	DeepSeek-V3-0324	1369	+10/-10	3526	DeepSeek	MIT
8	5	DeepSeek-R1	1358	+5/-5	14259	DeepSeek	MIT
9	14	Gemini-2.0-Flash-001	1354	+5/-5	20028	Google	Proprietary
9	4	o1-2024-12-17	1351	+5/-4	26722	OpenAI	Proprietary
12	14	Gemma-3-27B-it	1341	+5/-5	8420	Google	Gemma
12	14	Qwen2.5-Max	1340	+6/-3	18906	Alibaba	Proprietary
12	10	o1-preview	1335	+3/-4	33182	OpenAI	Proprietary
15	14	o3-mini-high	1325	+4/-4	15927	OpenAI	Proprietary
15	17	DeepSeek-V3	1318	+4/-5	22835	DeepSeek	DeepSeek
15	21	QwQ-32B	1315	+7/-9	5662	Alibaba	Apache 2.0

How can we build trust?

Gemma

Why openness matters

AI for everyone

Not just for those who can afford huge compute investments

Meet developers where they are

Getting worldwide feedback and partnering with the open ecosystem is a proven path to success

Why openness matters

AI for everyone

Not just for those who can afford huge compute investments

Meet developers where they are

Getting worldwide feedback and partnering with the open ecosystem is a proven path to success

Trust

Openness means you can trust because you do not have to