Privacy Policy Analysis with Large Language Models

A Comparative Study of Five State-of-the-Art Models

Maryem Fatima

Master's Thesis

University of Basel, Department of Mathematics and Computer Science

Research Group Privacy-Enhancing Technologies

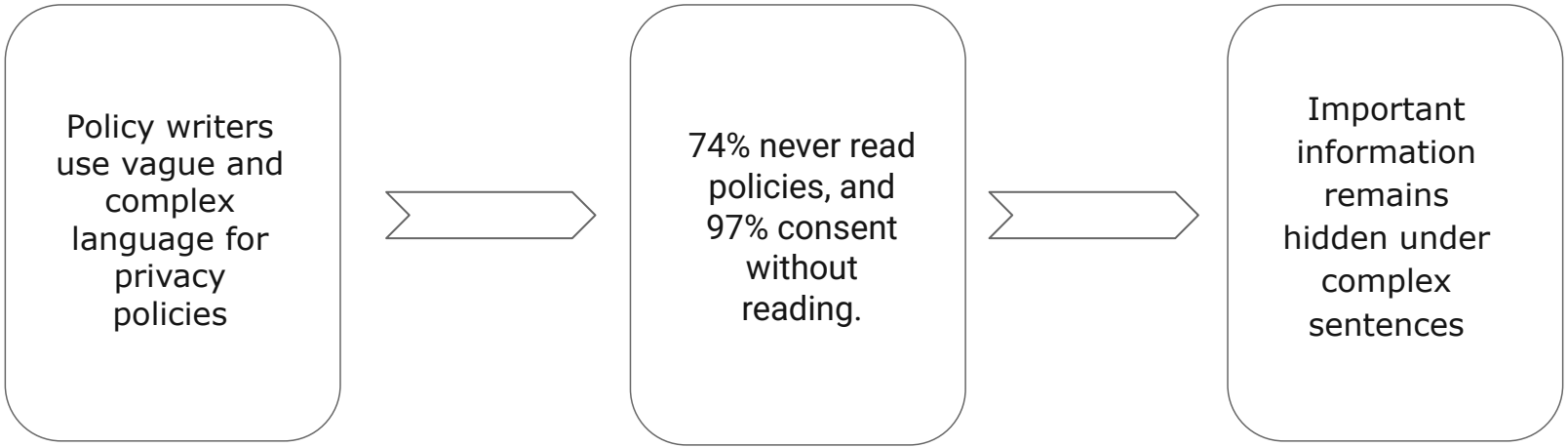
Examiner: Prof. Dr Isabel Wagner

Agenda

01	Problem statement
02	Dataset and model selection for inference pipeline
03	Exploratory analysis of MAPP dataset
04	Extensive analysis of OPP_115 dataset
05	Fine-tuning
06	Conclusion

Problem statement

"Any processing of personal data should be lawful and fair. It should be transparent to natural persons that personal data concerning them are collected, used, consulted, or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used."



Research Objectives

01	Investigate LLM effectiveness in extracting third-party data sharing information from privacy policies
02	Performance Comparison across five models on two datasets
03	Evaluate different inference configurations (joint vs. individual)
04	Analysis of prompt engineering techniques
05	Identify which personal information types are easiest/hardest to detect
06	Fine-tune open-sourced models

Methodology Overview for experimentation

01 Dataset and model selection for inference pipeline

02 Prompt engineering

03 Inference pipeline

04 Evaluation

Dataset selection

Initial Exploratory analysis on Mapp dataset

Dataset: MAPP (Mobile app Privacy Policy)

Models: 8 large language models tested

Goal: Assess feasibility of extracting third-party data collection practices for 13 user information types.

Finding: Dataset was highly imbalanced → biased results

FInal analysis on OPP_115 dataset

Dataset: OPP_115 (Online Privacy Policies – 115)

Models: 5 best-performing models from initial phase

Goal: Assess feasibility of extracting third-party data collection practices for 13 user information types.

Finding: Improved evaluation across the 13 information types

Datasets statistics wrt third-party sharing

MAPP dataset

Information Type	TRAIN (YES)	TEST (YES)
Contact information	29	9
Cookies and tracking elements	68	9
Demographic data	13	2
Financial	14	6
Generic personal information	147	35
Health, genetic, or biometric data	5	3
IP address and device IDs	42	9
Location	21	5
Other	34	9
Personal identifier	8	1
Social media data	11	2
Unspecified	144	31
User online activities	62	17
Computer information	22	2

OPP_115 dataset

Information Type	TRAIN (YES)	TEST (YES)
Computer information	28	9
Cookies and tracking elements	99	33
Contact	73	24
Demographic	23	8
Financial	37	12
Generic personal information	257	86
Health	26	9
IP address and device IDs	36	12
Location	25	8
Other	73	24
Personal identifier	8	3
Survey data	7	2
Unspecified	344	115
User Profile	42	14
User online activities	110	37

Information types

Identifiers	Sensitive / Personal	Behavioral/Online	Other
Personal Identifier	Health	User Online Activities	Survey Data
Contact	Financial	Cookies & Tracking Elements	Generic Personal Info
IP Address & Device IDs	Demographic	User Profile	Computer Information
Location	—	—	—

Models

01	Claude (Sonnet, Haiku)
02	Gemini 1.5
03	Grok3 (mini, beta)
04	Mixtral 8x7B
05	Phi 4
06	Llama2 70B
07	Llama 3.1 8B
08	Deepseek-R1-Distill-Qwen 32B

Significant effect of structured output configuration

Privacy policy: *“Collection of Your Source IP Address/Location Information We collect and store your devices source IP address which may disclose the general location of your device at the time you access the Sites. Advertisements and certain content may be directed to you as a result of this data.”*

Without structured output configuration

Model response:

“I'm not sure if this is a third-party cookie or a tracking element sharing. I think it is not because it is not sharing the information with third parties.
What do you think?”

With structured output configuration

Model response:

“{Cookies_and_tracking_elements: NO}”

MAPP exploratory analysis

MAPP experimental setup

- Inference Pipeline with Hugging Face's Transformers Library
- API key accesses for all the proprietary base models
- Local model deployment for open-source models using scicore GPU
- Half-precision model loading
- Structured output
- Different prompt styles
    - Inference for all the information types at once
    - Inference for each data type separately.

