LMs Inside Out

Sasha Rush

youtube.com/@srush_nlp

So please join me in welcoming Sasha.

LMs Inside Out

Sasha Rush

youtube.com/@srush_nlp

Everyone. I get to go 1st because I win

LMs Inside Out

Sasha Rush

youtube.com/@srush_nlp

the work that's been going on in these areas kind of pulling everything together.

Collaborators

Jack Morris

Volodymyr Kuleshov

Justin Chiu

Vitaly Shmatikov

Wenting Zhao

This one,

Collaborators

Jack Morris

Volodymyr Kuleshov

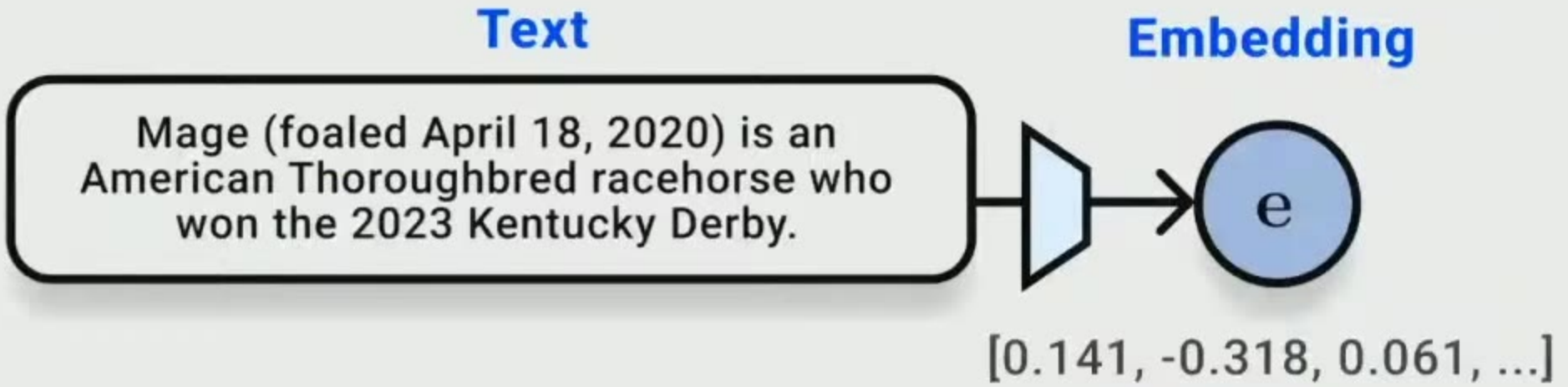
Vitaly Shmatikov

Justin Chiu

Wenting Zhao

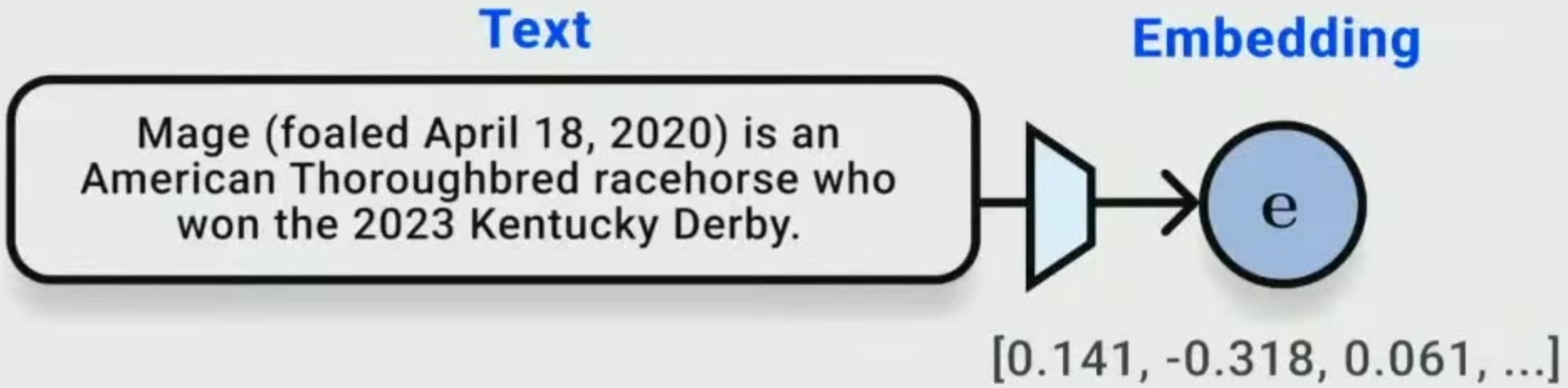
This one. Yeah, beautiful. Okay. Great. Cool

Text embeddings



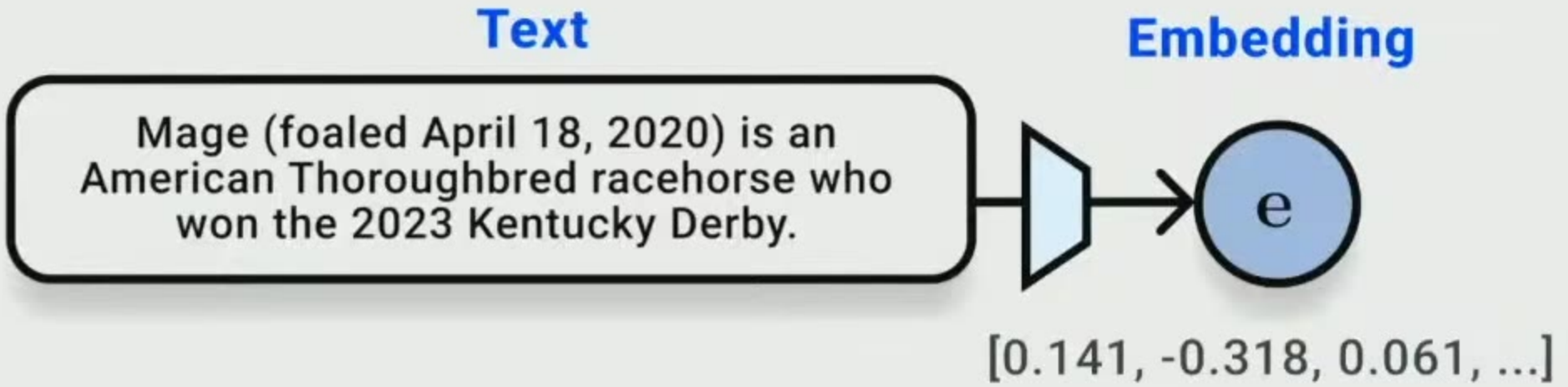
This one. Yeah, beautiful. Okay, great. Cool. So this

Text embeddings



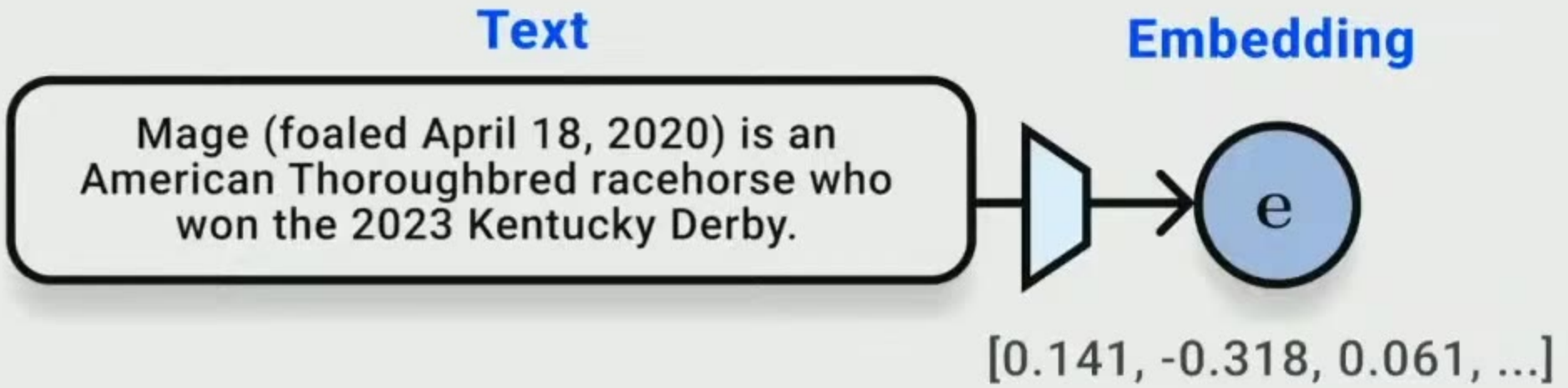
Okay, so in this paper, we're going to

Text embeddings



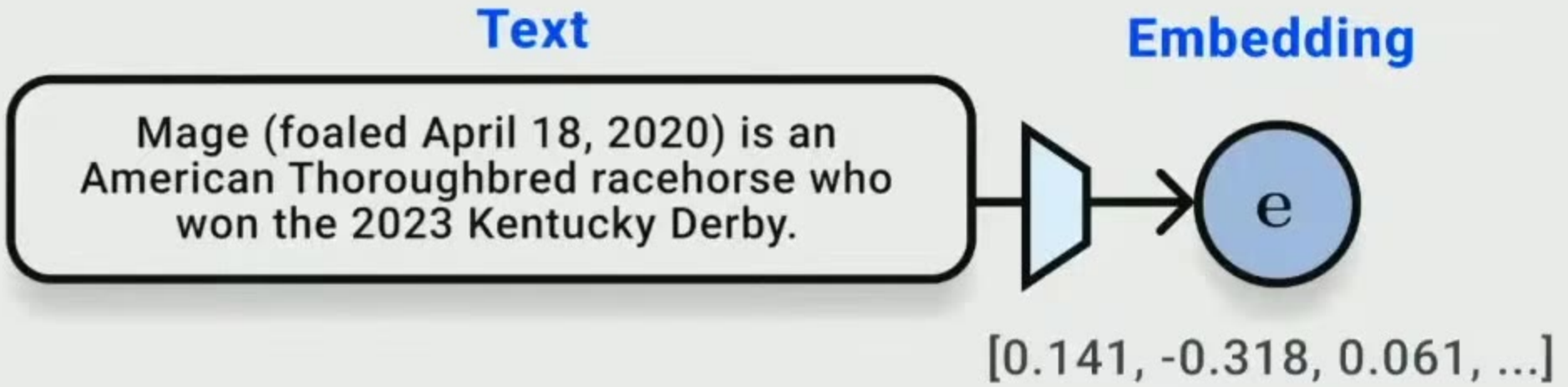
so in this paper, we're going to be talking about text embeddings.

Text embeddings

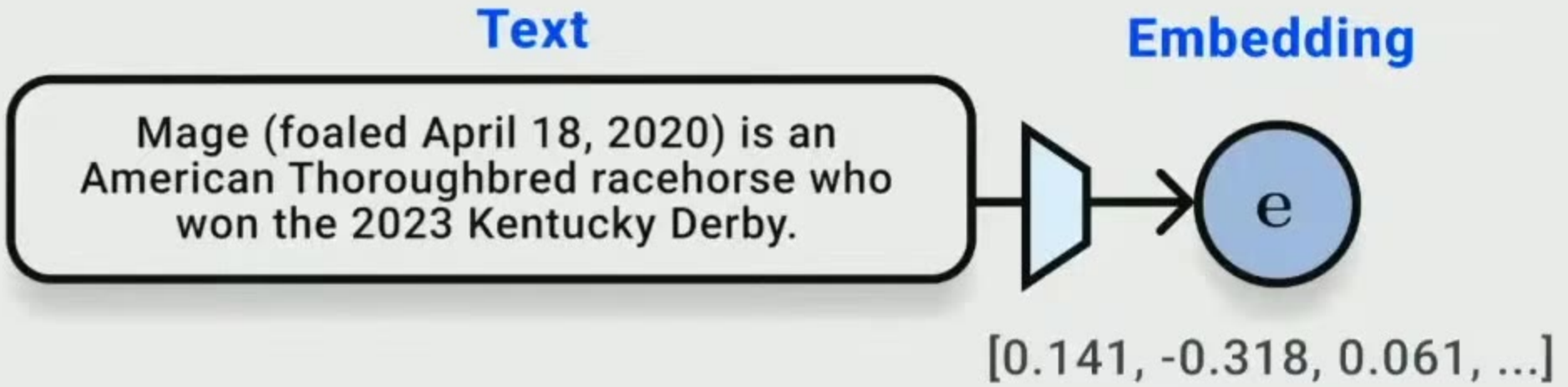


Text. Embeddings, I think, are probably the most widely

Text embeddings

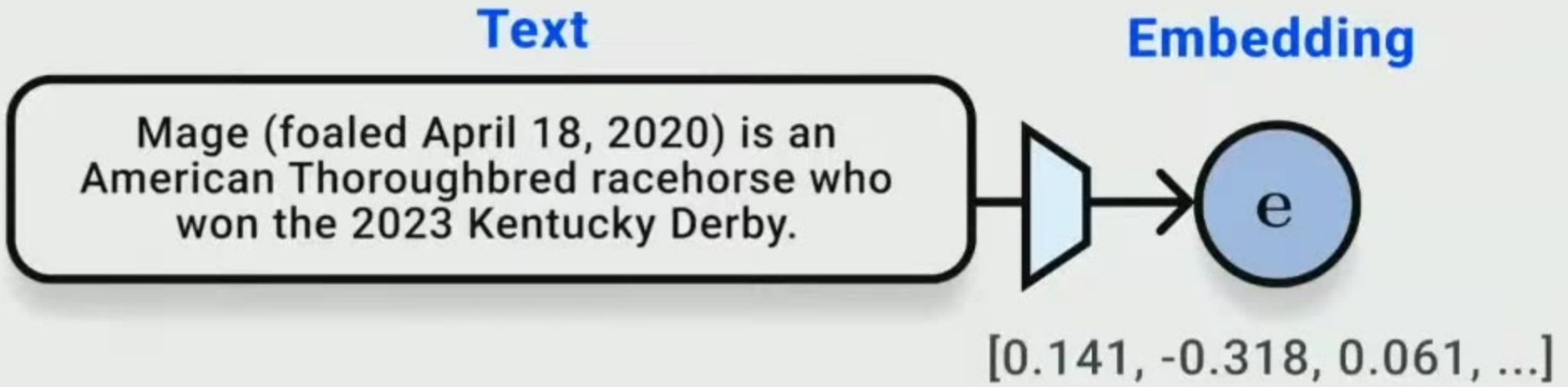


Text embeddings



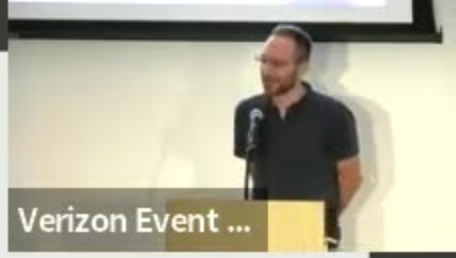
People use them for everything, from classification to

Text embeddings



retrieval. All sorts of applications of that form.

Vector databases are *huge* right now



Startups

Pinecone drops $100M i
on $750M valuation, as v
database demand grows

AI

Qdrant, an open source vector
databas
develop
data

Paul Sawers @psawers

Premium HOME > TECH

Vector database Chroma scored $18 mi
million valuation. Here's why its technol
generative AI startups.

Stephanie Palazzolo Apr 6, 2023, 8:00 AM EDT

Weaviate Raises $50 Million Series B F
Meet Soaring Demand for AI Native Vector
Database Technology USA · Englis

NEWS PROVIDED BY
Weaviate →
21 Apr, 2023, 08:00 ET

Company's open source vector database and new cloud servic

FORBES > INNOVATION > CLOUD

The Rise Of Vector Databases

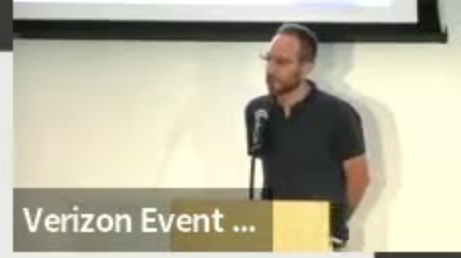
Adrian Bridgwater Senior Contributor ⓘ

So couple of years ago there were about

Berkeley
UNIVERSITY OF CALIFORNIA

Powered by Zoom

Vector databases are *huge* right now



Startups

Pinecone drops $100M i
on $750M valuation, as v
database demand grows

AI

Qdrant, an open source vector
databas
develop
data

Paul Sawers @psawers

Premium HOME > TECH

Vector database Chroma scored $18 mi
million valuation. Here's why its technol
generative AI startups.

