Berkeley Center for Decentraliz

Verizon Event ...

Session III: Foundations of Decentralized AI (II)

A Waterlog for for Large Language Models

Ian Miers
Assistant Professor, University of Maryland;
Lead Research Scientist, Aleo Foundation

Berkeley UNIVERSITY OF CALIFORNIA

Powered by Zoom

Berkeley Center for Decentraliz

Session III: Foundations of Decentralized AI (II)

A Waterlog for for Large Language Models

Ian Miers
Assistant Professor, University of Maryland;
Lead Research Scientist, Aleo Foundation

A Waterlog for Large Language Models

Ian Miers

University of Maryland

(Joint work with Brennon Brimhall, Orion Weller, and Matthew D Green)

Symmetric-key watermarks

- Goal: was this generated by my LLM?
- Basics: A Watermark for LLMs. (KGWK**M**G ICML 2024)
  - hash a secret key and the last *n* tokens
  - Generate a “red” and “green” list of tokens
  - Up sample green tokens mdown sample red
- To verify, run statistical test
- Issues:
  - Must trust LLM provider to verify
  - LLM provider can frame people
  - Not very robust to paraphrasing

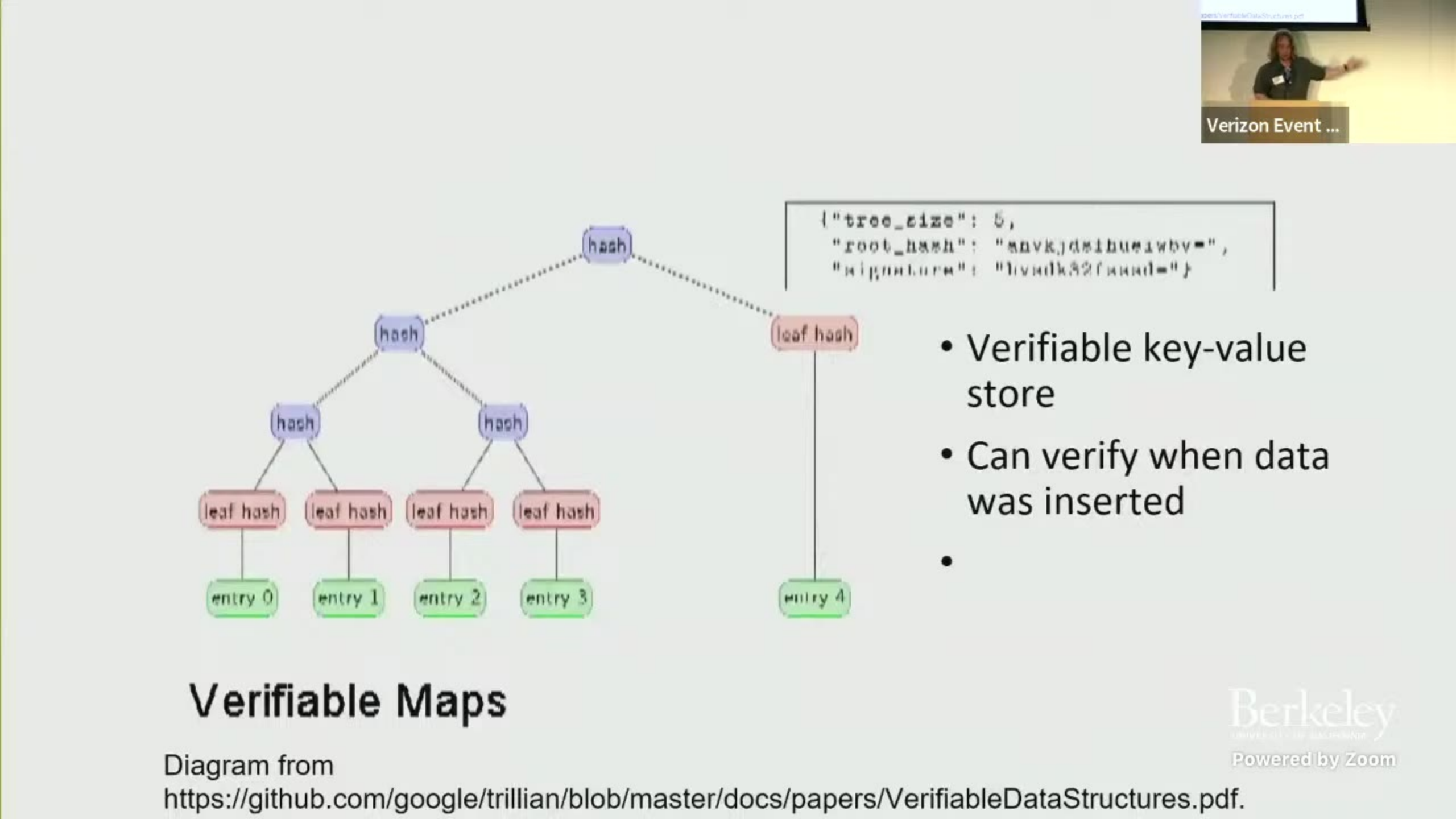
Asymmetric Watermarks

- Fairoze et al.
- Similar encoding to symmetric , but bits encode a digital signature
- Problem: you need a lot of bits to encode a signature => lots of rejection sampling
- Impractically slow (> 200 seconds for 1.5b pramater models)
- No metadata
- Operator can still frame people for

Waterlogs: a different approach

- What if we log LLM outputs instead of marking them?

Quality	Waterlog (this work)	Symmetric watermarks	Asymmetric watermarks
Metadata	✓	✗	✗
Non-repudiation	✓	✗	✓
Unframeability	✓	✗	✗
Federation	✓	✗	✓
Robust	✓	✓	✗
Efficient	✓	✓	✗

