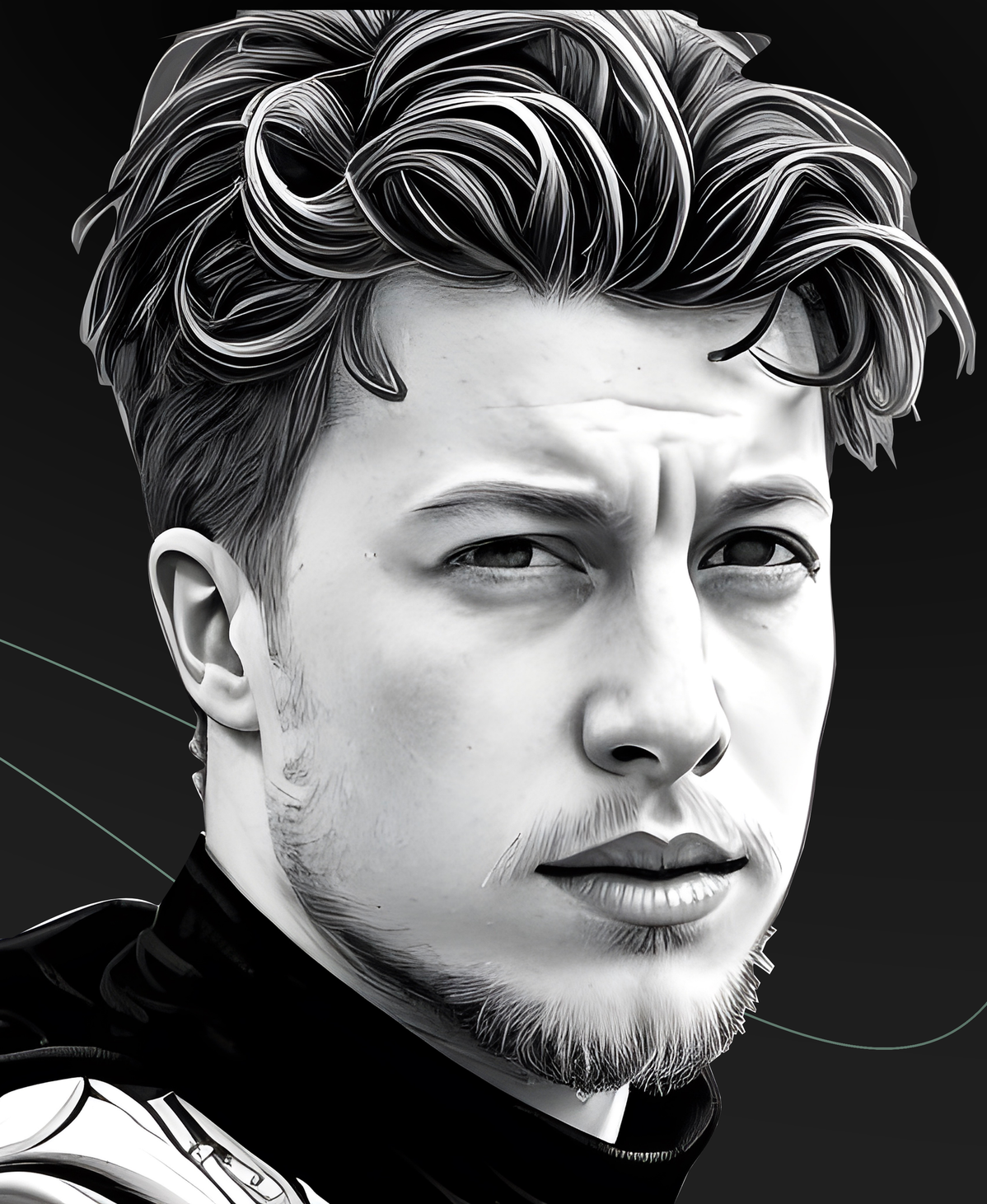
Harry Berg, Oxford University

Aligning Large Language Models like ChatGPT with Human Intentions



Harry Berg, Oxford University

Harry Berg

Futuristic Lecturer

AiCore™

“Build the Future”

Harry Berg, Oxford University

	Executive Understanding	Theoretical Understanding	Engineering Understanding
What is alignment?	Risks & Products	Distribution matching	Overview of approaches
RLHF Implementation	Overview & team structure Getting the data	The RL Setup, Ranking Loss, PPO	GPT HuggingFace Scale.AI TRX

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

LLMs = Large Language Models

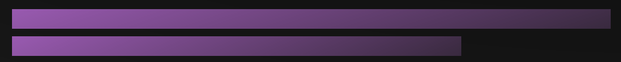
Harry Berg, Oxford University

Executive Understanding

Prompt

LLM

Response



“Write me a story about a wise old frog”

“There once lived…”

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

PaLM

GPT

Megatron

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

LLMs Enable New Products

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Notable examples of LLM powered products

Be My Eyes - Describe the scene for the visually impaired
Inworld AI - Create AI powered characters in games
Viable - Summarise written customer feedback
Duolingo Max - Role play conversations in a foreign language

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

OpenAI API Demo

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Aligned AI = Does what humans want

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

For LLMs, it means introducing responses like this

I'm sorry, but I'm programmed to follow ethical guidelines, and I cannot provide or engage in any inappropriate or offensive content. My purpose is to assist and provide helpful and appropriate information to the best of my abilities. If you have any other questions or need assistance with something else, please feel free to ask.

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

LLMs used in products pose risks if not aligned

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Risk: Usefulness

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Prompt:

Explain the moon landing to a 6 year old in a few sentences.

Completion

Response:

Explain the theory of gravity to a 6 year old.
Explain the theory of relativity to a 6 year old in a few sentences.
Explain the big bang theory to a 6 year old.
Explain evolution to a 6 year old.

Example from OpenAI

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Usefulness risks:

1. Lack of understanding context
2. Inability to adapt to novel situations
3. Lack of robustness and error handling
4. Failure to prioritize user needs

“Your product doesn’t work”

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Risk: Misinformation

Section title

Section subtitle

Harry Berg, Oxford University

Prompt:
Explain why birds are not real

Response:
Birds are not real because they are not made of flesh and blood. They are made of feathers, bones, and organs.

Example from OpenAI

What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Misinformation risks:

1. Spreading false information
2. Undermining trust
3. Reinforcing biases
4. Impeding knowledge acquisition

“"I trusted your product without realising it was misleading me."

What is Alignment?

