LLM Training and Prompt Engineering for Privacy Policies

Jeana Han, Arianna Hsu, Nicole Luo, Christine Tao

TABLE OF CONTENTS

01

Introduction

Privacy Policies for Library Vendors

The Shift to Digitization of Library Content

- Libraries increasingly rely on digital content from outside vendors
- These vendors may not follow the same privacy standards as libraries

Concerns with Privacy Policies

- Policies are unclear, inconsistent, and difficult to interpret
- Evaluating policies takes a long time
- Does a privacy policy actually protect users?

Fall 2024: Initial Keyword Approach

Approach: Students worked to use Natural Language Processing (NLP) to determine if policies were good or bad

- NLP: Field of AI to enable computers to understand human language

Idea:

1. Use keywords within policies and identify key categories of privacy policies
2. Score them as “Strong Encryption” OR “May not be secure” within that category

Challenge: Keywords do not provide context

- Example: “We do not share data” vs “We share data with third parties”

Fall 2025: New ML Approach

What is Machine Learning (ML)?

- Computers learn from data to find and recognize patterns to make decisions or predictions
- Considers the *meaning* of words within their context

What does "TRAIN a model" mean?

- We provide a computer program many **examples** so that it learns and recognizes **patterns**. Once trained, the computer can make predictions about *new* data it has not seen before

How to get a dataset of examples?

- No existing dataset of scored privacy policies
- Very tedious and time consuming for humans

Fall 2025: LLM Dataset Creation

What are Large Language Models (LLMs)?

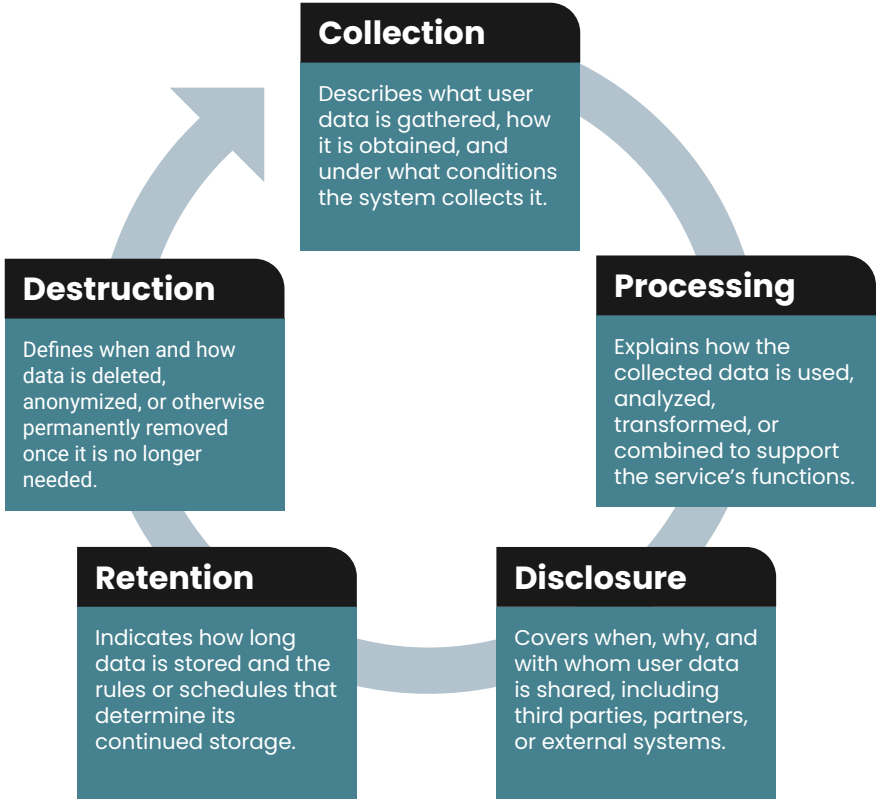
- Computers generate text by learning from sequences of text and recognize patterns to predict which word is most likely to appear *next* in a sequence
  - Ex: ChatGPT, Gemini, Claude, and Copilot

Goal: Create datasets by having LLM categorize and score privacy policies on its own

What do we need to provide an LLM to complete this task?

