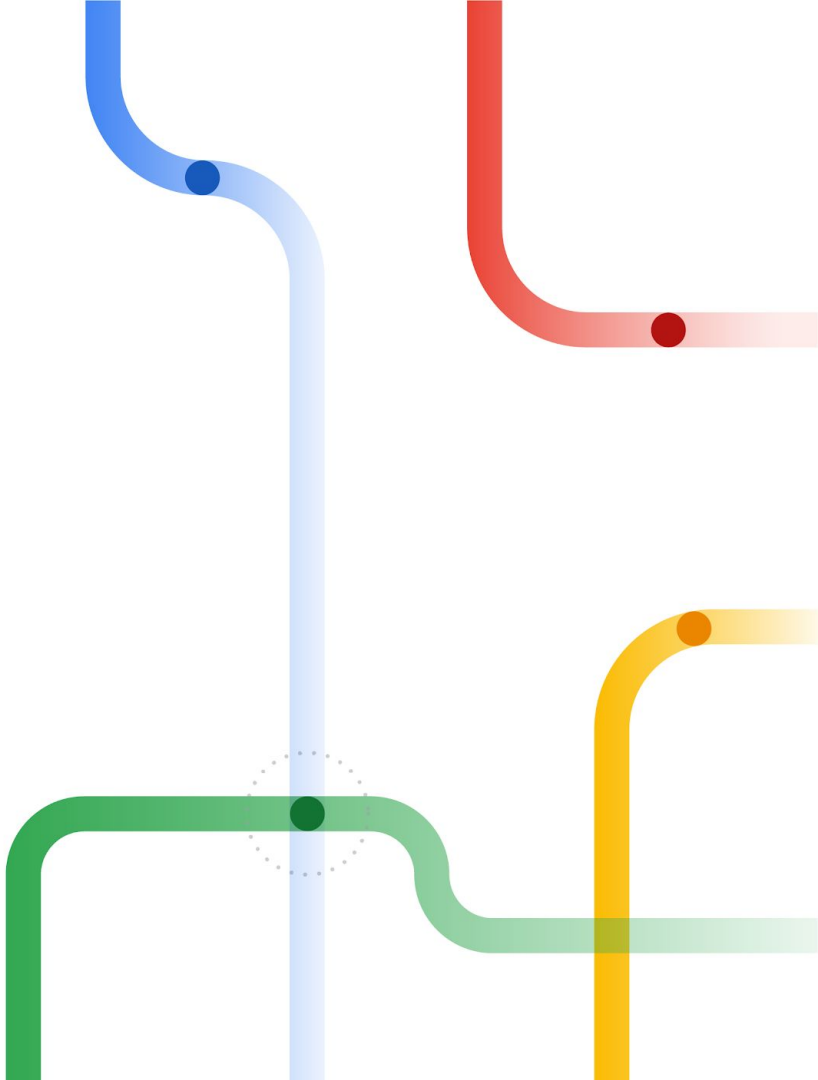
Towards Provably Private Insights into GenAI Use

Rakshita Tandon
rakshitatandon@google.com

Describing the work of many at Google

Google Research



Agenda

01 Motivation

02 Provable Private Insights

03 System Architecture

04 Real world application in Pixel Recorder

05 Scalability and Future Work

Google Research

01

Motivation

We've trained very powerful LLMs on the entire web.... now what?

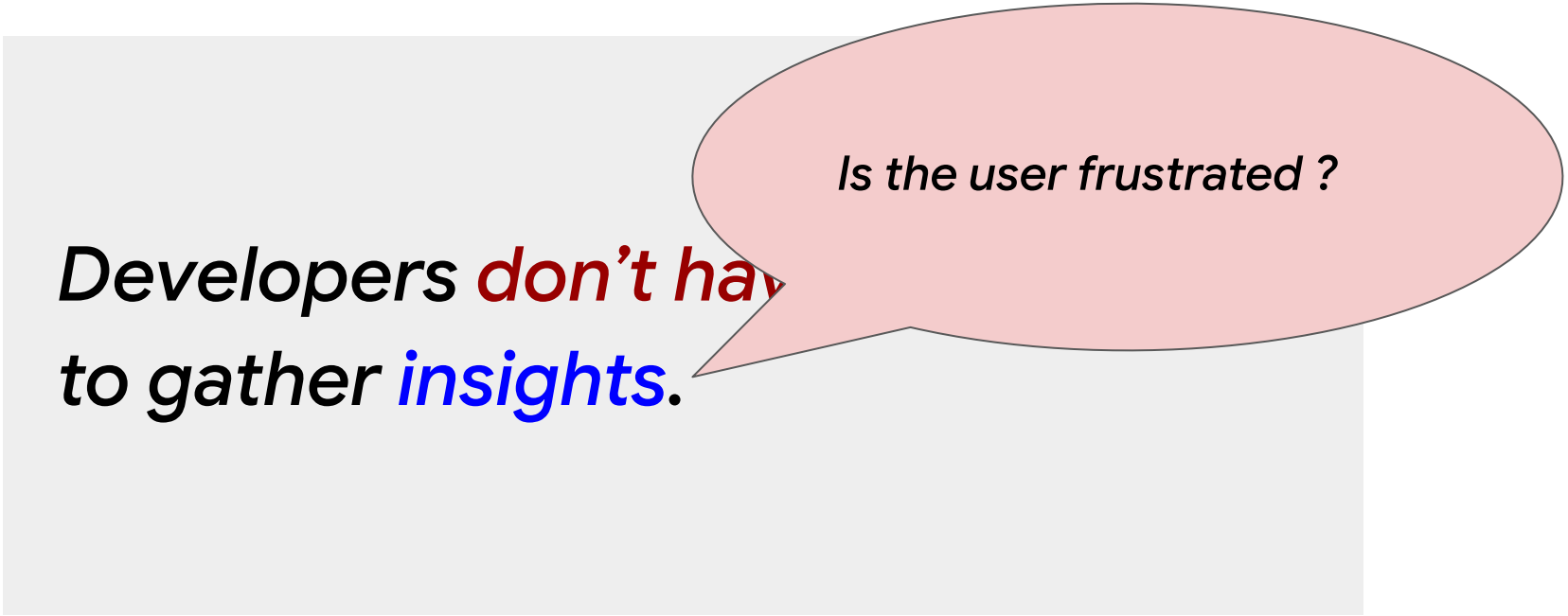
Google Research

We integrate them into products we use on a daily basis.... only to discover that.... they don't always work as well as we'd hoped!

Google Research

Developers don't have a good way to gather insights.

