Google

Hark:
A Deep Learning System for Navigating Privacy Feedback at Scale

Sai Teja Peddinti
psaiteja@google.com

Collaborators:

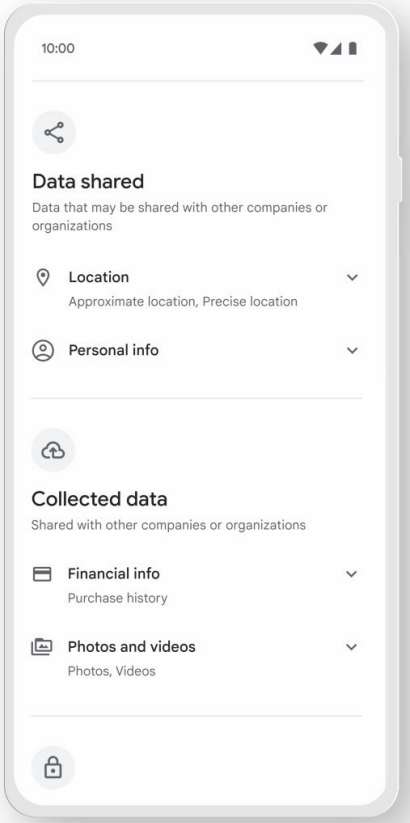
Hamza Harkous
harkous@google.com

Rishabh Khandelwal
rkhandelwal3@wisc.edu

Animesh Srivastava
sranimesh@google.com

Nina Taft
ninataft@google.com

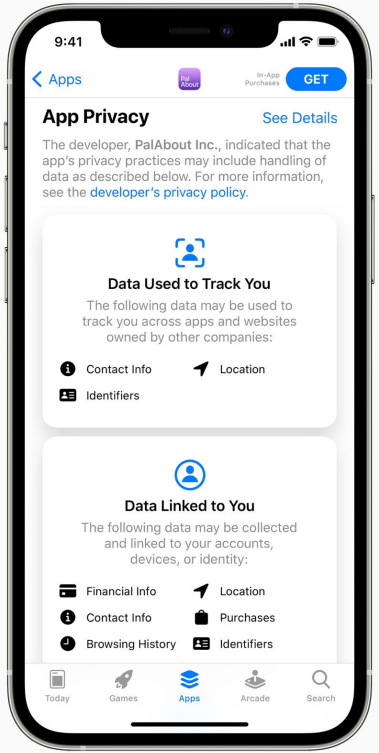
Efforts Targeting Developer-to-User Communication



Data Safety Section on Play Store

Get more information about how an app uses data

Efforts Targeting Developer-to-User Communication



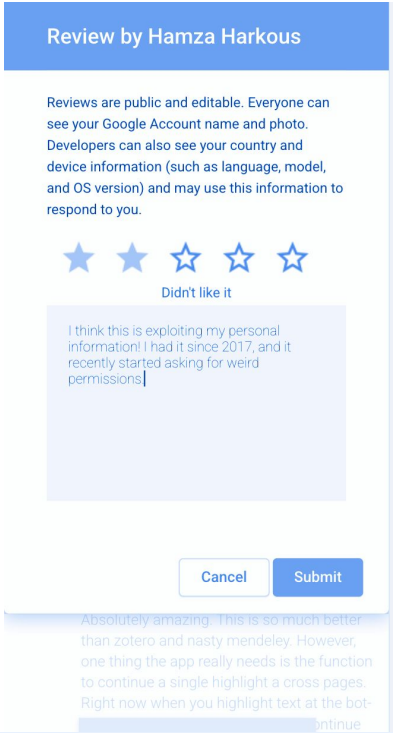
App Privacy Details on App Store

“

Q: How can we create a better **user-to-developer** communication channel of privacy concerns?

”

Hark in a Nutshell



Stop sharing my information!

This drains my battery

Hide my email in your app!

I've been hacked.

Undo incognito mode

App is offline.

I want to hide my name

Someone impersonating me

You are spying on me!

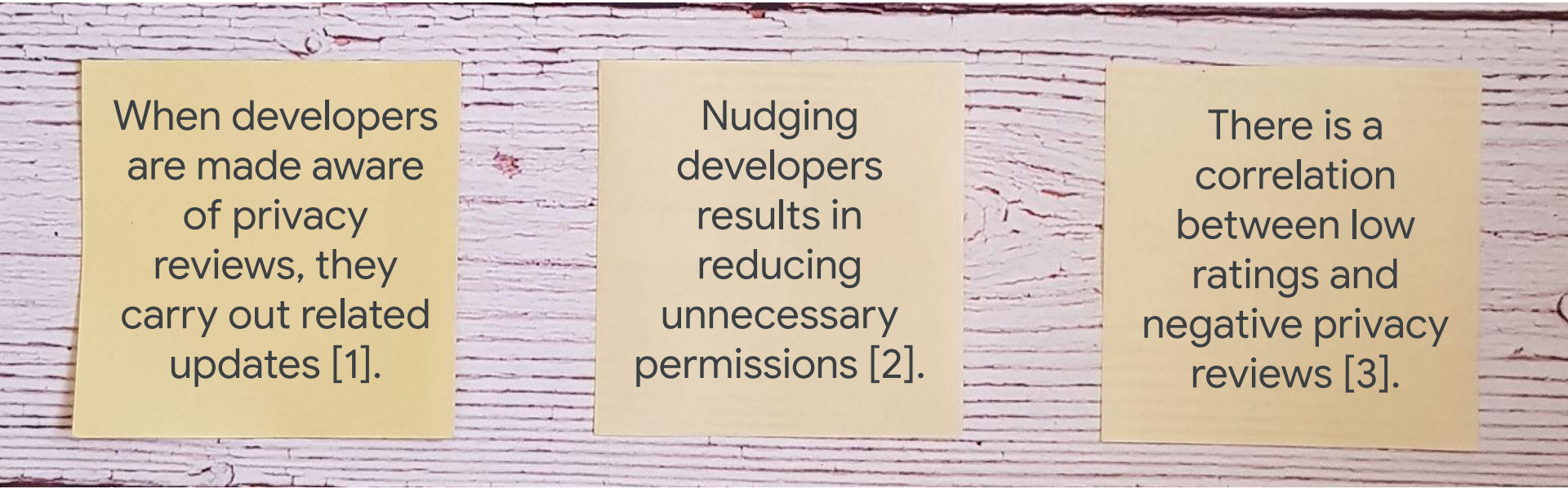
Please remove this device

Theme 1: Unneeded Access (546K)

- Issue 1: Unneeded Location Access (89K)
  - Top Review 1: … **(Emotion:** Anger**)**
  - Top Review 2: … **(Emotion:** Fear**)**
  - …
- Issue 2: Unneeded Contacts Access (80K)
  - Top Review 1: … **(Emotion:** Confusion**)**
  -
  - …
- …

Millions of Reviews Submitted → Fully Automated Pipeline → Top Privacy Themes

Importance of the User-to-Developer Channel

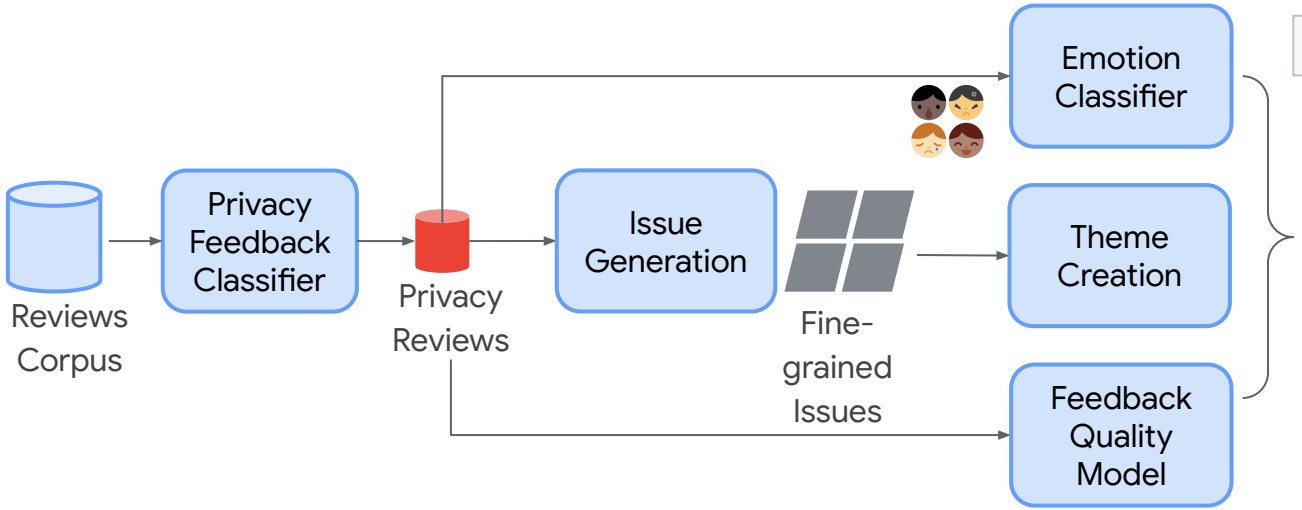


[1] Nguyen et al. “Short text, large effect: Measuring the impact of user reviews on android app security & privacy”. IEEE S&P 2019.
[2] Peddinti et al. “Reducing permission requests in mobile apps.” ACM IMC 2019.
[3] Besmer et al. “Investigating user perceptions of mobile app privacy: An analysis of user-submitted app reviews.” IJISP 2020.