- Privacy Policies
- A description of categories
- A rubric to score the policies
- A prompt with instructions

Fall 2025: Data Life Cycle Categories



(Breaux 2020)

Fall 2025: Rubric

Goal: Provide a consistent scoring framework that is clear to the LLM and interpretable for the library

Format: For each category, we defined the criteria for each score
(1 = Extremely Bad to 5 = Extremely Good)

Collection

- 5 (Extremely Good)
  - Clearly explains all personal & non-personal data collected and why.
  - Collects only necessary data for explicit, well-defined purposes.
  - Obtains informed and explicit consent before any data collection.
  - Users can easily opt in/out of non-essential data collection.
  - Does not collect or store biometric data.
  - Core features do not require personal data.
  - Shares no unnecessary data with third parties.
- 4 (Good)
  - States what data is collected and why, with specific, legitimate purposes.
  - Requests explicit consent for sensitive data, though some may be implied or bundled.
  - Provides clear privacy documentation and accessible opt-out options.
  - Does not require personal data for most core features.
  - No biometric data collection.
  - Third-party data sharing limited to essential authentication or authorization purposes.
  - Allows users some control over data sharing.
- 3 (Neutral)
  - Provides a general privacy notice but may lack detail or justification for all data types.
  - Business purposes for data use may be vague or overly broad.

Fall 2025: Prompt

After asking the LLM to read the rubric, this is the process we used to categorize and score policies:

Categorize Sentences

Using the data life cycle stages

Score Sentence

According to rubric for that category

Calculate Average

Across all sentences that fall in that category

The result from this prompt was a single numeric score (1-5) for each category for each policy.

Fall 2025: Verifying Accuracy

Method:

- To create a baseline, the librarian experts individually scored each privacy policy according to the same set of instructions and rubric as the LLM
- Tested on both real policies and student generated (good, mediocre, and bad) policies

Result:

- **Inconsistent with itself:** Despite students using the same prompt and rubric on the same LLM, students would receive different scores
- **Inconsistent with expert scores:** Even the librarian experts had different scores between themselves
- **Takeaway:** The rubric and prompt both need refining!

02

Spring 2026: LLM Training

Machine Learning Applications

Recap:

- **What is Machine Learning (ML):** Computers learn from data to find and recognize patterns to make decisions or predictions

- **What does "TRAIN a model" mean?** We provide a computer program many **examples** so that it learns and recognizes **patterns**. Once trained, the computer can make predictions about *new text* it has not seen before

→ Applying this to our project **(SP26)**:

- **Goal: Train the LLM** to score policies
- **Examples:** Human scored policies
  - Professor Ruihao Zhu in the SC Johnson College of Business. Professor Zhu conducted research regarding the usage of LLMs to solve large scale optimization problems.

Librarian Experts Score Policies

GOAL: TRAIN the LLM so that it can SCORE new privacy policies

HOW: Provide EXAMPLES of human scored policies so that it can learn

INFORMATION NEEDED FROM HUMANS:

Within EACH Category:

- A score in each category
- Justification
- Relevant Policy Statements that contribute to scoring decisions

Example of Librarian Scored Policy

Policy Name	Collection Score	Collection Justication	Collection Policy Statements	Processing Score	Processing Justification	Processing Policy Statements
harvard_20240216	45	Business purposes for data use may be vague or overly broad. | Provides a general privacy notice but may lack detail or justification for all data types.	Visitors to this website who have Javascript enabled are tracked using Google Analytics. |Our site uses cookies to track activity.|It does not include any personal information such as names, emails, or credit card numbers.|We may ask for email addresses to help facilitate orders or allow visitors to subscribe to our notification services.	26	Describes processing in generic terms ("to improve services") without specifics. |Provides no mechanism for users to review, contest, or opt out.	This data is primarily used to optimize our website