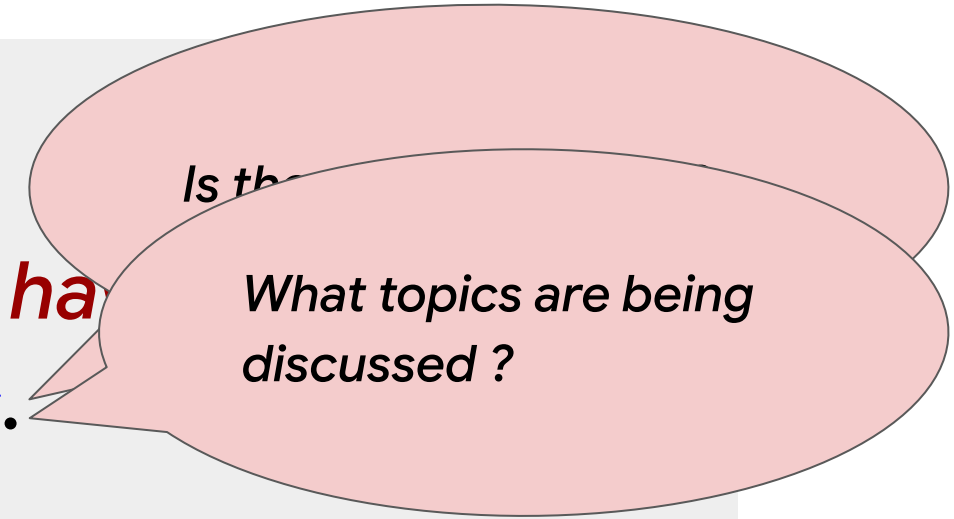
Google Research



Google Research

Developers don't ha

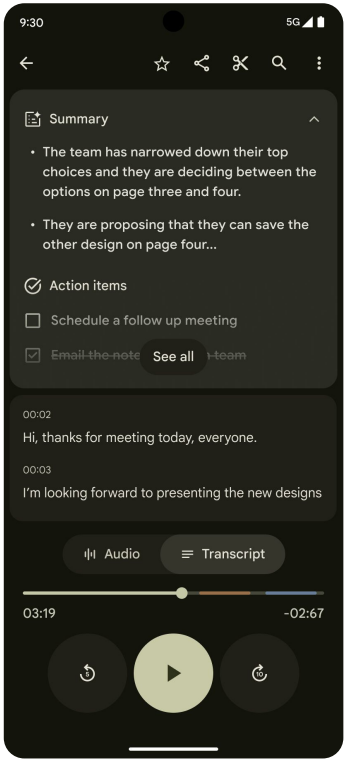
to gather insights.



Google Research

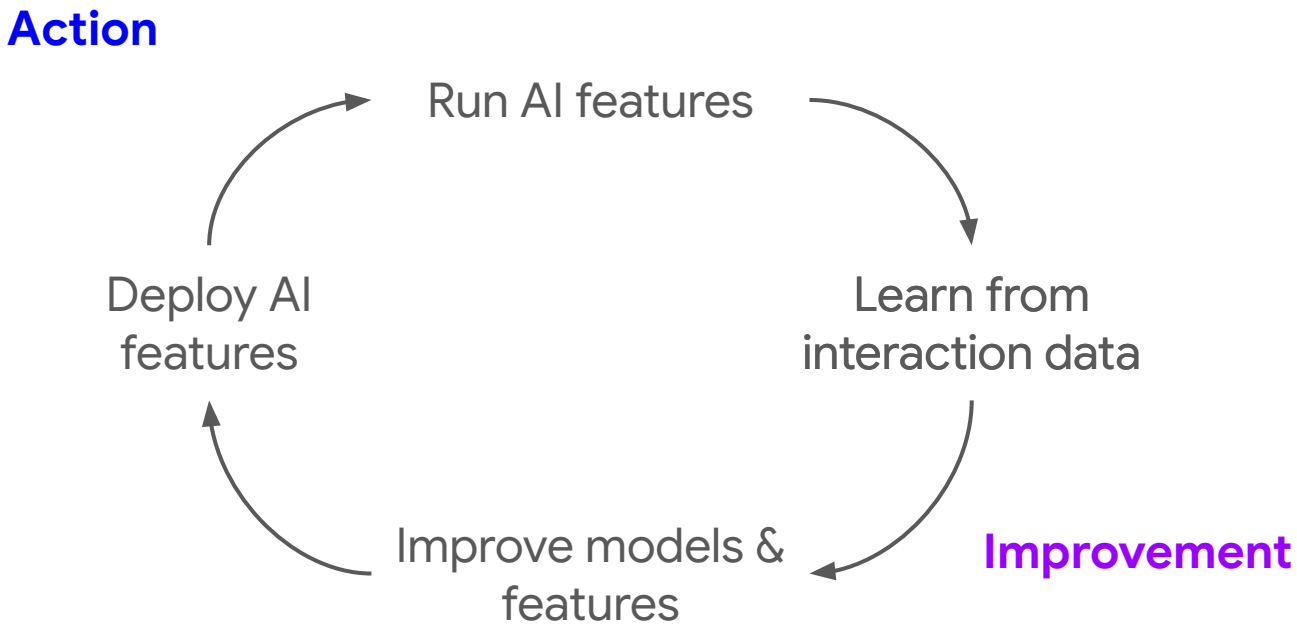
Motivation Example: Pixel Recorder

Pixel exclusive app with many GenAI capabilities such as summarization, speaker labels, action items generation.