Stephanie Palazzolo Apr 6, 2023, 8:00 AM EDT

Weaviate Raises $50 Million Series B F
Meet Soaring Demand for AI Native Vector
Database Technology USA - Englis

NEWS PROVIDED BY
Weaviate →
21 Apr, 2023, 08:00 ET

Company's open source vector database and new cl

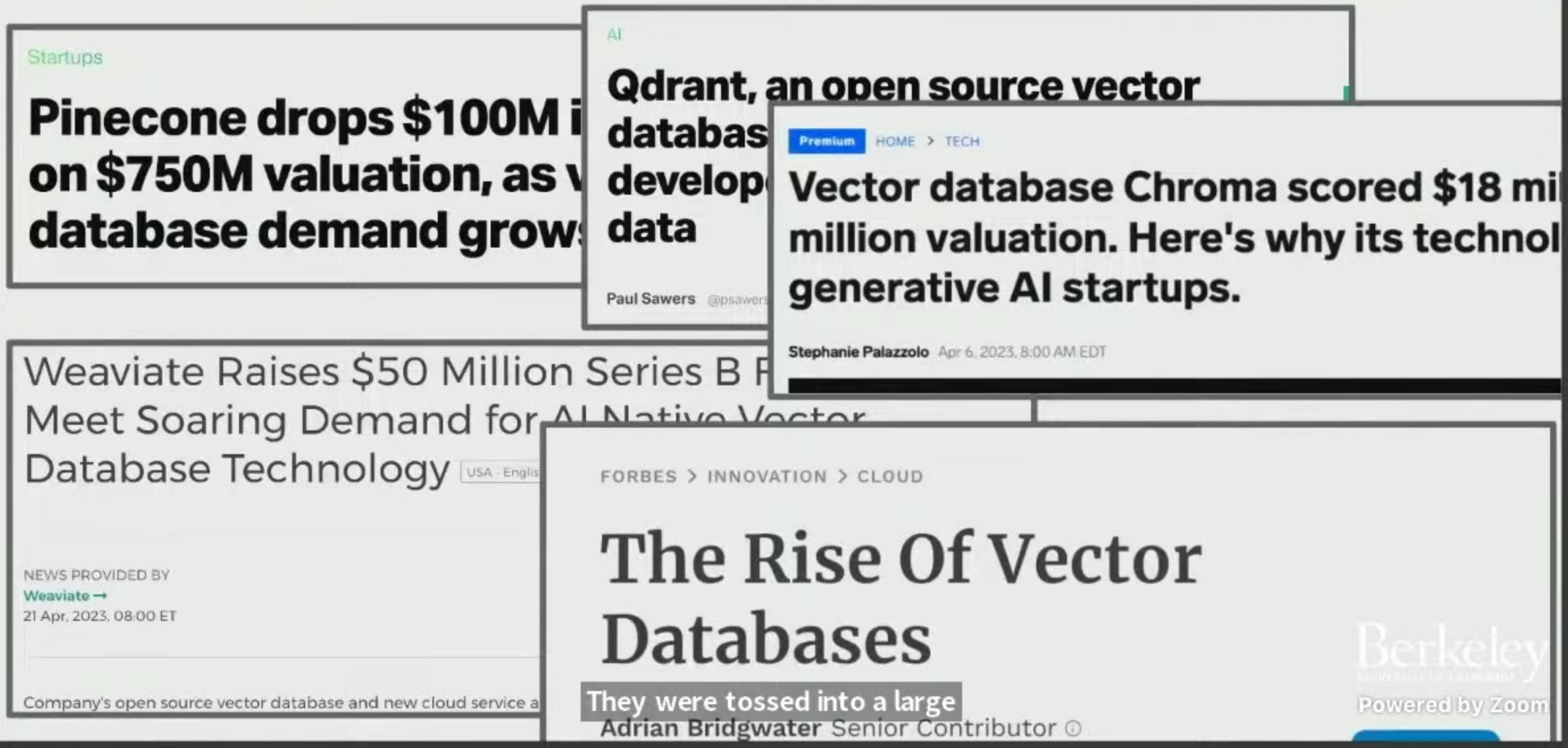
FORBES > INNOVATION > CLOUD

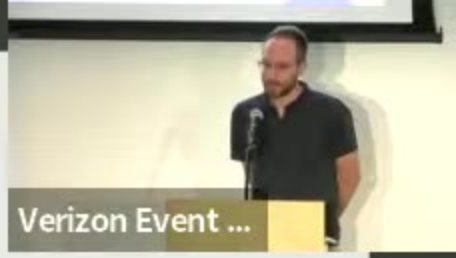
The Rise Of Vector Databases

companies that started that kind of offered a service

Adrian Bridgwater Senior Contributor

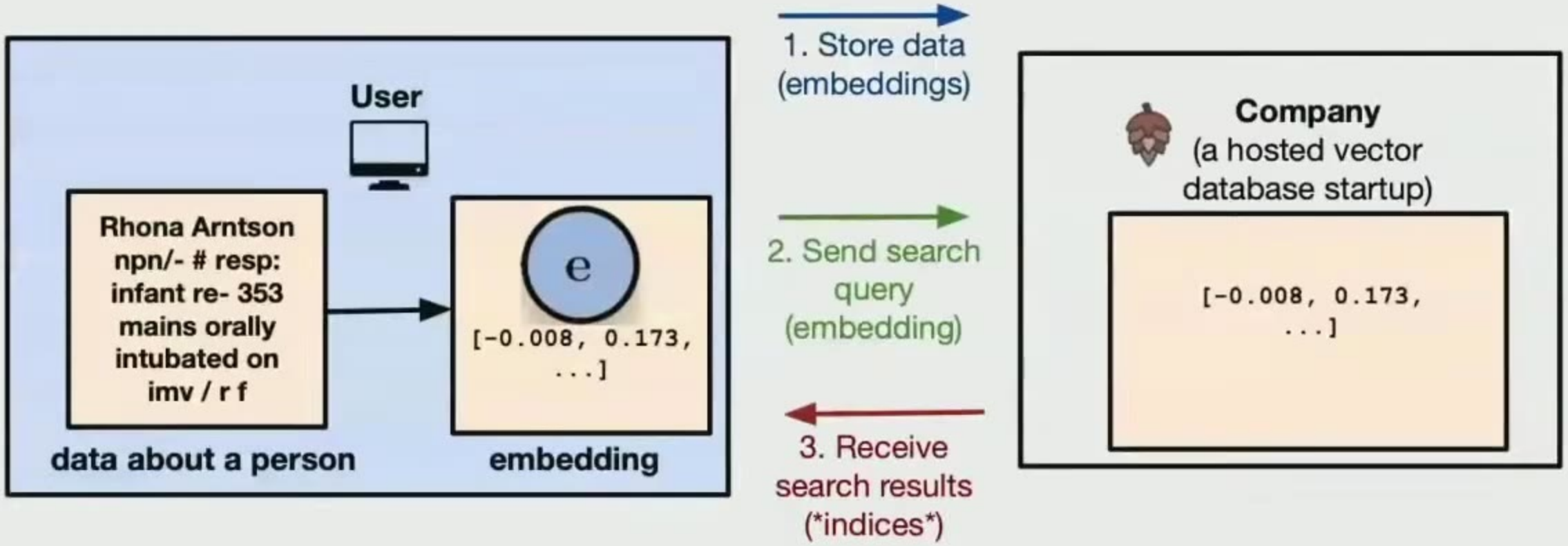
Vector databases are *huge* right now





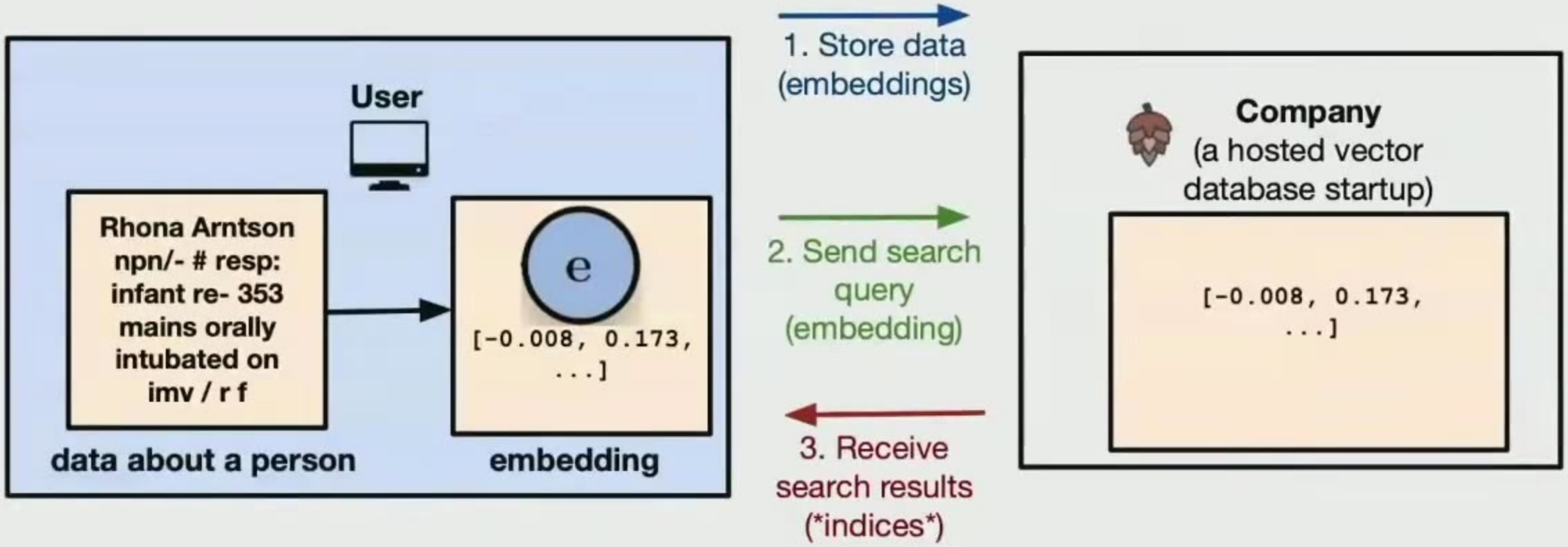
Powered by Zoom

Threat model



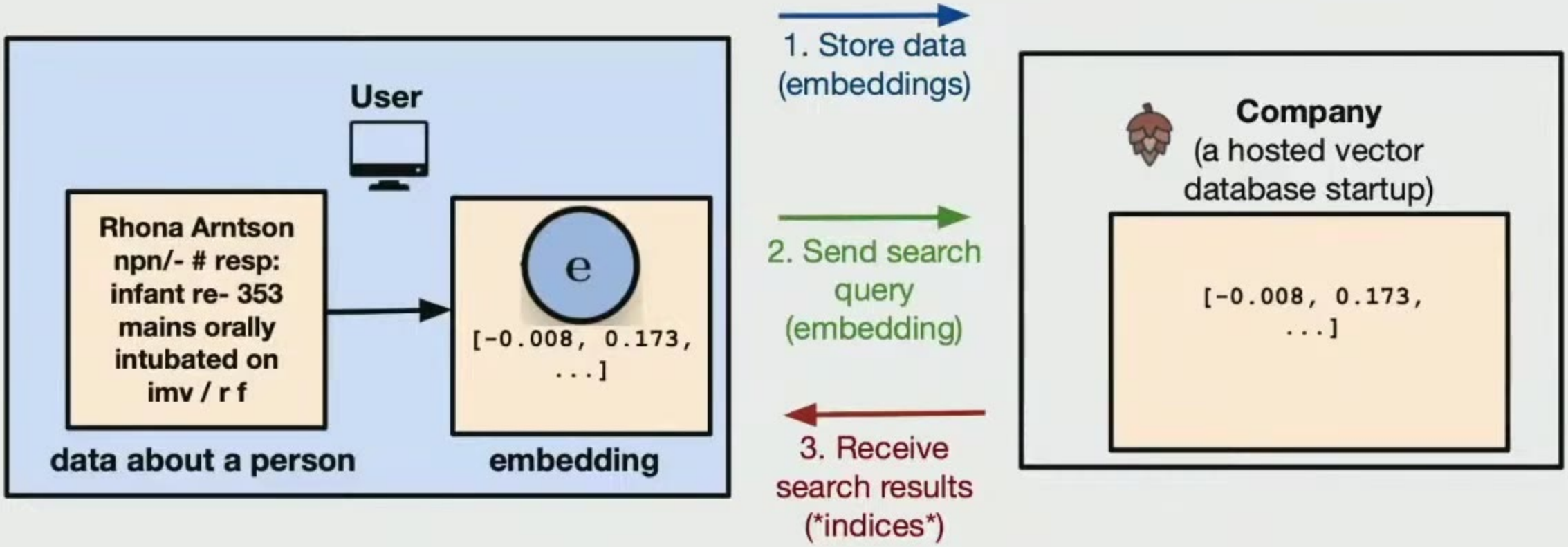
They were tossed into a large database, you'd be able to

Threat model



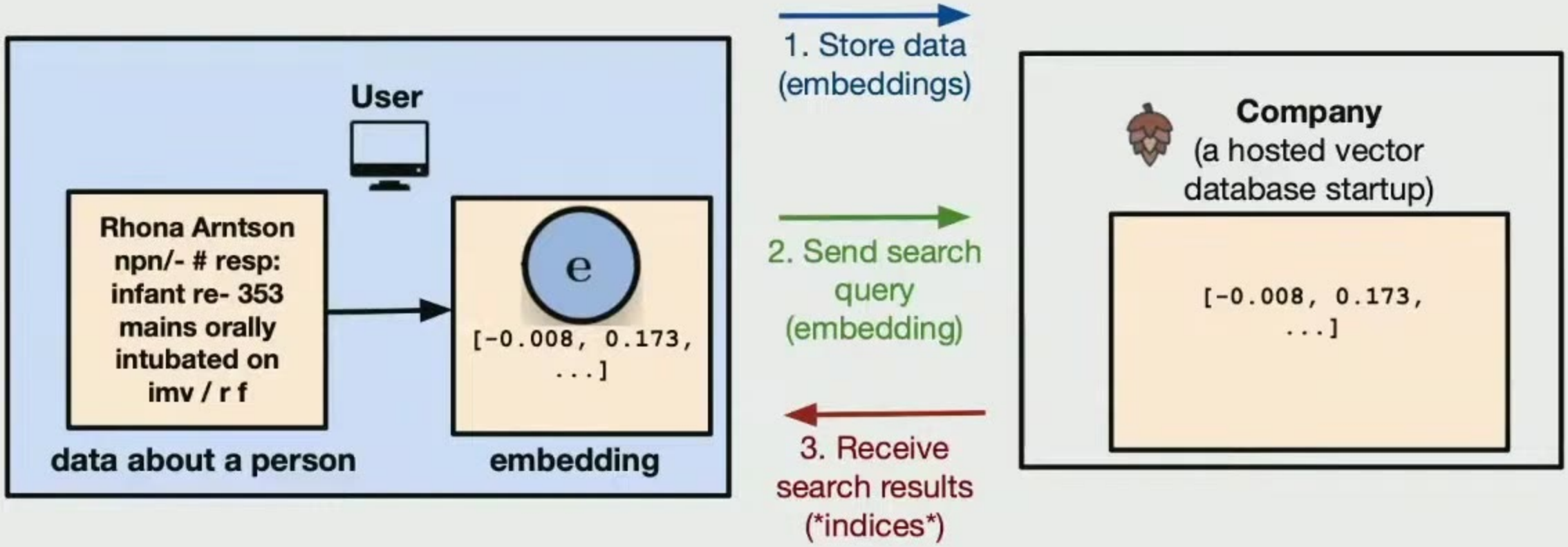
They were tossed into a large database, you'd be able to