Hark Overview

Hark ML Overview

Privacy Feedback Classification → User Issues Extraction & Grouping → Issues Summarization & Trend Analysis



Theme 1: Unneeded Access (546K)

- Issue 1: Unneeded Location Access (89K)
  - Top Review 1: ... **(Emotion:** Anger**)**
  - Top Review 2: ... **(Emotion:** Fear**)**
  - ...
- Issue 2: Unneeded Contacts Access (80K)
  - Top Review 1: ... **(Emotion:** Confusion**)**
  -
  - ...
- Trend Analysis

Original Feedback

This app automatically delete your important files from SD card

Why does it need access to Contacts now ?

WHY SHOULD IT NEED TO NO MY PHONE CALLS!!! STUPID PEOPLE WE NEED PRIVACY

Unable to delete account and unlink personal information

I Hope You Erase My ID number From Your Database

My account was hacked.

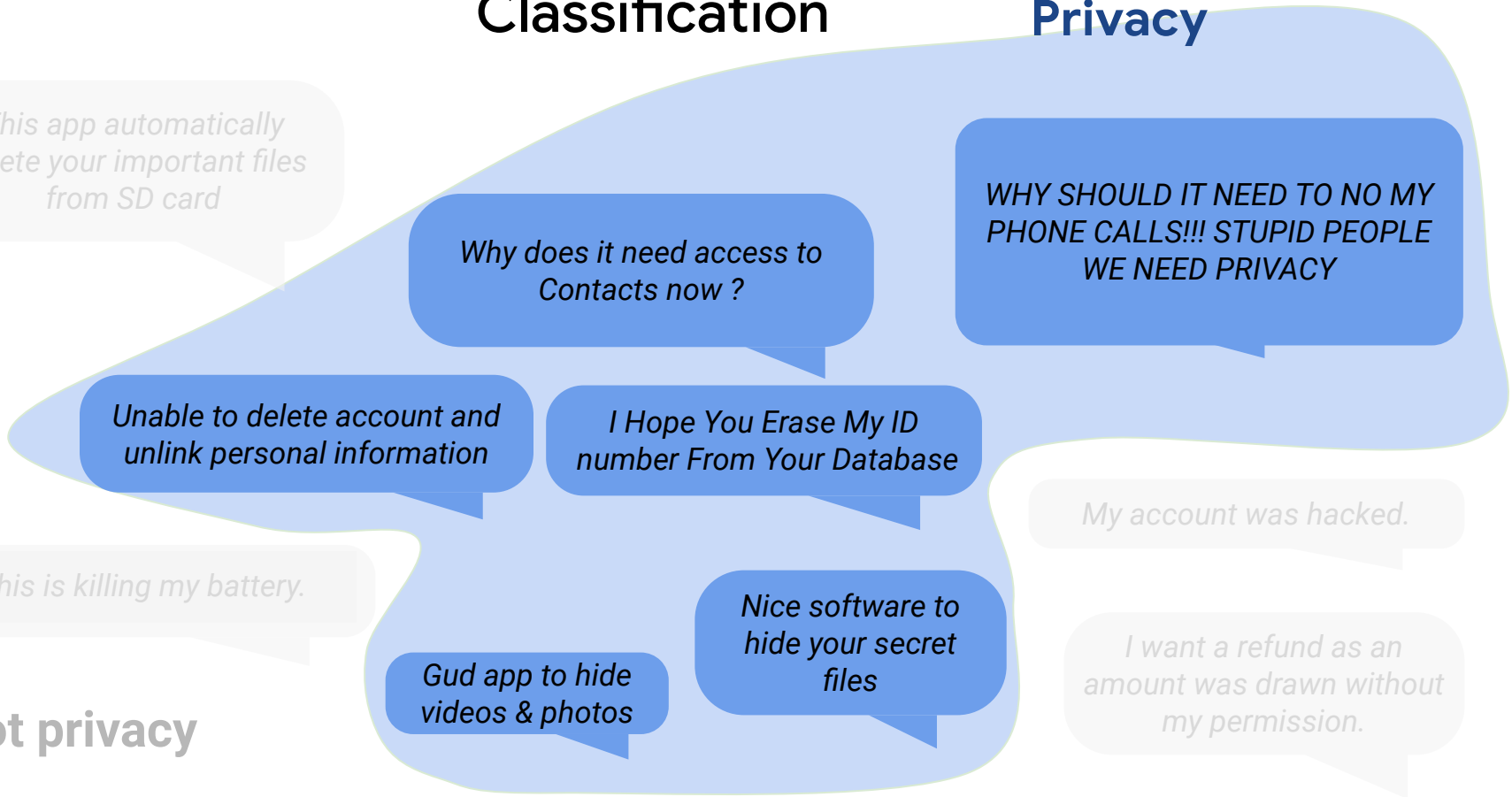
This is killing my battery.

Gud app to hide videos & photos

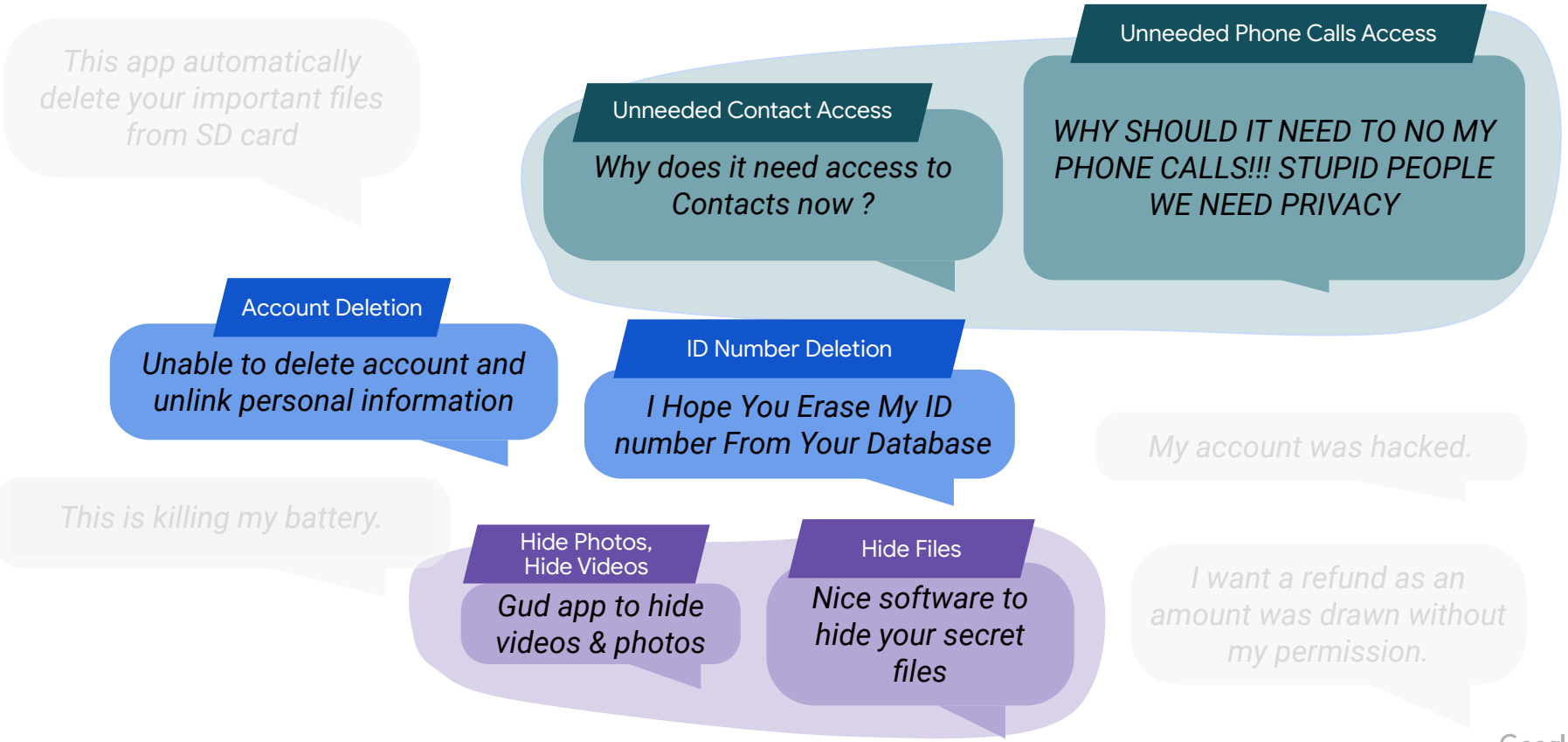
Nice software to hide your secret files

I want a refund as an amount was drawn without my permission.