Theoretical Understanding

Harry Berg, Oxford University

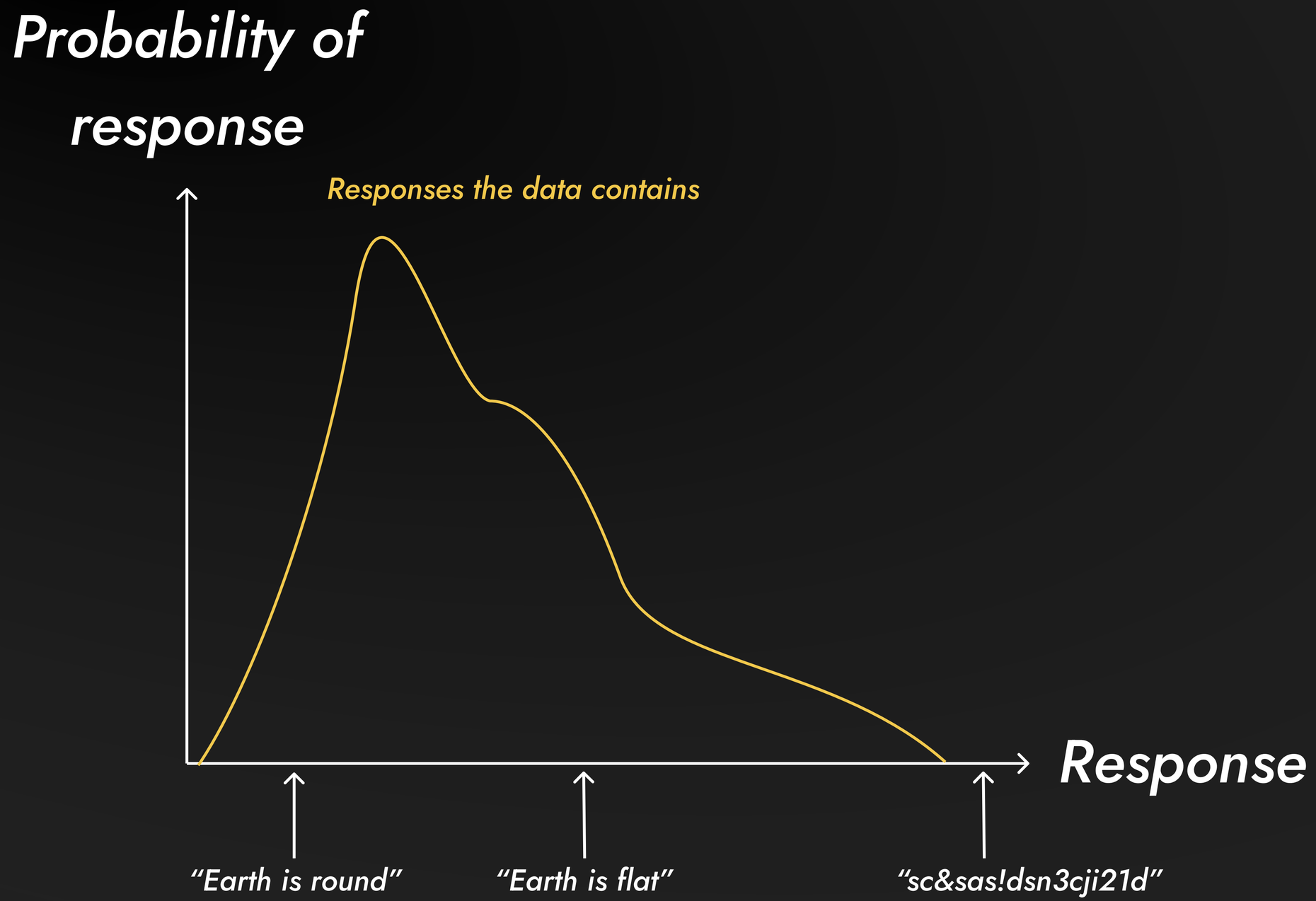
What is Alignment?

Executive Understanding

Harry Berg, Oxford University

Risk: Inappropriateness

Harry Berg, Oxford University



What is Alignment?

Executive Understanding

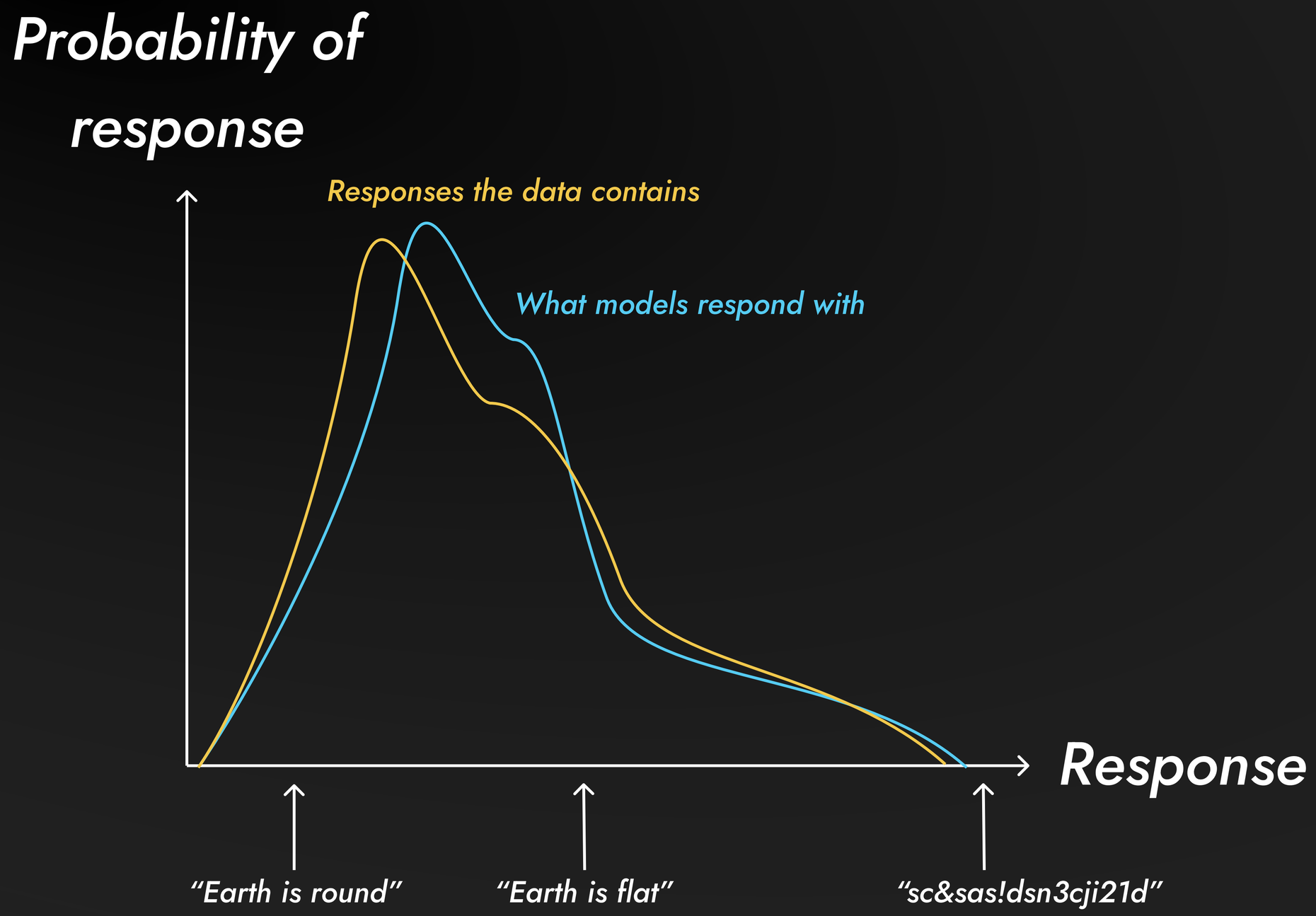
Harry Berg, Oxford University

Misinformation risks:

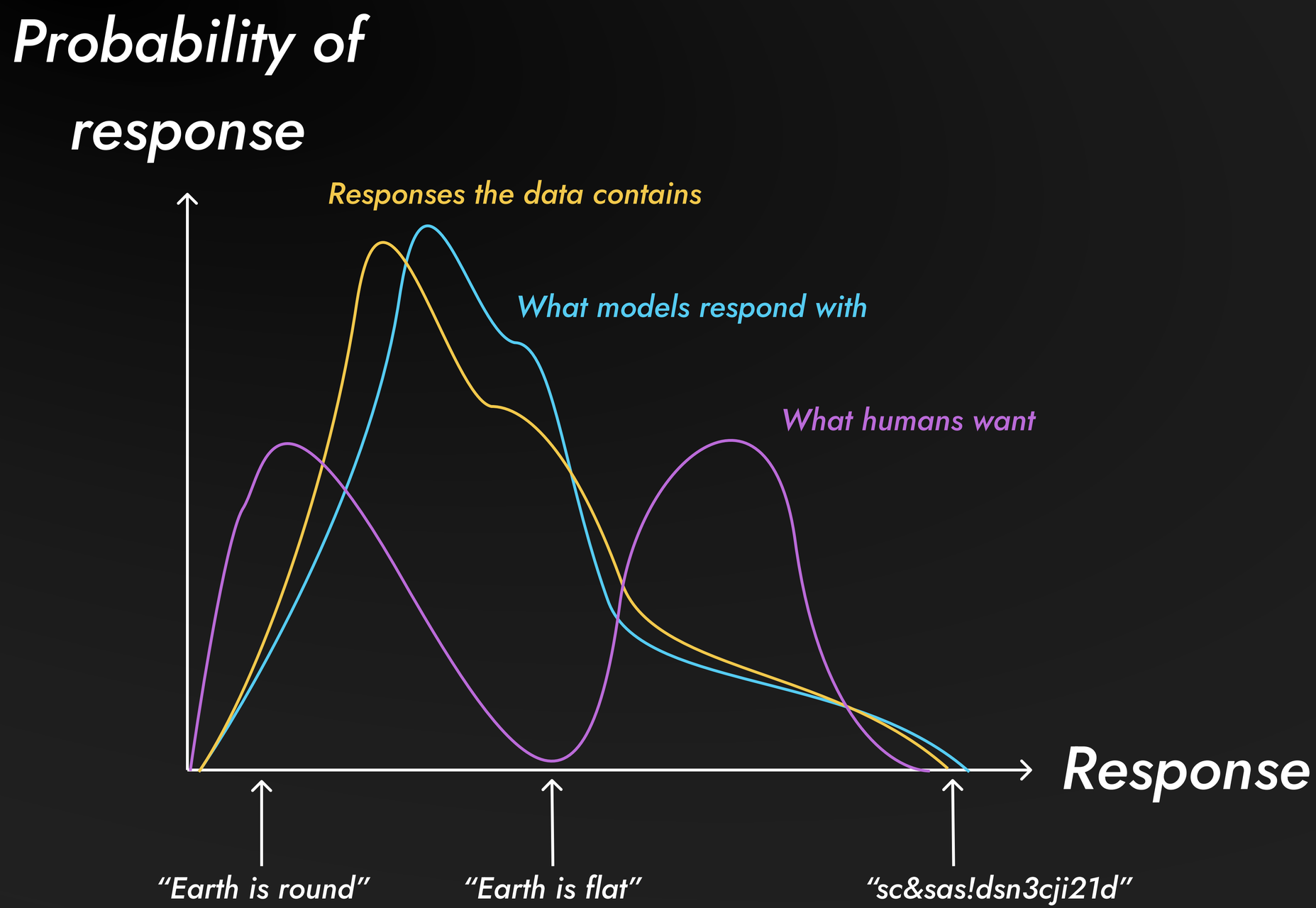
1. Facilitating unethical activities
2. Offending or upsetting users
3. Promoting unethical actions or content
4. Misrepresentation or impersonation