* For visual purposes, table does not show full content

Example of Librarian Scored Policy

Policy Name	Collection Score	Collection Justication	Collection Policy Statements
harvard_20240216	45	Business purposes for data use may be vague or overly broad. | Provides a general privacy notice but may lack detail or justification for all data types.	Visitors to this website who have Javascript enabled are tracked using Google Analytics. |Our site uses cookies to track activity. |It does not include any personal information such as names, emails, or credit card numbers. |We may ask for email addresses to help facilitate orders or allow visitors to subscribe to our notification services.

Collection

81-100 (Extremely Good)

- States what data is collected and why, with specific purposes.
- Collects only necessary data for explicit, well-defined purposes.
- Obtains informed and explicit consent before any data collection.
- Users can easily opt in/out of non-essential data collection.
- Core features do not require personal data.

61-80 (Good)

- Clearly explains all personal & non-personal data collected and why.
- Requests explicit consent for collection of sensitive data, though some may be implied or bundled.
- Provides accessible opt-out options for data collection.
- Does not require personal data for most core features.
- No biometric data collection.

41-60 (Neutral)

- Provides a general privacy notice but may lack detail or justification for data collection.
- Business purposes for data collection may be vague or overly broad.
- Opt-out options for data collection exist but may be difficult to find.
- Non-core features may require collection of personal data.

21-40 (Bad)

- Collects significant personal data without clear justification.
- Users defaulted into non-essential data collection with no opt-out mechanisms.
- Requires personal data for core functionality.
- Data collection may include some non-essential categories.
- Uses third party web beacons that collect data.
- Collects biometric or sensitive data with consent.

0-20 (Extremely Bad)

- Operates under a “collect everything” approach.
- Provides no transparency about data collected, purposes, or usage.
- Collects biometric or sensitive data without consent.
- Users have no control or ability to opt out of data collection.

The Process: Training the LLM

Step 1

Initialize Prompting

Provide rubric, policy, and instructions telling LLM to score the policy

Step 2

Correct the LLM

Provide the LLM with the "correct answers" → librarian expert scores

Step 3

Analyze, Compare, and Learn

Instruct the LLM to analyze the librarian scores, compare the scores against its own, and learn how the correct scores were determined

Step 4

Score a New Policy

Ask LLM to score a new policy using the knowledge it previously learned

LLM Output

What we ask the LLM to produce:

1) Score in each category
2) Justification
3) Relevant statements in the policy that influence scoring

Results of Training the LLM

UNWTO

	Collection Score	Collection Justification	Collection Relevant Statements	Processing Score	Processing Justification	Processing Relevant Statements	Disclosure Score	Disclosure Justification	Disclosure Relevant Statements	Retention Score	Retention Justification	Retention Relevant Statements	Destruction Score	Destruction Justification	Destruction Relevant Statements
Librarian Results	0	Users have no control or ability to opt out of data collection. |Provides no transparency about data collected, purposes, or usage.		0	Processing practices are concealed or misleads about processing activities. |No user rights to access, correct, or limit processing.		0	Disclaims all responsibility for third-party misuse or onward disclosure. |Users cannot restrict data disclosures.		0	Users unable to determine what data is stored, where, or for how long.		0	No evidence or policy indicating data is ever destroyed. |Users have no visibility into when, how, or whether data destruction occurs.	