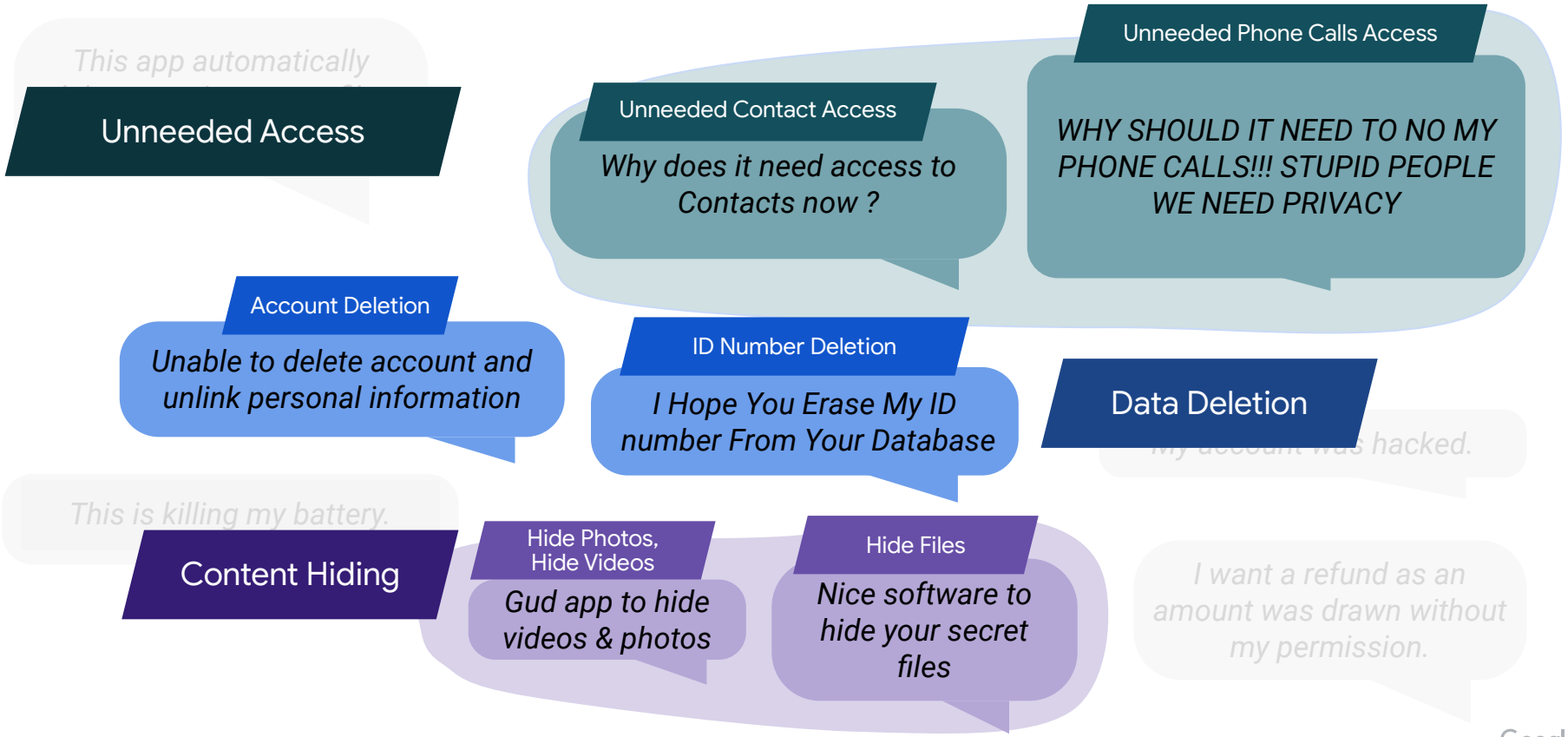
Classification



Automated Issue Generation (Existing & New)

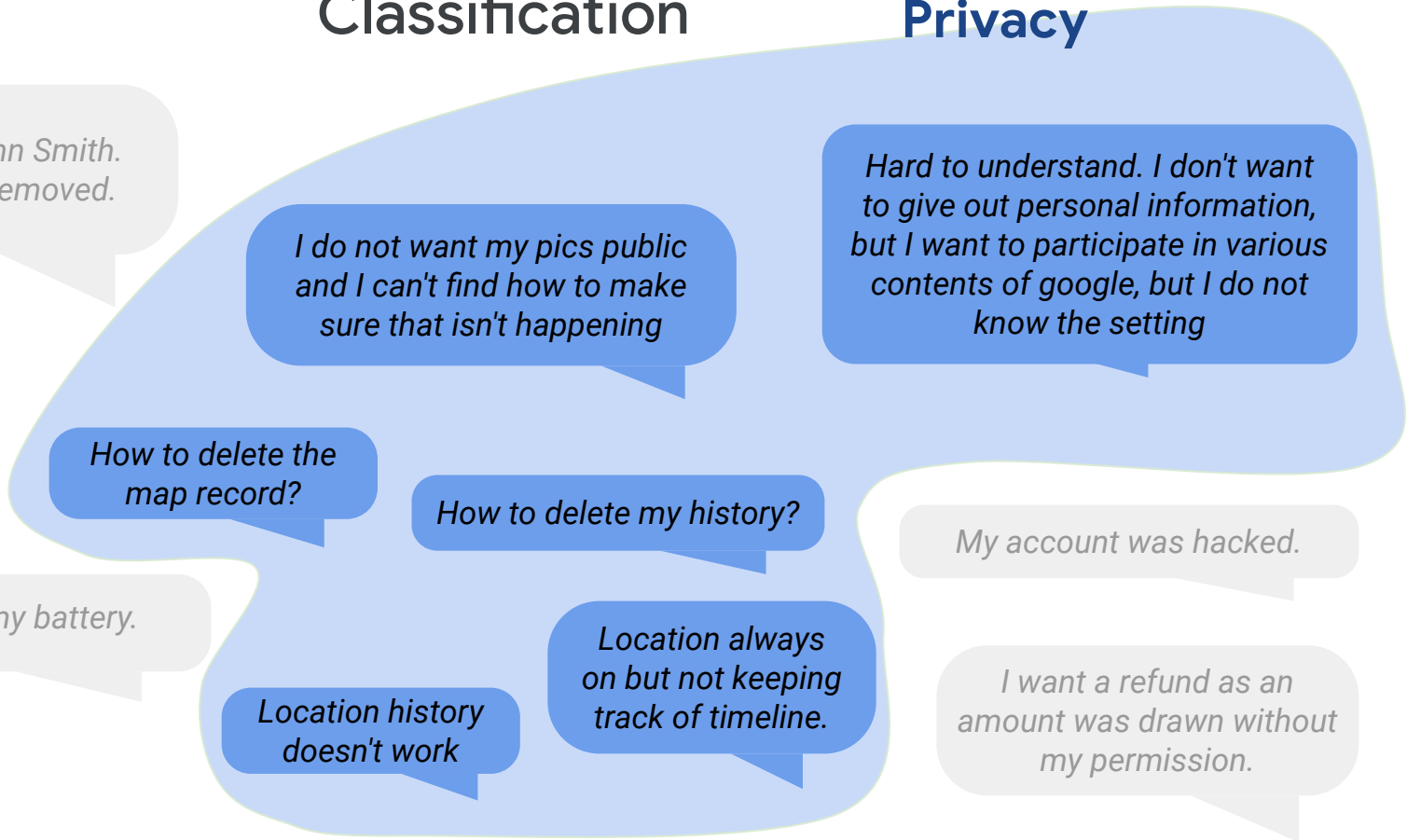


Summarizing the Top Themes

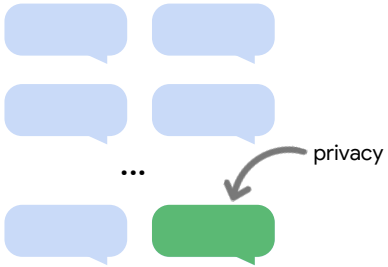


Privacy Classifier

Classification



Challenge: Tackling Data Imbalance



Imbalanced Data

Previous estimates: **<0.5%** of the reviews about privacy [1,2].

Previous approaches [1,2,3,4]**:**

- Search via keywords/regular expressions (e.g. for “privacy”, “permissions”, etc.).
- Then label data.

Limitations:

- Model will overfit on the presence/absence of these expressions.

Our approach:

- Use a Natural Language Inference (NLI) model for retrieving semantically relevant data.
- Guide the retrieval with concepts from a privacy taxonomy.

Advantages:

- Can generalize well beyond keywords.
- Ensures high coverage of a variety of privacy concepts

[1] Nguyen et al. IEEE S&P 2019; [2] Mukherjee et al., Arxiv 2019; [3] Nema et al. ICSE 2022; [4]; Besmer et al. IJISP 2020.