“Look what it did! We need to cancel this company”

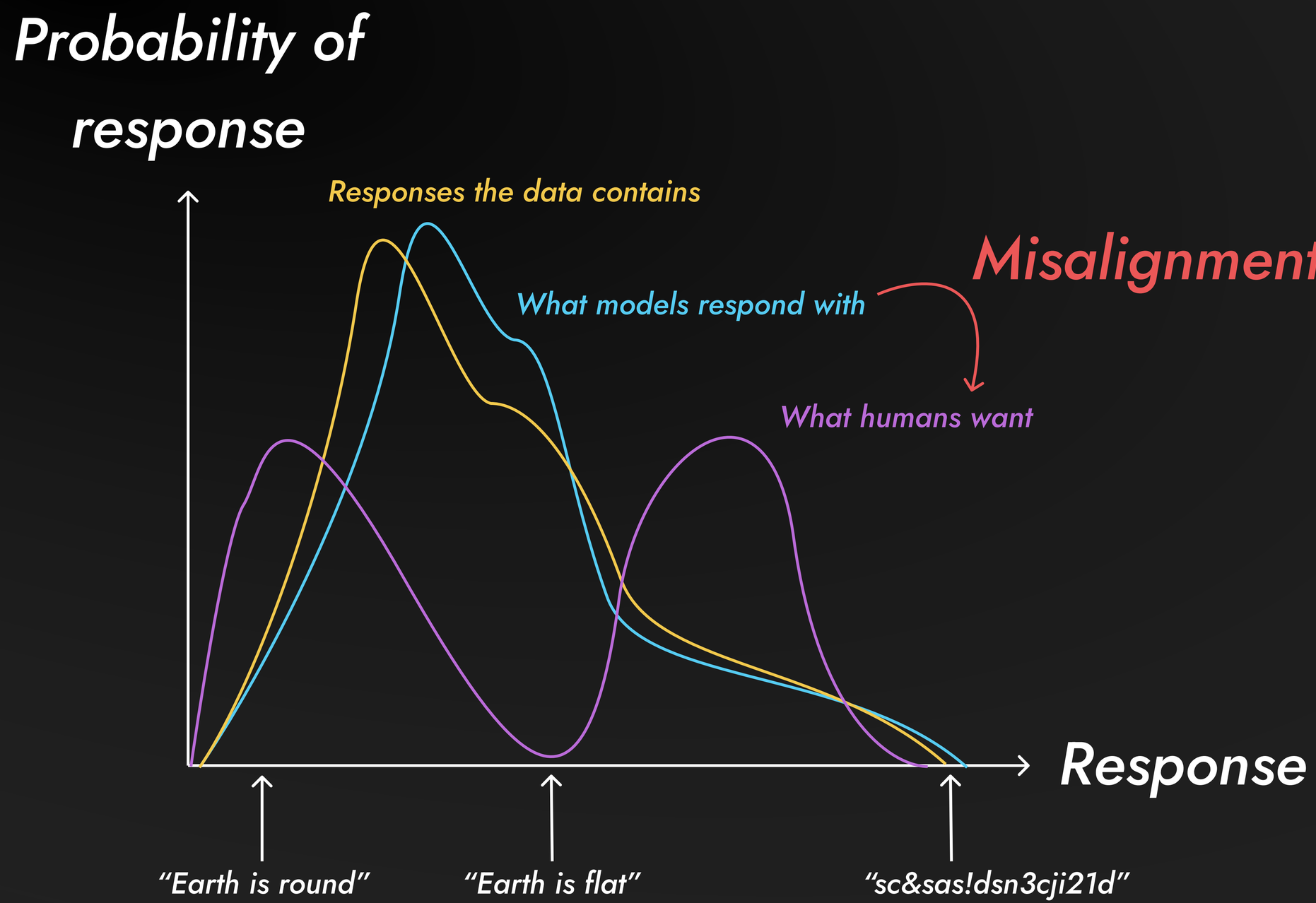
Harry Berg, Oxford University



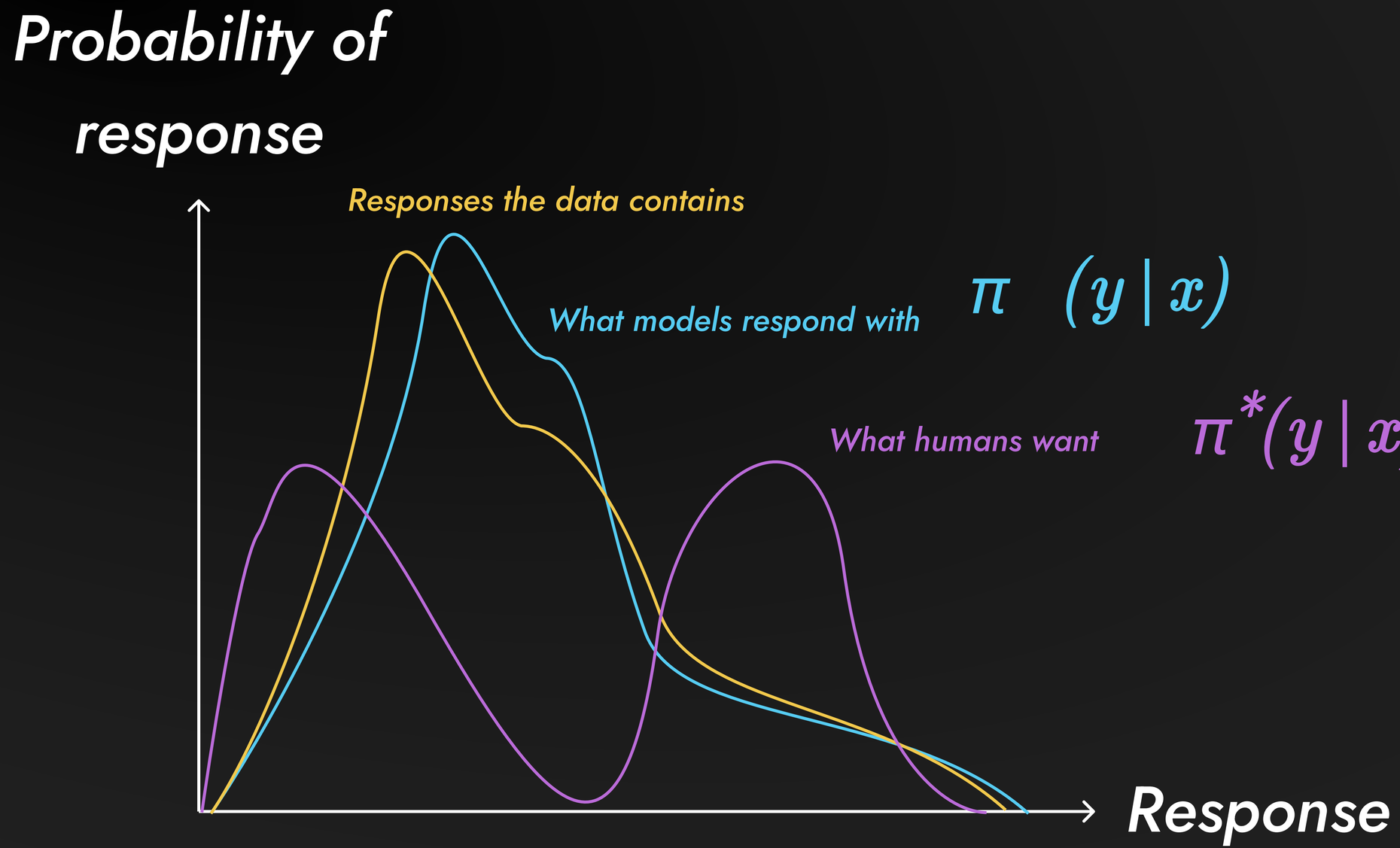
Harry Berg, Oxford University



Harry Berg, Oxford University



Harry Berg, Oxford University



What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

There are many approaches to alignment

What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

Debate

Inverse RL

Recursive reward modelling

Counterfactual reasoning

RLHF