Threat model



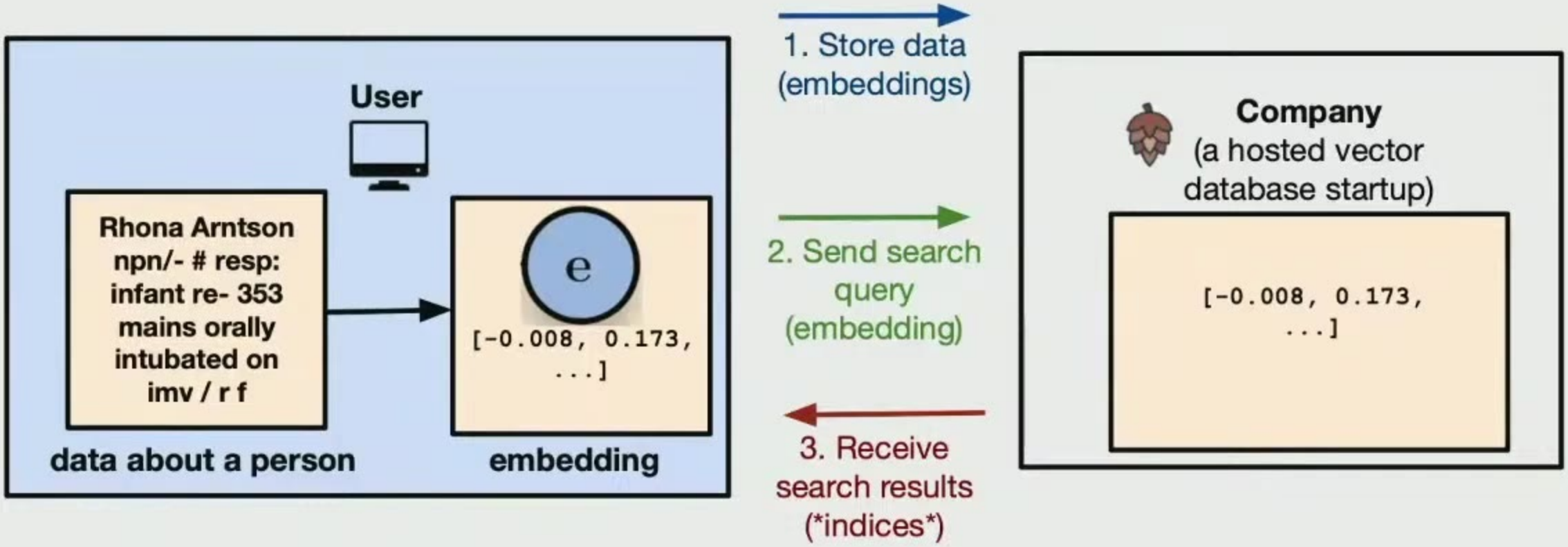
We got pretty interested in,

Threat model



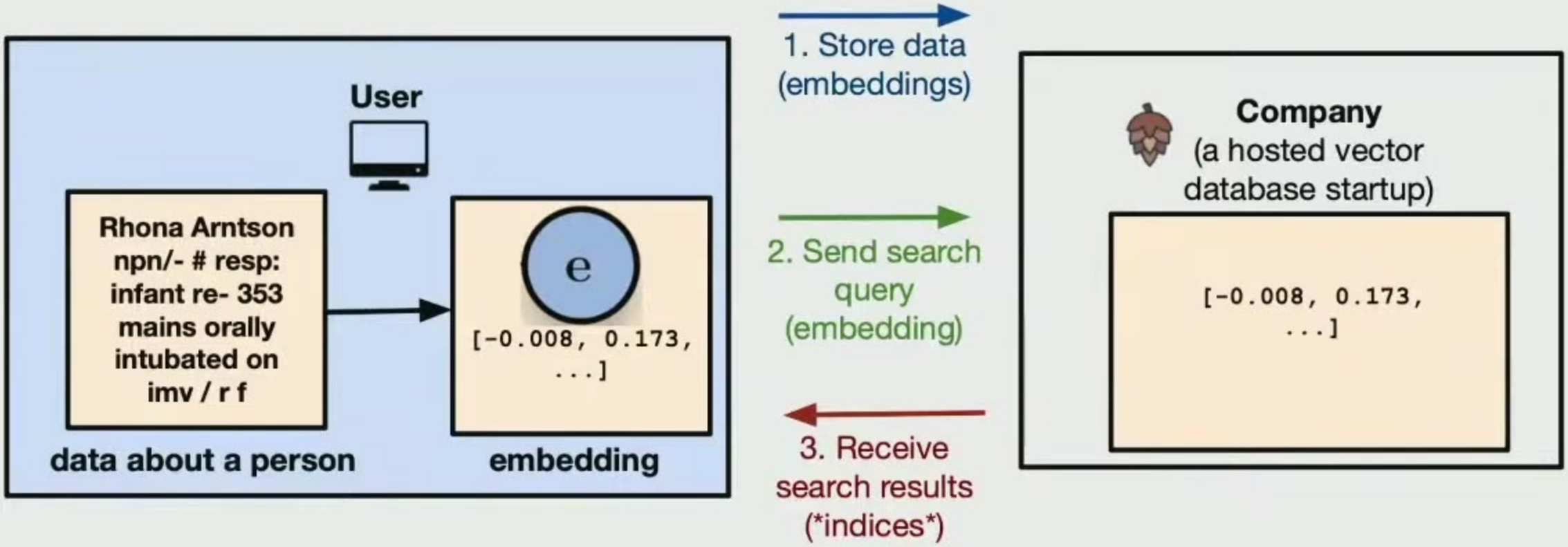
for these sort of databases, which is we didn't really

Threat model



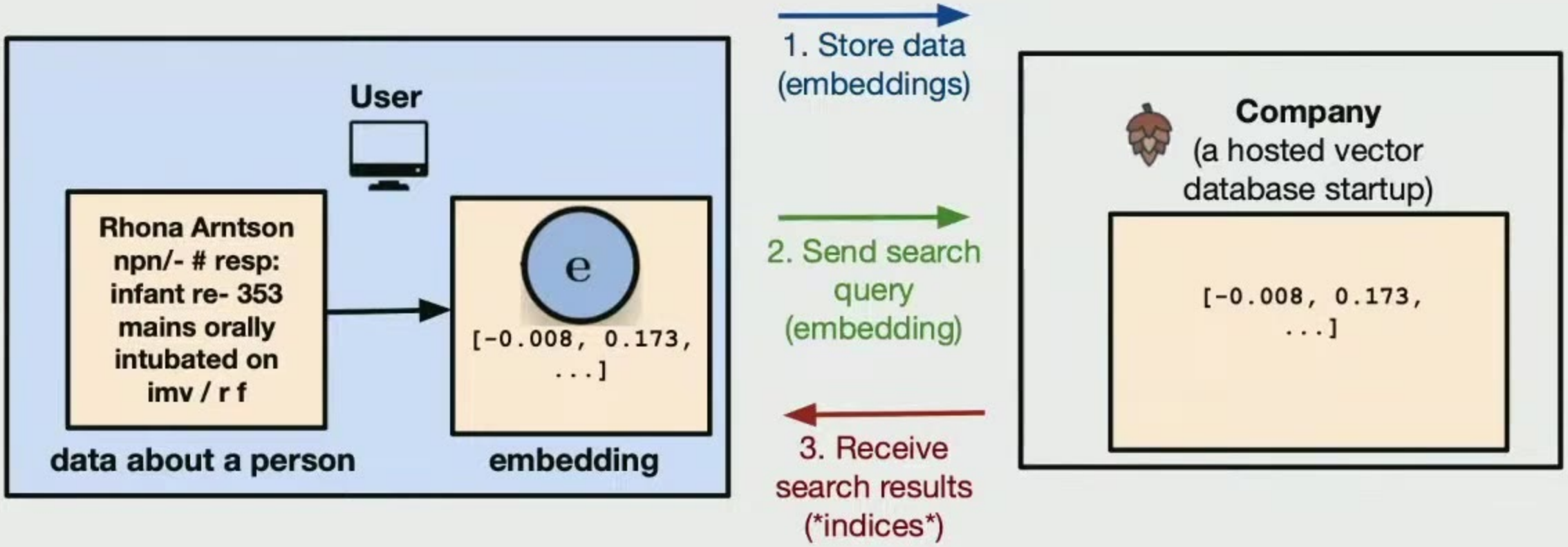
Which is,

Threat model

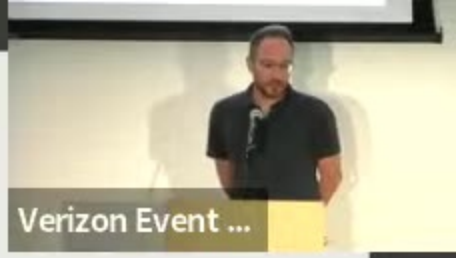


is, imagine that a user got access to an embedding database.

Threat model

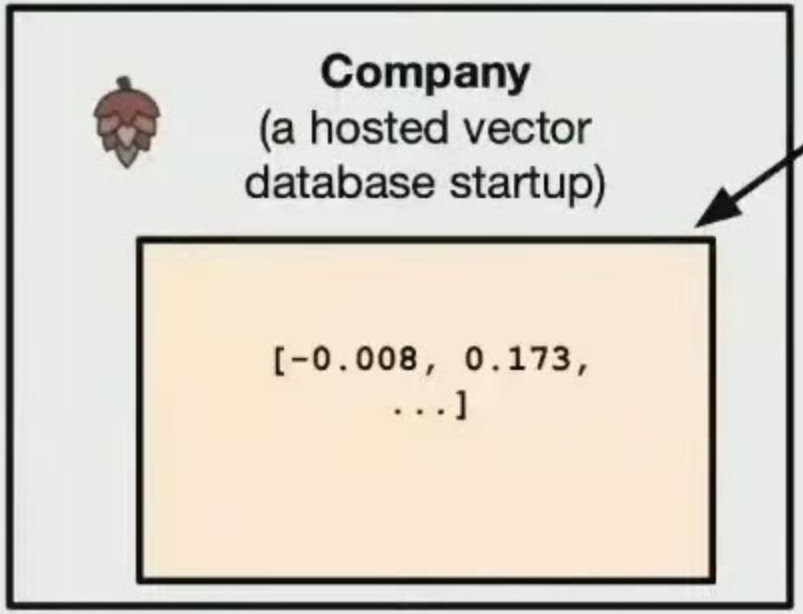


was going to take all their data, send it over to the



Threat model

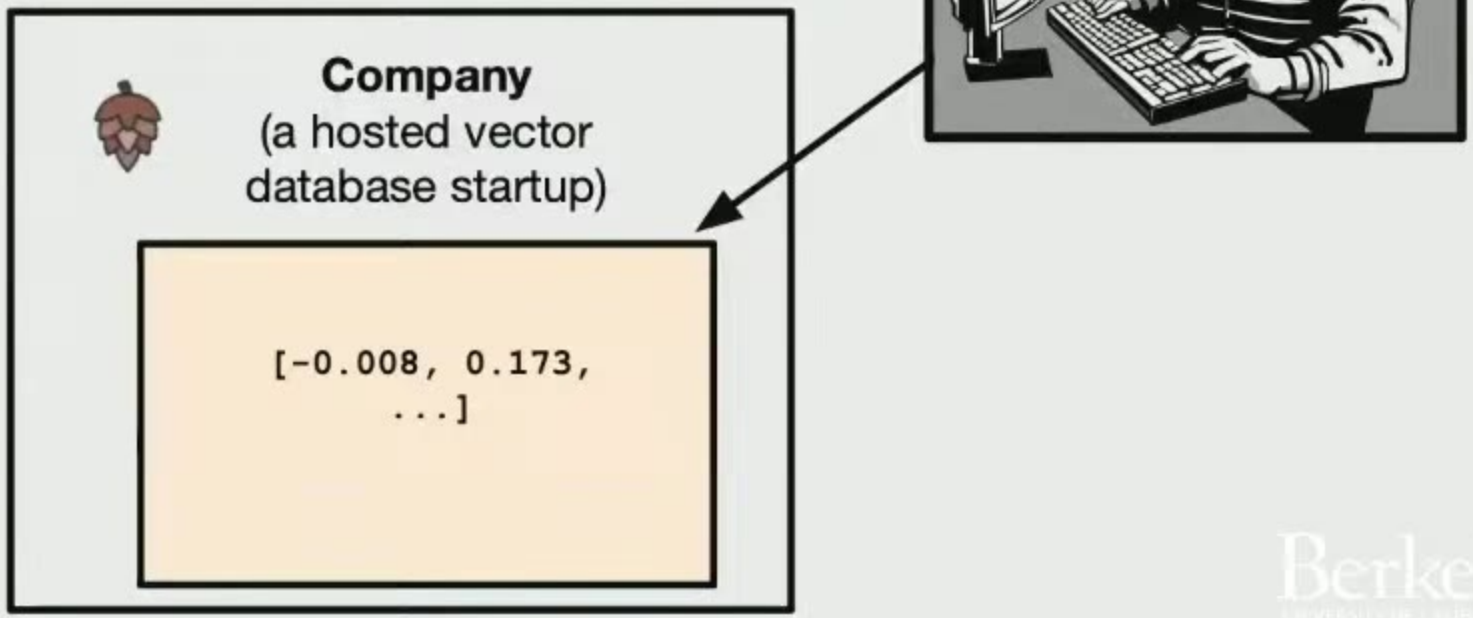
What can a *bad actor* learn from just looking embeddings of text?