Natural Language Inference

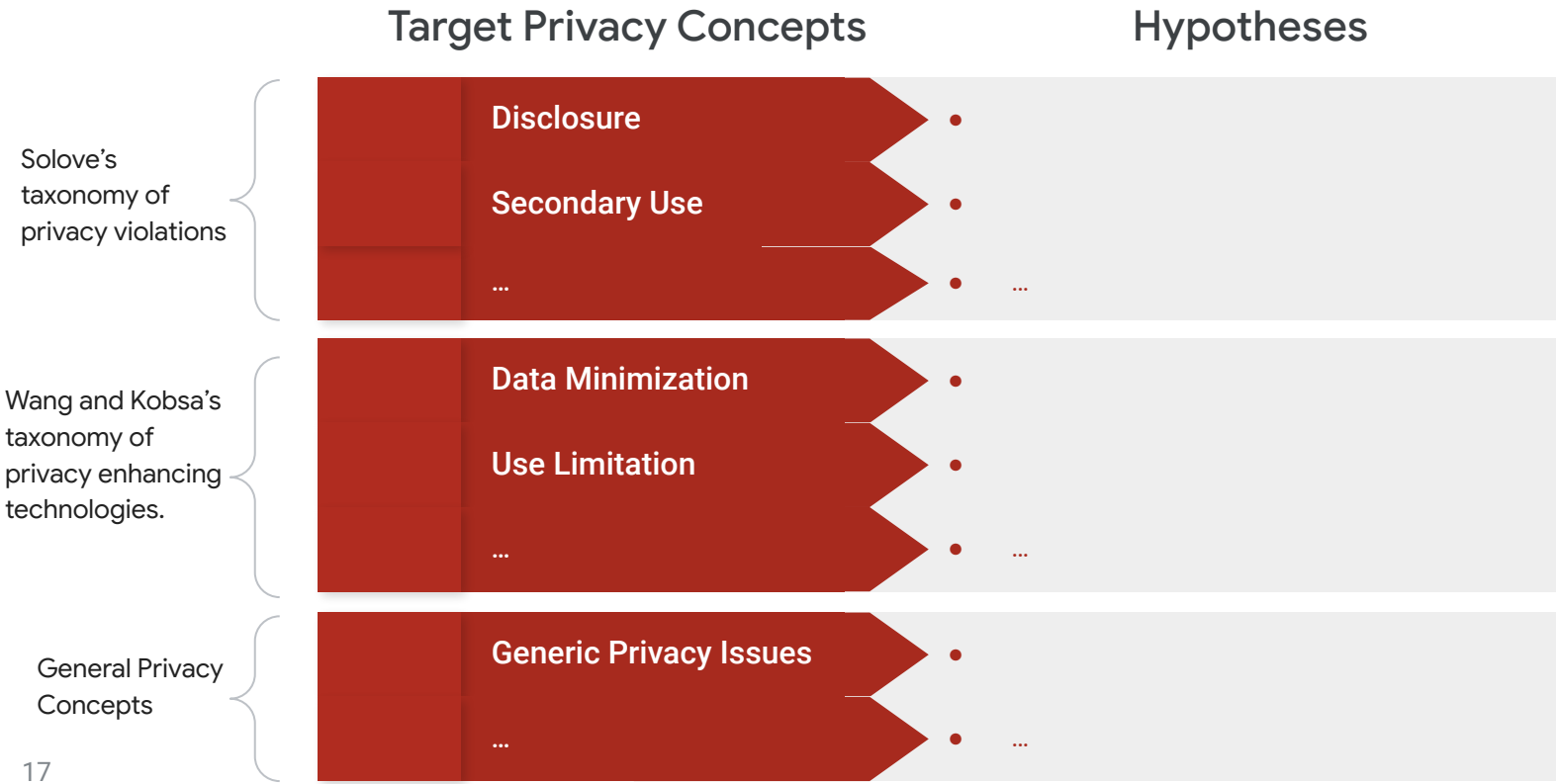
Natural language inference (NLI) is the task of determining whether a "**hypothesis**" is true (entailment), false (contradiction), or undetermined (neutral) given a "**premise**".

Generic NLI Dataset

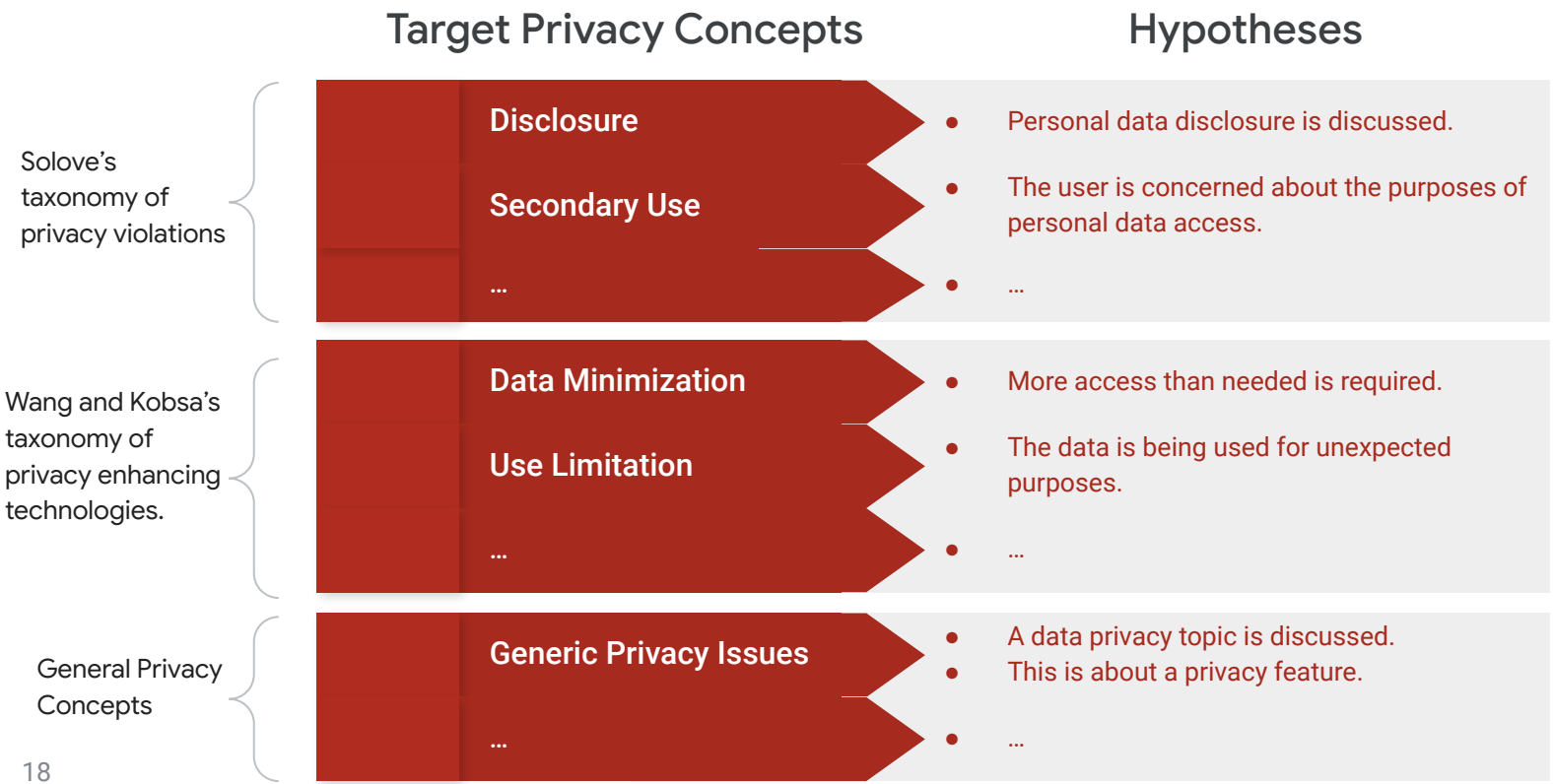
Premise	**Hypothesis**	**Verdict**
Houston is freezing and dry right now.	The city is really **humid** now.	contradiction
Houston is freezing and dry right now.	The city is really **cold** now.	entailment
Houston is freezing and dry right now.	The city is really **crowded** now.	neutral

Idea*: Use a generic model trained on NLI datasets to sample data matching privacy hypotheses.

How Do We Guide Our Data Sampling?



How Do We Guide Our Data Sampling?



How Do We Guide Our Data Sampling?

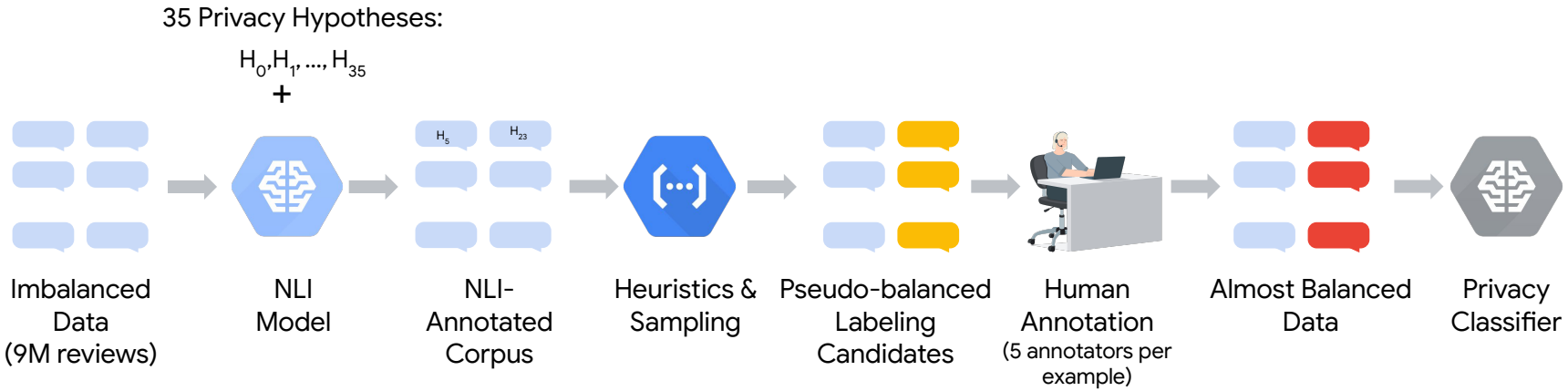
Target Privacy Concepts	Hypotheses
Disclosure	• Personal data disclosure is discussed.