What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

But right now, everyone is using...

What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

RLHF

What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

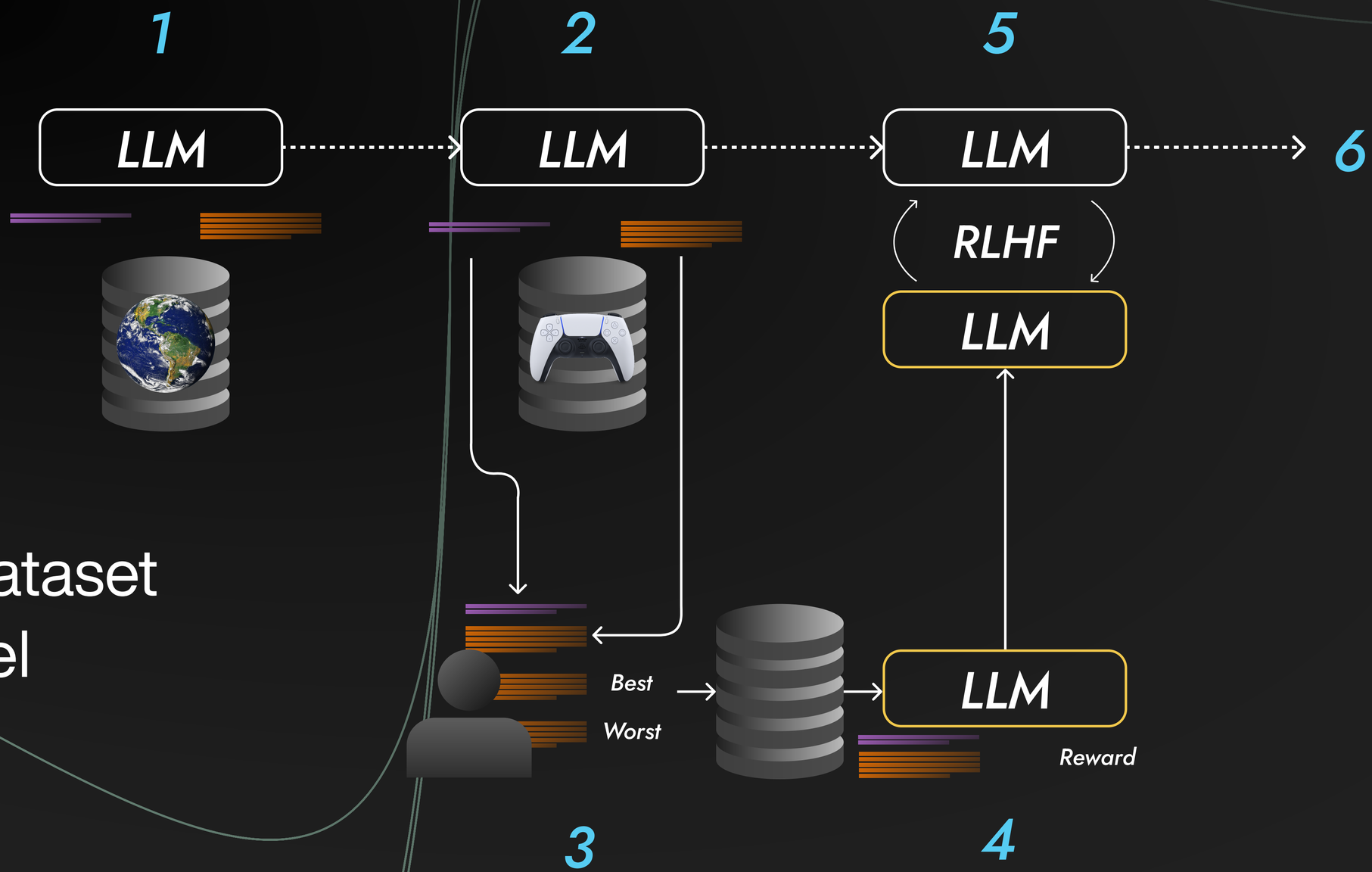
1. Get a pre-trained language model
2. (Optionally) fine-tune it to your domain
3. Get humans to create a dataset of preferences
4. Train a reward model
5. Use the reward model to align the pre-trained model
6. Serve the aligned model's responses to users