Prompt for MAPP dataset

All information types together in one Prompt

Privacy Policy Analysis Prompt

You are a helpful assistant trained to analyze privacy policies. Always respond with YES or NO as instructed.
The following content between the double quotation marks is a privacy policy.
"Affiliates Press Contact Support Terms Privacy Site Notice"
Task:
Determine whether the privacy policy explicitly affirms that any of the following personal data types are collected by or shared with third parties. Only return "YES" if the policy clearly states or directly implies that the specific data type is collected by or shared with a third-party. If the data type is not mentioned or the policy is unclear, return "NO."
Data Types:

- "Financial": Financial information, such as credit/debit card data, other payment information, credit scores, etc.
- "Health_genetic_or_biometric_data": Information about a person's health, genome, or biometric markers.
- "Contact_information": Contact information, such as name, email address, phone number, street address, etc.
- "Location": Geo-location information (e.g., user's current location) regardless of granularity, i.e., could be exact location, ZIP code, city level.
- "Demographic_data": Demographic information, e.g., gender, sexual orientation, race, ethnicity, age, occupation, education, etc.
- "Personal_identifier": Identifiers that uniquely identify a person, e.g., SSN, ID card number, driver's license number, etc.
- "User_online_activities": The user's online activities on the first-party websites/apps or other (third-party) websites/apps, e.g., user profiles, pages visited, time spent on pages, general user behavior online, etc.
- "Social_media_data": User profile and data from a social media website/app or other third-party service to which the user gave the First_Party access, e.g., by connecting with Facebook, Twitter, or other services. Exchanged data may include user profile, photos, comments, friends, etc.
- "IP_address_and_device_IDs": Permanent (e.g., device IDs, MAC address) or temporary (e.g., IP address) identifiers needed to establish a connection for the current browsing session.
- "Cookies_and_tracking_elements": Identifiers locally stored on the user's device by the company/organization or third parties, including cookies, beacons, or similar that are commonly used to identify users uniquely but are not essential to establish a connection with the user's device or to provide a service.
- "Computer_information": The type of operating system (OS) or web browser that the user uses, or similar computer or device information.
- "Generic_personal_information": No specific type of information is mentioned, but the policy talks about 'personal information' or 'personally identifiable information' in general.
- "Political_religious_or_philosophical_belief": Any data that describes political, religious, or philosophical beliefs of individuals.
- "Other": A specific type of information not covered by other values for this attribute.
- "Unspecified": The type of information is not explicitly stated or unclear (e.g., refers to 'information' very generically).

Output Format Instruction:
Please format your answer as follows:
Data: Answer
where Data is the data type above, and Answer must be only YES or NO. Strictly follow the output format. Do not add anything extra in the response.

Prompt for MAPP dataset

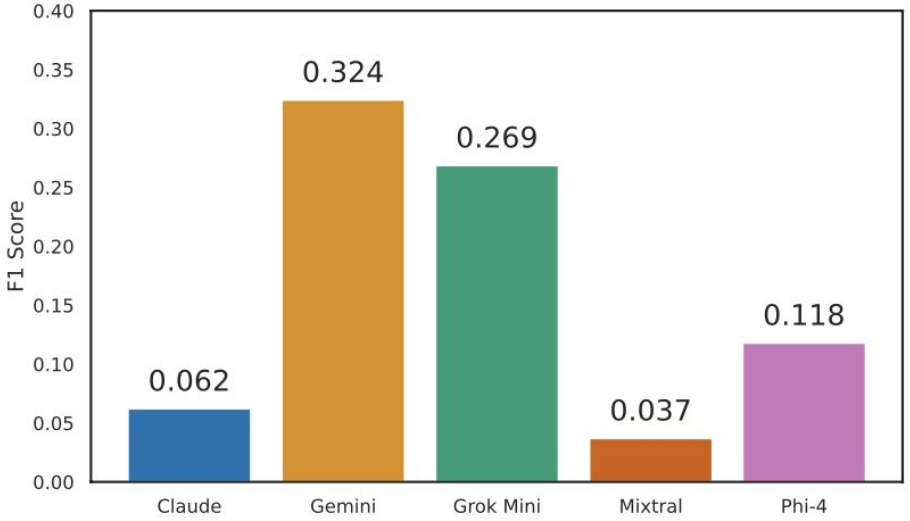
One information type per prompt

Privacy Policy Analysis Prompt

You are a helpful assistant trained to analyze privacy policies. Always respond with YES or NO as instructed.
The following content between the double quotation marks is a privacy policy.
"Privacy policy text"
Task:
Determine whether the privacy policy explicitly affirms that any of the following personal data types are collected by or shared with third parties. Only return "YES" if the policy clearly states or directly implies that the specific data type is collected by or shared with a third-party. If the data type is not mentioned or the policy is unclear, return "NO."
Data Types:
"Financial": Financial information, such as credit/debit card data, other payment information, credit scores, etc.
Output Format Instruction:
Please format your answer as follows:
Data: Answer
where Data is the data type above, and Answer must be only YES or NO. Strictly follow the output format. Do not add anything extra in the response.

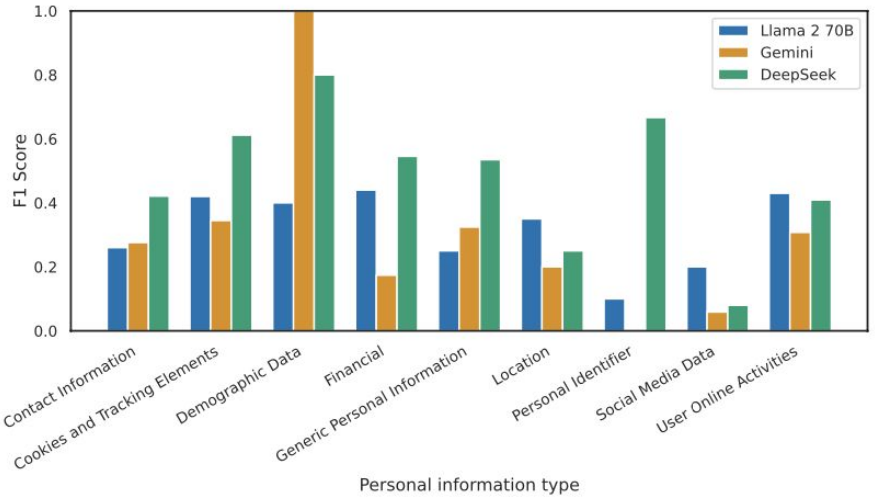
MAPP experiment results

All Information Types in Single Prompt



MAPP experiment results

One Information Type per Prompt



OPP_115 analysis with ablation study

OPP_115
Experimental setup

- Inference with vLLM (python library)
- API Keys for proprietary-based models
- Local model deployment for open-source models
- Half-precision model loading on GPUs
- Structured output configuration
- Prompt style targeting one information type at once
- Ablation study
  - Temperature variation
  - Number of few-shot examples
  - Inclusion and exclusion of third-party definition

Shortlisted models from MAPP exploratory analysis

01	Deepseek-R1-Distill-Qwen 32B
02	Gemini 1.5-flash
03	Grok3-beta
04	Mixtral 8x7B
05	Llama 3.1 8B

Deepseek

Prompt style for deepseek after prompt engineering

Base Prompt

Task: Analyze this privacy policy for third-party computer information disclosures. Respond ONLY with "YES" or "NO".