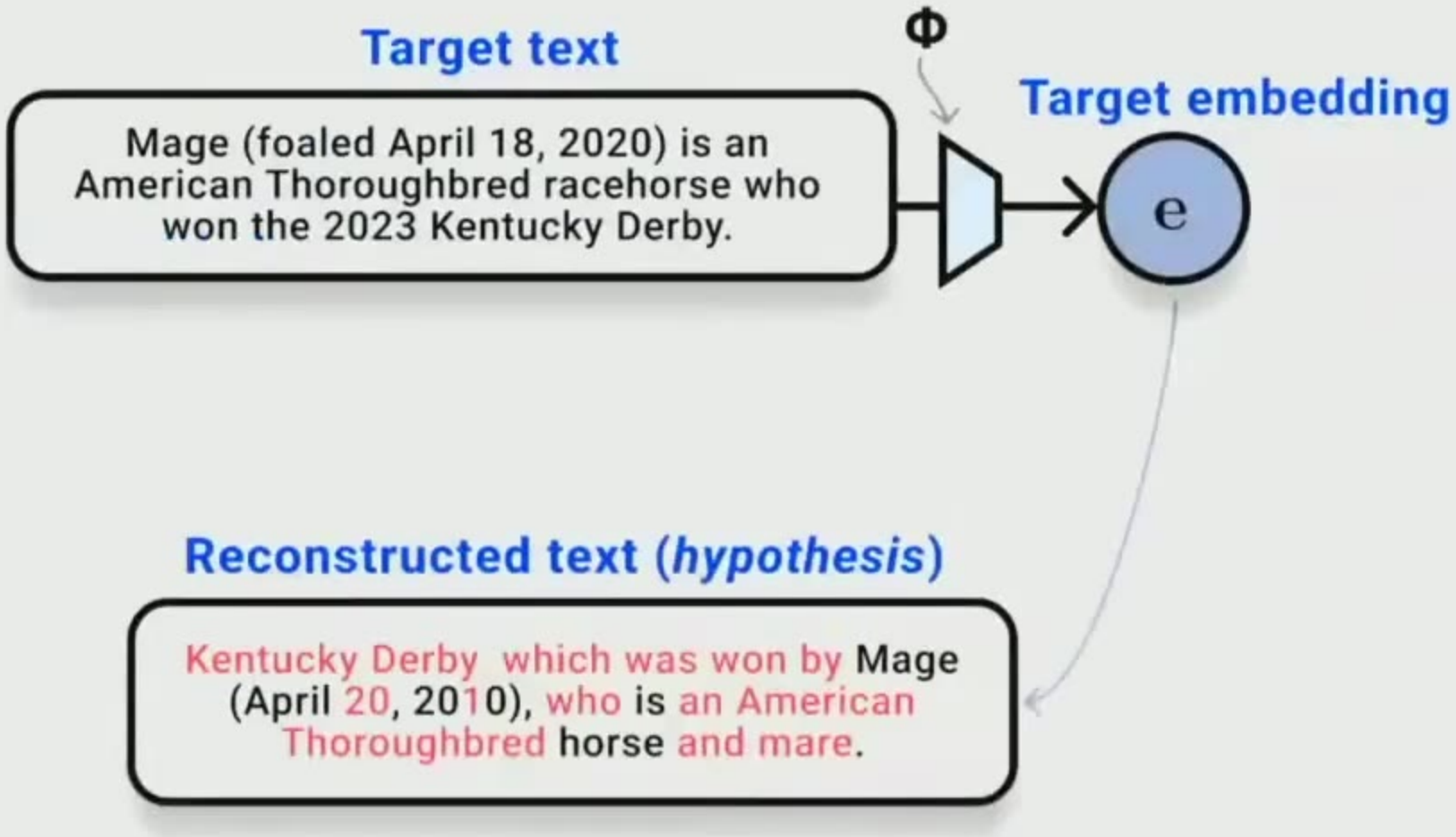
And what could you do

Threat model

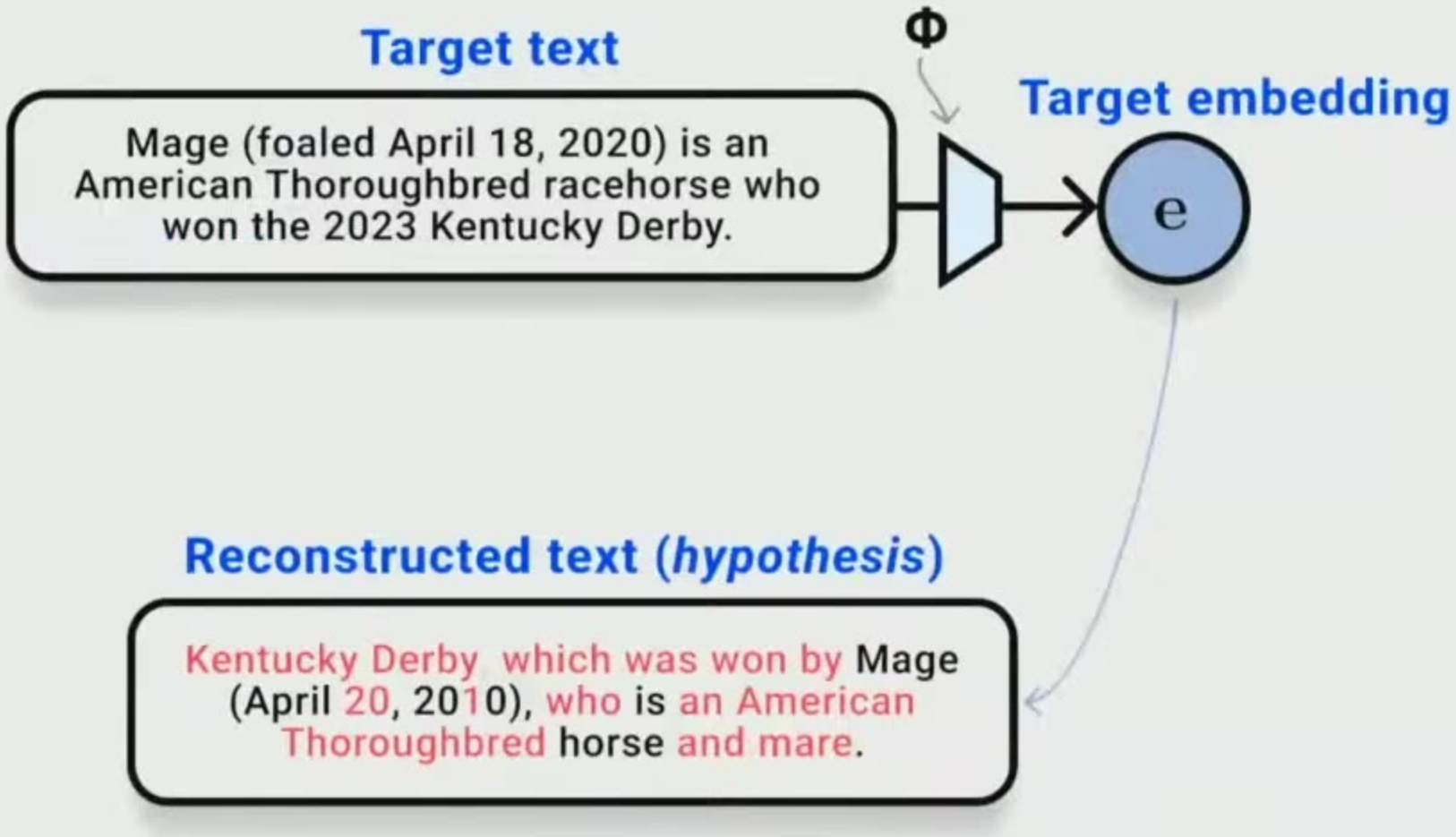
What can a *bad actor* learn from just looking embeddings of text?



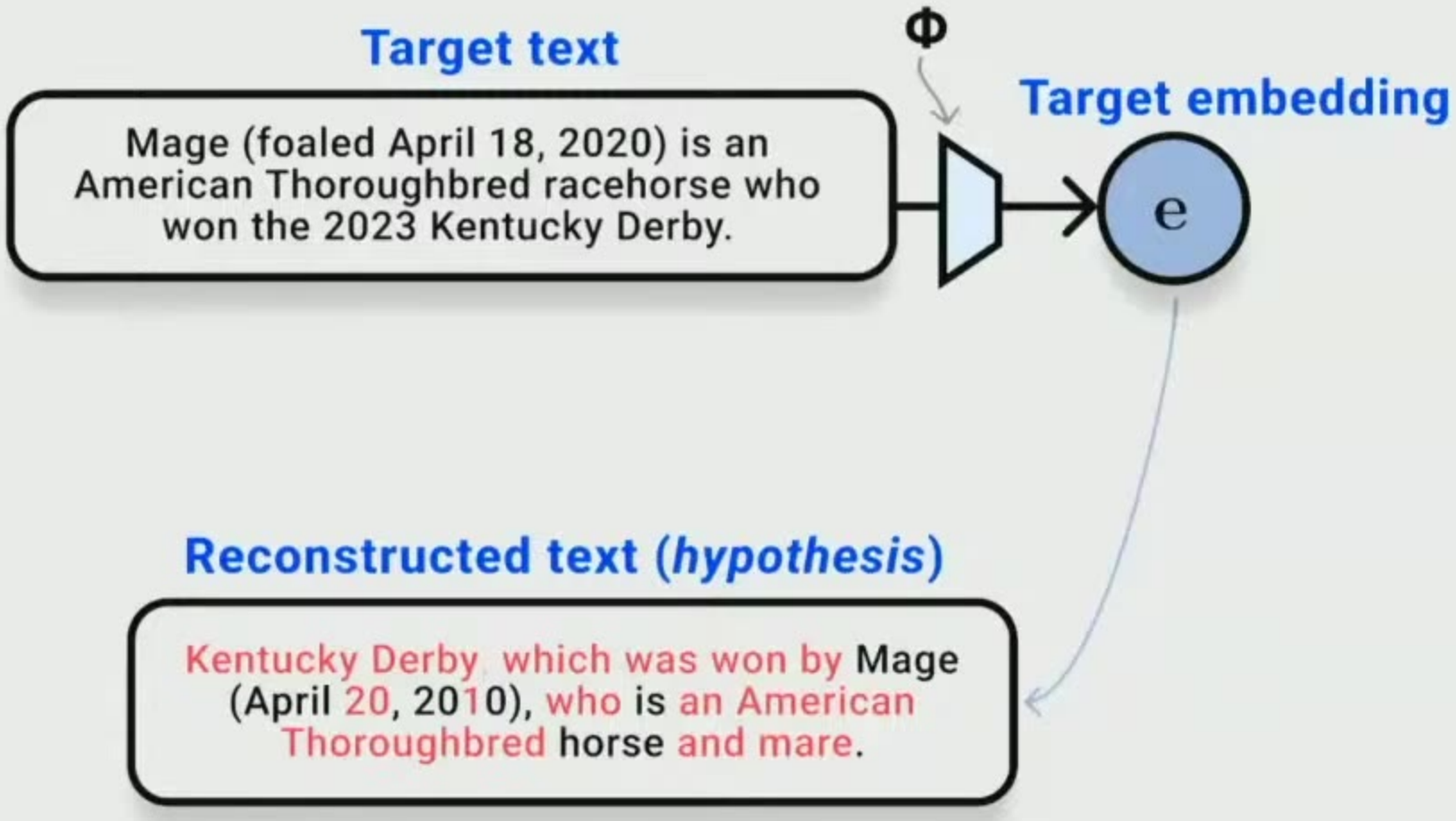
Inverting text embeddings



Inverting text embeddings



Inverting text embeddings



Why is it hard to invert text embeddings?

- Embedding models are trained to have maximum similarity between two similar pieces of text

- But we still find small numerical differences between different paragraphs' embeddings, even with just one word substituted with a synonym

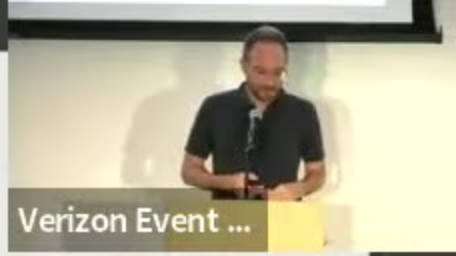
Why is it hard to invert text embeddings?

- Embedding models are trained to have maximum similarity between two similar pieces of text

- But we still find small numerical differences between different paragraphs' embeddings, even with just one word substituted with a synonym

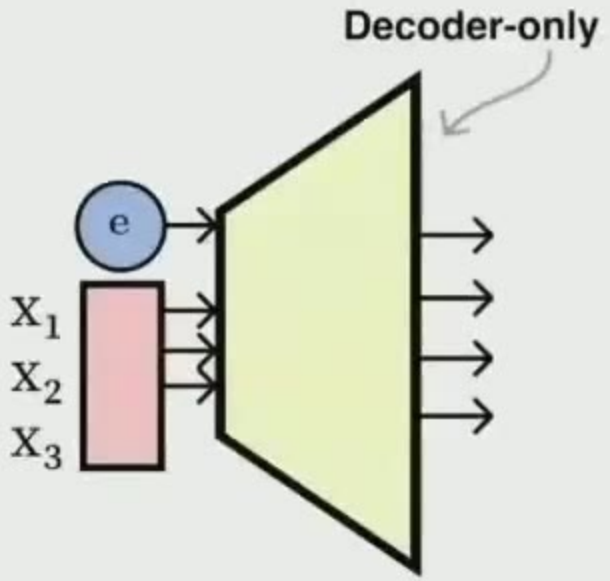
Baseline approach: conditional generation

> How can we generate text from an embedding?



Baseline approach: conditional generation

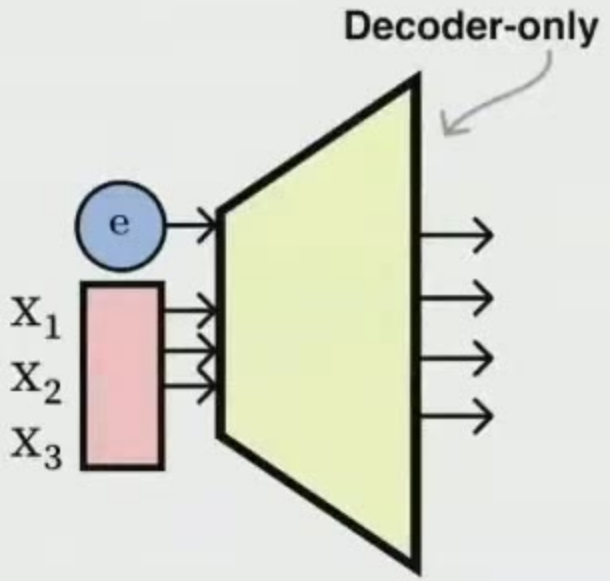
Prior architecture



(Adolphs et al., 2022; Li et al., 2023)

Baseline approach: conditional generation

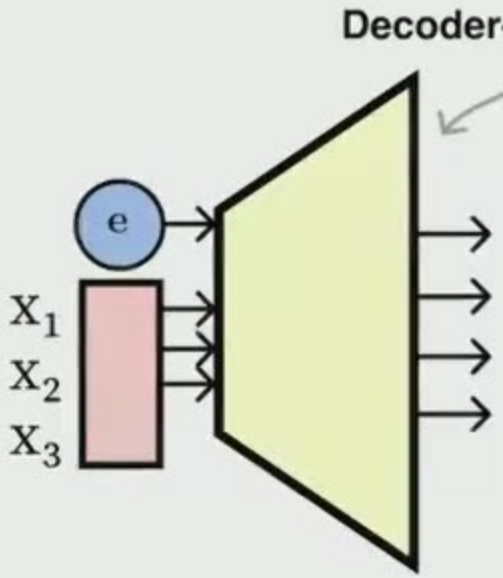
Prior architecture



(Adolphs et al., 2022; Li et al., 2023)

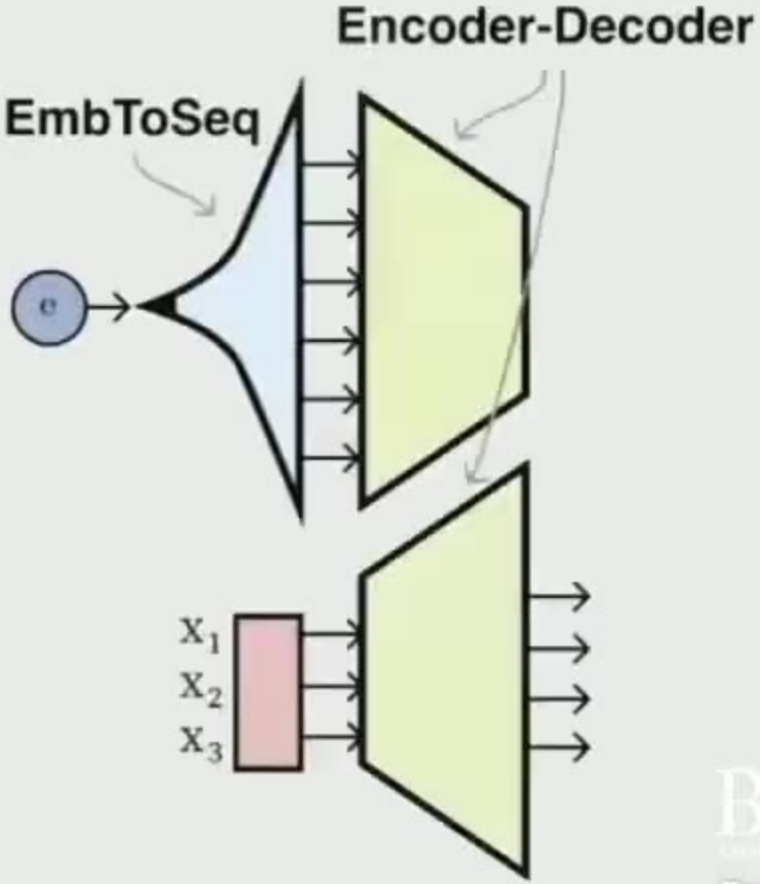
Baseline approach: conditional generation

Prior architecture

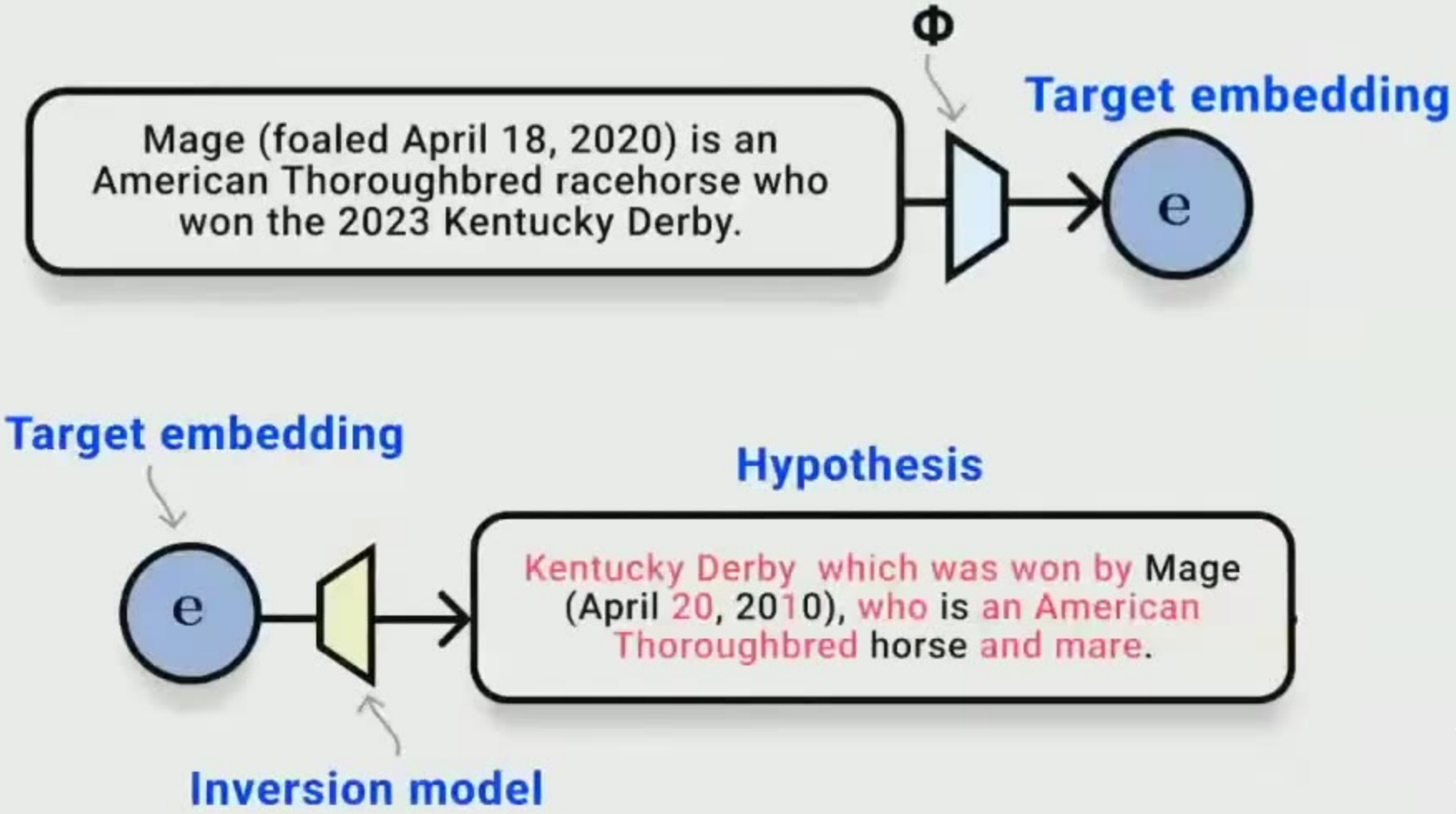


(Adolphs et al., 2022; Li et al., 2023)