Premises (User Reviews)	Entailment Probability
this game will NOT open unless you agree to them sharing your information to advertisers	0.89
and doesn't ask for access to unneeded personal data permissions. Well done developers 5Stars	0.75
super easy to use it removed my virus . . perfect	0.01

Approach*: Sample a pseudo-balanced set for labeling containing:

- reviews with high entailment scores for *some* hypotheses (maybe-privacy)
- reviews with low entailment scores for *all* hypotheses (maybe-not-privacy).

Privacy Classifier Building Pipeline Summary



Model Training with T5

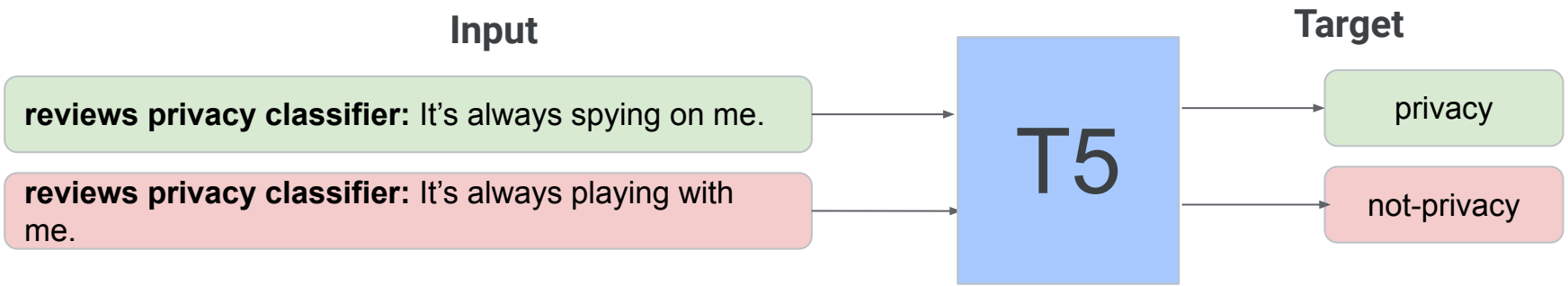
Model choice: **T5**

- a unified text-to-text model where the input and output are always text strings
- leading NLU benchmarks
- simplifies multi-task learning

Model Training with T5

Model choice: **T5**

- a unified text-to-text model where the input and output are always text strings
- leading NLU benchmarks
- simplifies multi-task learning



Evaluation Setup

Training Data Choices:

- **Hark Data:** constructed as described before (2.75K training set, 300 testing set)
- **ICSE Data:** obtained from the work by Nema et al. (ICSE 2022). Data is sampled via a set of regexes, guided by a set of privacy concepts.

Training Model Choices:

- **T5-11B:** constructed as described before
- **RoBERTa-large:** a 24-layer encoder model, achieving strong results on various classification tasks.
- **SVM:** SVM classifier based on bag of words (using 3-5 character n-grams), reproducing the model used in Nguyen et al. (IEEE S&P 2019).

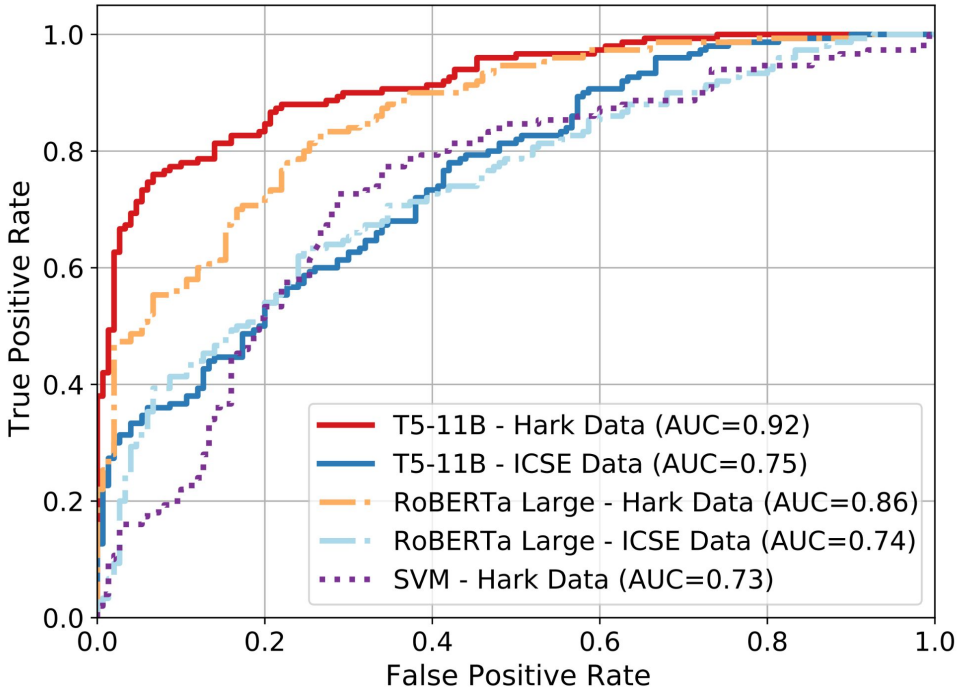
Privacy Classifier Evaluation (Hark Test Set)

Data sampling matters:

Training on a well-sampled dataset boosts AUC by 0.17 with T5-11B on Hark's test set.

Model choice matters:

SVM fails to learn the nuances. Leveraging transfer learning, larger models improve quality with the same training data.



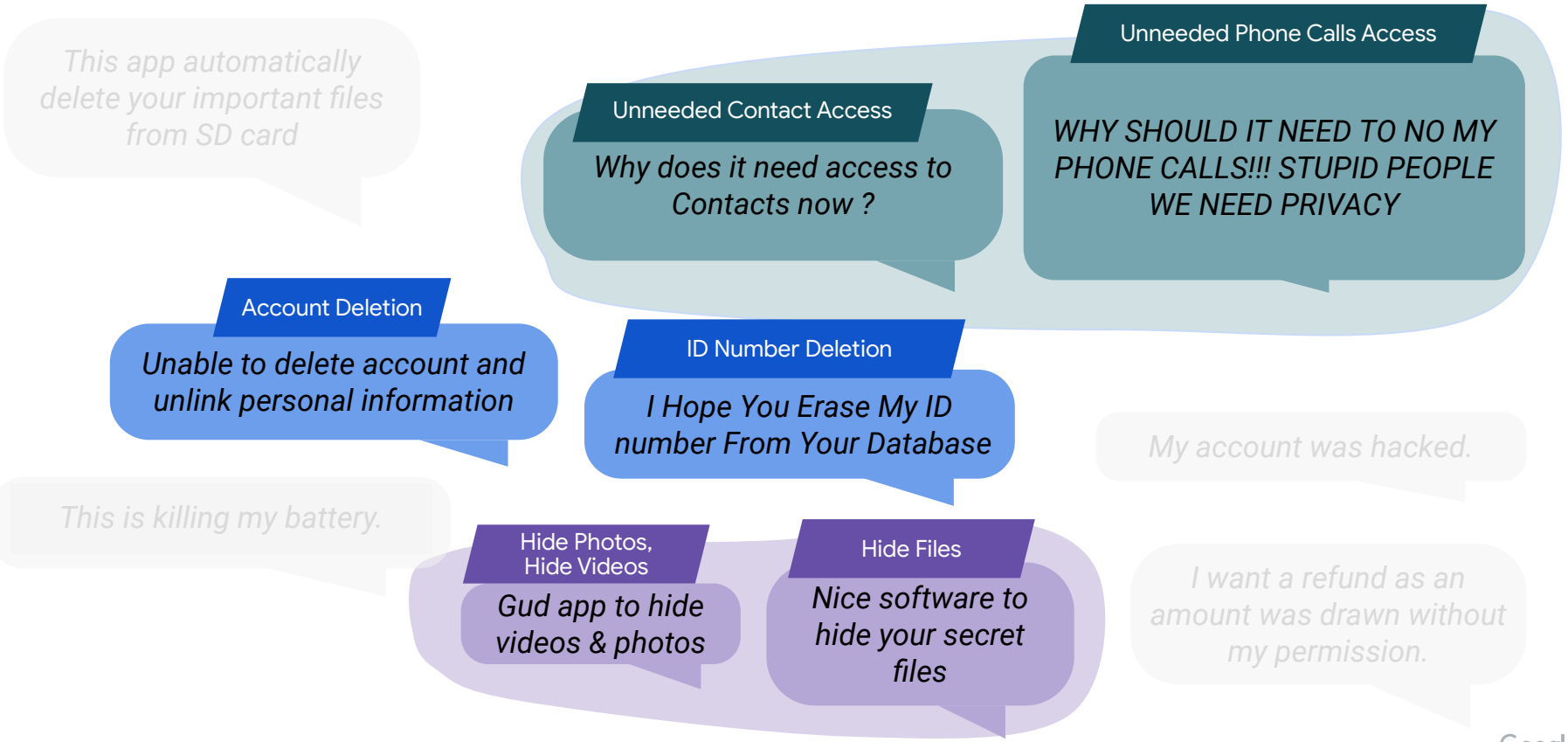
Privacy Classifier Evaluation (ICSE Test Set)

NLI-based sampling has better generalization power:

- Hark T5-11B had an AUC=0.98 on the ICSE test set.
- It matches the reported SOTA ensemble model trained on ICSE training set in the paper.