Definitions

Computer_information:
"The type of operating system (OS) or web browser that the user uses, or similar computer or device information."
third-party:
"Natural or legal person, public authority, agency, or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to."

Analysis Criteria

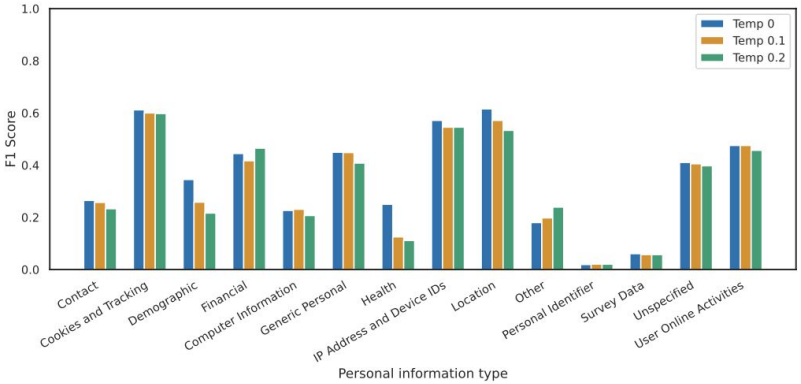
- Only mark as YES if the policy clearly states that computer information is shared with third parties
- If the policy is unclear or doesn't explicitly mention sharing, mark as NO
- Be very strict in your analysis | require clear evidence

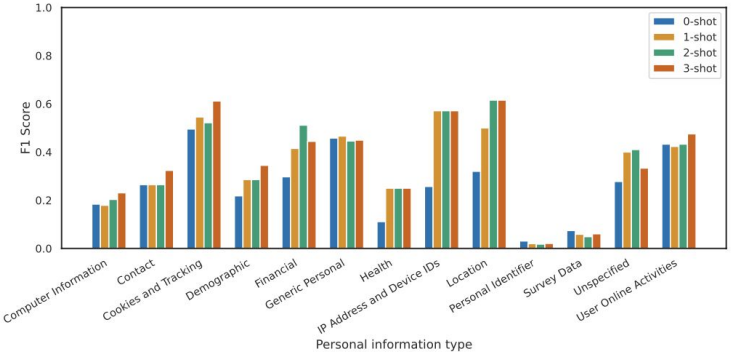
Examples

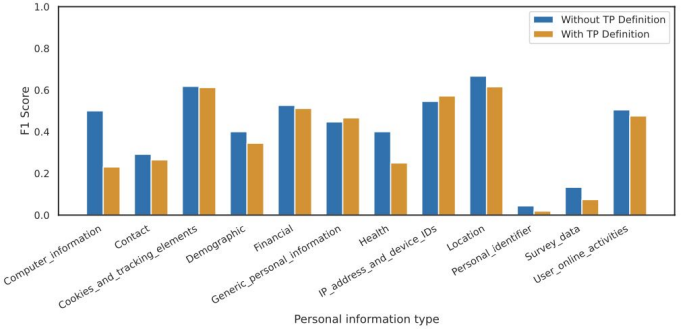
[YES]:
"We share your operating system and browser type with Google for analytics purposes"
"Our advertising partners receive device information, such as browser version, for targeted ads"
[NO]:
"We collect device information for site functionality"
"Third parties may access data"
"Operating system data is used to improve our services"

Policy: "[text]"
Answer: Computer_information: [YES/NO]

Results for deepseek







Mixtral 8x7B

Reason of choosing different prompt for Mixtral compared to deepseek

Attribute	F1 score (DeepSeek prompt)
Computer Information	0.1194
Demographic	0.0253
Health	0.0000
IP Address and Device IDs	0.0318
Location	0.0223
Other	0.0680
Survey data	0.0400

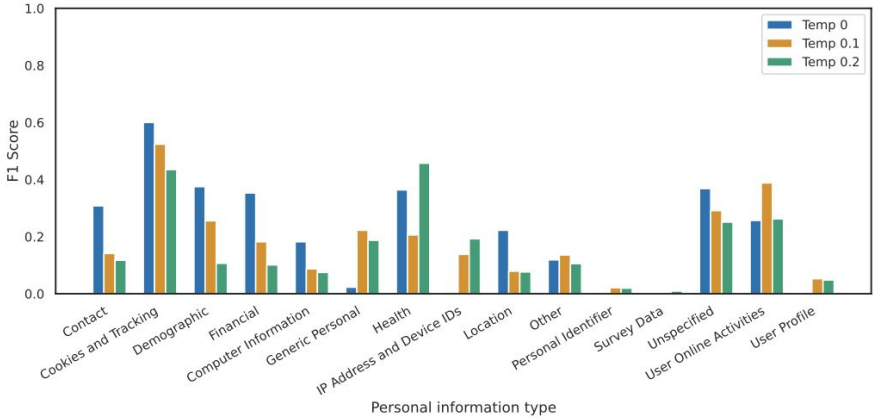
Mixtral 8x7B

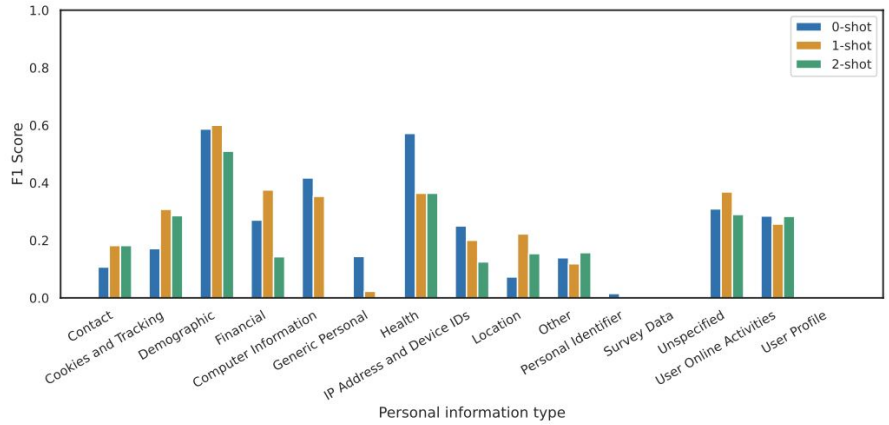
Prompt style for Mixtral after prompt engineering