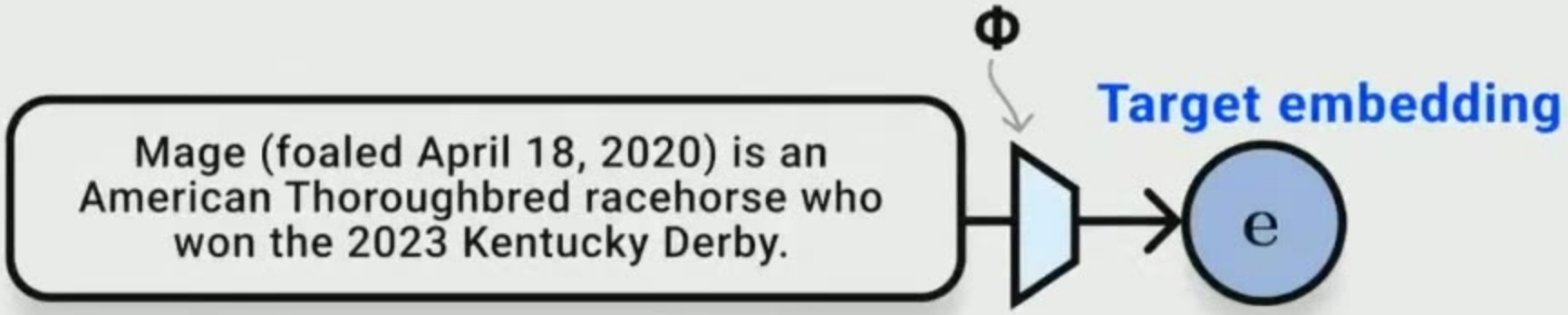
Initial architecture (ours)

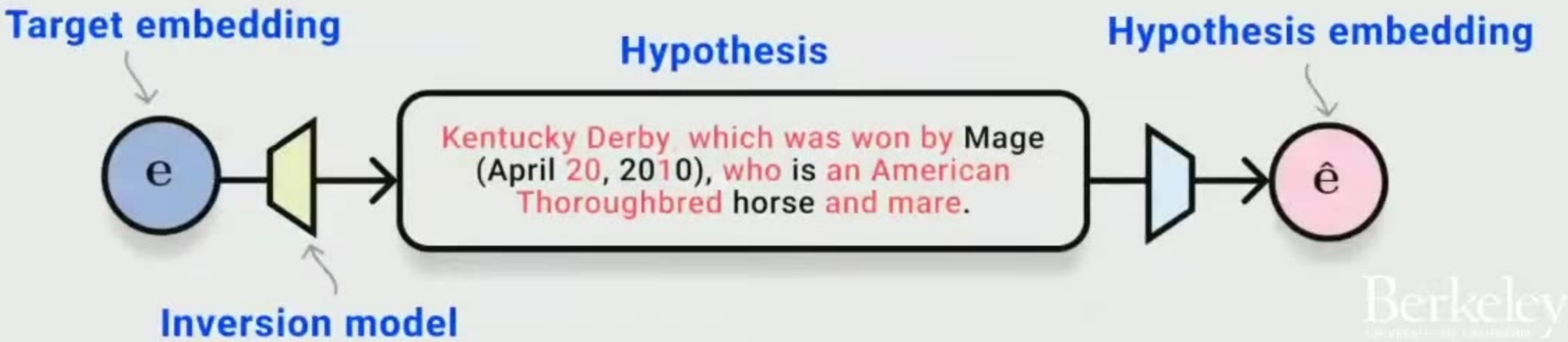


Baseline approach: conditional generation

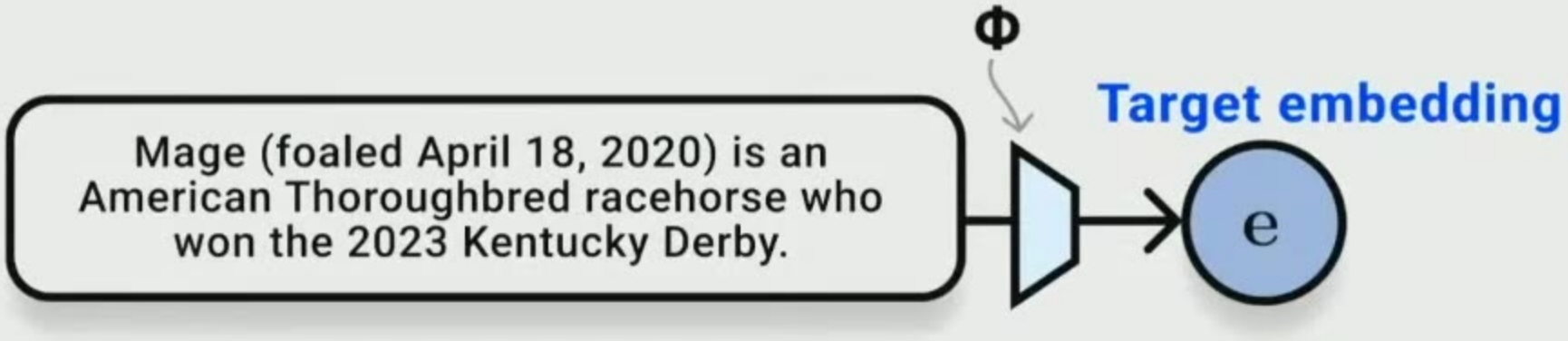


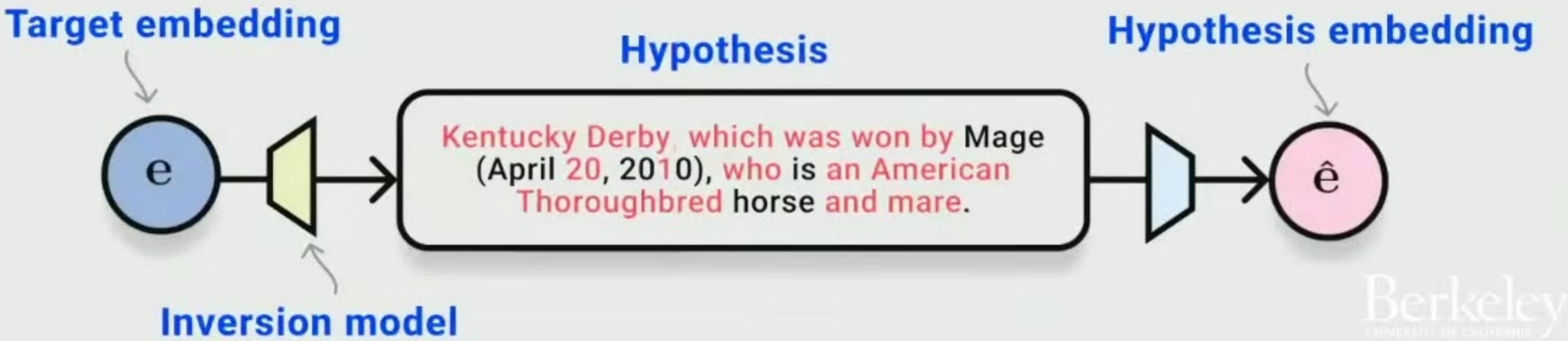
Observation: generated text has a *different* embe



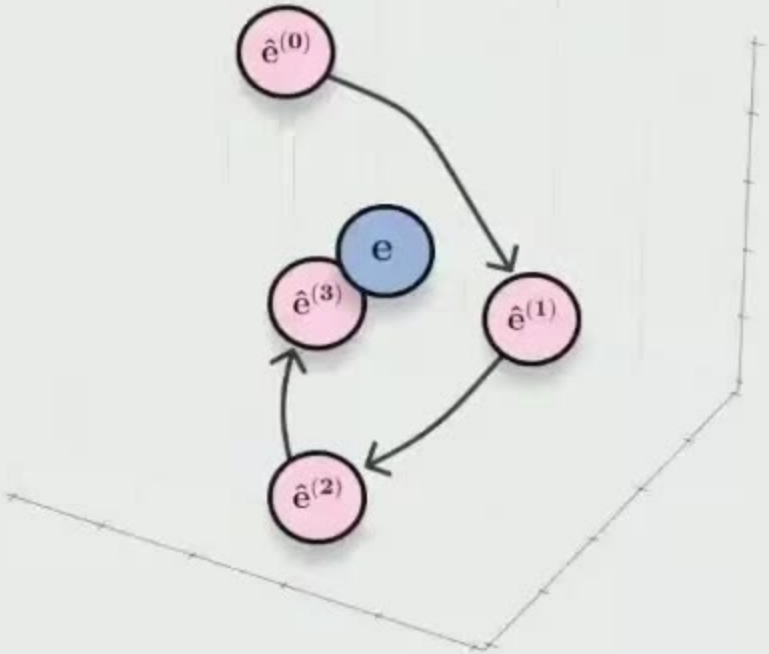


Observation: generated text has a *different* embe



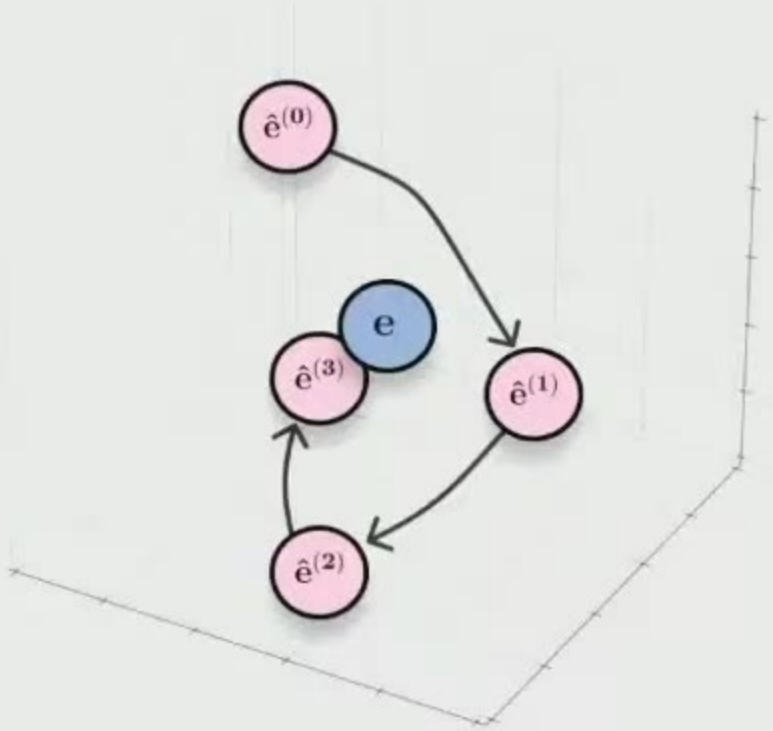


Idea: iteratively refine and re-embed



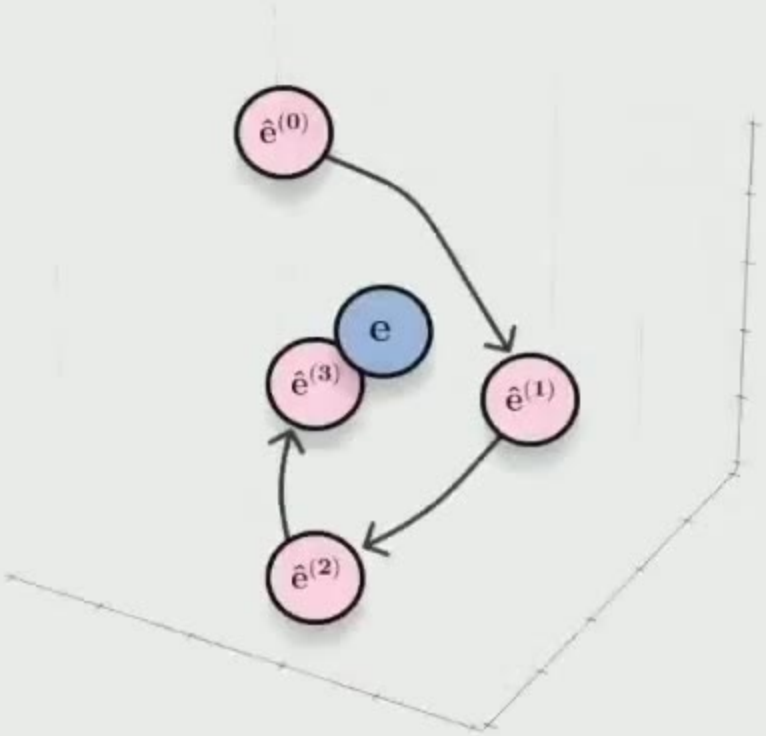
Inspired by approaches to *Iterative Refinement* (Lee et al., 2018; Ghazvininejad et al., 2019; Welleck et al. 2022)

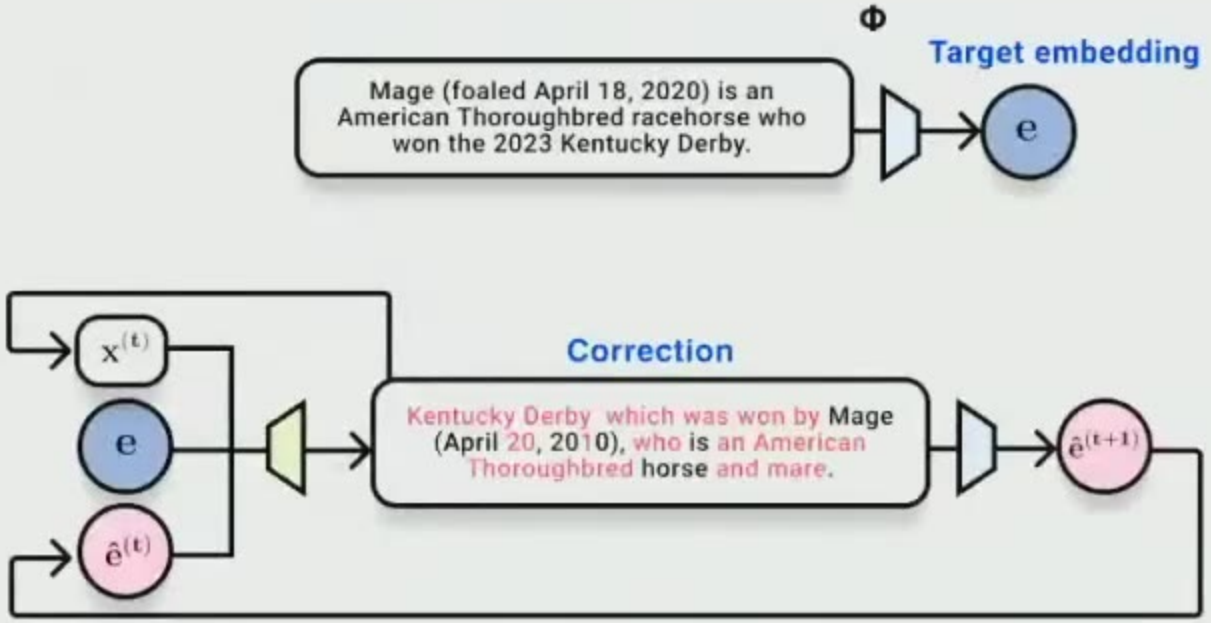
Idea: iteratively refine and re-embed



Inspired by approaches to *Iterative Refinement* (Lee et al., 2018; Ghazvininejad et al., 2019; Welleck et al. 2022)

Method: iteratively refine and re-embed (vec2text)