	Collection Score	Collection Justification		Processing Score	Processing Justification		Disclosure Score	Disclosure Justification		Retention Score	Retention Justification		Destruction Score	Destruction Justification	
Claude's Results	0	The policy provides zero transparency about what personal data is collected, how it is collected, or for what purposes. No categories of data are named, no collection methods are described, and no consent mechanism of any kind is presented. The rubric's Extremely Bad band explicitly applies when there is "no transparency about data collected, purposes, or usage" and users have "no control or ability to opt out of data collection" — both of which are true here. The lone reference to personal data is a generic pledge to protect it, which does not constitute disclosure about collection practices.	"The UNWTO has implemented technologies and security policies to protect the stored personal data of our users from unauthorized access…" [acknowledges personal data exists but discloses nothing about what is collected or why] "UNWTO employees and processors who have access to personal data are obliged to respect the privacy of our visitors…" [confirms data is accessed internally but provides no collection detail] No statements identifying data types, collection methods, purposes, or consent mechanisms found in the policy.	0	No processing purposes, methods, or limitations are described anywhere in the policy. There is no mention of how personal data is used after collection, no indication of whether data is processed for analytics, marketing, or operations, and no user rights to access, correct, or limit processing. The policy's complete silence on processing practices — combined with the acknowledgment that employees and processors do access personal data — means users have no visibility into what is done with their information, placing this squarely in the Extremely Bad band.	"UNWTO employees and processors who have access to personal data are obliged to respect the privacy of our visitors and the confidentiality of their personal data." [confirms processing occurs but describes no purposes, methods, or limits] No statements about processing purposes, consent, user controls, or secondary uses found in the policy.	0	The policy provides no information whatsoever about whether personal data is shared with third parties, under what circumstances, or to whom. No service providers, partners, or disclosure categories are named. No contractual safeguards for third parties are described, and no user controls over disclosure are offered. The rubric's Extremely Bad band applies when there is "no transparency for how data is disclosed, for what purpose, or to whom" — exactly the case here. The only third-party language in the policy is a disclaimer about external website links, which is not a disclosure statement.	"[T]his website may contain links to other sites. Please be aware that we are not responsible for the privacy practices of such other sites." [link disclaimer only — not a disclosure statement about UNWTO's own sharing practices] "This privacy statement applies solely to information collected by this website." [scope limitation, not a disclosure commitment] No statements about third-party sharing, service providers, disclosure categories, or user controls over disclosure found in the policy.	0	The policy contains no retention language of any kind. No retention periods, schedules, or triggers are stated for any category of personal data. Users are entirely unable to determine what data is stored, where, or for how long. There is no mention of deletion-on-request, no distinction between active-use and archive periods, and no indication that data is ever removed. This is the precise condition the rubric identifies as Extremely Bad: "Users unable to determine what data is stored, where, or for how long."	No retention periods, schedules, deletion triggers, or user rights to request deletion found anywhere in the policy.	0	There is no evidence anywhere in the policy that personal data is ever destroyed. No destruction methods, timelines, or commitments are described. No user rights to request deletion are mentioned, no backup or archive clearing is referenced, and no obligations on third parties to destroy data are present. The security language ("protect… from… accidental destruction") is oriented toward preventing data loss — not ensuring data destruction — and cannot be counted as a destruction commitment. This is a textbook Extremely Bad score of 0.	"[T]echnologies and security policies to protect the stored personal data of our users from unauthorized access, improper use, alteration, unlawful or accidental destruction and accidental loss." [security against data loss — not a destruction commitment] No destruction methods, deletion timelines, backup purging, user deletion rights, or third-party destruction requirements found in the policy.

HARVARD

	Collection Score	Collection Justification	Collection Relevant Statements	Processing Score	Processing Justification	Processing Relevant Statements	Disclosure Score	Disclosure Justification	Disclosure Relevant Statements	Retention Score	Retention Justification	Retention Relevant Statements	Destruction Score	Destruction Justification	Destruction Relevant Statements
Librarian Results	45	Business purposes for data use may be vague or overly broad. | Provides a general privacy notice but may lack detail or justification for all data types.		26	Describes processing in generic terms ("to improve services") without specifics. |Provides no mechanism for users to review, contest, or opt out.		21	May share for clearly defined original business purposes or analytics without user consent. |Discloses user data with external entities for business or marketing gain. |Consent is implied with no clear path to contest data disclosures.		0	Users unable to determine what data is stored, where, or for how long.		0	No evidence or policy indicating data is ever destroyed. |Users have no visibility into when, how, or whether data destruction occurs.	