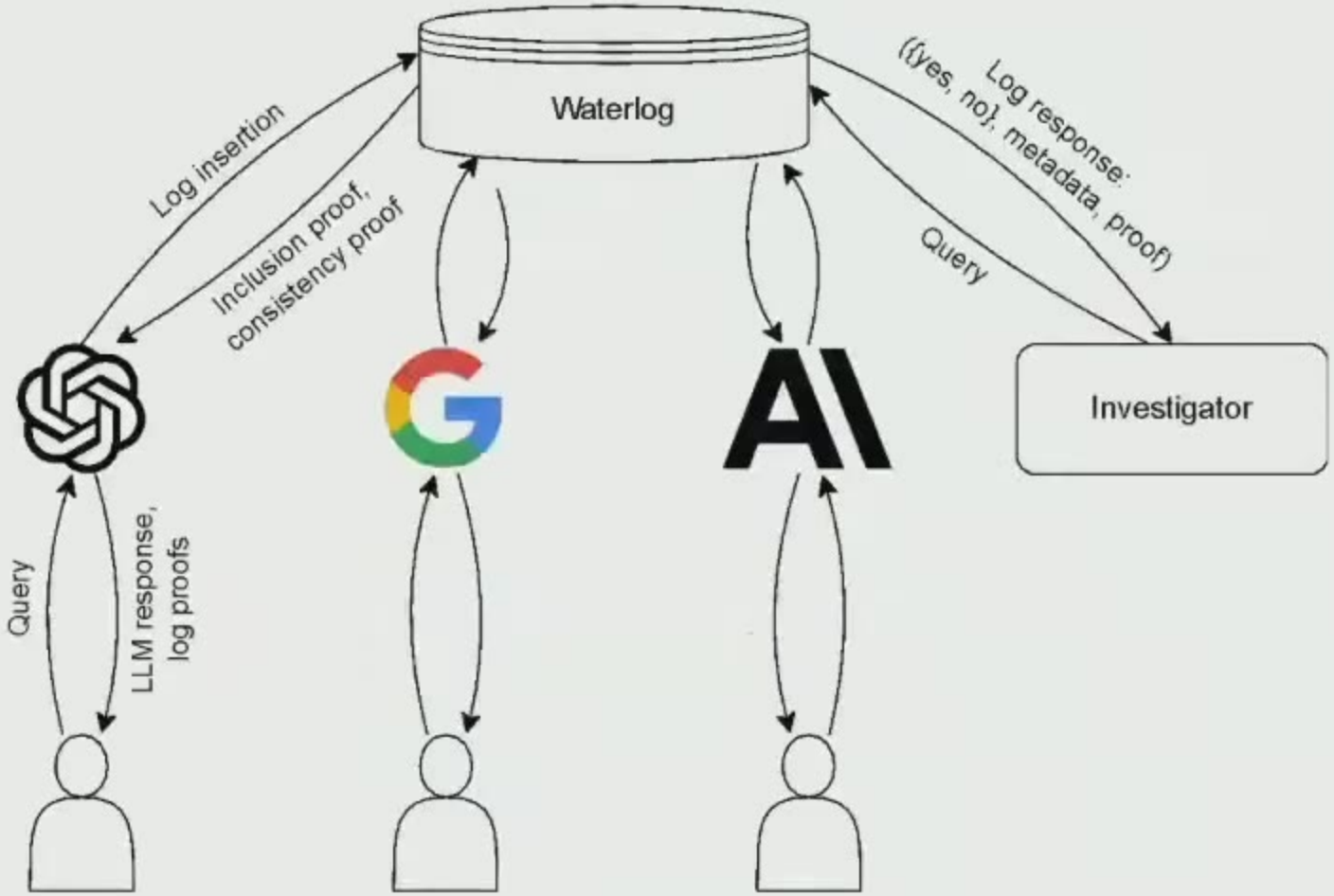


- Verifiable key-value store
- Can verify when data was inserted
-

Verifiable Maps

Diagram from https://github.com/google/trillian/blob/master/docs/papers/VerifiableDataStructures.pdf.