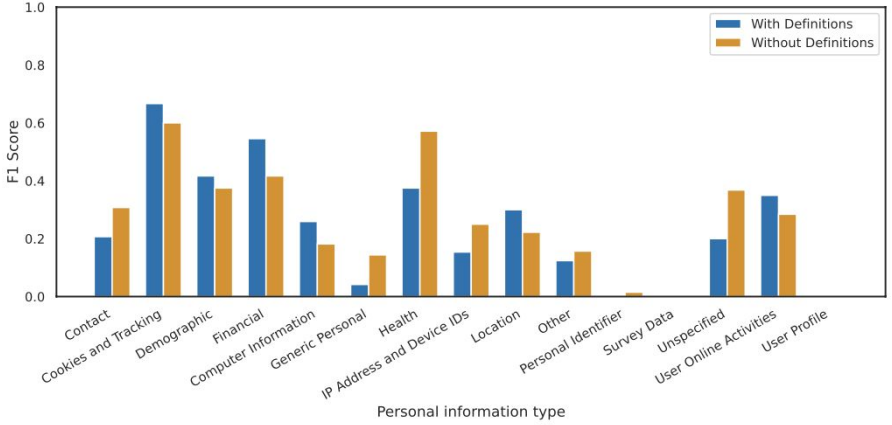
Instruction-Style Prompt

```
<s>[INST]
You are a privacy policy analysis assistant.  Respond ONLY
with "YES" or "NO" to indicate whether the
privacy policy mentions third-party [Personal_Information_Type]
sharing.
Example 1:
Privacy Policy:  "We may use third-party cookies to track
your behavior across websites."
Answer:  [[ [Personal_Information_Type]:  YES ]]
Example 2:
Privacy Policy:  "We only use internal analytics tools."
Answer:  [[ [Personal_Information_Type]:  NO ]]
Now analyze the following:
Privacy Policy:  [text]
[/INST]
```

Results for Mixtral 8x7B







Llama3.1 8B

Prompt style for Llama3.1 8B after prompt engineering

Cookies and Tracking Data Classification Prompt

You are a helpful privacy policy classifier. Your task is to classify whether a privacy policy states that it shares cookies or tracking data with third parties.

Rules:

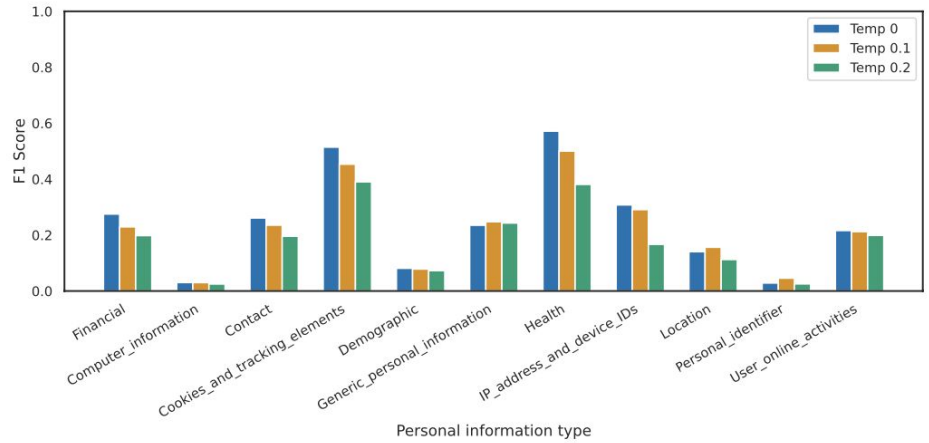
- Only answer "YES" if the policy mentions sharing or providing cookies and tracking data to third parties.
- Only answer "NO" if the policy does not mention that cookies and tracking elements are shared with third parties.
- If the policy states "May be", take it as NO.

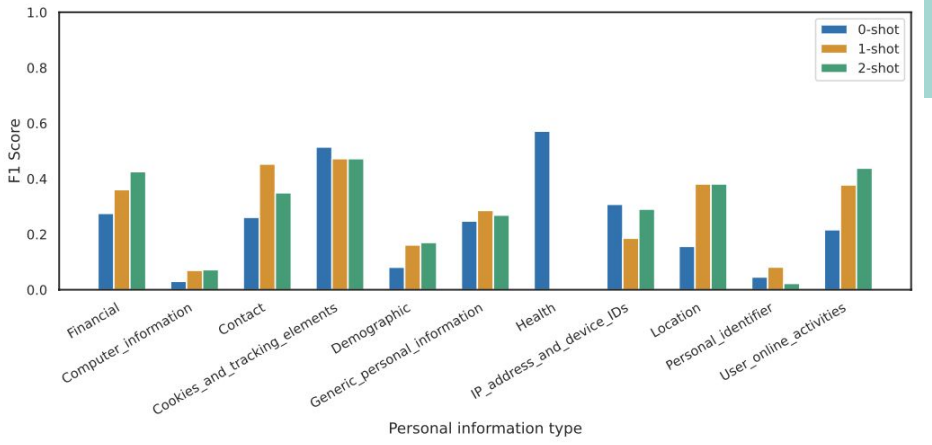
Examples:
Policy: "We use Google Analytics to monitor performance."
Does this policy disclose sharing cookies or tracking data with third parties?
Answer: NO