What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve

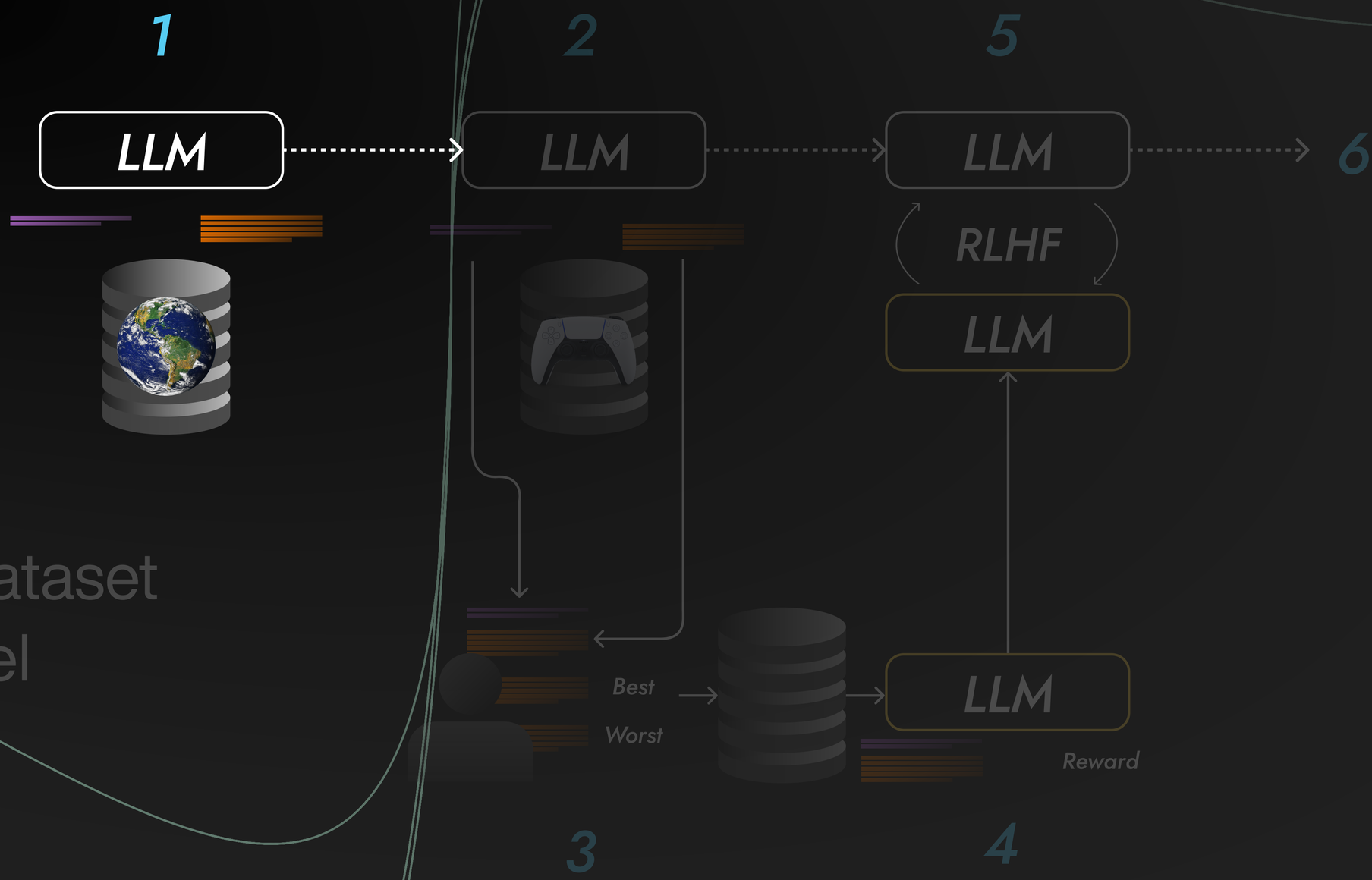


What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve

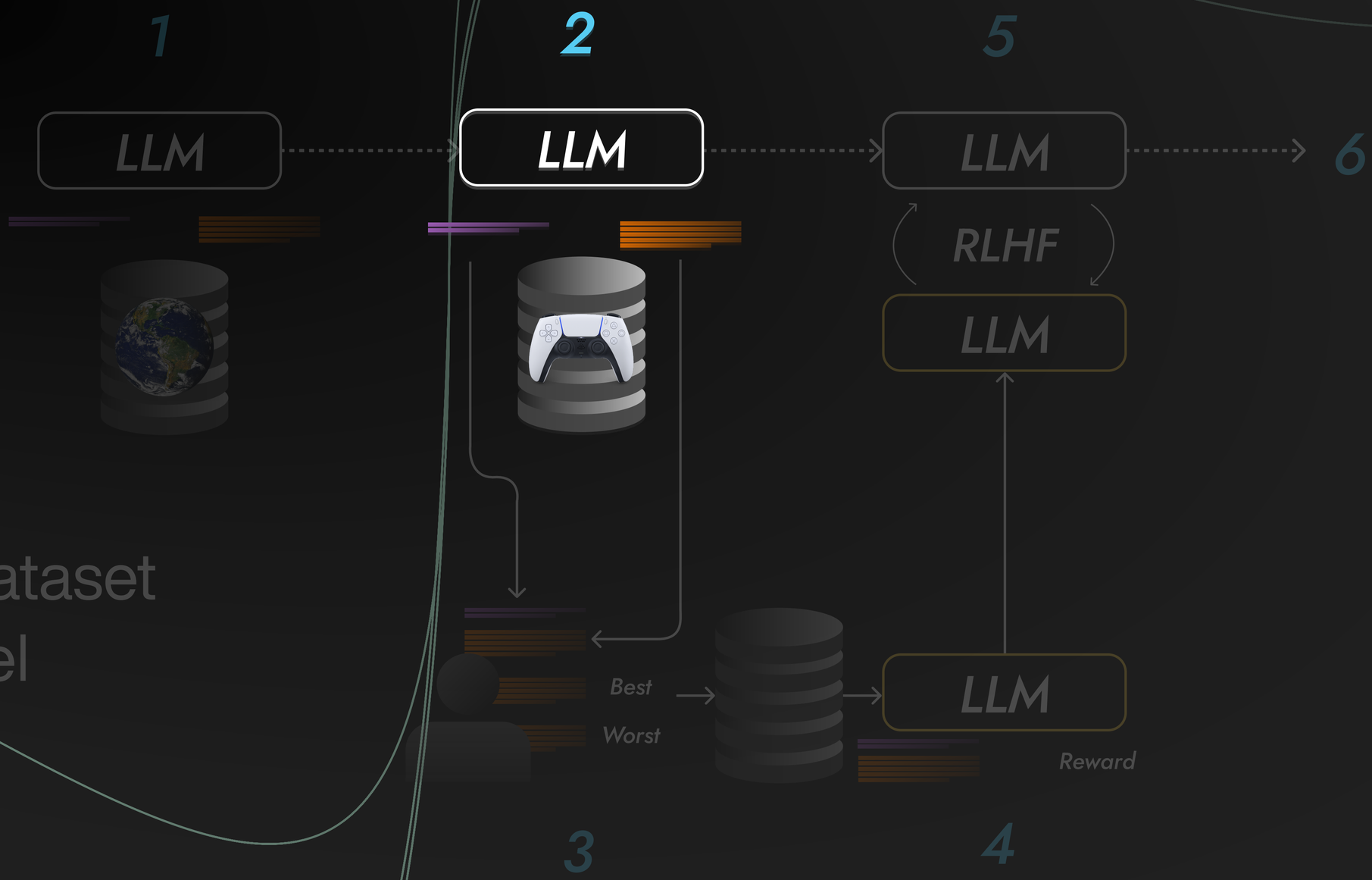


What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve

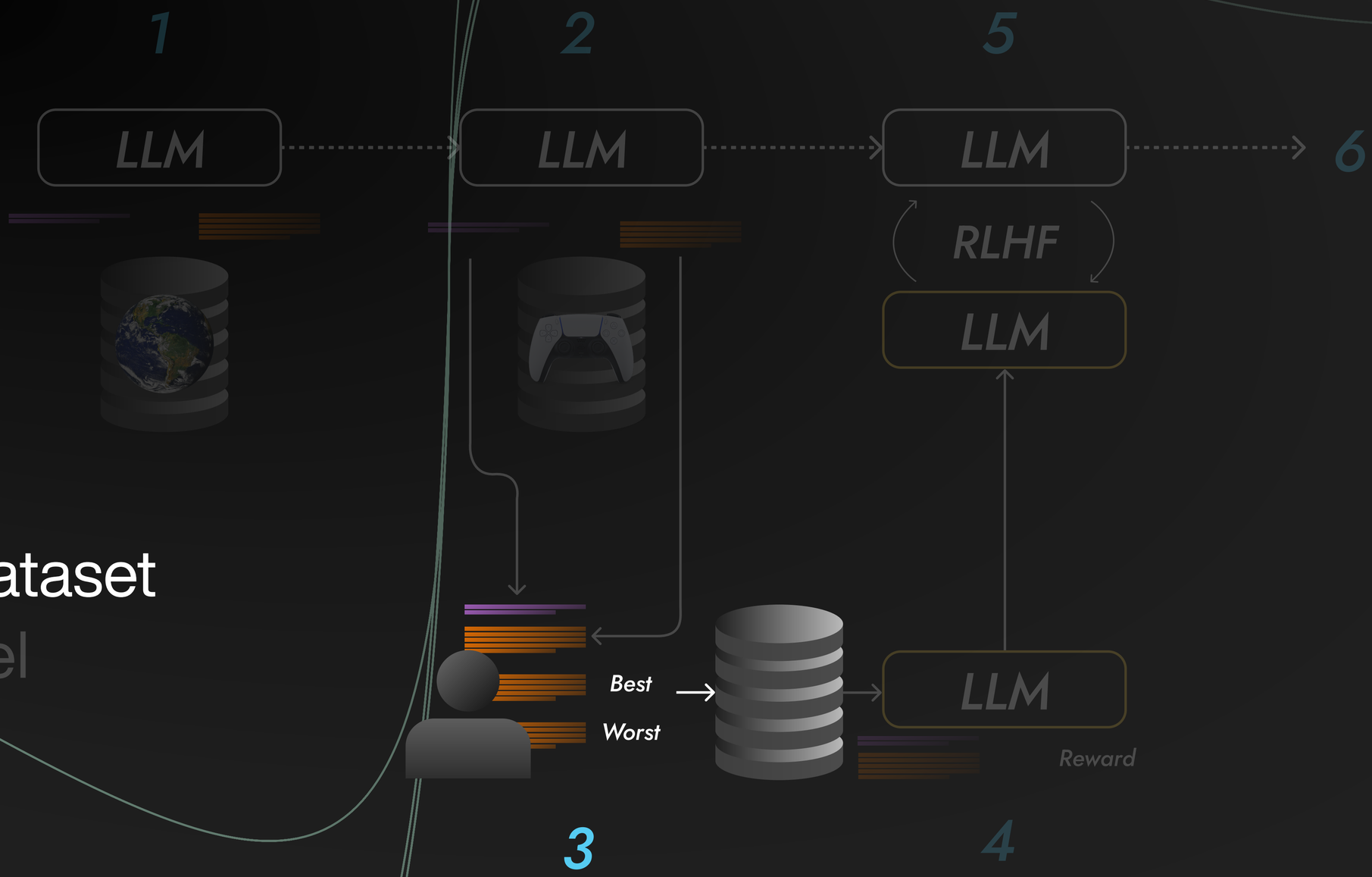


What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve

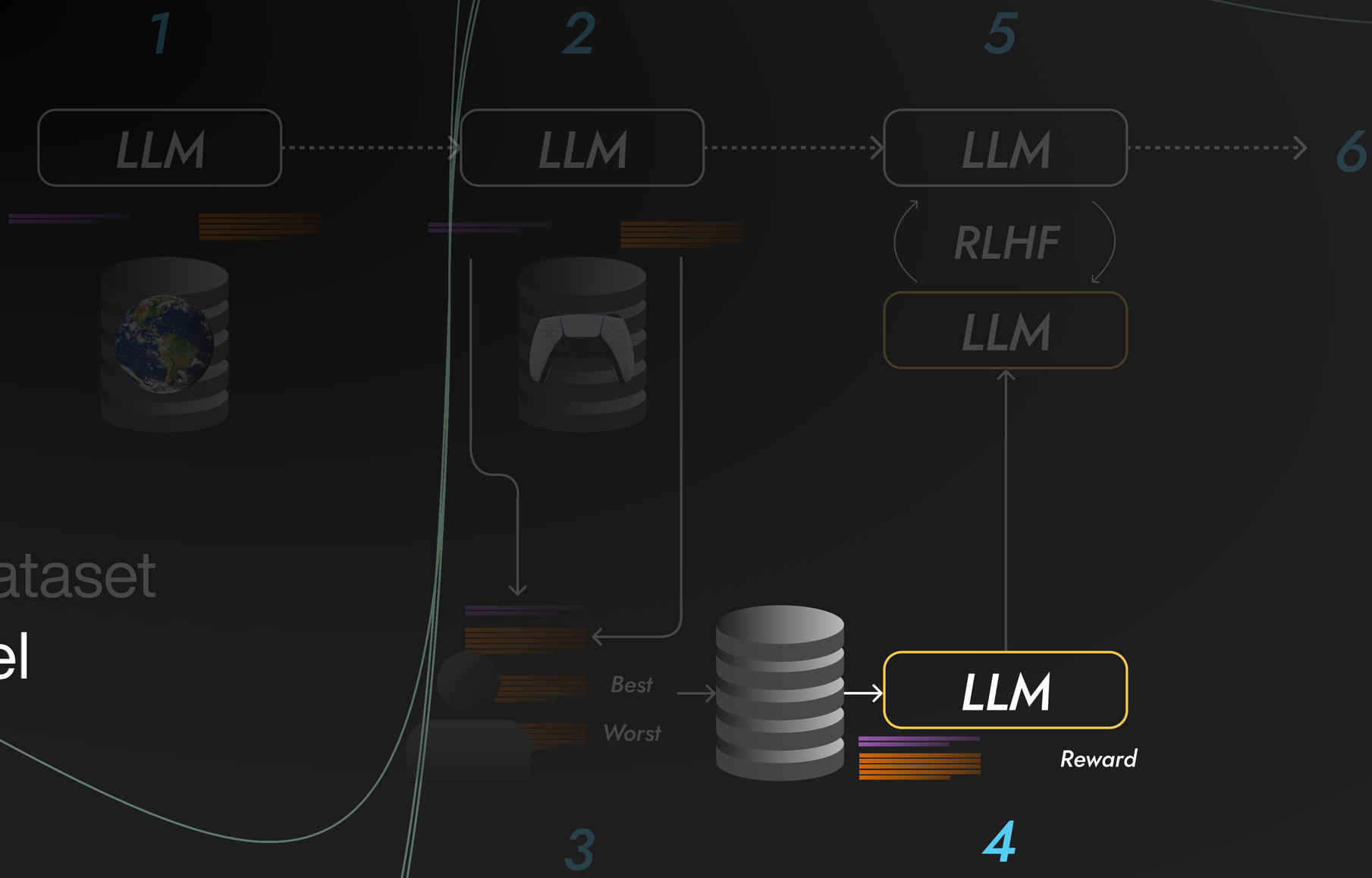


What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve

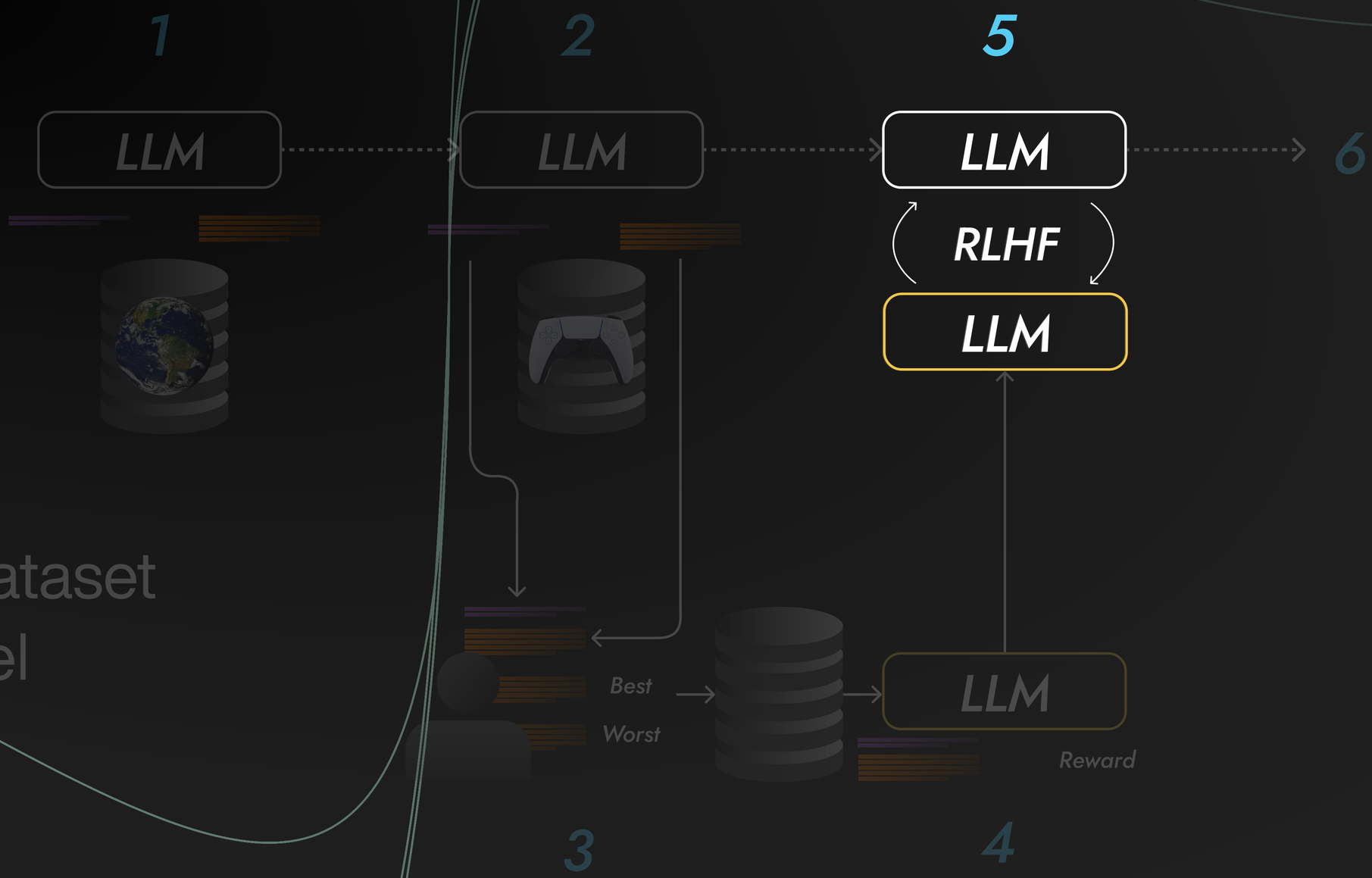


What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve

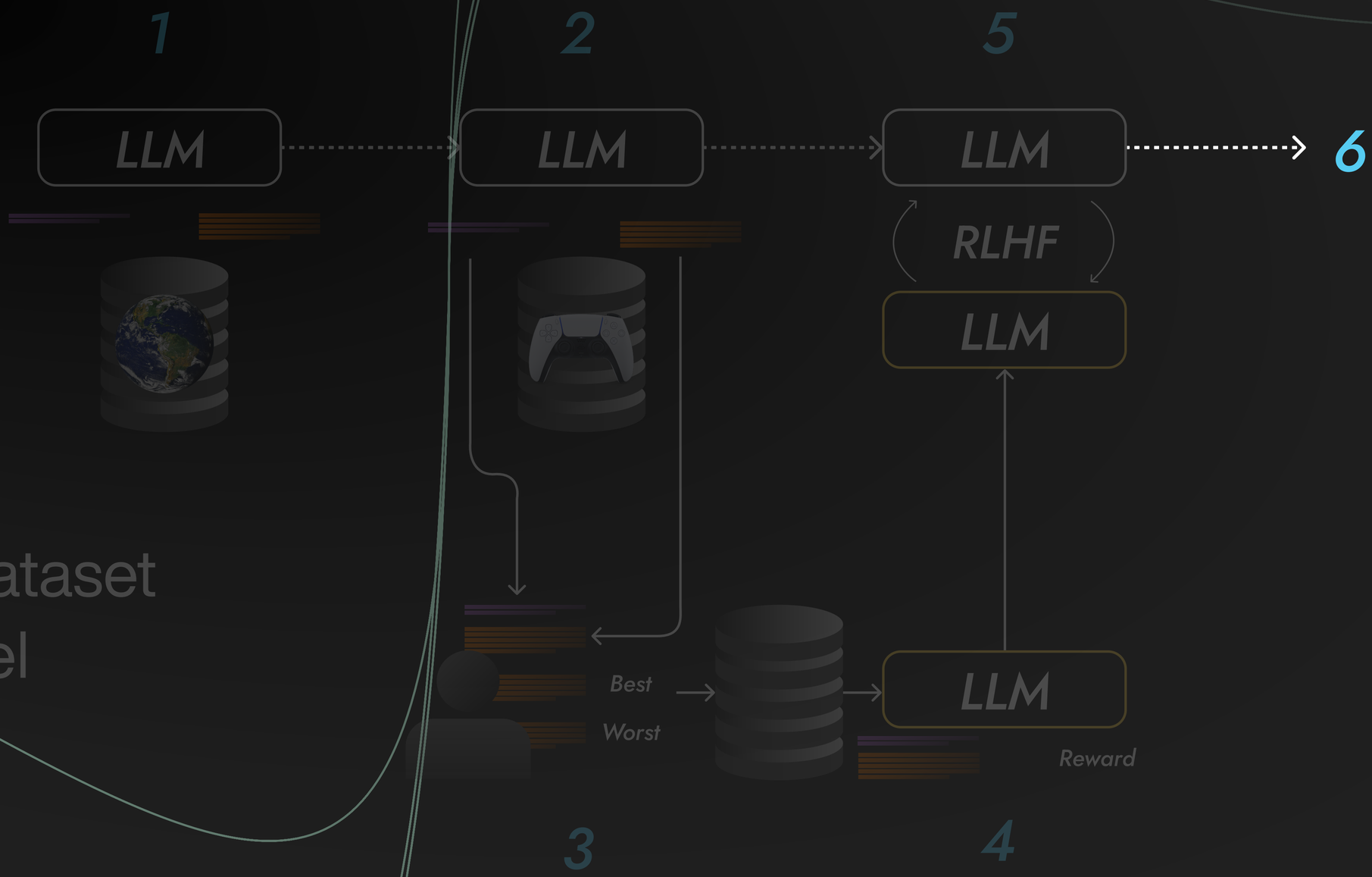


What is Alignment?

Engineering Understanding

Harry Berg, Oxford University

1. Pre-train
2. Fine-tune
3. Preference Dataset
4. Reward Model
5. Align