	Collection Score	Collection Justification	Collection Relevant Statements	Processing Score	Processing Justification	Processing Relevant Statements	Disclosure Score	Disclosure Justification	Disclosure Relevant Statements	Retention Score	Retention Justification	Retention Relevant Statements	Destruction Score	Destruction Justification	Destruction Relevant Statements
Claude's Results	5	Users have no control or ability to opt out of data collection. The policy acknowledges three collection mechanisms — Google Analytics tracking, cookies, and email address collection — but provides almost no specificity about what each captures, why, or under what conditions. No consent mechanism is presented before collection begins, no opt-out is offered for cookie or analytics tracking, and the cookie statement ("our site uses cookies to track activity") provides no detail about what activity is tracked or for what purpose. A marginal score above 0 reflects that the policy at least names the collection mechanisms and confirms that Google Analytics data excludes certain personal identifiers, which is a minimal degree of transparency absent entirely in the UNWTO policy.	"Visitors to this website who have Javascript enabled are tracked using Google Analytics." "It does not include any personal information such as names, emails, or credit card numbers." [partial transparency — states what is excluded but not what is included] "Our site uses cookies to track activity." [no detail on what activity, what data, or for what purpose] "We may ask for email addresses to help facilitate orders or allow visitors to subscribe to our notification services."	5	Processing practices are essentially concealed. The only stated purpose — "primarily used to optimize our website" — is vague and admits secondary uses without identifying them ("primarily" implies other uses exist). Email addresses are said to facilitate orders and subscriptions, but no processing limitations, user rights, consent mechanisms, or controls are described for any data type. No user rights to access, correct, or limit processing are mentioned. A minimal score above 0 reflects that at least one original purpose is named, however vaguely.	"This data is primarily used to optimize our website." [vague, and "primarily" implies undisclosed secondary uses] "We may ask for email addresses to help facilitate orders or allow visitors to subscribe to our notification services." [purpose stated for email only; no controls or rights described] No user rights to access, correct, or limit processing; no consent mechanisms; no secondary use disclosures found in the policy.	10	The policy explicitly states email addresses are never sold, traded, or shared with third parties — a meaningful positive commitment. However, the policy simultaneously discloses user tracking data to Google Analytics, a third-party advertising and analytics platform, with no discussion of what Google does with that data, no contractual safeguards described, and no opt-out offered. This represents an active, undisclosed third-party data relationship that contradicts the spirit of the non-sharing pledge. The non-sale of email addresses lifts this marginally above 0, but the undisclosed Google Analytics data flow and complete absence of any other disclosure framework keeps this firmly in the Extremely Bad band.	"Email addresses are never sold, traded, or shared with third parties in any way." [strong positive — but applies only to email] "Visitors to this website who have Javascript enabled are tracked using Google Analytics." [active third-party disclosure with no safeguards, no opt-out, no contractual obligations described] "Our site uses cookies to track activity." [no disclosure of whether cookie data is shared with third parties]	0	The policy contains no retention language of any kind. No periods, schedules, or deletion triggers are stated for any data type — not for email addresses, not for analytics data, not for cookie data. Users are entirely unable to determine what data is stored, where, or for how long.		0	There is no evidence anywhere in the policy that personal data is ever destroyed. No destruction methods, timelines, or commitments are described. No user rights to request deletion are mentioned, no backup or archive clearing is referenced, and no obligations on third parties to destroy data are present. Users have no visibility into whether their data is ever destroyed.	

Results of Training the LLM

→ **QUESTION:** Can the LLM actually learn how policies are scored based on the information it is provided?