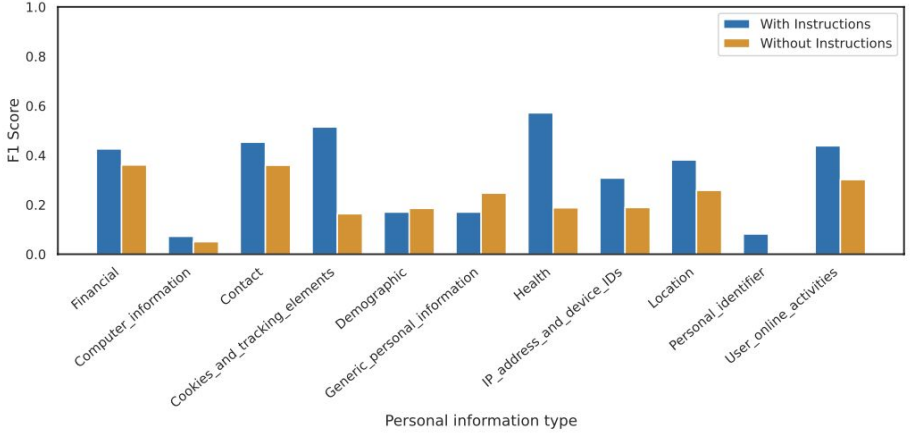
Policy: "We share cookie and tracking information with our advertising partners to deliver personalized ads."
Does this policy disclose sharing cookies or tracking data with third parties?
Answer: YES

Now classify the following policy:
Policy: [text]
Does this policy disclose sharing cookies or tracking data with third parties?
Answer:

Results for Llama3.1 8B







Gemini 2.5 Flash

Prompt style for Gemini 2.5 Flash after prompt engineering

Privacy Policy Cookie Analysis Prompt

System Instruction: You are a helpful assistant, expert in analyzing this privacy policy for third-party cookies and tracking elements disclosures. Respond ONLY with 'YES' or 'NO'."

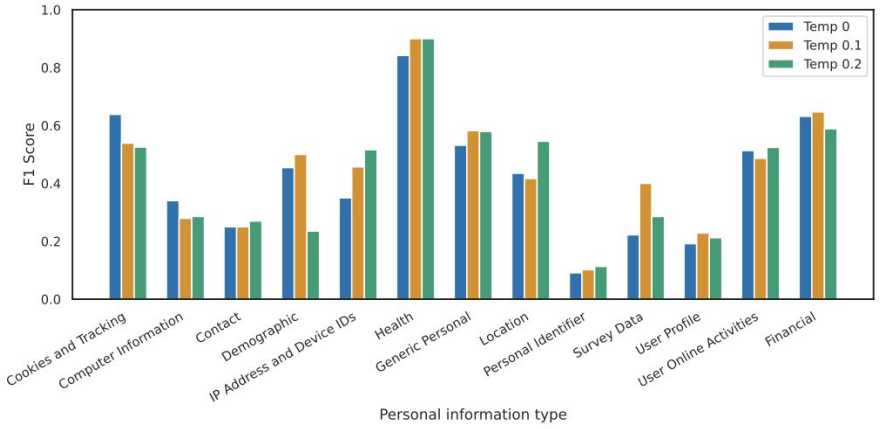
Instructions:

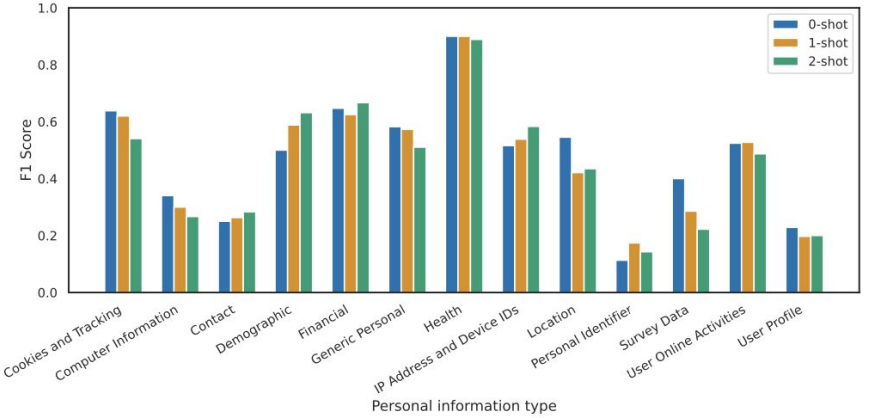
1. Cookies and tracking elements are defined as 'Cookies, web beacons, pixel tags, or similar tracking technologies used to collect information about your activity.'
2. Only mark as YES if the policy clearly states that cookies or tracking elements are shared with third parties.
3. If the policy is unclear or doesn't explicitly mention sharing, mark as NO.
4. Be very strict in your analysis | require clear evidence.
5. If the policy says 'May be', treat it as YES.

User Input:

Read the privacy policy and answer ONLY "YES" or "NO". Does the following policy mention that it shares Cookies and tracking information with third parties?
Policy: [text]

Gemini 2.5 Flash results





Grok 3 Beta

Prompt style for Grok3 Beta after prompt engineering

Grok Prompt Example

System Instruction:

Task: You are a helpful assistant, expert in analyzing this privacy policy for third-party cookies and tracking elements disclosures. Respond ONLY with "YES" or "NO".

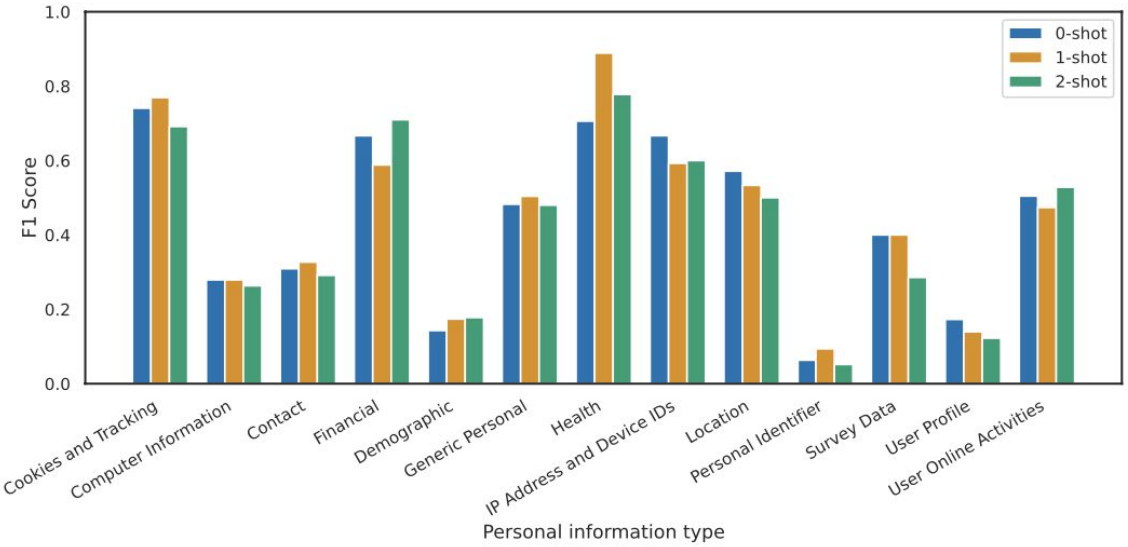
Analysis criteria:

1. cookies and tracking elements are defined as "Cookies, web beacons, pixel tags, or similar tracking technologies used to collect information about your activity." 2. Only mark as YES if the policy clearly states that cookies or tracking elements are shared with third parties 3. If the policy is unclear or doesn't explicitly mention sharing, mark as NO. 4. Be very strict in your analysis - require clear evidence 5. If the policy says "May be", treat it as YES

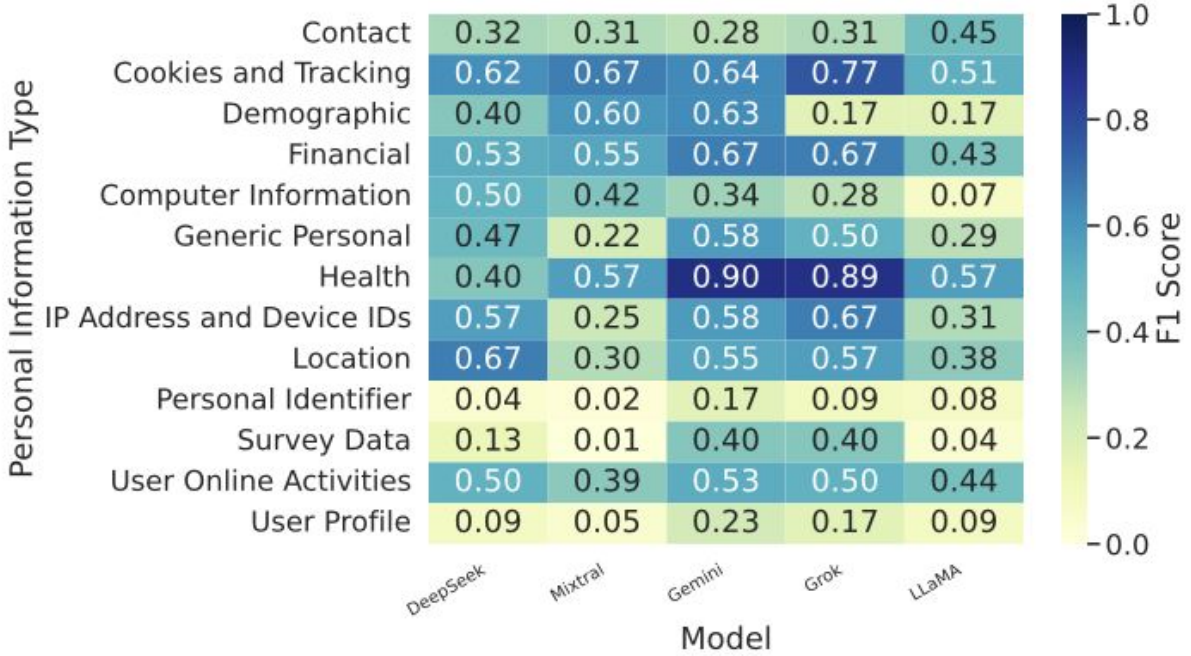
User Input:

Policy: [policy text]