Issue Generation

Automated Issue Generation (Existing & New)



Challenge: Generating Unseen Issues

How can we generate issues (not necessarily seen before) while being:

- concise
- consistent
- fine-grained

Previous approaches:

- Either rely on classification approaches [1].
- Or rely on extracting key phrases [2,3].

Limitations:

- Classifiers are limited to a small set of predetermined issues.
- Key phrases don't repeat across reviews.

Our approach:

- Develop a generative model that produces issues in the style of succinct labels (new or seen before).

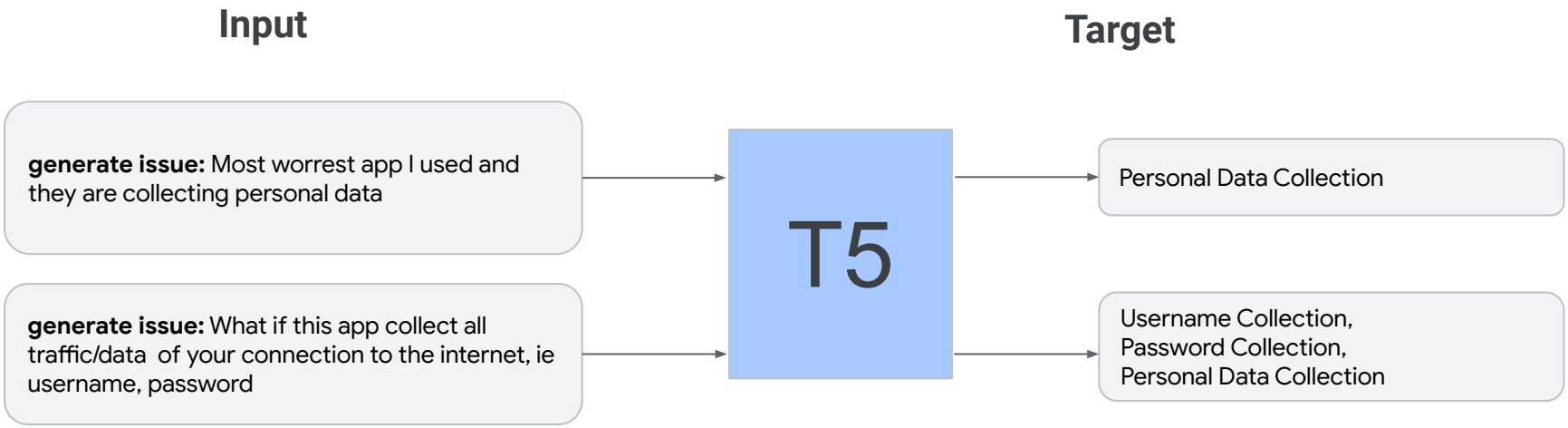
Advantages:

- Outputs consistently worded issues that repeat across reviews.
- Accounts for newly emerging issues.
- Issues are easily glanceable.

[1] Ciurumelea et al.SANER 2017; [2] Ferreira de Araújo et al., ACM SAC 2021; [3] Johann et al.RE 2017.

Abstractive Issue Generation Approach

Fine-tune the T5 model to generate issue tags given the review text.



Dataset Creation Strategy

Data sampling: We sampled reviews such that:

- We cover a variety of privacy concepts (using our NLI-Annotated Corpus).
- We balance the different app categories and review lengths.

Training Data Labeling:

- Two authors labeled each review with a set of issues.

Training Data Restrictions:

- Trick: limit the unique issues in the training data to encourage the model to be creative.
- 613 training examples and 447 validation examples.

Evaluation Setup

Compared models:

- **Hark Issue Gen:** T5-11B model trained on our dataset **(abstractive, our data)**
- **T5 Wikihow:** T5-11B model trained on an existing public dataset for abstractive summarization of article sections into section titles. **(abstractive, public data)**
- **RE-BERT:** SOTA model for extracting the key phrases representing software requirements in reviews (**extractive, previous work's data**).

Evaluation data:

- Sample a set of reviews diversified across concepts, text lengths, app categories (600 reviews)
- Generate the outputs from each model.
- Create a crowdsourced evaluation experiment of these outputs.
- Measure:
  - **Issue Accuracy:** how precise each issue is in capturing the intent of the review
  - **Issues Coverage:** how comprehensive a set of issues is in capturing the main topics mentioned in the review.

Example Outputs

Review: “*I love this app but dose my journey and my photos in the map can seen by other users and google map or it is private* 🤔.”

Hark Issue Gen: Journey Visibility, Photo Visibility

T5 Wikihow: Using the Google Maps App

RE-BERT: Photos

Instructions for Human Evaluation of Issue Accuracy

(7 annotators per review, 267 annotators in total)

Consider the following Android app review:

(Reviews are texts submitted by users to share their experience or concerns)

"I love this app but dose my journey and my photos in the map can seen by other users or it is private 🤔"

Which of the following best describes each label?

- **Journey Visibility**
  - The **topic is discussed** in the review
  - Contains keywords present in the review, but is **not a topic**
  - **Unrelated** to the review
- **Using the Google Maps App**
  - ...
- **Photos**
  - ...

Accuracy at N: % of reviews where a majority of N annotators agreed on **Topic_Discussed** being the rating.

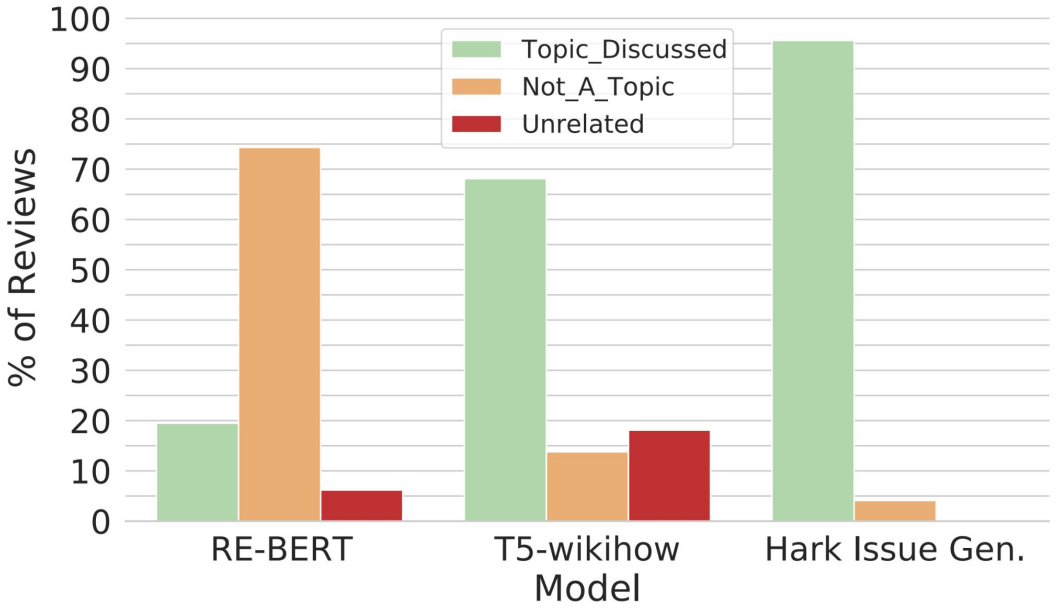
Issue Generation Accuracy (when N≥5 annotators agree)

Abstractive models do better.

Even if not targeted for review issue generation, abstractive models achieve better accuracy.

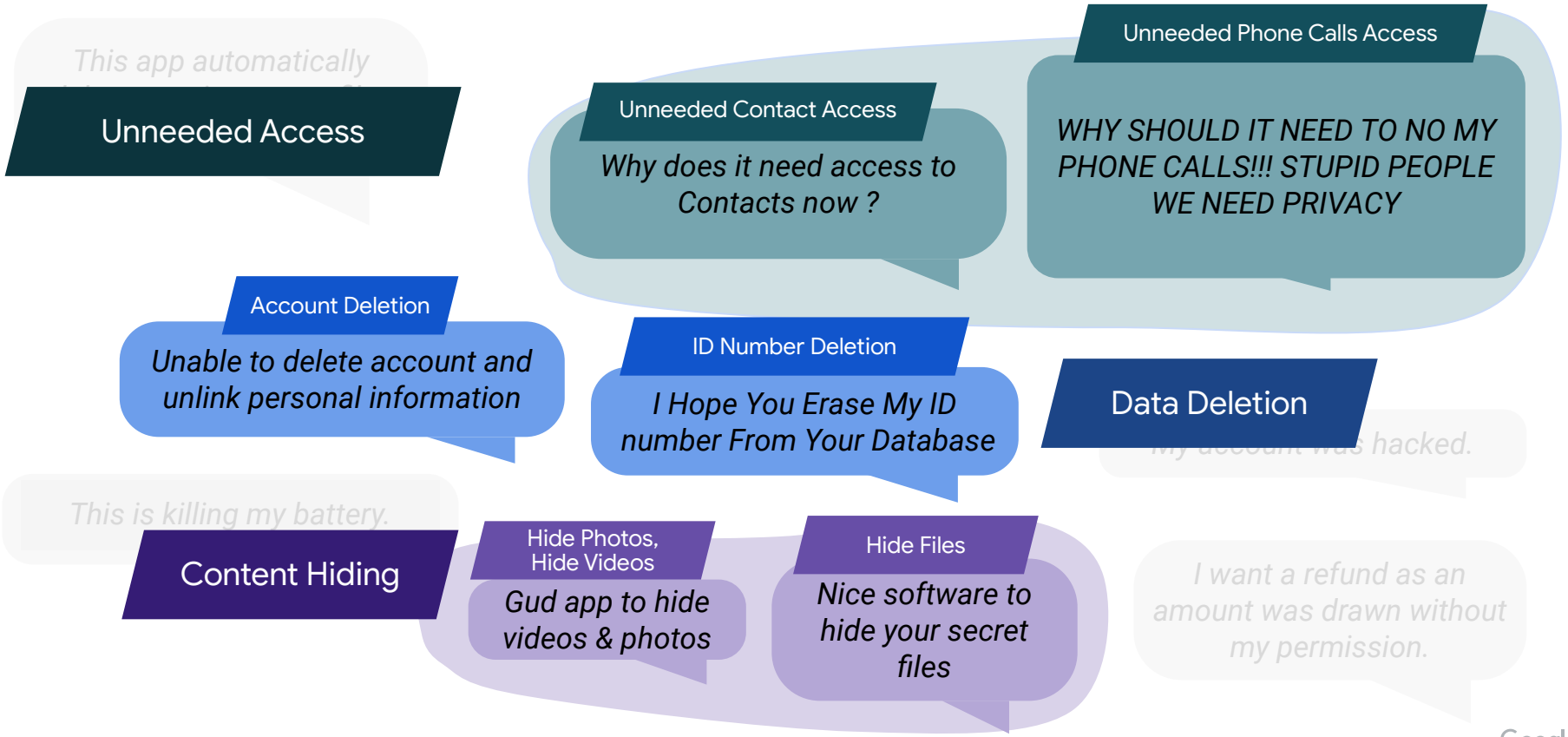
A custom dataset is a must.