Waterlog setting

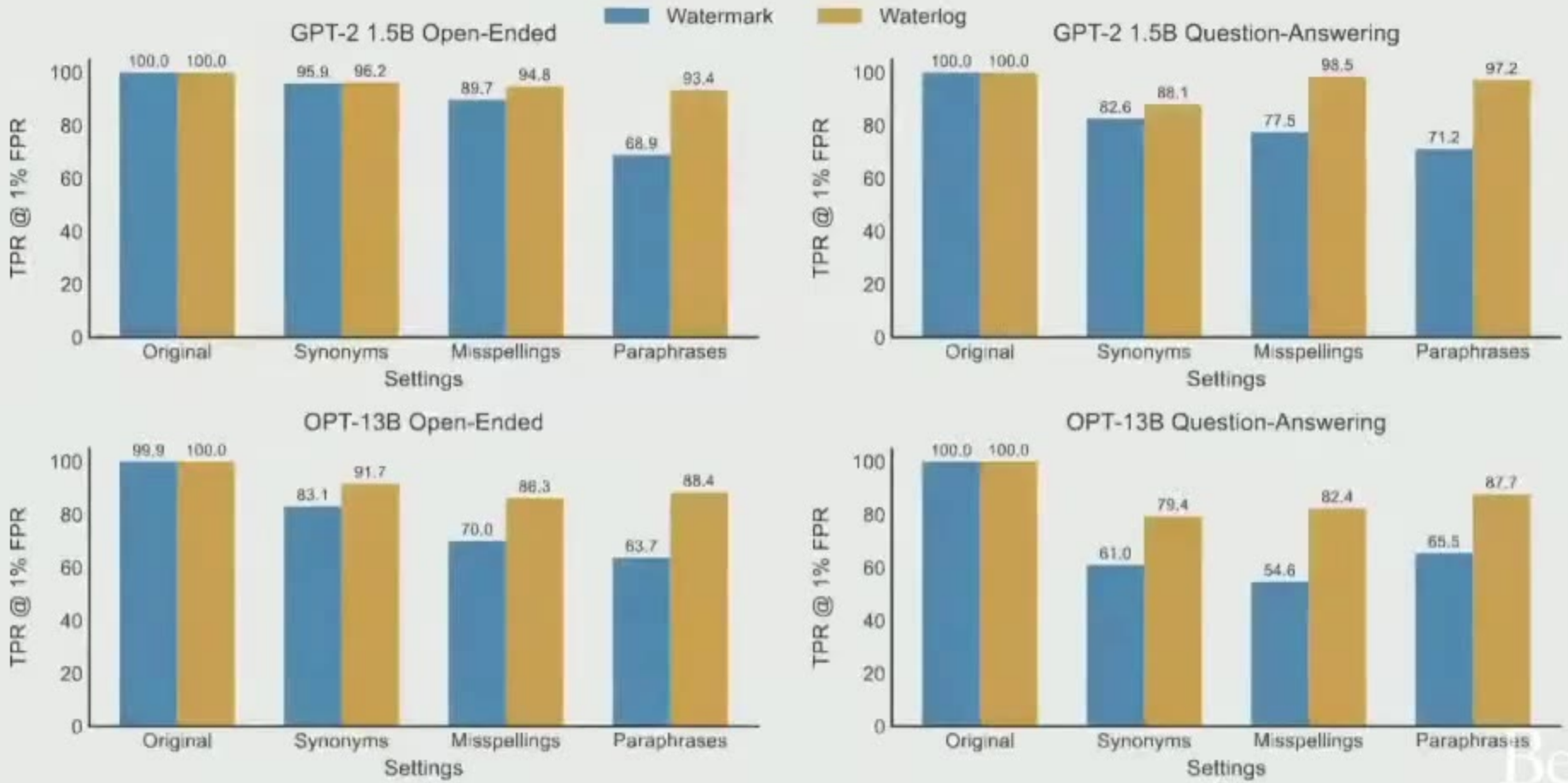


- Log can be a 3rd party
- LLM will honestly log digests
- Queries are made later, do not require trust
- Challenges
  - What do we log?
  - How do we handle inexact matches?