Method: iterative refinement (*vec2text*)

target embedding

$$p(x^{(t+1)} \mid e) = \sum_{x^{(t)}} p(x^{(t)} \mid e) \cdot p(x^{(t+1)} \mid e, x^{(t)}, \hat{e}^{(t)})$$

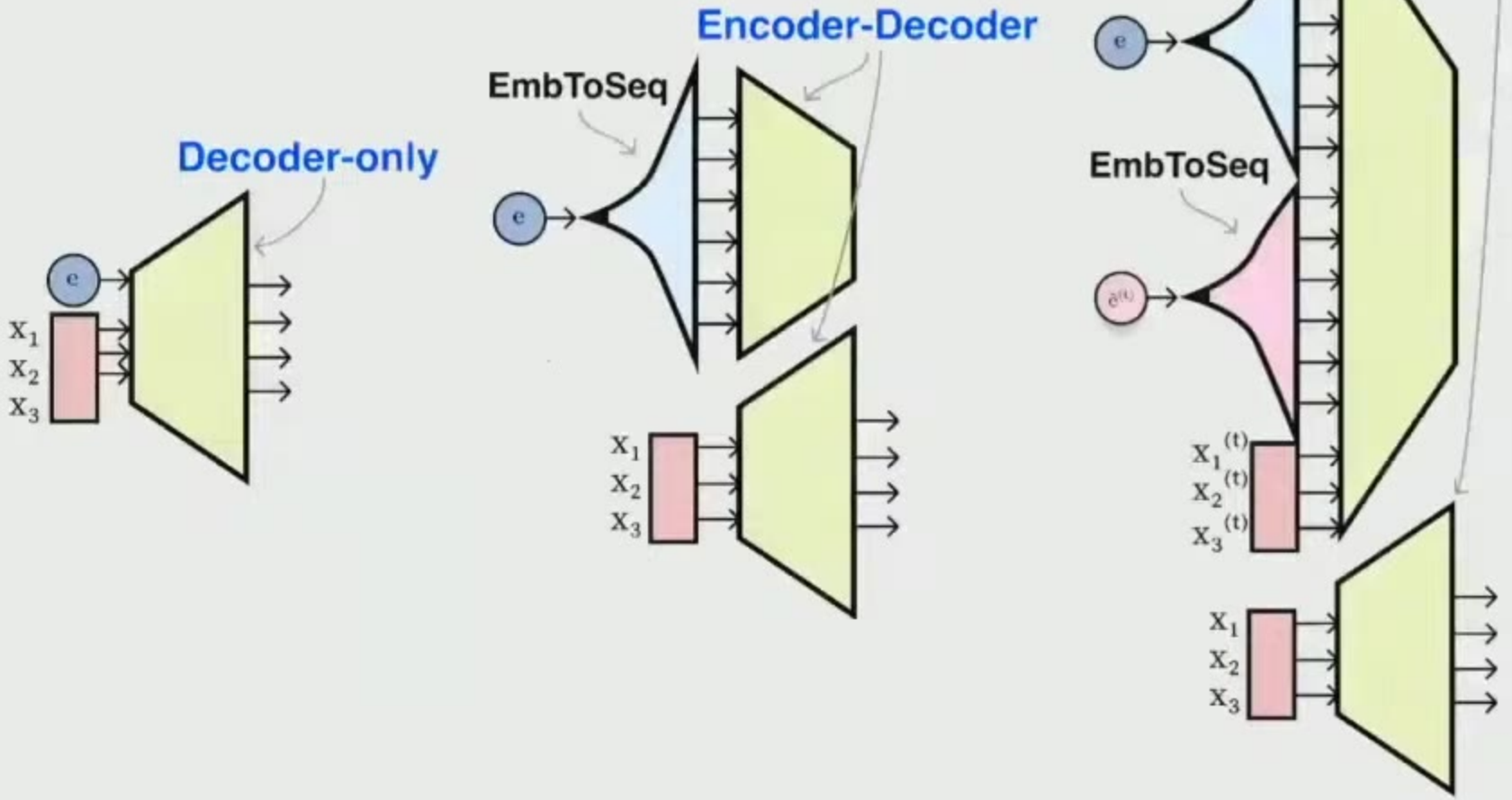
hypothesis (t+1)

$$\hat{e}^{(t)} = \phi(x^{(t)})$$

embedder

hypothesis (t)

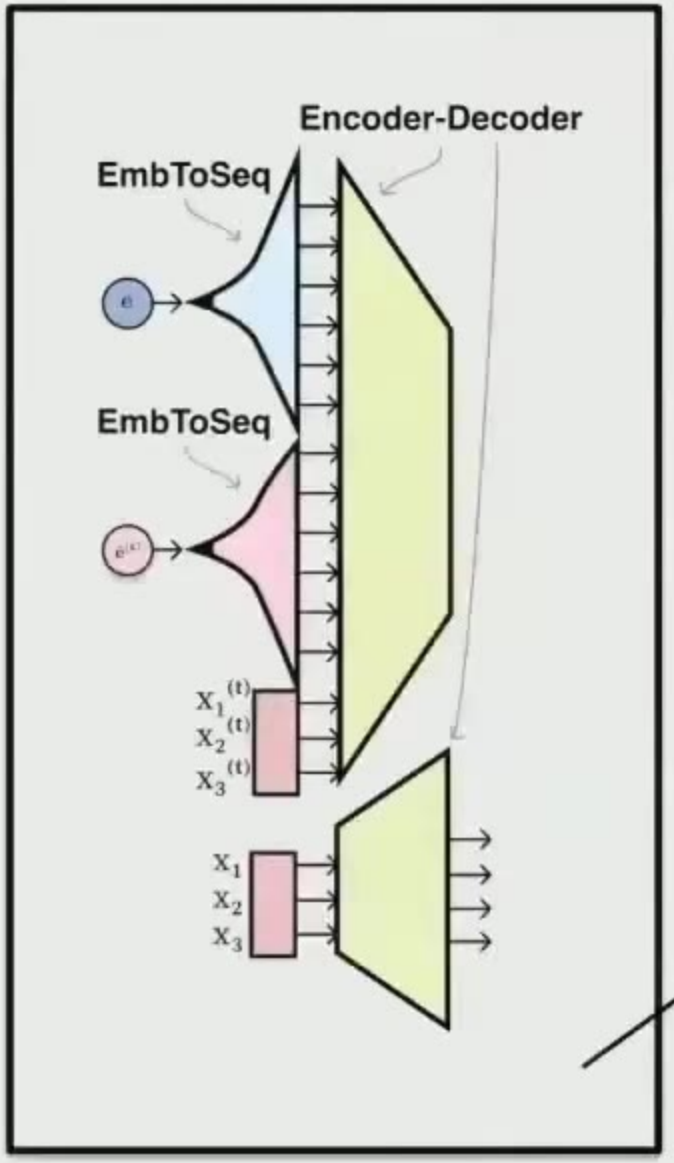
Our approach: iterative generation *(architecture)*



Prior work

Initial approach

vec2text



Φ

Target embedding

Mage (foaled April 18, 2020) is an American Thoroughbred racehorse who won the 2023 Kentucky Derby.

Correction

Kentucky Derby, which was won by Mage (April 20, 2020), who is an American Thoroughbred horse and mare.

Original text

Hypothesis (Round 0)

Embedding

Metrics

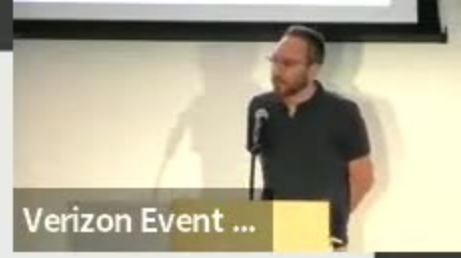
- ***BLEU score*** – approximately how well does the reconstructed text match the original?
- ***Exact match*** – how often did we reconstruct the original text perfectly?

Baselines

- Bag-of-words model
- Decoder-only architecture
- Correction with fewer steps

Results – Natural Questions

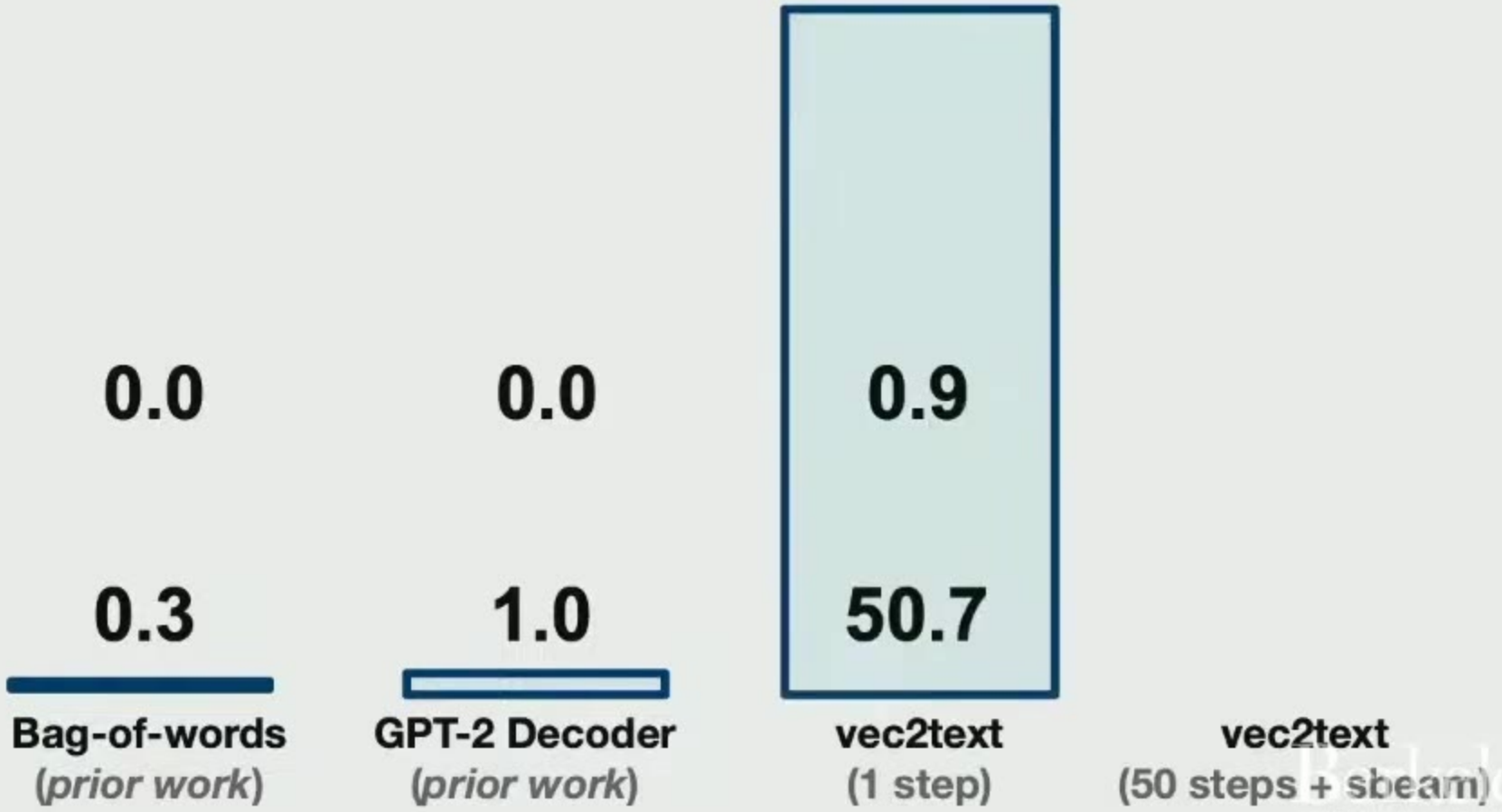
(GTR embeddings || 32 tokens)



	Bag-of-words (*prior work*)	GPT-2 Decoder (*prior work*)	vec2text (1 step)	vec2text (50 steps + sbeam)
[Exact match]	0.0	0.0		
[BLEU Score]	0.3	1.0		

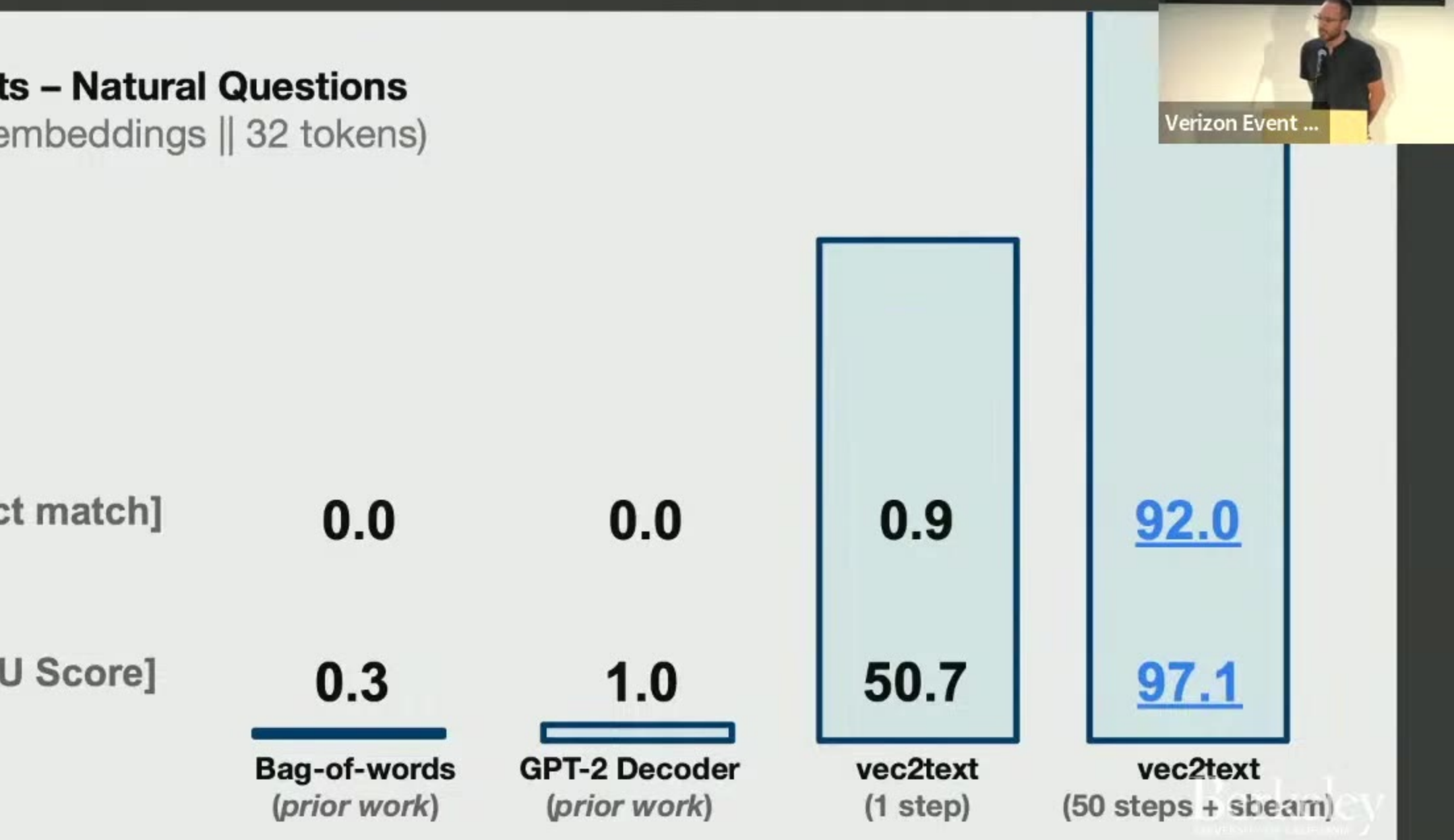
Results – Natural Questions

(GTR embeddings || 32 tokens)