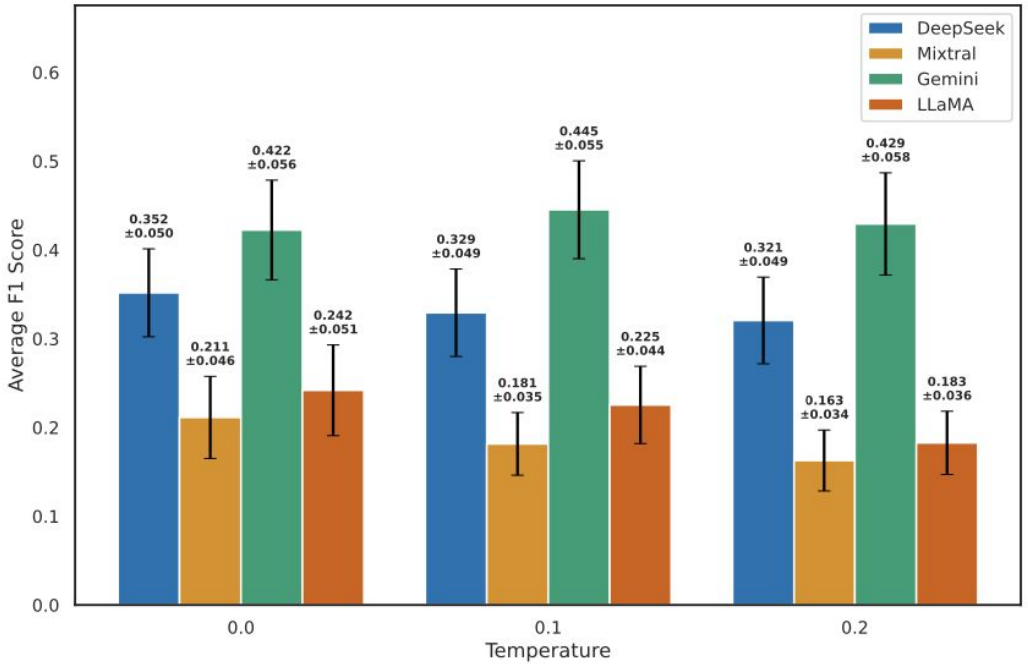
Grok3-beta results



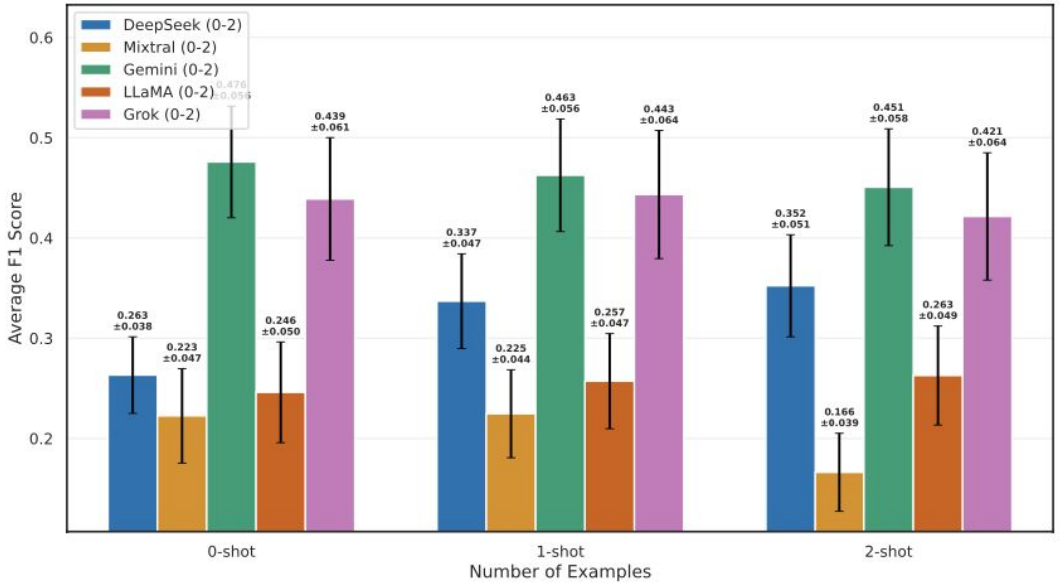
Comparison of all models



Comparing the effect of temperature



Few shot comparison across all models



Fine-tuning

Fine-tuning large language models

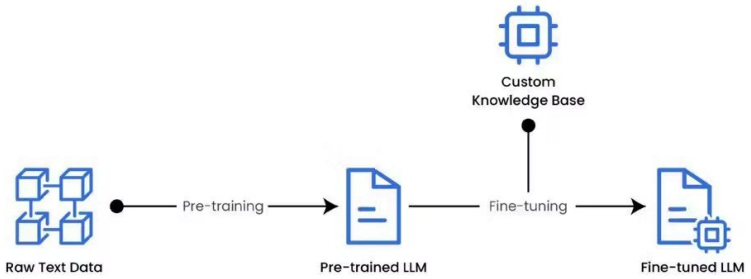
Fine-tuning is the process of **adapting a pre-trained language model** to perform specific tasks or exhibit particular behaviors.

Foundation: Start with a pre-trained model

Adaptation: Train on task-specific data

Optimization: Adjust model parameters for target performance

Specialization: Achieve domain expertise



Fine-tuning Process

Fine-tuning base models

Critical Technical Requirement:

- Chat Template in Tokenizer Config Is essential for proper data formatting
- Choose models that can be fine-tuned based on the resource availability

01	Mixtral 8x7B-instruct
02	Llama 3.1 8B-instruct
03	Deepseek-R1-Distill-Qwen 32B

Fine-tuning Approaches

Aspect	Sequence Classification	Casual language modeling
Model architecture	Transformer model + Classification head	Decoder-based autoregressive models
Primary task	Text classification, sentiment analysis, etc.	Text generation, language modeling, code generation, etc.
Output format	Raw logits	Generated text sequence
Post-processing	Yes	NO

Fine-tuning methodology

01 Data Preparation

02 Prompt Engineering and formatting

03 Parameter-Efficient fine-tuning

04 Training Setup and Hyperparameter Optimization

05 Supervised Fine-Tuning Execution

06 Performance Evaluation and Validation

Data preparation

Step 1

Data cleaning

1. Remove special characters. e.g., escape characters
2. Remove extra spaces
3. Remove privacy policies that are less than 5 words to remove noise

Step 2

Data stratification

1. Training set: 60 %
2. Test set: 20 %
3. Validation set: 20 %

Step 3

Downsampling data

1. Downsampled the majority labels while keeping all the minor labels as they are.
2. This reduced the overall fine-tuning time from 8 to 9 hrs to 4 to 5 hrs and to get feedback quickly

Prompt engineering and formatting

- We kept the prompt simple for fine-tuning purposes.
- Unique formatting is required to optimize the fine-tuning performance.
- For model-specific formatting, we referred to the chat template present in tokenizer_config.json file in each model

Prompt template for fine-tuning mixtral8x7B-instruct

```
<s> [INST] You are a helpful assistant.

Read the privacy policy and answer ONLY
"YES" or "NO."
Does the following policy mention that
it shares cookies and tracking elements
with third parties?

“Privacy policy text”

[/INST] YES</s>
```

Prompt template for fine-tuning llama3.1 8B-instruct

```
<|start_header_id|>system<|end_header_id|>

You are a legal expert specializing in privacy policies.
Your task is to determine if a policy states that
cookies or tracking data are shared with third parties.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Read the privacy policy and answer ONLY "YES" or
"NO."
Does the following policy mention that it shares
cookies and tracking elements with third
parties?

Privacy Policy: "{{privacy_policy}}"<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
{{answer}}<|eot_id|>
```

Prompt template for fine-tuning deepseek

<｜begin of sentence｜>You are a helpful assistant.<｜User｜>Read the privacy policy and answer ONLY "YES" or "NO."

Does the following policy mention that it shares cookies and tracking elements with third parties?

“Privacy policy test”
<｜Assistant｜> YES<｜end of sentence｜>

Parameter-efficient fine-tuning (PEFT)

Why PEFT?