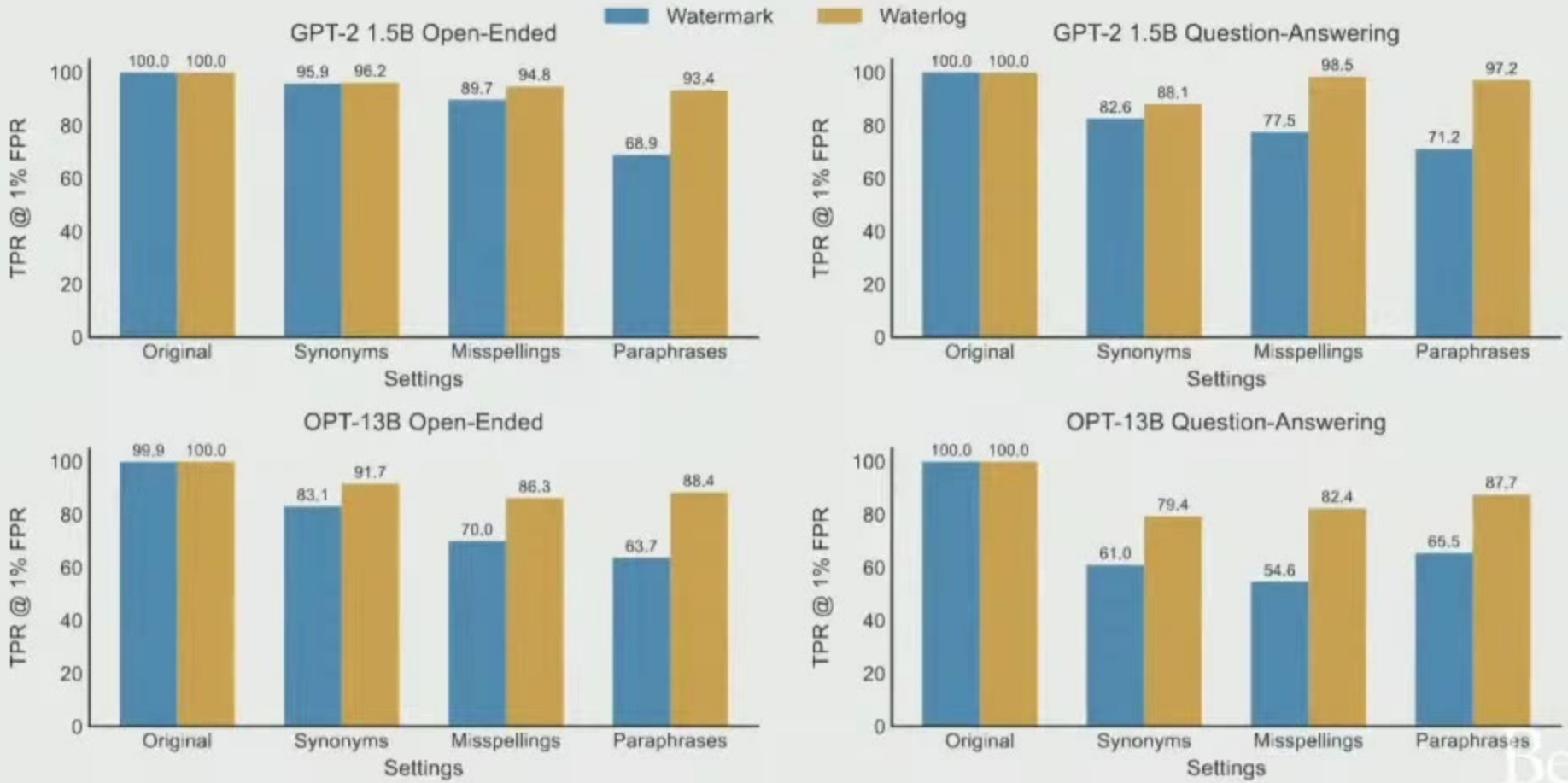
Waterlog: a proof of concept implementation

- Transparency log stores
  - (Commitment(metadata), SimHash(Embedding of LLIM Output))
  - Similarity Hash + Embedding handles document modeling
- *Verifiable hamming distance index* supports lookups
- Integrity of index assured by random audits