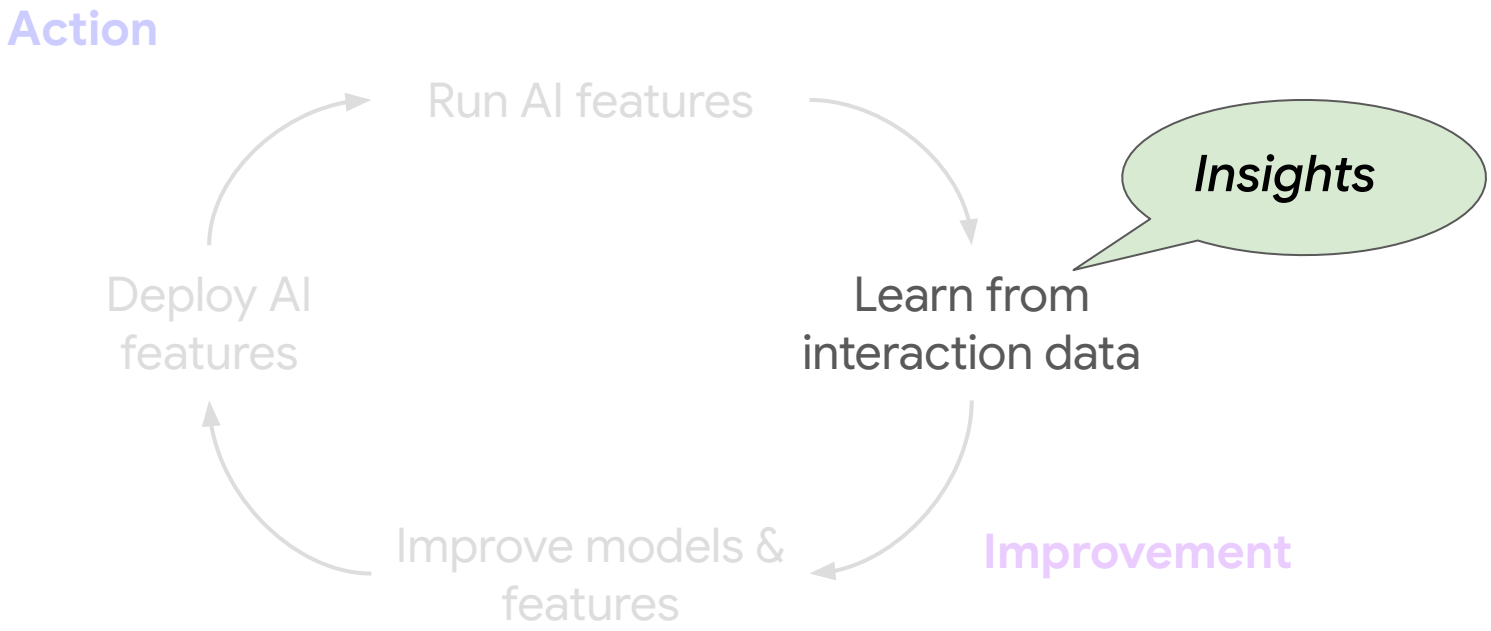
Google Research

Can GenAI systems safely improve from user interaction?



Google Research

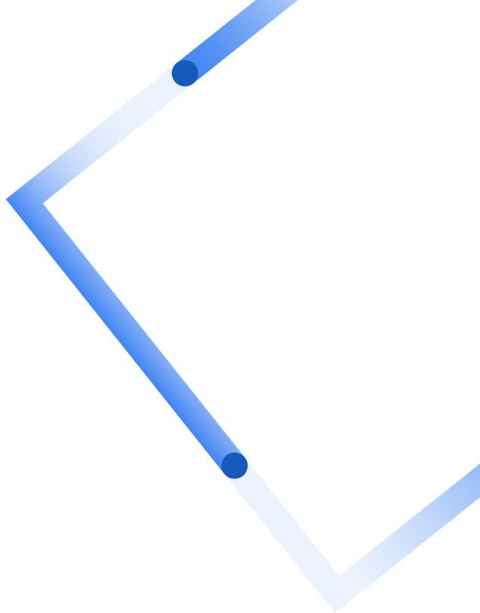
Can GenAI systems safely improve from user interaction?



Google Research

02

Provable Private Insights



Google Research

Goal

A system where unaggregated user data is

Private ***never visible to humans (e.g. service providers)***

Insight and ***released outputs*** always have

Provable ***formal verifiable privacy guarantees.***

Google Research

3 Pillars for Provable Private Insights

01

TEEs, OSS code, and reproducibly buildable binaries prove the identity of privacy-critical components

02

"Data Expert" LLM inside TEEs acts as an automated analyst.

03

Differential Privacy inside TEEs for formal privacy guarantees

Google Research

Trusted Execution Environments (TEEs)

Confidentiality*

An executing workload's state remains secret, even from root / the system operator.

Integrity

A workload's execution cannot be unexpectedly altered, even by root / the system operator.

Verifiability

Processes (e.g. clients) communicating with the workload receive cryptographic proof of workload identity, confidentiality, and integrity

* With known attacks on current hardware

Google Research

Provable Privacy via TEEs

- TEEs allow us to prove that we run a specific binary securely on a specific CPU.

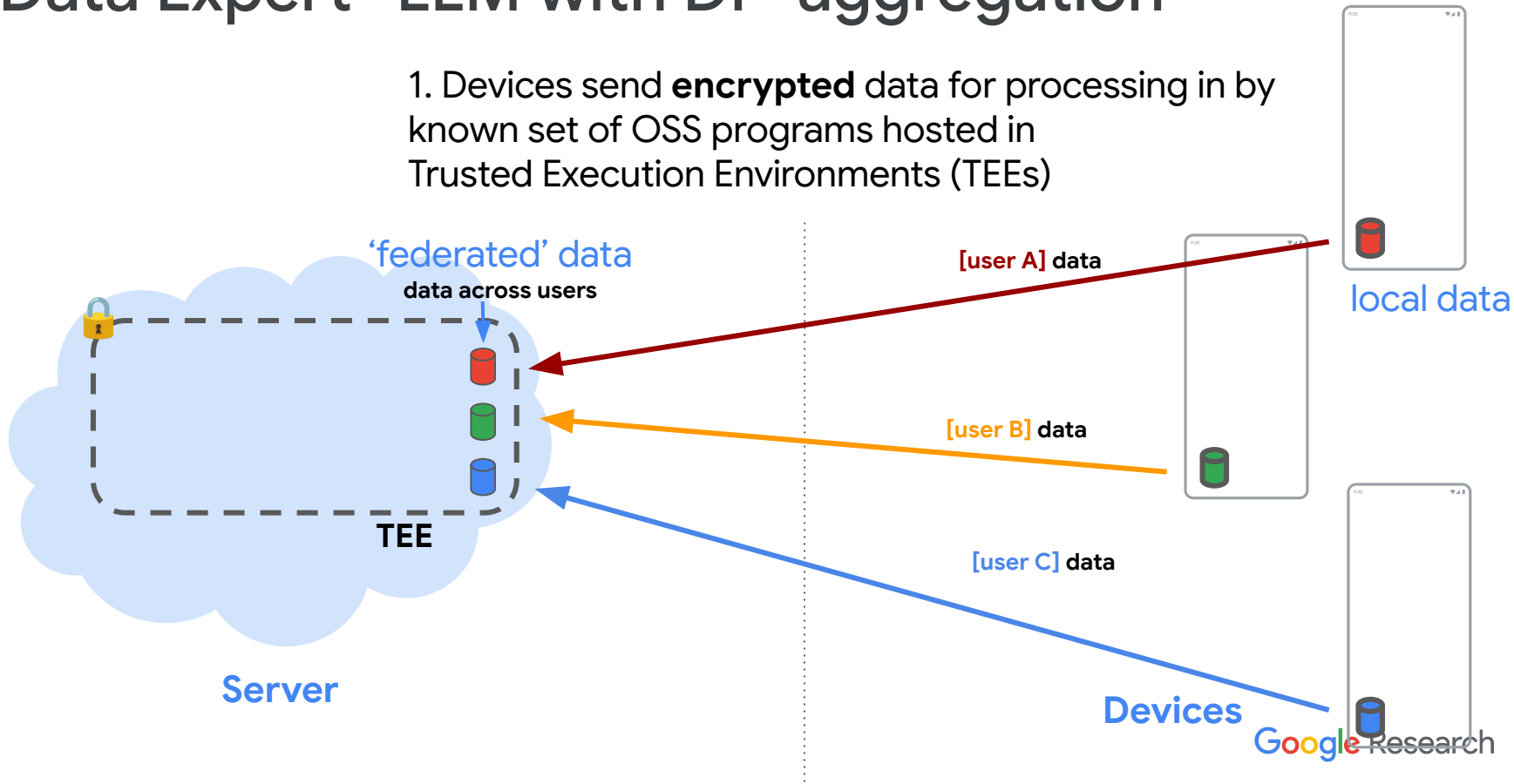
- We can create a chain of trust extending from client devices to server-side TEEs, enabling us to prove statements such as:

  "*Data from individual devices will not be visible to the service provider until it has been anonymously aggregated by the server (subject to TEE correctness).*"

Google Research

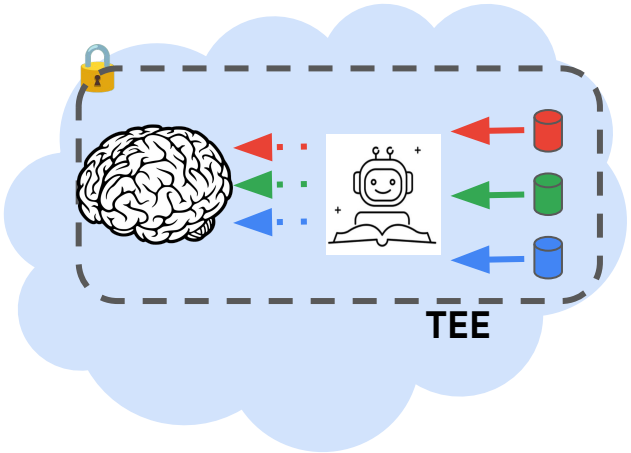
“Data Expert” LLM with DP-aggregation

1. Devices send **encrypted** data for processing in by known set of OSS programs hosted in Trusted Execution Environments (TEEs)