→ **EXPLORING THE ANSWER:** Humans (other than librarians) try scoring policies

Students Attempt Scoring Policies

- Students attempted to score 2 policies together using all the information an LLM is provided
    - **Step 1:** Individually determine a score and/or range a policy fits under
    - **Step 2:** Come together and discuss individual scores
    - **Step 3:** Come up with a final range of 10 (e.g. 21-31)

→ Takeaways:

- Students generally had agreements on the range (Extremely Good, Bad, etc.) but had difficulty determining lower/middle/higher range

Students Scores vs. Librarian Scores

TEXTILE

	Collection Score	Collection Justification	Collection Relevant Statements	Processing Score	Processing Justification	Processing Relevant Statements	Disclosure Score	Disclosure Justification	Disclosure Relevant Statements	Retention Score	Retention Justification	Retention Relevant Statements	Destruction Score	Destruction Justification	Destruction Relevant Statements
Librarian Results	41	Provides a general privacy notice but may lack detail or justification for data collection. | Business purposes for data collection may be vague or overly broad. | Data collection may include some non-essential categories. | Uses third party web beacons that collect data.	We profile web site usage by tracking the textilehive.com pages viewed, the previous and following sites visited, the browser used, and the computer's address. This information is used to analyze trends, track use, and optimize site administration. We use small bits of data called cookies stored on users' computers to simulate a continuous connection.	26	Describes processing in generic terms ("to improve services") without specifics. | Combines data for multiple purposes (e.g., marketing + analytics) without consent. | Provides no mechanism for users to review, contest, or opt out.	This information is used to analyze trends, track use, and optimize site administration. We use small bits of data called cookies stored on users' computers to simulate a continuous connection. | We employ other companies to perform functions on our behalf. Examples include sending email, analyzing data and providing marketing assistance.	23	Discloses user data with external entities for business or marketing gain. | Consent is implied with no clear path to contest data disclosures. | Disclaims all responsibility for third-party misuse or onward disclosure. | Users cannot restrict data disclosures. |	We employ other companies to perform functions on our behalf. Examples include sending email, analyzing data and providing marketing assistance. They have access to personal information needed to perform their functions, but may not use it for other purposes. | Our Website provides users the opportunity to interact and post content through various social media sites including, but not limited to, Twitter, Facebook, Google+, linkedin, Instagram, Tumblr, Linkedin. Your interactions with such tools, and any information you provide in connection with these interactions are publicly viewable outside of the Website domain. Each social media site has its own privacy policies which may be more or less protective than this Policy.	0	Users unable to determine wha t data is stored, where, or for how long.		0	No evidence or policy indicating data is ever destroyed. |User s have no visibility into when, how, or whether data destruction occurs.	

	Collection Score	Collection Justification	Collection Relevant Statements	Processing Score	Processing Justification	Processing Relevant Statements	Disclosure Score	Disclosure Justification	Disclosure Relevant Statements	Retention Score	Retention Justification	Retention Relevant Statements	Destruction Score	Destruction Justification	Destruction Relevant Statements
Student Results	50-60	No-opt out options to stop tracking website usage. Mentions use of third parties that use data. Broad description of collection. Doesn't require data collection to access features.	We profile web site usage by tracking the textilehive.com pages viewed, the previous and following sites visited, the browser used, and the computer's address. This information is used to analyze trends, track use, and optimize site administration. We use small bits of data called cookies stored on users' computers to simulate a continuous connection. Disabling cookies does not prevent you from accessing our website though some features of the website may be degraded.	21-31	Users do not have rights to access or limit processing.	textilehive.com is the sole owner of the information collected on this site. We also communicate with users in who sign up for email updates and provide updates on new functionality or content introduced on textilehive.com.	30-40	Trusted third parties have access to user information. Users cannot restrict data discourse. Lists some examples data can be used for.	We employ other companies to perform functions on our behalf. Examples include sending email, analyzing data and providing marketing assistance. They have access to personal information needed to perform their functions, but may not use it for other purposes.	0-10	Users unable to determine what data is stored, where, or for how long.	No applicable statement.	0-10	No evidence or policy indicating data is ever destroyed. Users have no visibility into when, how, or whether data destruction occurs.	No applicable statement.