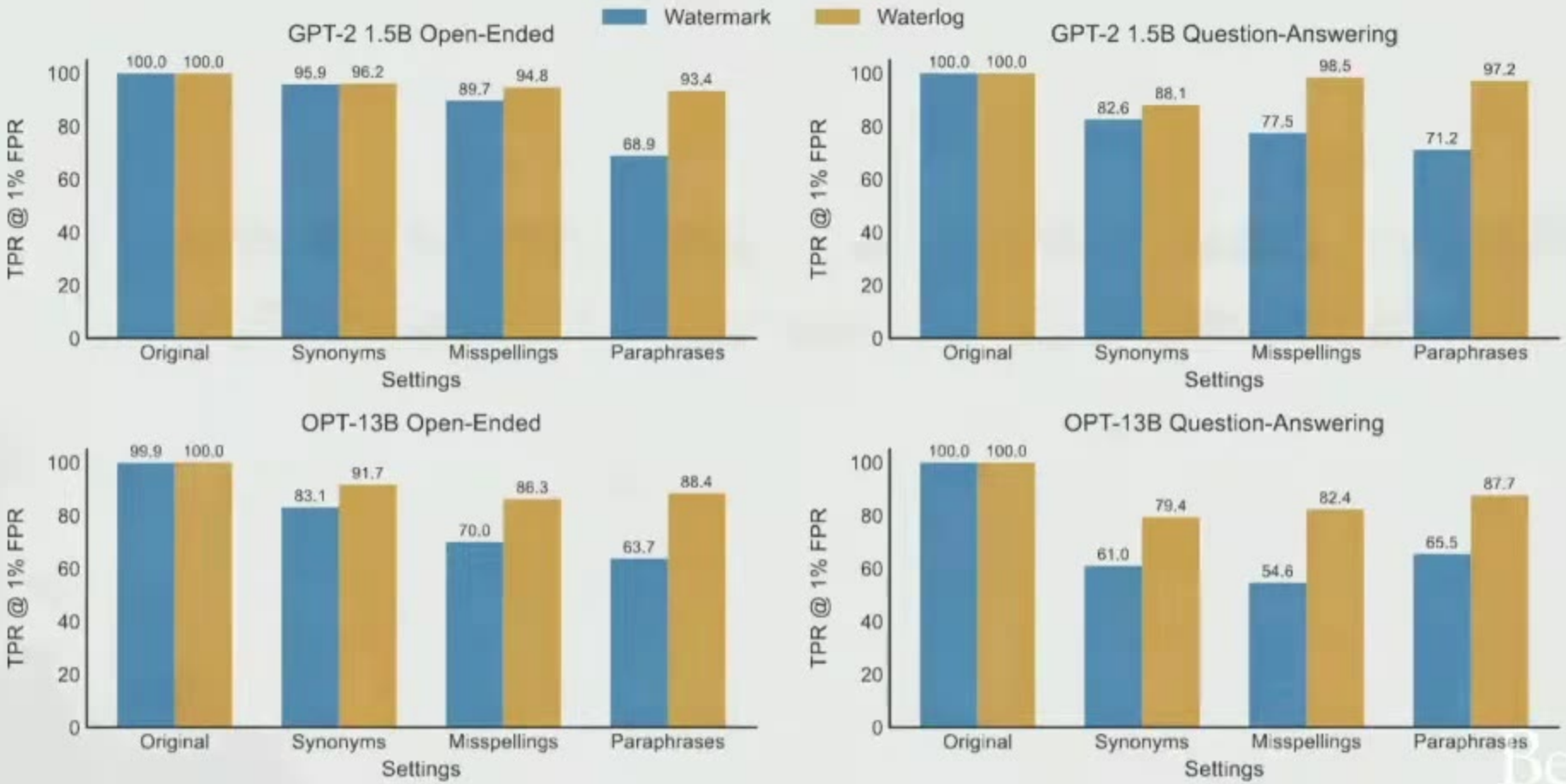
Waterlog: a proof of concept implementation