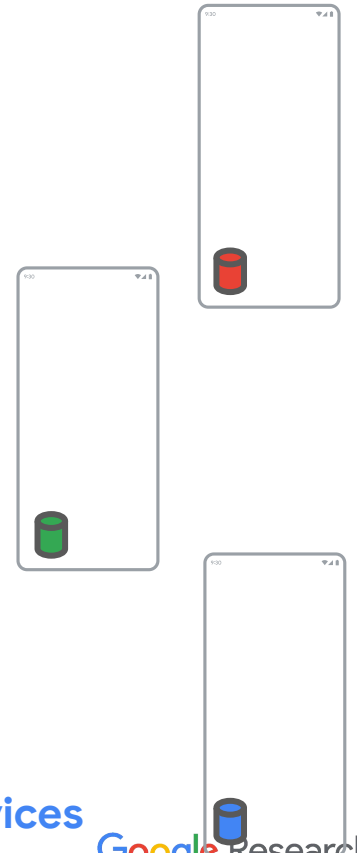


“Data Expert” LLM with DP-aggregation

2. Inside the TEE, “Data Expert” LLM is used to gain insights which are then combined across users using DP aggregates.



Server

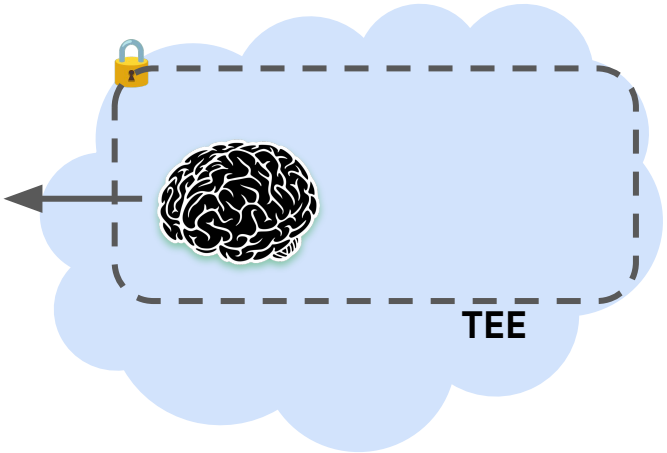


Devices

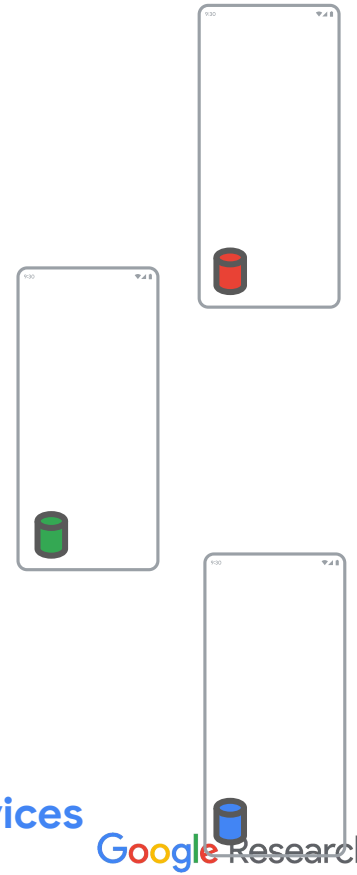
Google Research

“Data Expert” LLM with DP-aggregation

3. Aggregate statics are released from TEEs only if anonymization conditions are met.



Server



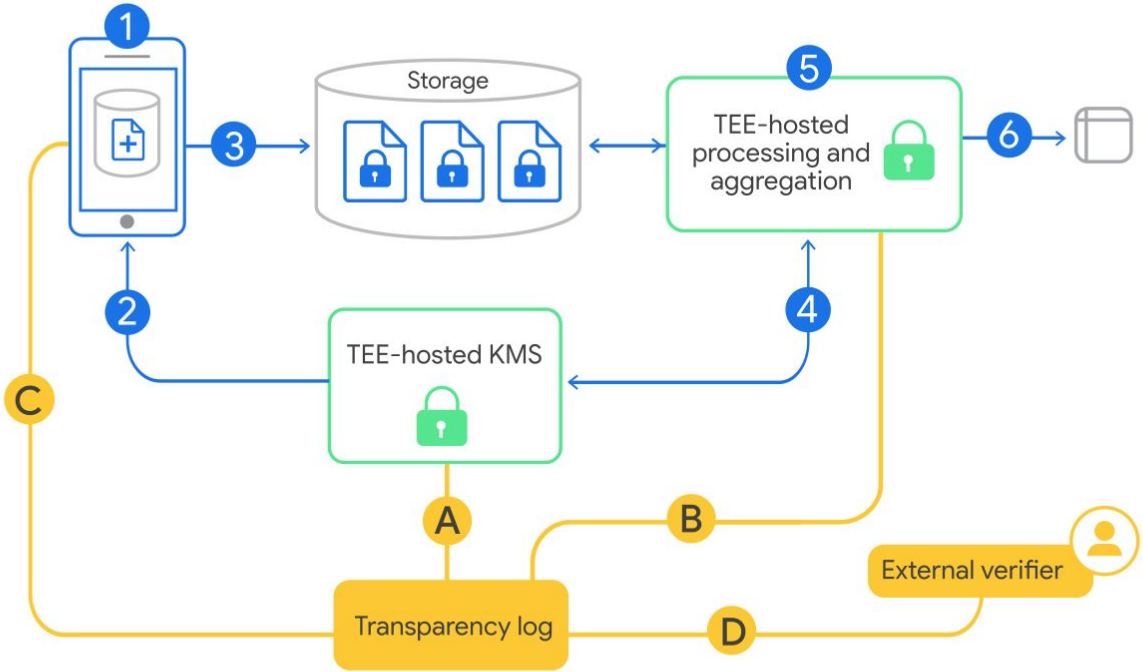
Devices

Google Research

03

System Architecture

Confidential Federated Analytics



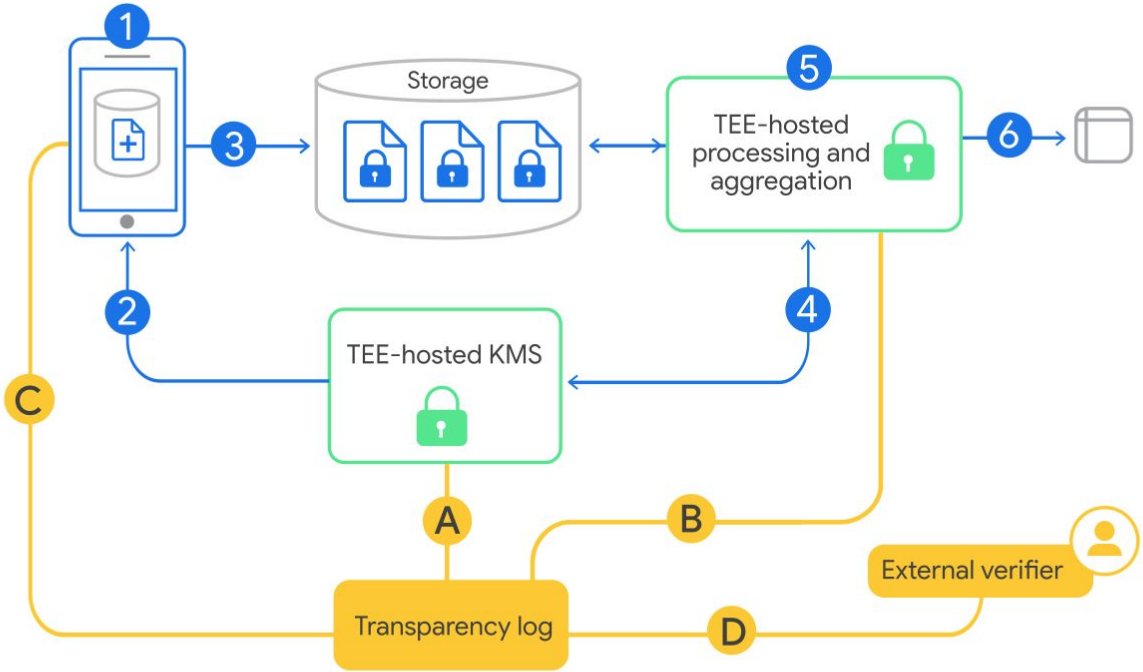
Data Flow

1. Devices compute minimal messages for selected tasks.
2. Devices fetch and verify keys signed by a TEE-hosted KMS.
3. Devices encrypt and upload messages, tagged with access policy.
4. KMS grants decryption keys to access-policy-approved processing steps only.
5. Uploads are processed and combined with privacy preserving (DP) algorithms.
6. Anonymous aggregates are released to data analyst.

AMD SEV-SNP

Project Oak

Confidential Federated Analytics



Verification Flow

- A KMS binaries are endorsed and published to the transparency log.
- B Processing steps described in access policies are endorsed and published to the transparency log.
- C Device checks that KMS and access policy endorsements are present in transparency log.
- D External verifier can monitor endorsed KMS and processing steps along with corresponding OSS code.

sigstore

Open source at **Google Parfait**

github.com/google-parfait

Private **a**ggregation **&** **r**etrieval**,** **f**ederated **a**nalytics**,** **i**nference**, &** **t**raining from **Google**

Repositories for defining analytics workflows and ML workflows with strong privacy claims consistent with users' privacy expectations.

- Libraries and frameworks for simulation
- Reference architectures for production systems
- Components that power Google's production deployments of federated compute infrastructure

TensorFlow Federated, Federated Computation Platform, and Confidential Federated Compute, and more.

Note: there is an alternate approach to inspecting the ledger attestation evidence and data access policies that a given client device accepts, which involves instrumenting the device. See inspecting_attestation_records.md for more details on this approach.

Inspecting ledger binary transparency log entries

The endorsement keys client devices use to validate the ledger application can be found in the /reference_values/ledger directory.

To find transparency log entries for these endorsement keys you can use the rekor-monitor tool. For example, the following configuration lists the endorsement key fingerprints of each of the ledger TEE binary layers, and will find a recent endorsement log entry for the application layer TEE binary.

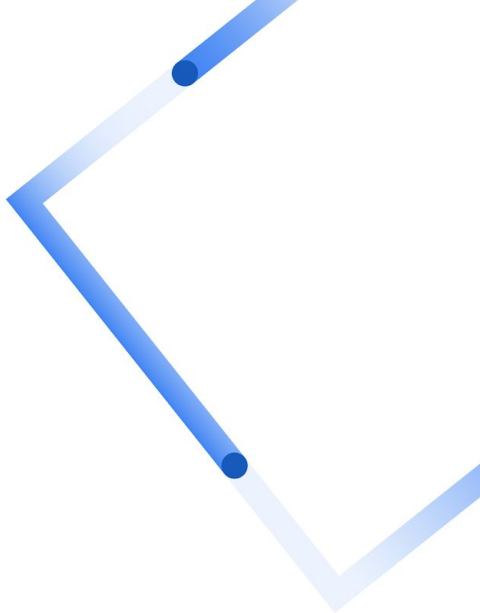
```
$ cat << EOF
startIndex: 175178225
endIndex: 175178226
monitoredValues:
  fingerprints:
    - 98fc8ad40908f6c079c5d1677b85f261acdf08262c73f448f04bd4e9a090c8bb # stage0
    - 6052f352eac71f16947815eb34010f49ea2f1284a46b61777fb8c2accfa26d29 # kernel
    - 5f884b699bb66fe0b0ab07e2ee9ed9c221109ffdb2d13f470ed964952271d867 # init_ram_fs / orchestrator
    - ea0d1f8ffed9512019a2ec968790263d88ea3324c6a6a782114e5ea1be4fd38f # application
EOF > config.yml

$ go run github.com/sigstore/rekor-monitor/cmd/rekor_monitor@6248cd70ec4f0c18e4d23901041caea126da36bc \
    --config-file config.yml
...
Found ea0d1f8ffed9512019a2ec968790263d88ea3324c6a6a782114e5ea1be4fd38f 175178226 108e9186e8c5677a5d4bc53a7212664ffa2
```

Note that the above configuration limits the tool's search to a log index range, but you can also omit the start and end indices and run the tool periodically (say, every 10 minutes), in which case the tool will scan all new log entries added since the last run, and print out any that match the configured endorsement keys. Rekor also exposes a REST API and a GCP Pub/Sub event stream which you can use to monitor for new log entries instead.