→ After attempting to score policies, we realized that we need to change the prompt/instructions and rubric to make it more systematic

03

Prompt Engineering

Prompt Engineering Overview

Prompt engineering refers to the practice of refining and optimizing input queries to large language models (LLMs).

Instead of casually conversing with an LLM to produce output, prompt engineering creates a systematic approach in order to produce consistent and reliable outputs.

Our project
Over the course of the semester, we explored different prompting methods in the effort to create a workflow that could reliably score privacy policies in agreement with librarians.

Prompt Engineering Relevance

The librarians we worked with indicated that the current scoring process has a lot of variance. Hopefully, automating the process with an LLM would result in more consistent scoring.

Other applications of prompt engineering

- Privacy policy scoring in other fields
- General document summarization and analysis
- Software development

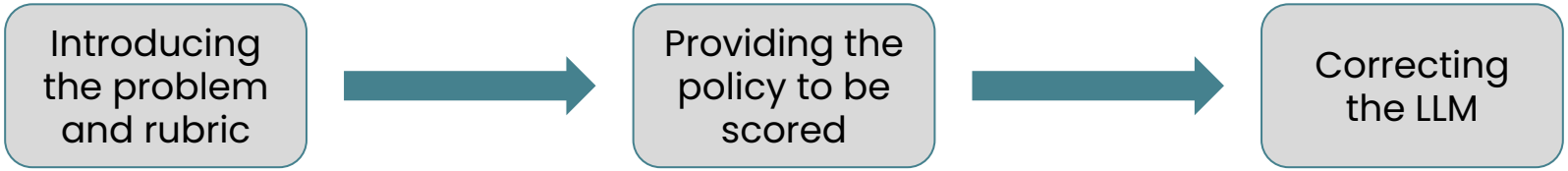
Our Approach

Our primary challenge this semester was generating reliable and consistent policy scores.

Topics Explored

- Correction and training
  - After each attempt by the LLM, we tried providing correct scores and reasonings in an effort to help the LLM learn
- Ranges and Intervals
  - Instead of asking for a singular integer score, we allowed the LLM to provide a range to offer more flexible scoring

Prompting Workflow



Prompting Workflow (Introduction)

MESSAGE #1 — Initial Scoring

[Insert Full Rubric Here]

You are evaluating privacy policies using a detailed rubric that assesses five categories:

1) Collection
2) Processing
3) Disclosure
4) Retention
5) Destruction

Task:
Read the provided privacy policy. Evaluate each of the five categories using the rubric. Note that the descriptions in the rubric that are noted with a 'Y' are statements that may have more influence in the scoring of categories than others.

Prompting Workflow (Correction)

MESSAGE #2: Comparison with Human (Librarian) Scores

[INSERT HUMAN SCORES + JUSTIFICATIONS + STATEMENTS]

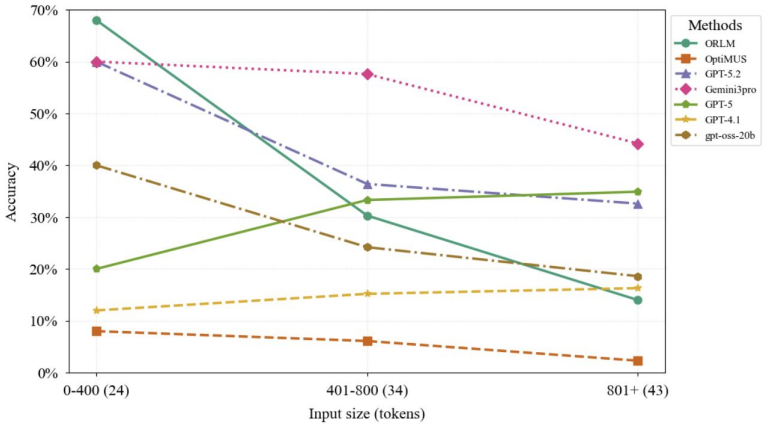
You will now be given human-generated (librarian) score ranges, justifications, and supporting statements for the same policy.