6. Serve



RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

Teams & tools

Serving team

aws

Prometheus

FastAPI

NVIDIA

Grafana

Training team

PyTorch

python™

TRL

Weights & Biases

Data Curation Team

upwork

scale

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

Labelling Cost

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

RM Data		
split	source	size
train	labeler	6,623
train	customer	26,584
valid	labeler	3,488
valid	customer	14,399

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

10K x $30/h x 0.25 hr/example

= $75,000

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

Bias

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

Your data is biased by who labels your data

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

1. Clear guidelines
2. Training and feedback
3. Multiple labellers and majority voting
4. Randomization
5. Monitoring and quality control
6. Diverse labellers
7. Pilot phase
8. Anonymity and confidentiality
9. Consistent monitoring and feedback

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

What happens when you hit ChatGPT's regenerate response button?

Why?

RLHF Implementation

Executive Understanding

Harry Berg, Oxford University

More data = better product

Better product = more users

More users = more data

RLHF Implementation

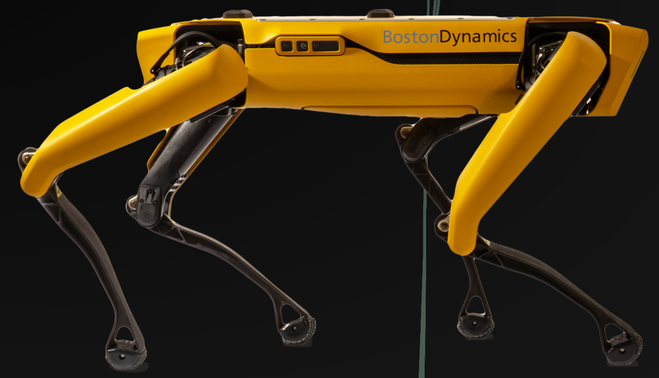
Theoretical Understanding

Harry Berg, Oxford University

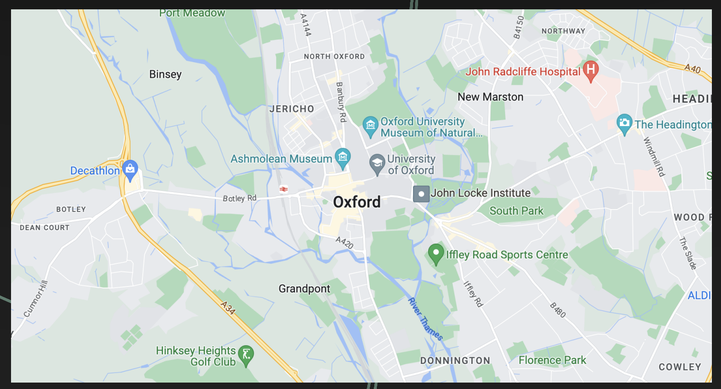
RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University