T5-11B trained on Hark's issue generation dataset significantly outperforms that trained on off-the-shelf dataset (96% vs 79% accuracy).

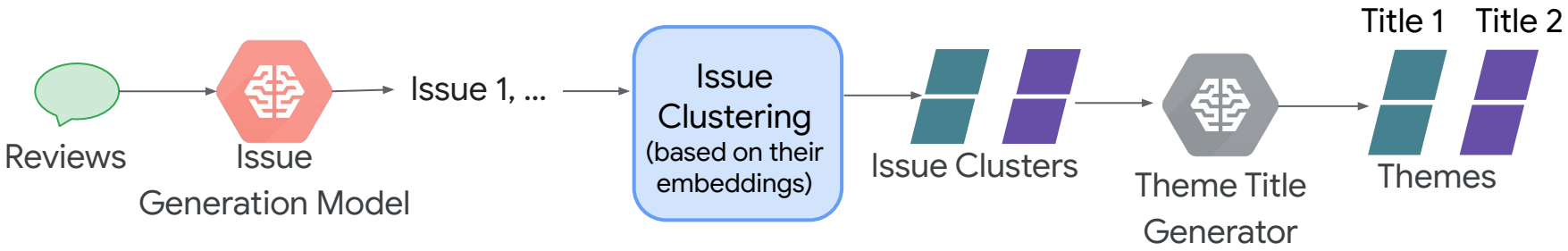


Theme Generation

Summarizing the Top Themes

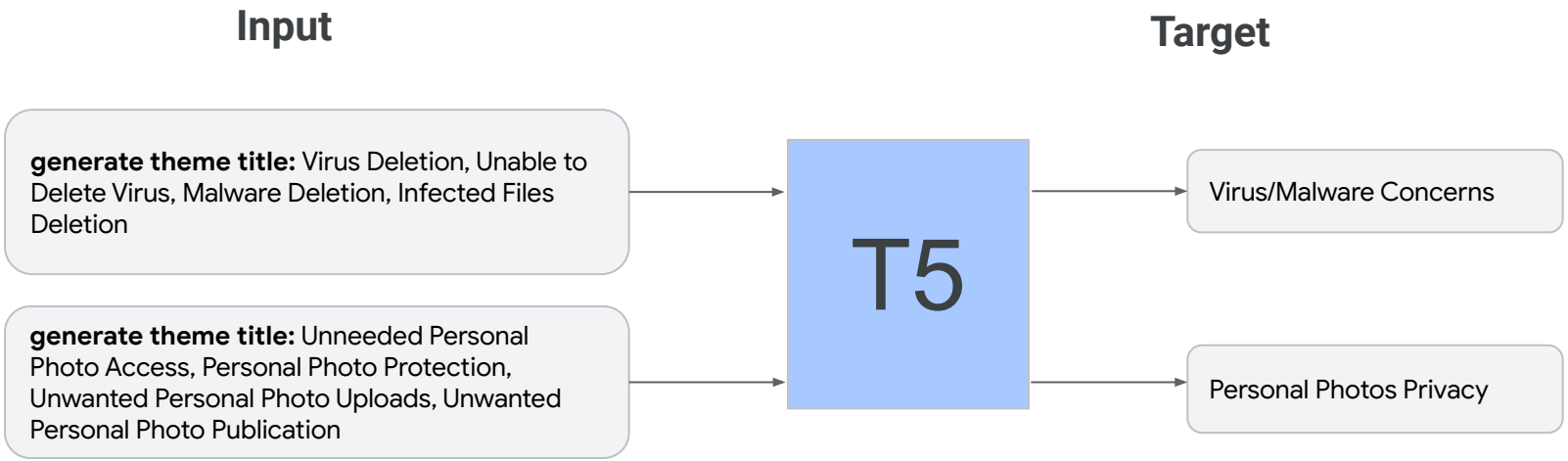


Theme Generation Overview

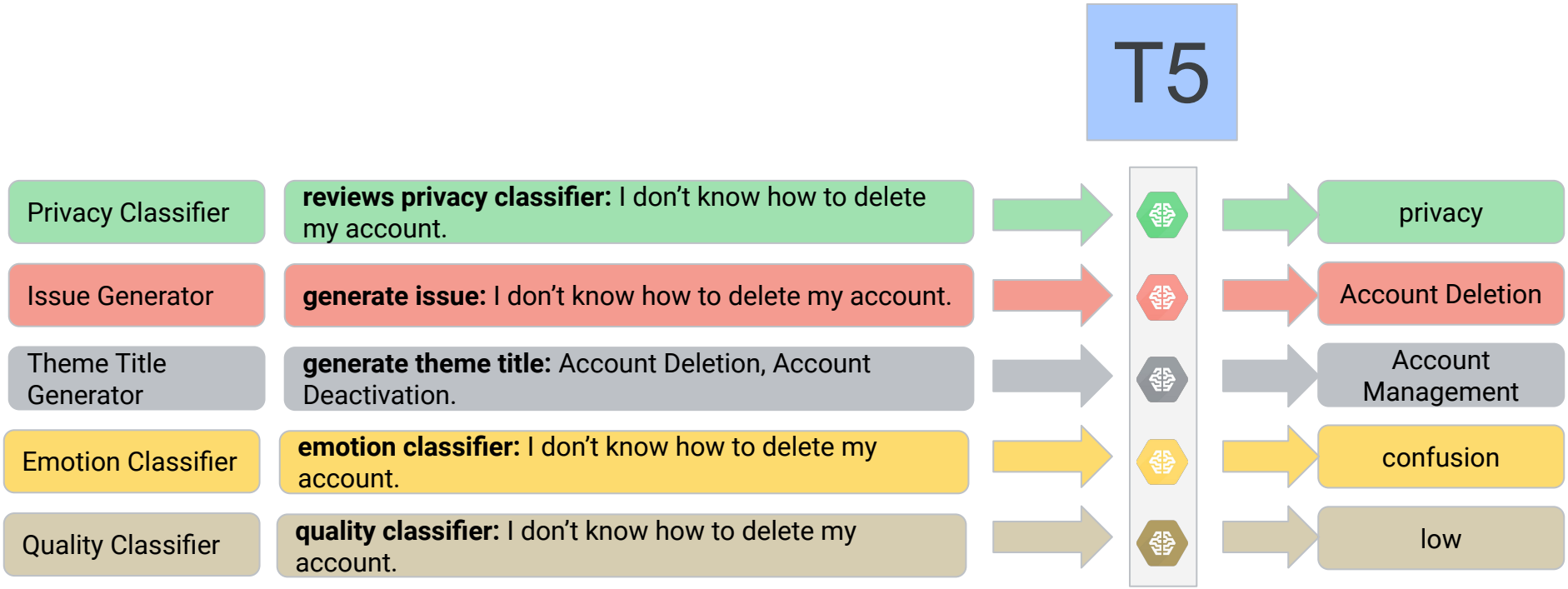


Abstractive Theme Title Generation Approach

Fine-tune the **T5 model** on a **custom dataset** to generate theme titles from a group of issues.



Hark's Unified Modeling Paradigm



Applying Hark Pipeline at Scale

Analyzing a Large Reviews Dataset

- Applied our methodology to the largest dataset of Play Store reviews to date.
- Illustrated how the Hark methodology allows easily navigating reviews at that scale.

626M

Reviews analyzed

1.3M

Apps with >10K installs and >1K reviews.