For each category, compare the correct score against your score. If the two scores in each category do not fall within the same range as defined in the rubric, provide an explanation as to why the correct score aligns with the rubric more.

Policy Length and Accuracy

Across the board, we have found that increased input size results in significantly reduced accuracy.

Professor Ruihao Zhu's Research



1 **Modeling accuracy of existing methods drops substantially as input size grows**

Student Prompting Results

Category	Unwto (178 words)	Teton (87 words)	Harvard (105 words)	Umich (1047 words)	Uchicago (1031 words)
Collection					
Processing					
Disclosure					
Retention					
Destruction					

Segmenting Prompting

Since the LLMs struggled with longer text files, we wanted to figure how we could reduce workload when scoring lengthy policies. One approach was asking to score different categories in different prompts. Originally we had the LLM score all 5 categories in one prompt.

Before and after comparison when using segmented prompting on Uchicagopress.txt

Category	LLM Score	Expert Score	Previous LLM Score
Collection	45	65-75	85
Processing	50	38-48	55
Disclosure	35	47-57	35
Retention	20	20-30	15
Destruction	15	0-10	30

Intermediary “Thinking” Step

We noticed that the LLMs would often skim through the policies and skip key lines to score the policy faster. To mitigate this, we tried adding a prompt asking the LLM to make observations about the policy without converging to a score. The LLM was then asked to provide scores after.

Before and after comparison when using the intermediary thinking step (harvard.txt).

LLM Score	Librarian Score	LLM Score (Previous)
30-40	42-52	65
25-35	21-31	60
15-25	16-26	50
0-10	0-10	5
0-10	0-10	5

Editing the Rubric

Towards the end of the semester, we also began experimenting with different rubrics. It seemed as if the LLM was having a hard time applying the original rubric due to subjectivity.

We used a rubric that had more objective scoring guidelines. Overall, we saw improved results across 7 different policies.

	uchicago		umich		unwto		harvard		teton		jhu		textile	
	Week 6-7	Week 11	Week 6-7	Week 11	Week 6-7	Week 11	Week 6-7	Week 11	Week 6-7	Week 11	Week 6-7	Week 11	Week 6-7	Week 11
Collection														
Processing														
Disclosure														
Retention														
Destruction														

04

Wrapping Up

Challenges

Establishing consistent rubrics and prompts —hard to measure LLM consistency and improvement

Standardized output formatting — LLM responses, rubric criteria, and scores weren't always structured the same way, making comparisons difficult

Ambiguity in longer policies — the more complex a policy, the harder it was to get stable, reliable scores from the model

Next steps

- Convert qualitative rubric criteria into binary yes/no checkpoints to reduce subjectivity

- Study how performance degrades as policies get longer or more complex to identify where the process breaks down

- Explore having the model explain its reasoning before assigning a score. Early results suggest this leads to more consistent and justifiable outputs

05

Takeaways

Why this matters

The goal is not to replace librarian judgment, but to make it more powerful, giving libraries a practical tool tool to leverage.

The framework developed here could extend beyond privacy policies to help Cornell evaluate a wide range of vendor and institutional documents.

Through this process, we gained firsthand experience with the challenges of applied AI, fom prompt engineering to evaluating model reliability. These skills extend well beyond this project!

What the Students the Learned

QUESTIONS?

Recap:

- **Introduction**
  - FA25: The start of ML training
- **Training the LLM**
  - Librarians and students score policies
- **Prompt Engineering**
  - Editing our prompt/instructions and rubric
- **Wrapping Up**
  - Challenges and Next Steps