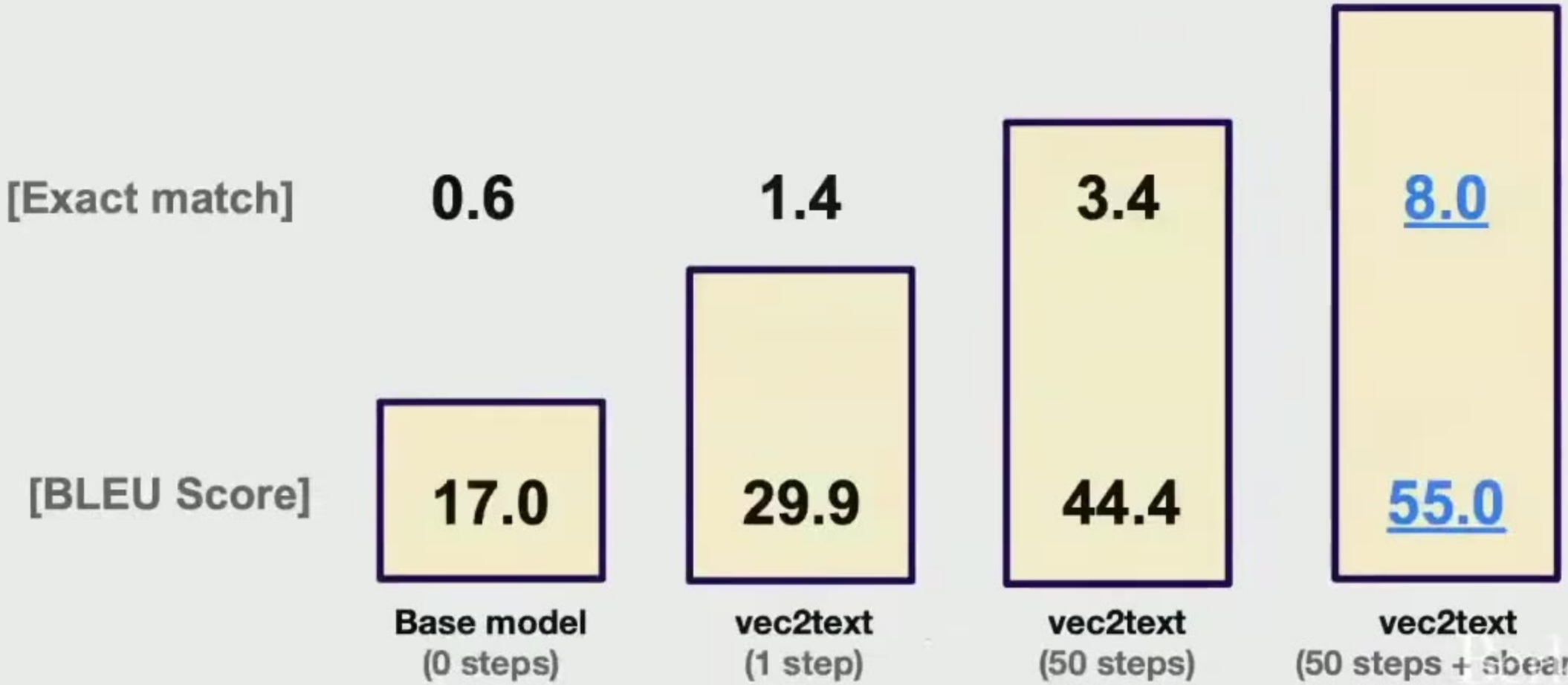
Results – Natural Questions

(GTR embeddings || 32 tokens)

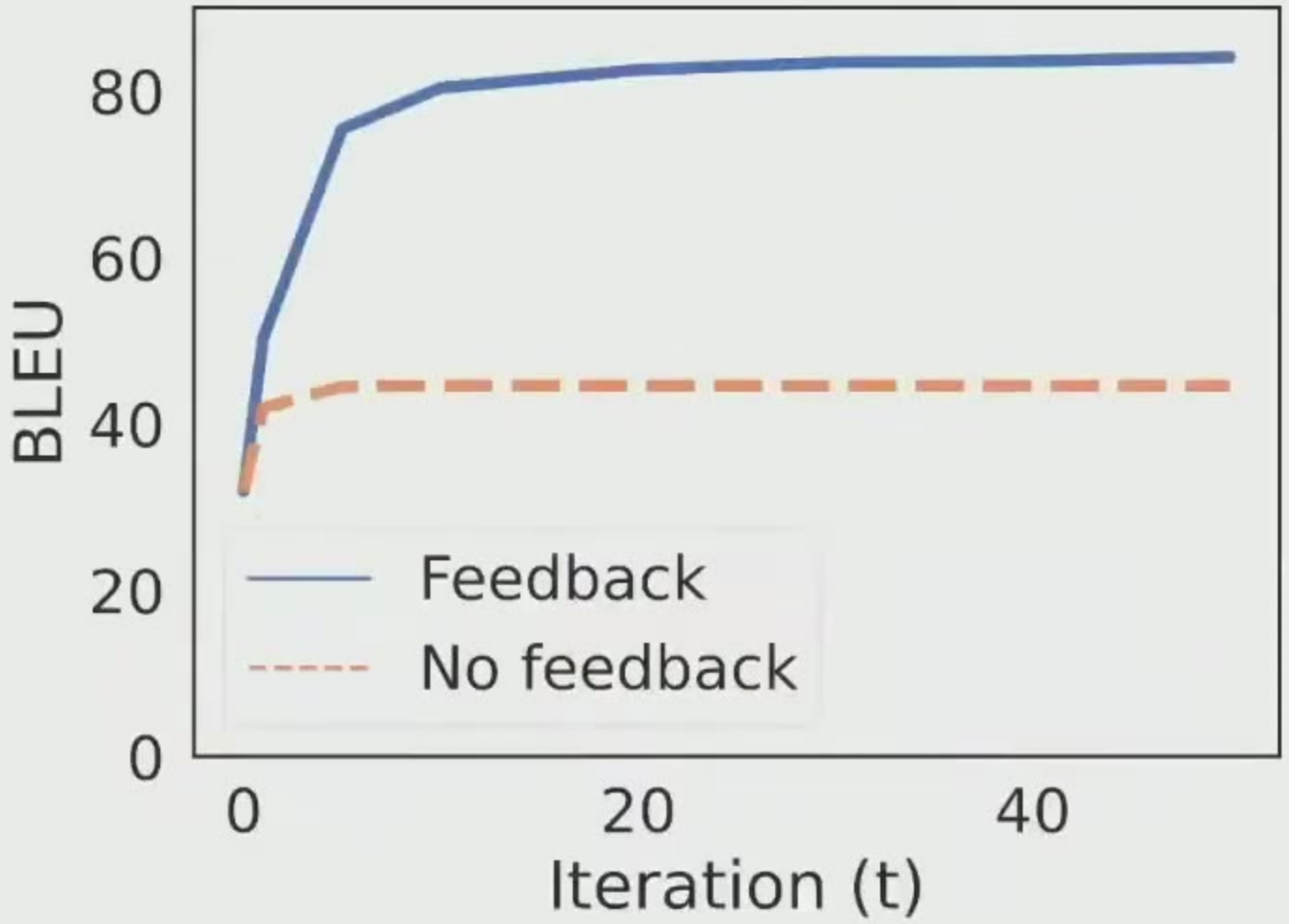


Results – MSMARCO

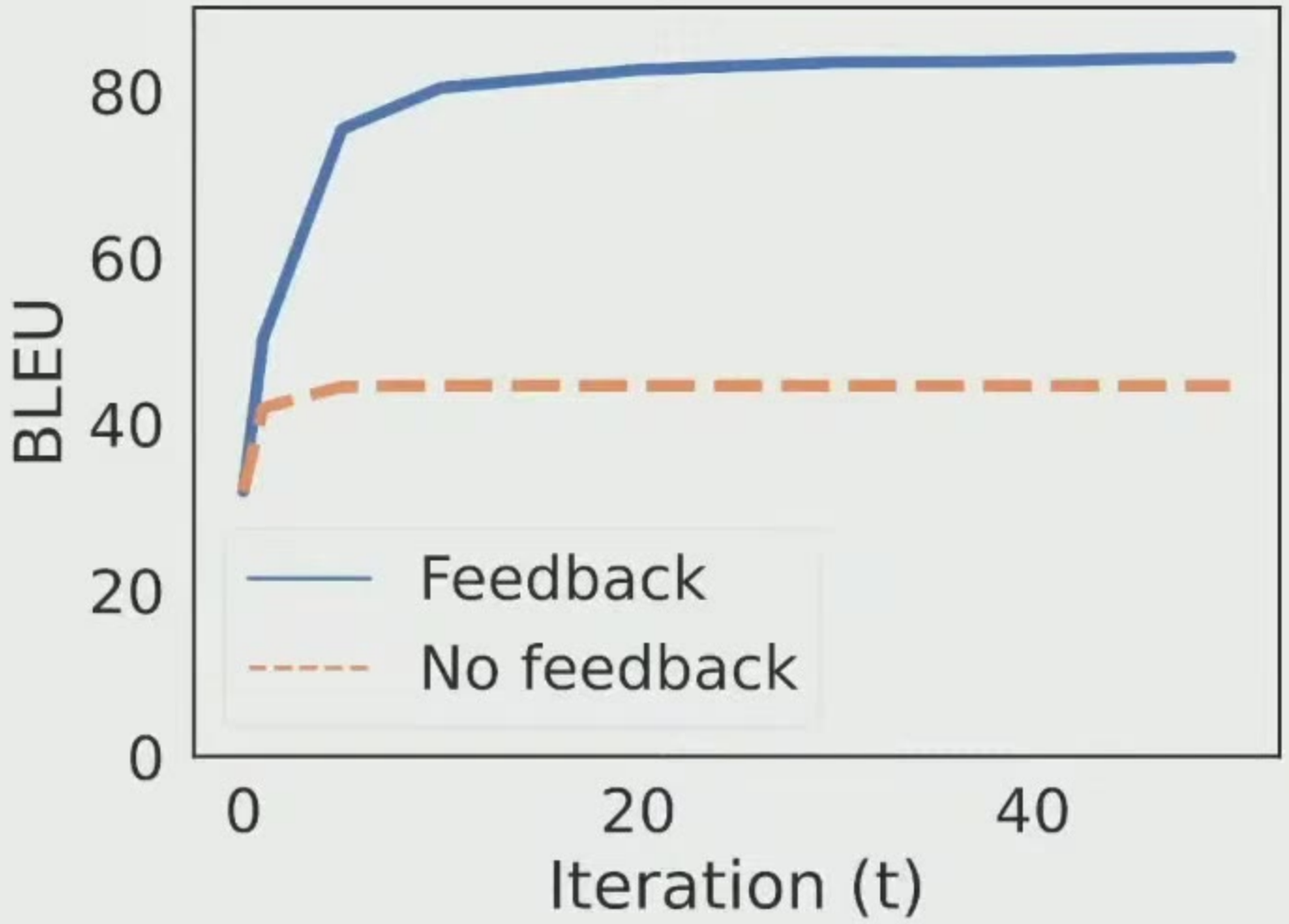
(OpenAI embeddings || 128 tokens)