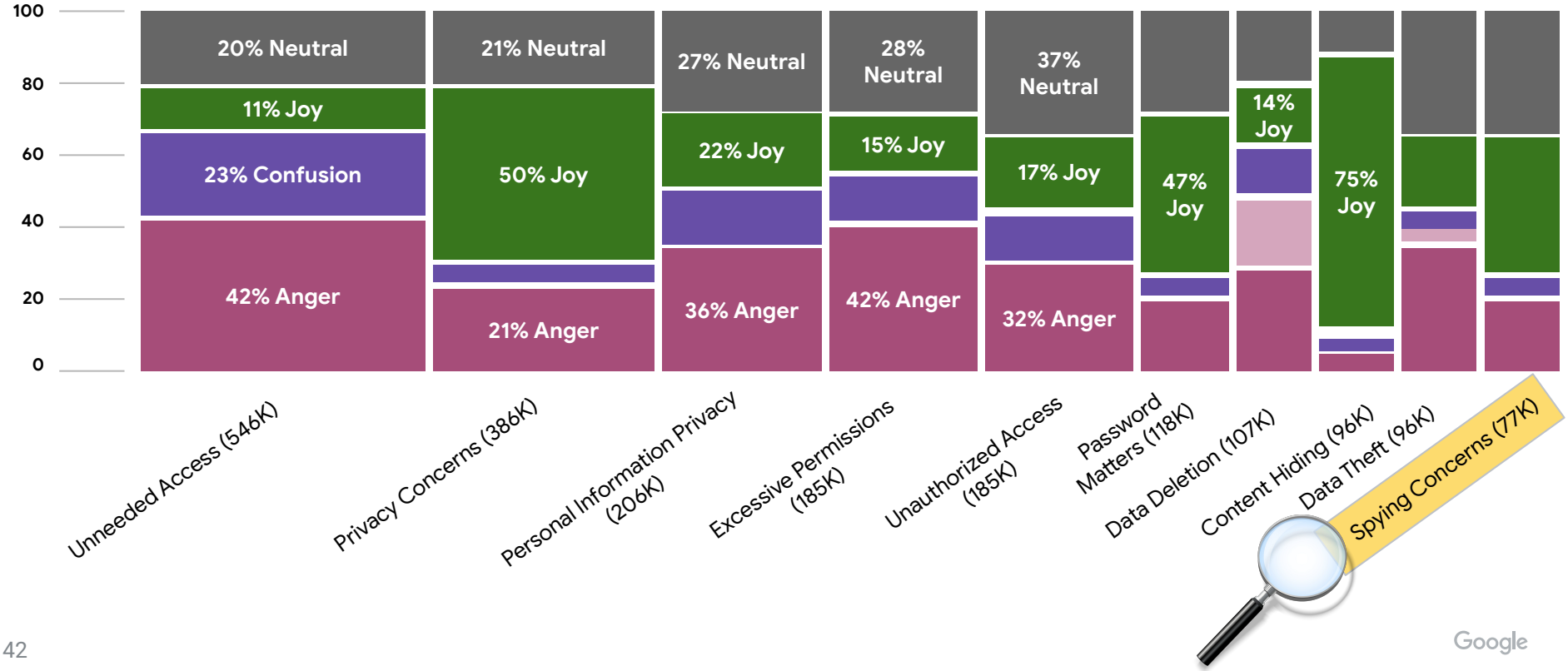
6M

Reviews classified as privacy.

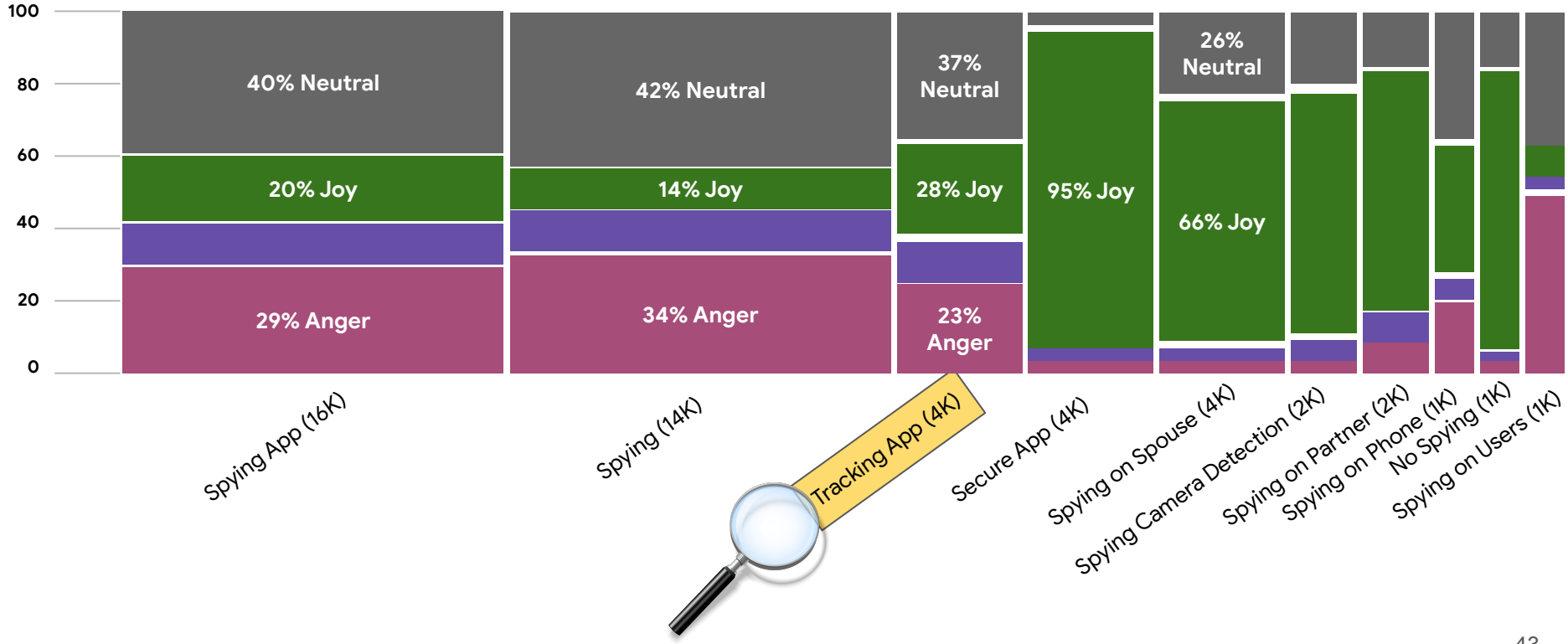
>300

High-level themes with >1K reviews identified

Top 10 Themes



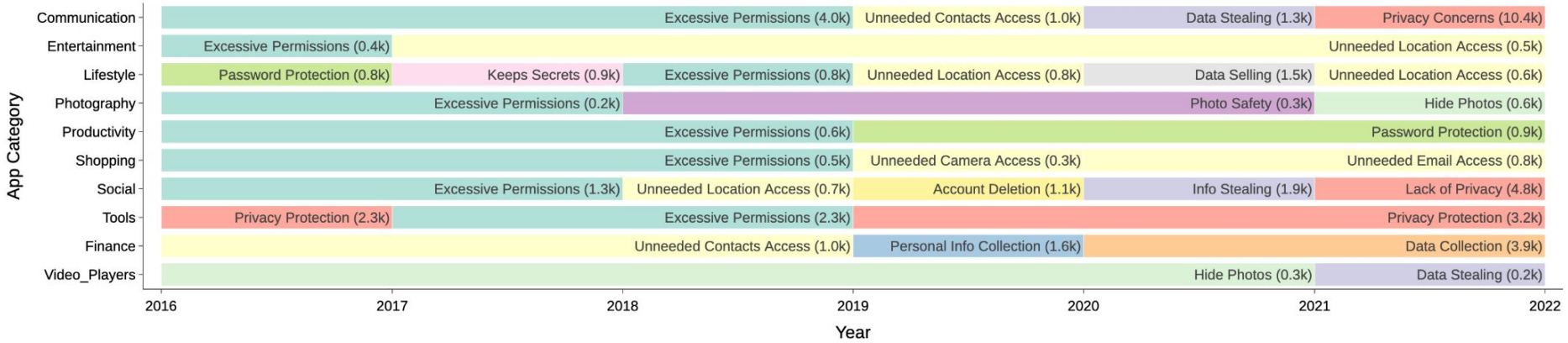
Spying Concerns Theme: Top 10 Fine-grained issues



Diversified Quotes Within the “Tracking App” Issue

Neutral	This app tracks everything u do on your smartphone.. Banking the whole nine yards... It has a 25 page privacy disclosure..
Joy	It has been very accurate. I love how it will give u a history of every time someone has left work/school/home, for how long, and the route they took to get to their destination.
Confusion	It wants to track and report the sites you visit, and which apps you use—which seems unnecessary and somewhat creepy (...) not sure what these folks have in mind (...)
Anger	IT’S @@@@ING SPYWARE! The app has become over-reaching in permissions and has become a tracking bot. Everytime I drive by a Walgreens I get coupon notifications.

Top-5 fine-grained privacy issues in last 5 years



Summary

Hark is an end-to-end system for extracting and analyzing privacy reviews. It achieves:

- **Topical diversity:** High coverage of the privacy concepts
- **Glanceability:** Developers can understand the gist of the topics without reading all reviews.
- **Navigability:** Developers have a high level picture and the ability to delve deeper into issues.

Questions?