04

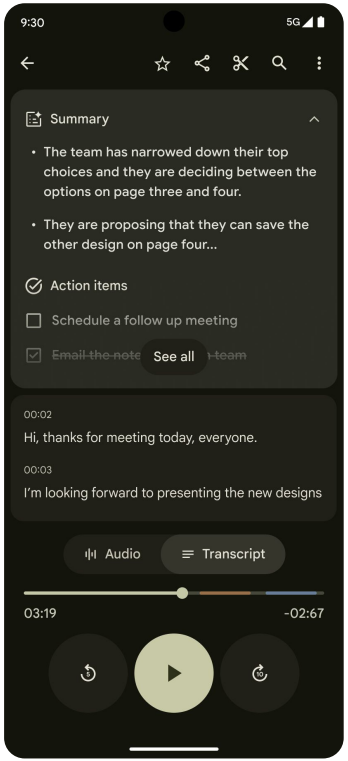
Real world application in Pixel Recorder



Google Research

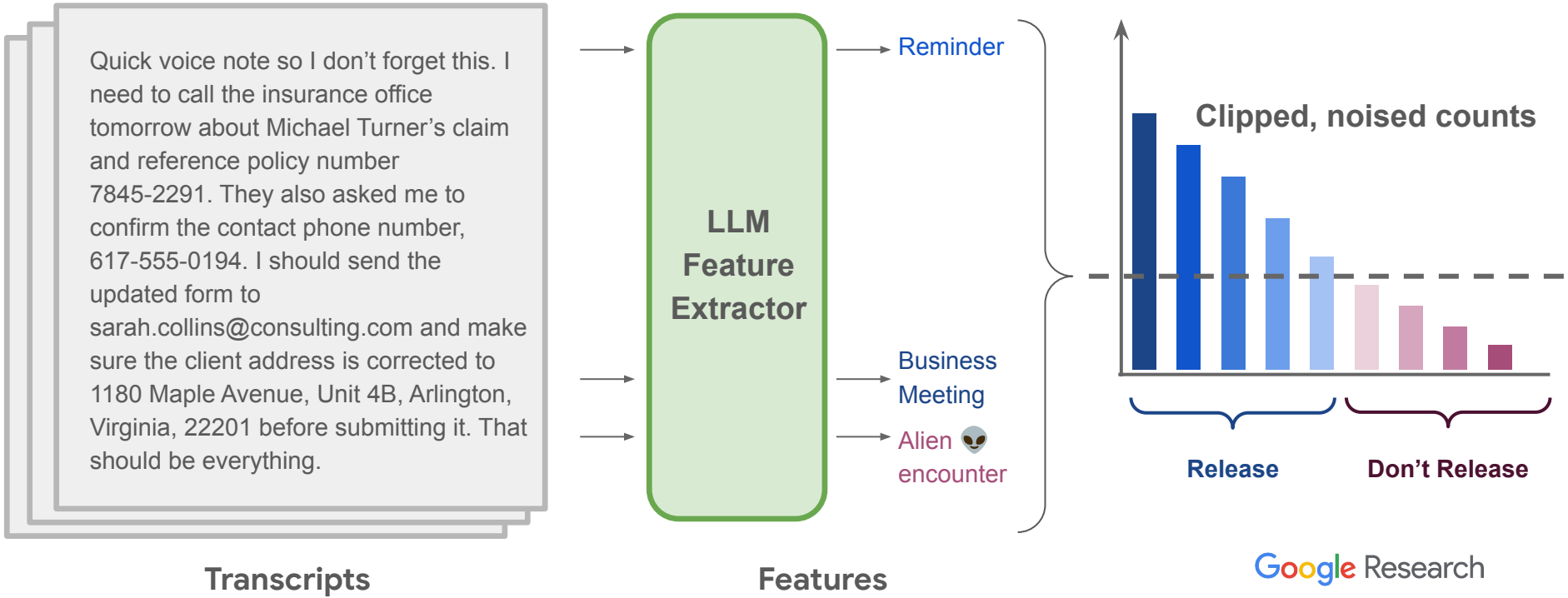
Pixel Recorder

Provable Private Insights (PPI) enables transcript classification from unstructured raw data into various categories of interest.