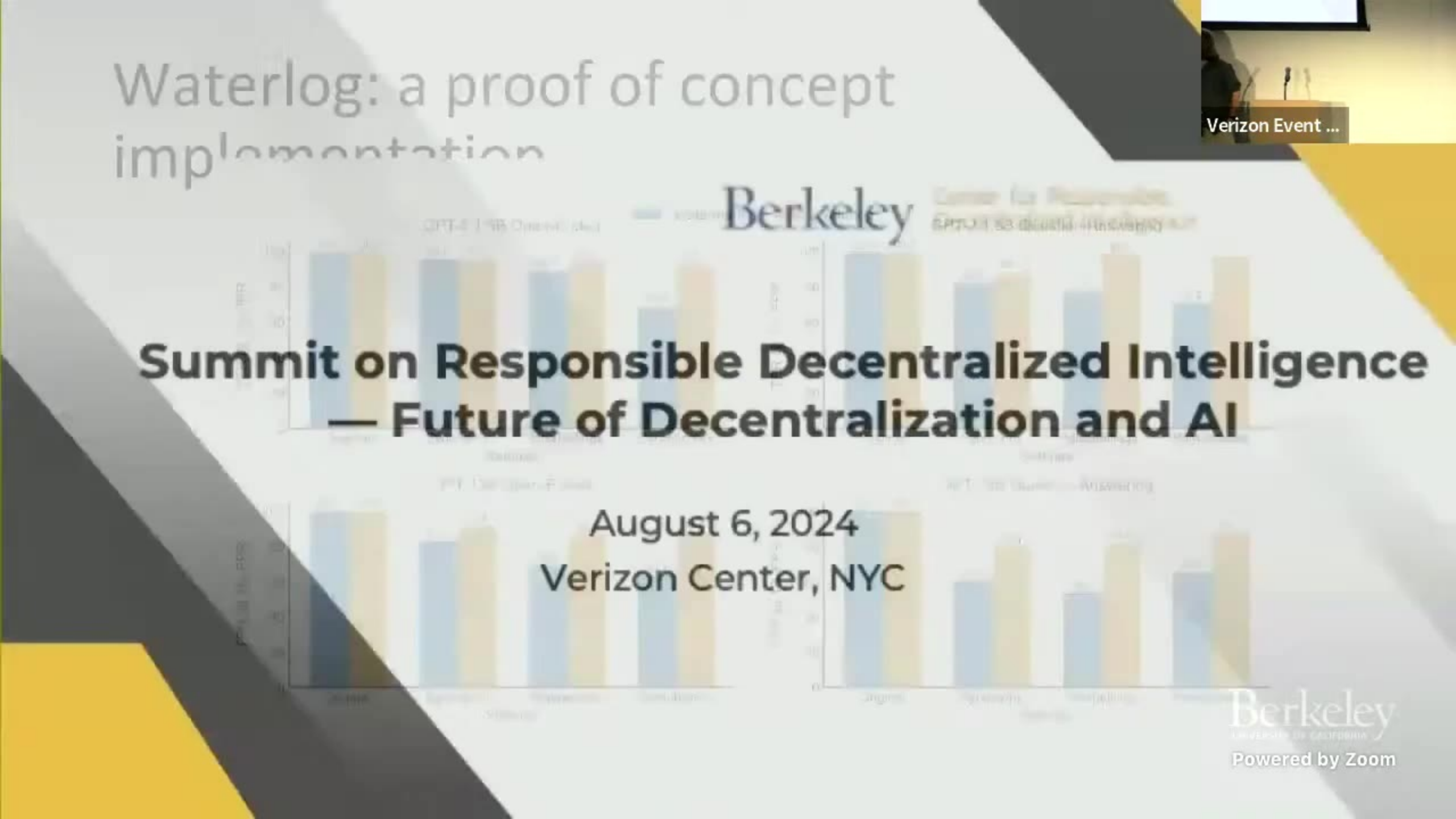


Waterlog: a proof of concept implementation



Waterlog: a proof of concept implementation




Waterlog: a proof of concept implementation
Berkeley
Verizon Event ...
Summit on Responsible Decentralized Intelligence
— Future of Decentralization and AI
August 6, 2024
Verizon Center, NYC
Berkeley
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 III: Foundations of Decentralized AI (II)

zkML: GKR Based Solution for Scalable and Verifiable ML

Tiancheng Xie
CTO
Polyhedra Network

Berkeley Center for Decentraliz

Session III: Foundations of Decentralized AI (II)

zkML: GKR Based Solution for Scalable and Verifiable ML

Tiancheng Xie
CTO
Polyhedra Network

Berkeley Center for Decentraliz

Verizon Event ...

Session III: Foundations of Decentralized AI (II)

zkML: GKR Based Solution for Scalable and Verifiable ML

Tiancheng Xie
CTO
Polyhedra Network

Berkeley

Powered by Zoom