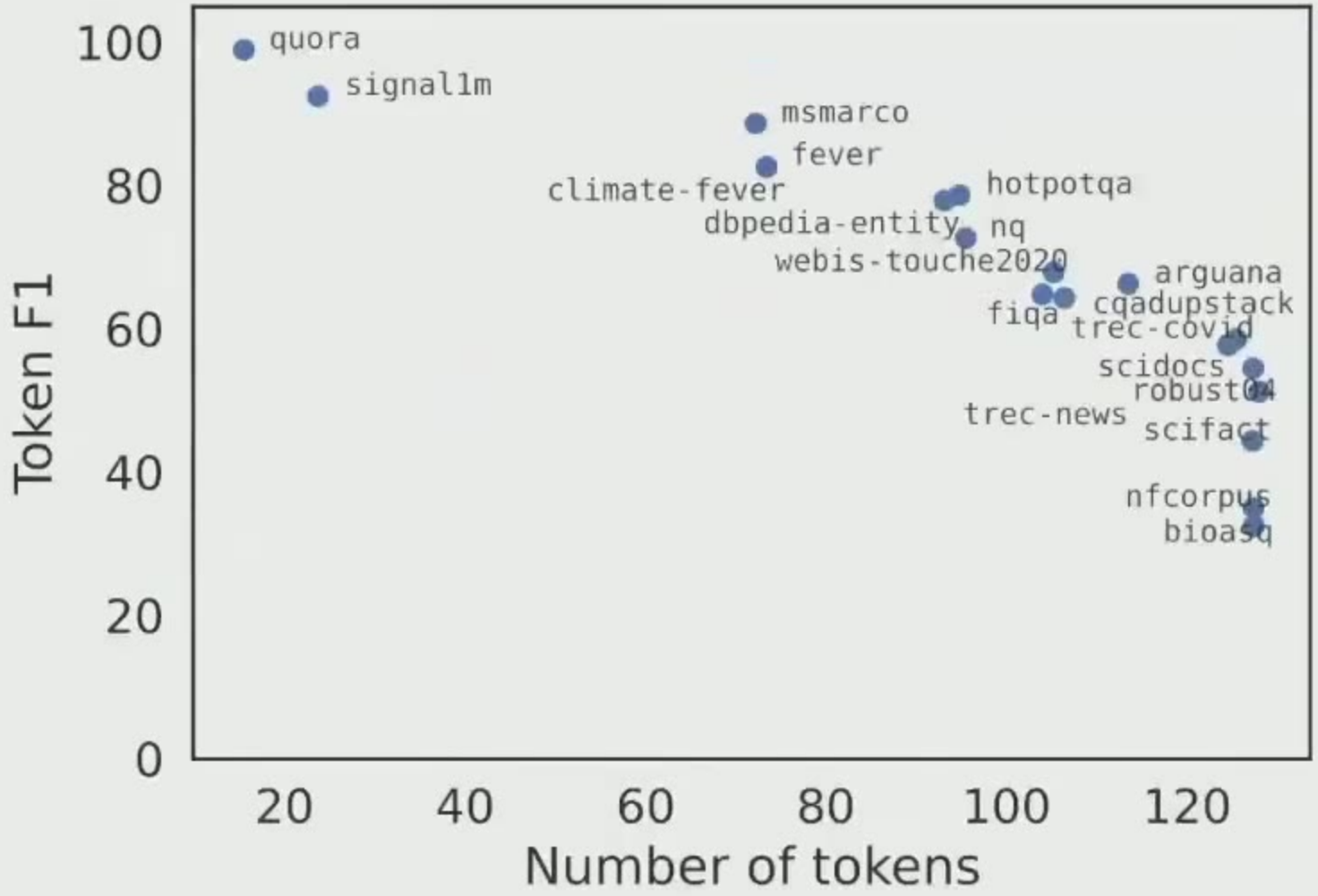
Inversion with and without feedback



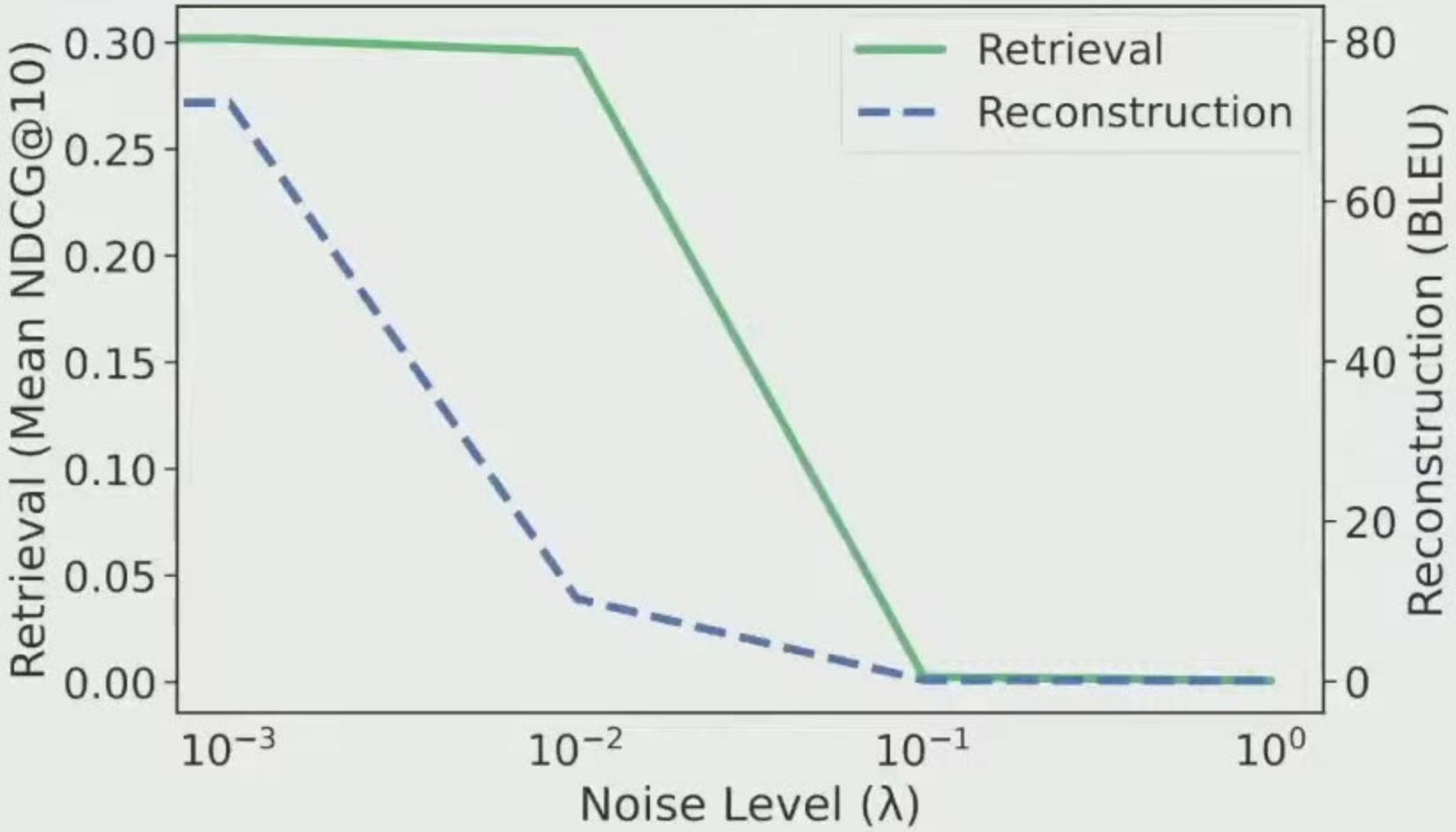
Inversion with and without feedback



Inversion performance vs. sequence length – BEI

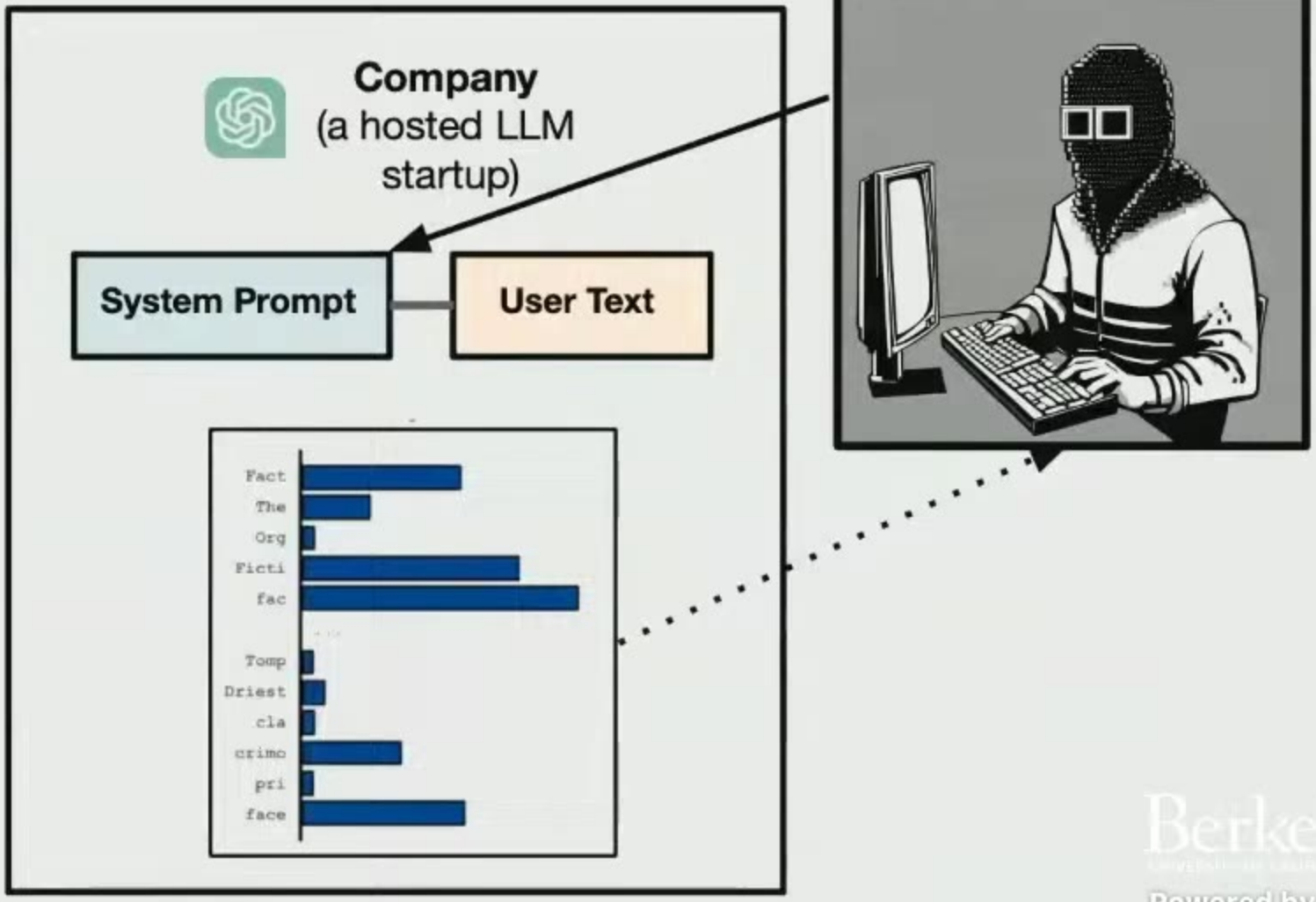


Defending against inversion

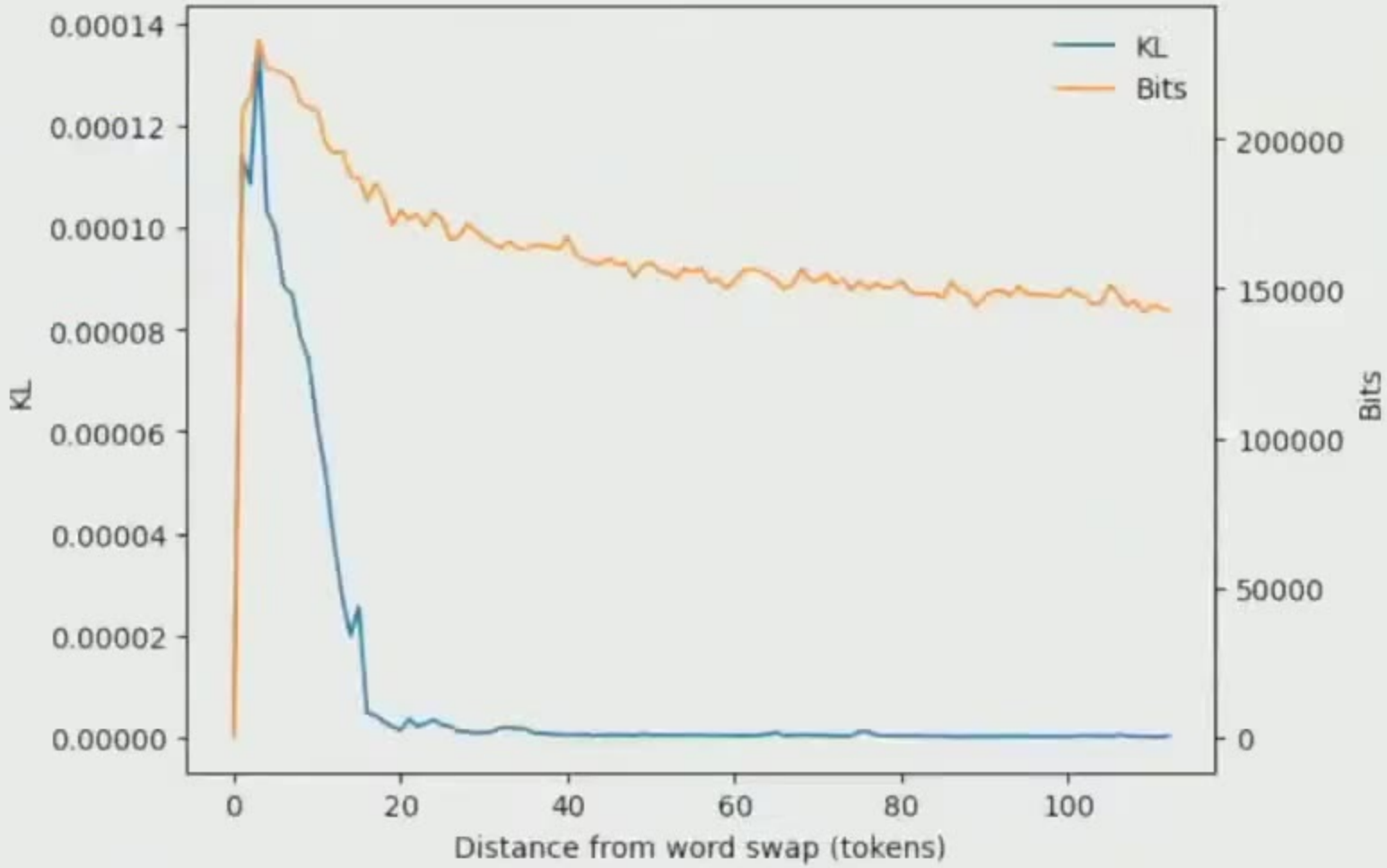


Threat model

What can a *bad actor* learn from just next word distribution of text?



Distributions are similar. Vectors are differen

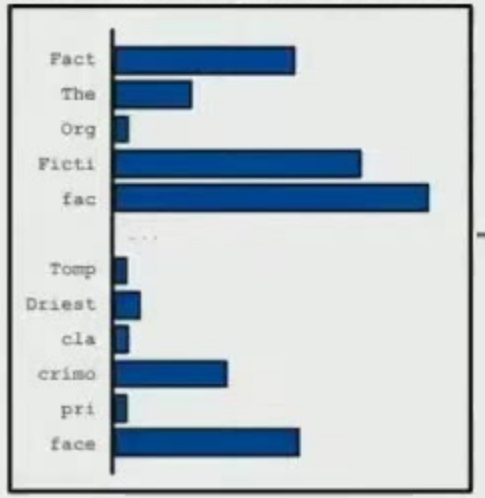


Prompt *(hidden)*

This is a conversation between a curious user and an artificial intelligence assistant. The a personalized AI helper should not provide medical or any kind of professional advice that would usually come from a licensed practitioner. Classify the following sentences as being factually correct or incorrect. The Earth is the third planet from the Sun.

Language model

Next-token probabilities



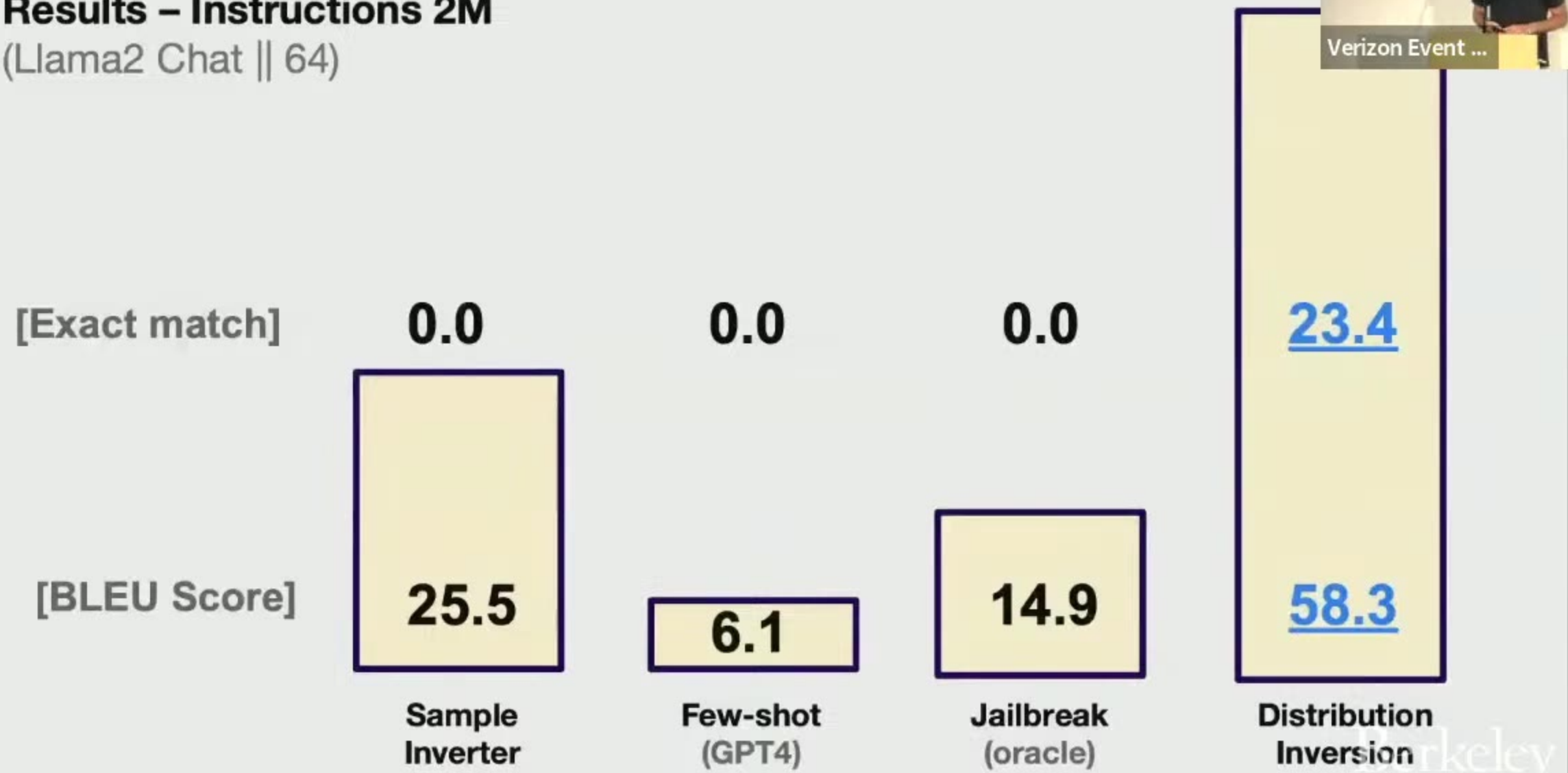
Inversion model

Reconstruction

This is a conversation between a curious user and an AI assistant. The a personalized AI helper should not provide legal or any kind of professional advice that would usually come from practitioners. Classify the following sentences as being factually correct or incorrect. The third planet from the Sun is called Earth.

Results – Instructions 2M

(Llama2 Chat || 64)



LMs Inside Out

- Iterative methods can recover embeddings
- Embeddings are not compressive

Language models

- LM distributions are information dense
- Jailbreak protected models can still be recovered

jxmorris12/vec2text

Berkeley
UNIVERSITY OF CALIFORNIA

LMs Inside Out

- Iterative methods can recover embeddings
- Embeddings are not compressive

Language models

- LM distributions are information dense
- Jailbreak protected models can still be recovered

jxmorris12/vec2text

Berkeley
UNIVERSITY OF CALIFORNIA

LMs Inside Out

- Iterative methods can recover embeddings
- Embeddings are not compressive

Language models

- LM distributions are information dense
- Jailbreak protected models can still be recovered

jxmorris12/vec2text

Berkeley
UNIVERSITY OF CALIFORNIA

Berkeley Center for Responsible, Decentralized Intelligence

Summit on Responsible Decentralized Intelligence — Future of Decentralization and AI

August 6, 2024

Verizon Center, NYC

Berkeley UNIVERSITY OF CALIFORNIA

Powered by Zoom

Berkeley Center for Responsible, Decentralized Intelligence

Summit on Responsible Decentralized Intelligence — Future of Decentralization and AI

August 6, 2024

Verizon Center, NYC

Berkeley UNIVERSITY OF CALIFORNIA

Powered by Zoom

Berkeley

Center for Responsible, Decentralized Intelligence

Summit on Responsible Decentralized Intelligence — Future of Decentralization and AI

August 6, 2024

Verizon Center, NYC

Berkeley
UNIVERSITY OF CALIFORNIA

Powered by Zoom

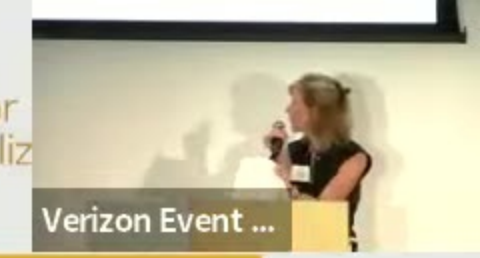
Berkeley Center for Decentraliz

Session I: Open Source AI

Building a Global AI Safety Benchmark

Mala Kumar
Director of Program Management
AI Safety
MLCommons

Berkeley Center for Decentraliz



Session I: Open Source AI

Building a Global AI Safety Benchmark

Mala Kumar
Director of Program Management
AI Safety
MLCommons

Berkeley

Powered by Zoom