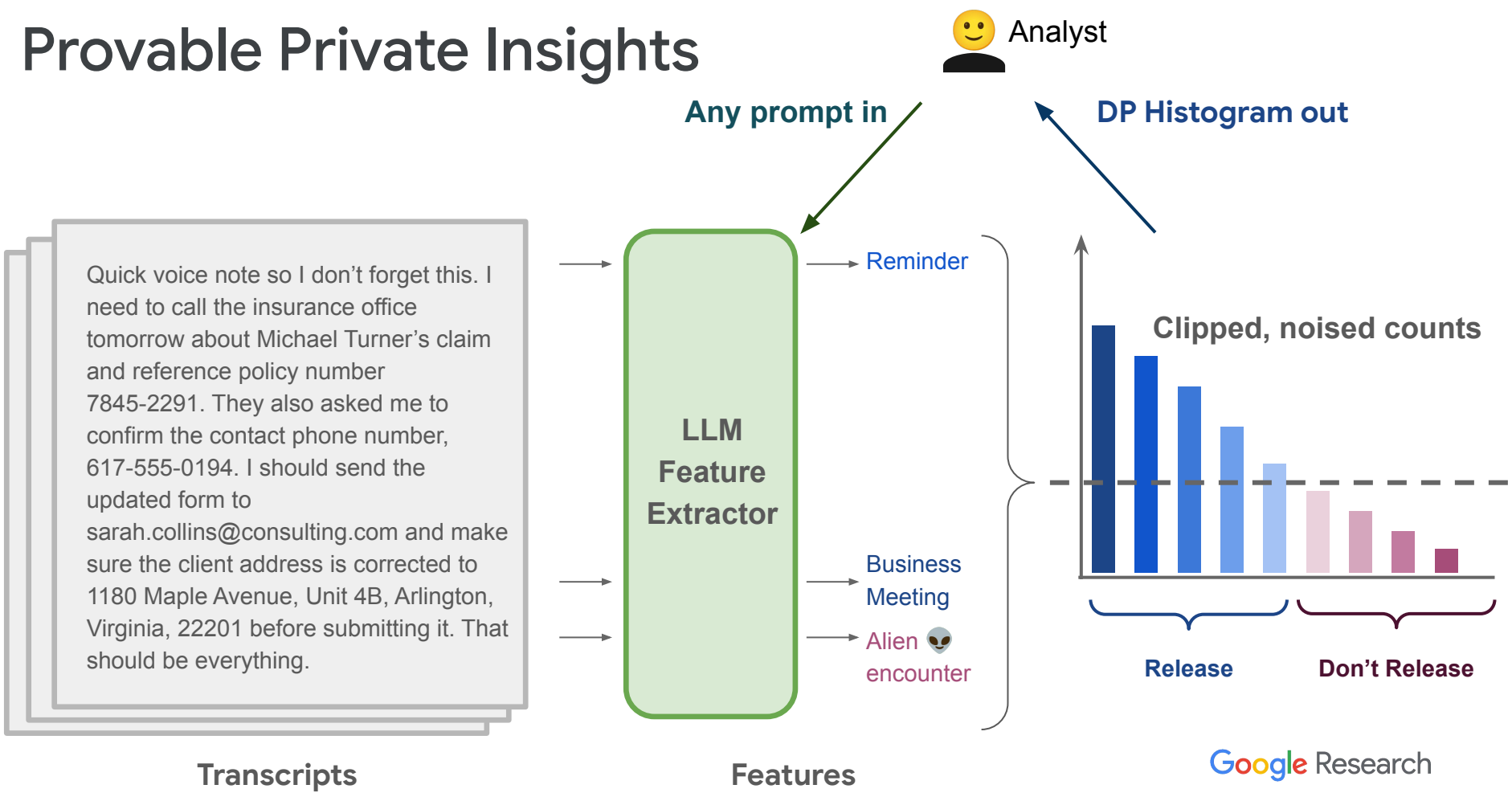
Google Research

Provable Private Insights

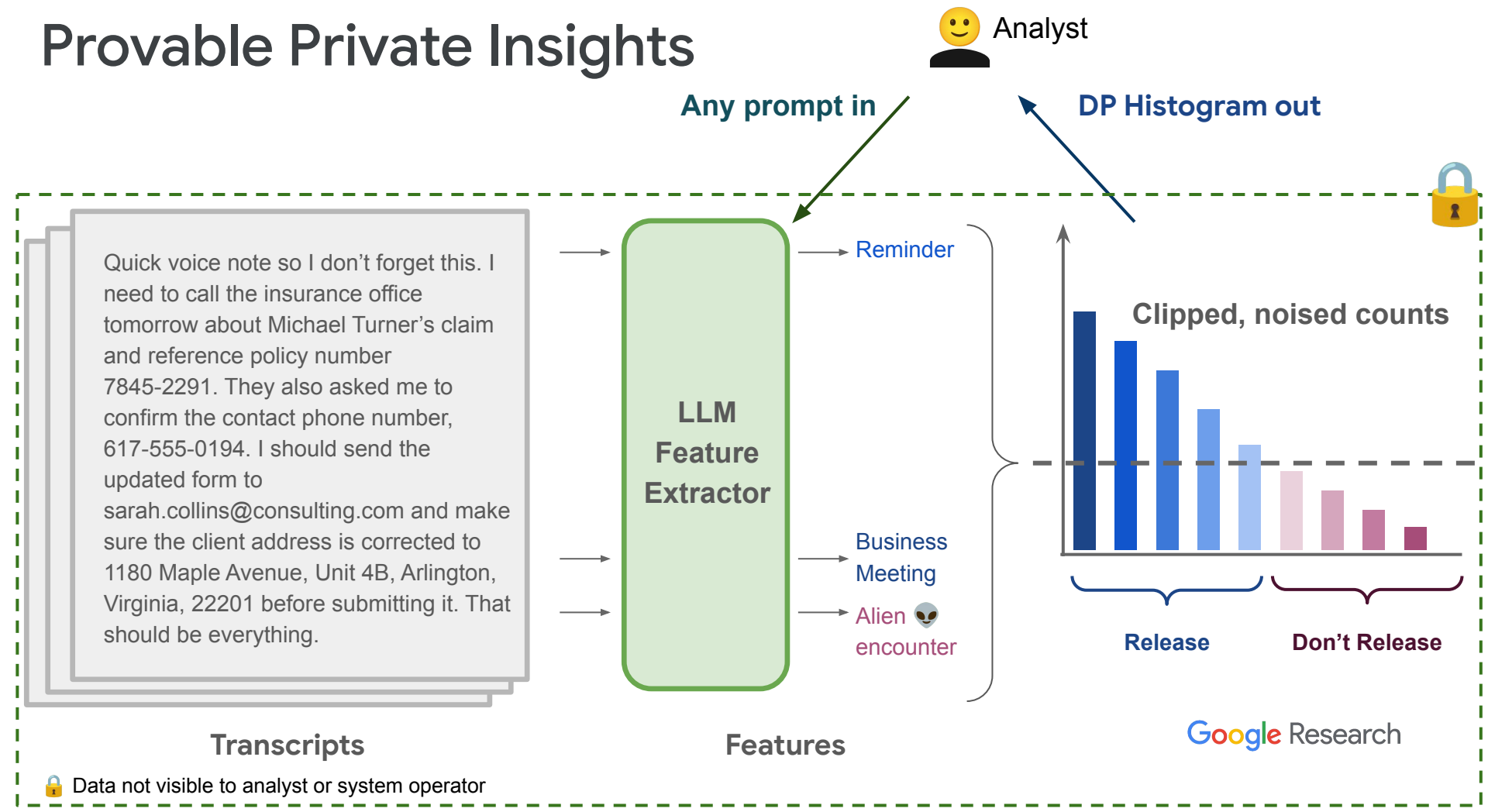


Google Research

Provable Private Insights

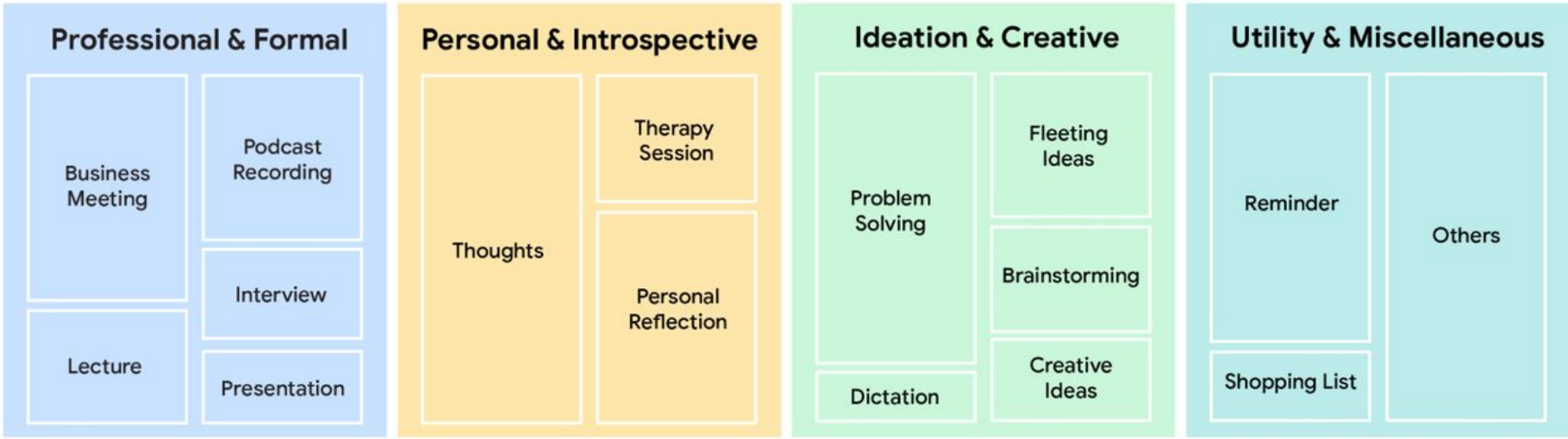


Provable Private Insights



Provable Private Insights

user-level ε=1 for label discovery

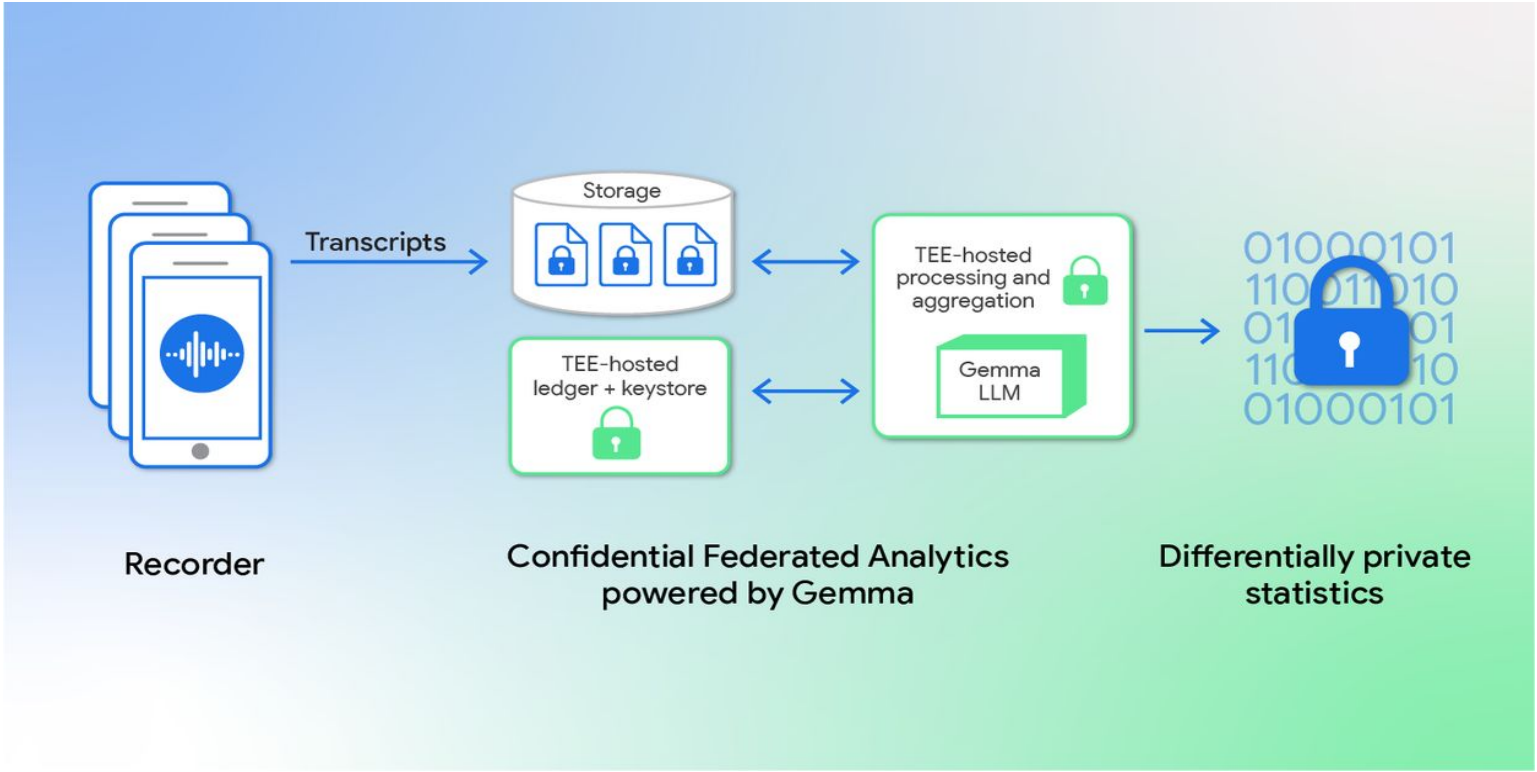


Toward provably private analytics and insights into GenAI use (arXiv:2510.21684)

Blogpost: https://research.google/blog/toward-provably-private-insights-into-ai-use/

Google Research

Provable Private Insights



Google Research

05

Scalability and Future Work

Scalability Challenges

- Current production workloads at Google run exclusively on CPUs within TEEs using Gemma 3 4B model, processing upto O(100k) inputs in a single batch-processing pipeline.

- Speed and memory-constraints are a significant challenge for scaling to larger scale workloads and larger models exceeding 20B parameters.