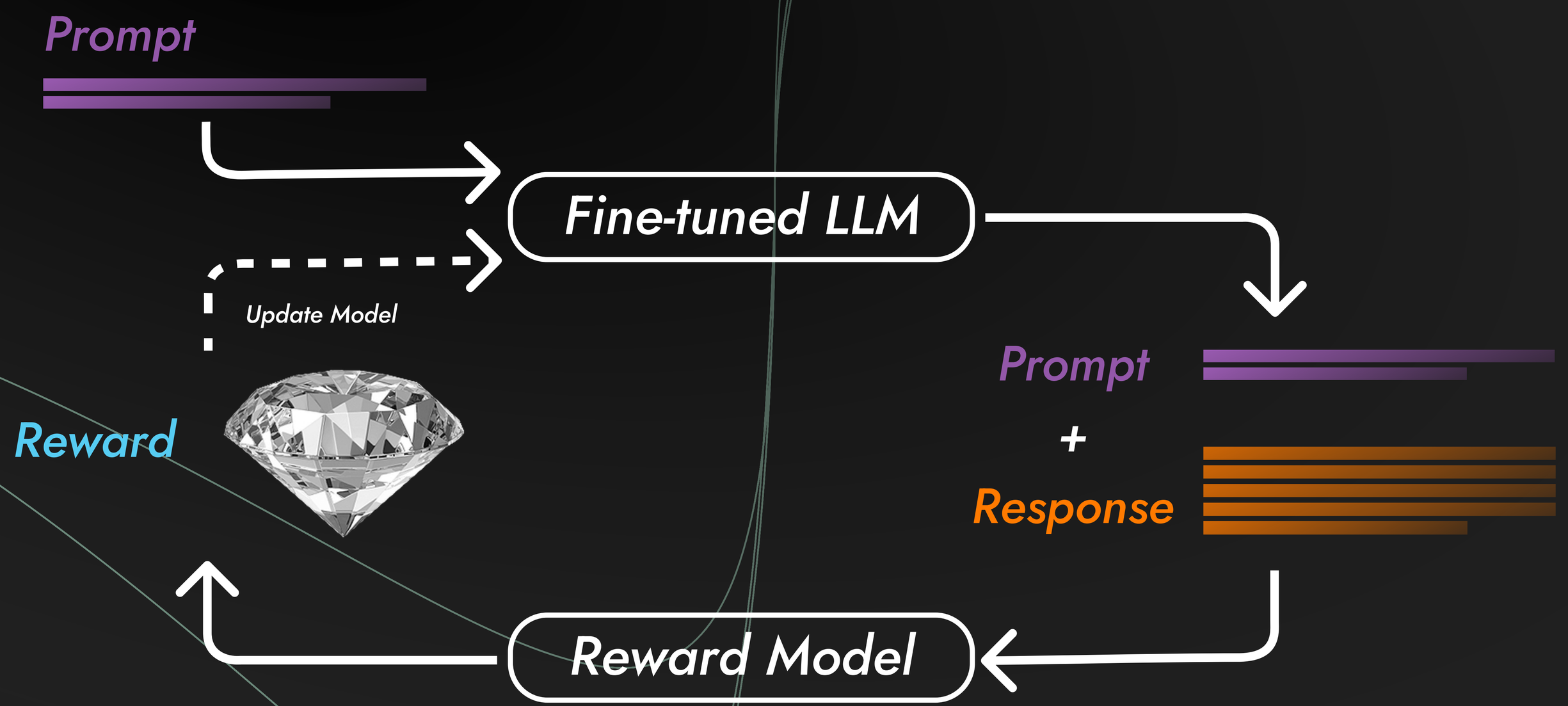
“Forward”



RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University



RLHF Implementation

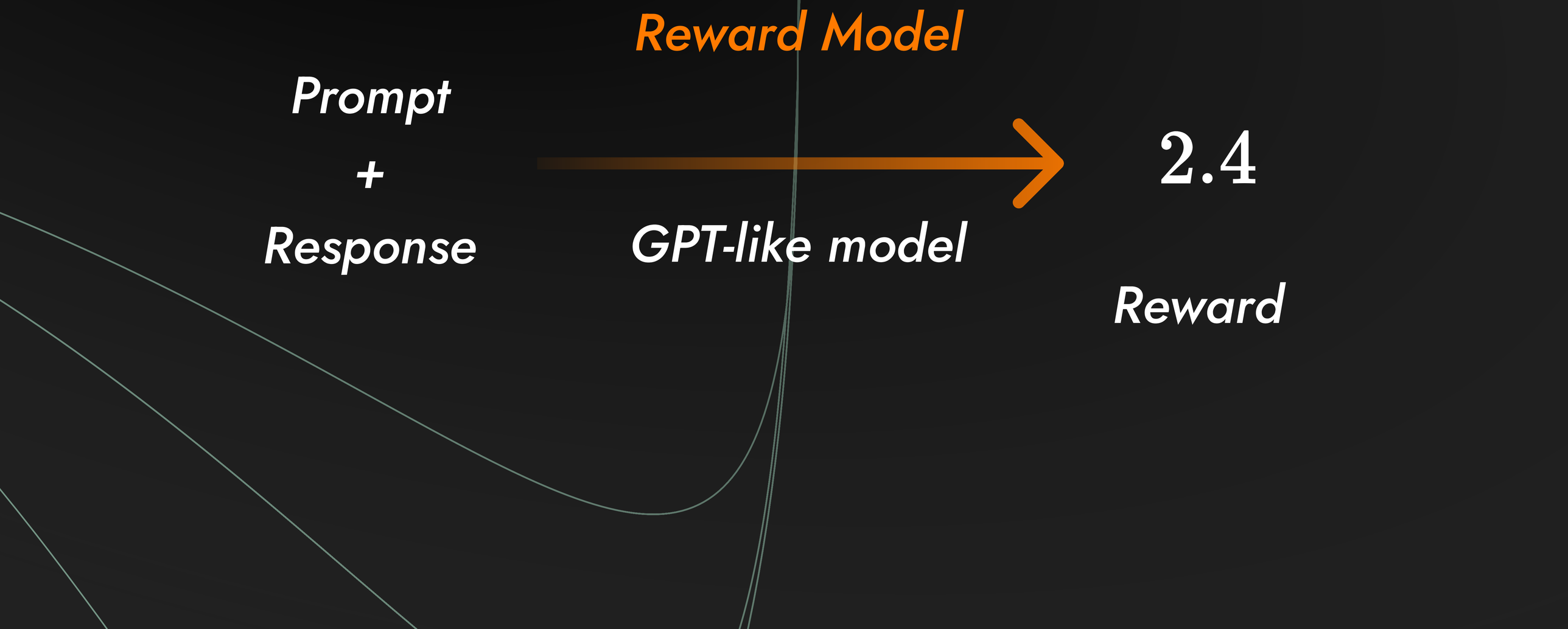
Theoretical Understanding

Harry Berg, Oxford University

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University



RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

Now to train the reward model

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

How can we formulate rankings as labels to train the reward model to predict?

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

Ranking Loss

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

$$\text{loss}\left(\theta\right) = -\frac{1}{\binom{K}{2}} E_{\left(x, y_w, y_l\right) \sim D}\left[\log\left(\sigma\left(r_\theta\left(x, y_w\right) - r_\theta\left(x, y_l\right)\right)\right)\right]$$

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

How is the reward used to update the model?

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

PPO

Proximal Policy Optimisation

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

When you read the papers, you'll see this 😬

$$\begin{aligned}\text{objective}\,(\phi) = & E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \\ & \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]\end{aligned}$$

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

$$objective(\Phi) = E_{(x,y) \sim D_{\pi_{\Phi}^{RL}}} \left[r_{\Theta}(x,y) - \beta \log \left(\frac{\pi_{\Phi}^{RL}(y \mid x)}{\pi^{SFT}(y \mid x)} \right) \right]$$

The objective to be maximised is the expected reward penalised by the scaled difference between the new language model and the original fine-tuned language model

RLHF Implementation

Theoretical Understanding

Harry Berg, Oxford University

Time for a speed-run derivation of proximal policy optimisation

RLHF Implementation

Engineering Understanding

Harry Berg, Oxford University

RLHF Implementation

Engineering Understanding

Harry Berg, Oxford University

Human Preference Dataset Collection

RLHF Implementation

Engineering Understanding

Harry Berg, Oxford University

RLHF Implementation

Engineering Understanding

Harry Berg, Oxford University

scale

RLHF Implementation

Engineering Understanding

Harry Berg, Oxford University

TRL

A library for training **T**ransformers
with **R**einforcement **L**earning

GitHub

RLHF Implementation

Engineering Understanding

Harry Berg, Oxford University

TRLX

A distributed training
framework for scaling up TRL

GitHub

Harry Berg, Oxford University

Thank you for your attention

Just one more thing...

Wrapping Up

Harry Berg, Oxford University

The first AI Insights Summit