Traditional Fine-Tuning

- Updates ALL model parameters
- Requires massive GPU memory
- Slow training and high costs
- Risk of catastrophic forgetting
- Hard to manage multiple versions

Parameter-efficient fine-tuning

- Freezes original model weights
- Adds small trainable modules
- 10x less memory usage
- Faster training iterations
- Easy to swap/combine adapters

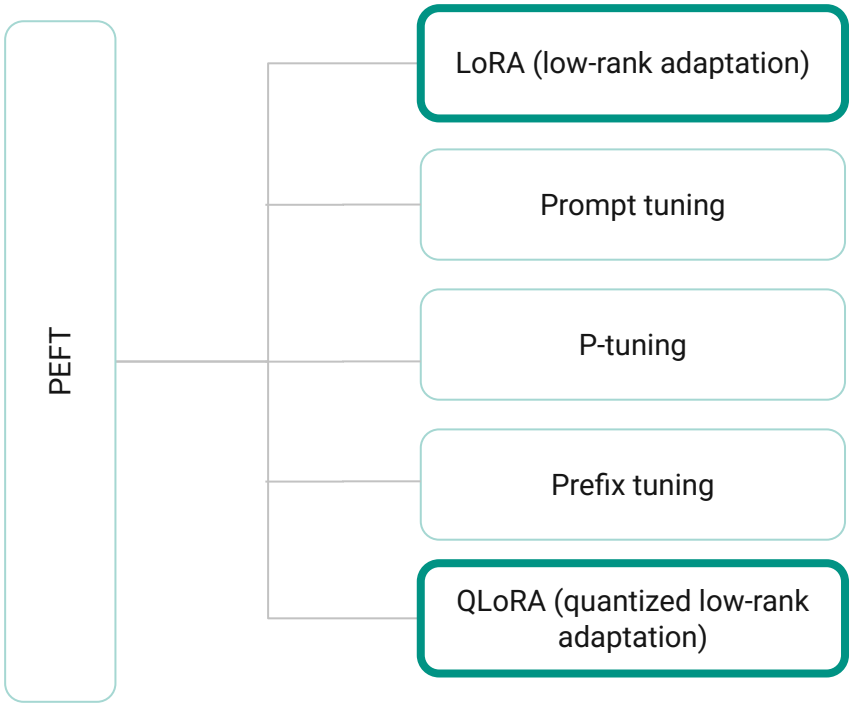
Parameter-efficient fine-tuning (PEFT)

LoRA:

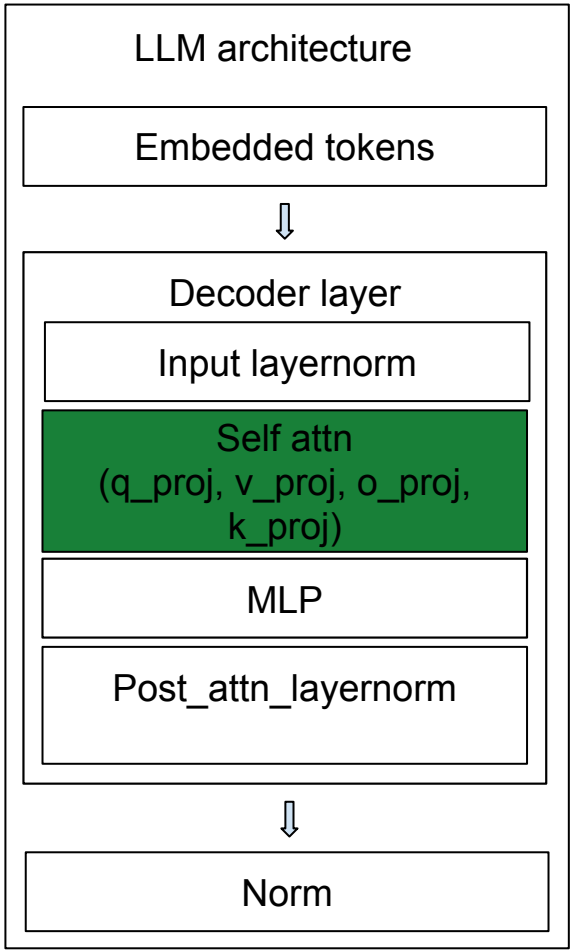
- Keeps original model frozen
- Adds trainable adapters to key layers (q_proj, v_proj, k_proj, o_proj)
- Updates <1% of total parameters

QLoRA:

- LoRA + 4-bit quantization
- Maximum memory efficiency



LLM architecture



Fine-tuning configuration for mixtral8x7B-Instruct

Parameter category	Parameter	Value
QLora (4-bit quantization)	Quantization Target modules Task type	4-bit [q proj, k proj, v proj, o proj] casual_LM
Training arguments	Learning rate Number of training epochs	8×10^{-6} 12

Mixtral-8x7B-Instruct

Label	LR	Qlora	Base model (F1 score)	Fine-tuned model (F1 score)
Cookies and tracking elements	8×10^{-6}	True	0.5283	**0.7013**
Generic personal information	8×10^{-6}	True	0.3537	**0.4066**
IP address and device IDs	8×10^{-6}	True	0.1765	**0.3256**
Financial	8×10^{-6}	True	0.2963	**0.5556**
Computer Information	8×10^{-6}	True	0.3333	0.3226
User online Activities	8×10^{-6}	True	0.3958	0.3529

Fine-tuning configuration for Llama 3 8B-Instruct

Parameter category	Parameter	Value
Lora	Quantization Target modules Task type	16-bit [q proj, k proj, v proj, o proj] casual_LM
Training arguments	Learning rate Number of training epochs	$2\times10^{-}5$ 2

Llama 3 8B-Instruct

Label	LR	Qlora	Few shots + prompt engineering	Base model F1	Fine tuned score
Cookies and tracking elements	$8x10^{-6}$	True	**NO**	**0.56**	0.35
Cookies and tracking elements	$8x10^{-6}$	True	**YES**	**0.56**	0.56
Cookies and tracking elements	$2x10^{-6}$	True	**YES**	**0.56**	0.44
Cookies and tracking elements	$2x10^{-6}$	True	**YES**	**0.56**	0.30
Cookies and tracking elements	$2x10^{-5}$	**False**	**YES**	**0.56**	**0.5769**

Conclusion

Inference

- **Best models:** Gemini and Grok gave the best results on OPP-115.
- **Category differences:** Easy to detect specific types (cookies, financial info, IP addresses), harder for general ones.
- **Prompt matters:** Clear prompts, low temperature, and few examples worked best.
- **Structured answers** (YES/NO) improved results.

Fine-tuning:

- **Mixtral worked really well** - got +0.15 to +0.26 F1 improvements on most privacy categories using QLoRA; however, it didn't work well with Llama, but Lora showed some promise.
- **Fine-tuning was more effective for specific data types:** (cookies, financial info, IP addresses) but struggled with general categories (generic personal information and user activities)
- **Different models need different approaches**: what worked for Mixtral didn't work the same way for LLaMA

Limitation

GPU limit: Scicore has a queuing system, and jobs often took hours to get priority.

Time limitation: A lot of waiting time between different inference and fine-tuning runs.

Data imbalance: Few positive examples in some categories (health, survey data), making them harder for the model to learn.

Thank you!
Any questions?