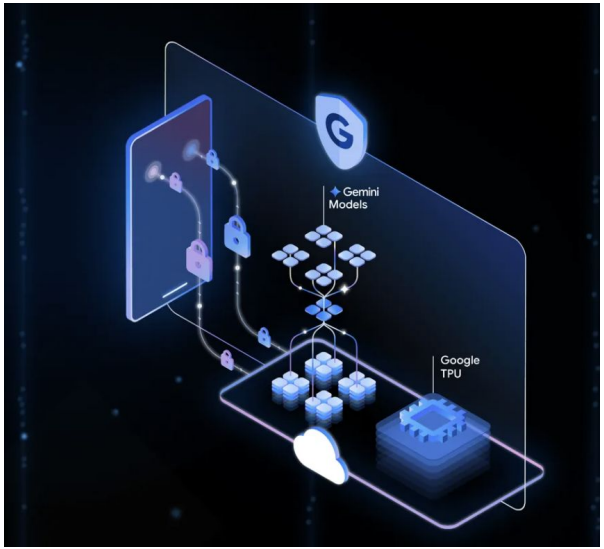
Google Research

Future Work: Hardware-based accelerators

With future work to enable confidential use of hardware-based accelerators, using more complex models and higher throughput will become possible.

Examples:

- Google TPUs (via Google's Private AI Compute (PAIC) stack)
- GPUs such as NVIDIA H100 (via Google Cloud Platform (GCP) Confidential Compute)



Google Research

Future Work: Hardware-based accelerators

This will further unlock richer analyses such as:

- Detailed **transcript analysis** and **auto-rating**
- New capabilities like **differentially private clustering** and **synthetic data generation**, all with similar levels of verifiability and confidentiality

Google Research

Thank You

Rakshita Tandon
Software Engineer, Google

Google Research