SP26 ORIE 4999:
LLM Training and Prompt Engineering for LibraryPrivacy Policies

Jeana Han (jjh358), Arianna Hsu (ach282), Nicole Luo (nl545), Christine Tao (ctt44)

Table of Contents

1. Introduction

As information shifts online, the Cornell Library has been acquiring digital content through different vendors. Since each vendor has different data protection policies, evaluating the privacy policies has become a time-consuming and tedious challenge for library staff. These policies can be lengthy, inconsistent in structure, and difficult to interpret.

This project, carried out across three semesters (FA24, FA25, SP26), explores how machine learning (ML) and large language models (LLMs) can be used to automate the scoring of privacy policies in a consistent and reliable way that aligns with librarian/human judgement.

This report focuses specifically on the Spring 26 (SP26) semester, during which the team explored LLM training and worked to develop a systematic approach to prompt engineering.

2. Background and Prior Work

2.1 Fall 2024: Keyword-Based NLP

In Fall 2024, a team of students used Natural Language Processing (NLP) to classify privacy policies as “good” or “bad” based on the presence of keywords. The idea was to identify key privacy categories within each policy and flag words that indicate a strong or weak policy. The issue with this approach was that since keywords lack contextual meaning, it led to incorrect identification. For example, the phrase “We do not share data” and “ We share data with third parties” both contain the word “share” but have completely opposite meanings. This approach did not take into account the context of keywords, showing the need for a new approach.

2.2 Fall 2025: LLM-Based Dataset Creation

Recognizing the limitations of keywords and NLP, the Fall 2025 team pivoted to focus on a machine learning (ML) approach that would better capture the semantic meaning of privacy policies. The end goal of this approach was to train a model to categorize and score privacy policies based on a structured rubric covering five data lifecycle categories:

1) Collection: What user data is gathered, how it is obtained, and under what conditions
2) Processing: How the collected data is used, analyzed, or transformed
3) Disclosure: When, why, and with whom user data is shared
4) Retention: How long data is stored and the rules governing continued storage
5) Destruction: When and how data is deleted or permanently removed

After many iterations, a rubric was developed to provide a scoring framework on a 0-100 scale, with criteria defined at each level (e.g. “Extremely Good”, “Bad”, etc.).

In order to train a model, datasets were needed. Since human-created data would be tedious and time-consuming, the team decided to use prompt engineering and have LLMs create datasets. LLMs (ChatGPT, Claude, Gemini. Copilot) were asked to read a privacy policy, categorize its sentences into the five data lifecycle categories, and produce an average score in each category.

Two issues and challenges were identified with this approach. First, even when multiple students used the same prompt on the same LLM, they received different scores for the same policy. Second, when librarians were similarly asked to score privacy policies against the same rubric, they arrived at different scores from each other. These findings revealed the need to refine the rubric and prompting strategy to reduce ambiguity on how policies should be scored before further progress could be made.

3. Spring 2026: LLM Training

3.1 Approach

Using the documents, information, and experience gathered from Fall 2025, the goal for Spring 2026 shifted. Rather than focusing on dataset creation to train an ML model, the goal was to directly train an LLM to score policies by providing it with human-scored examples and asking it to learn from them. This mirrors the supervised learning idea used in machine learning, where a model is shown labeled examples so that it can recognize patterns and generalize to new inputs.

To generate those examples, Cornell librarians scored a set of privacy policies using the same rubric and instructions provided to the LLM. For each policy and category, the librarians provided a numeric score, a written justification, and the specific policy statements that influenced their scoring decision. These human-generated scores served as the "ground truth" the LLM would be trained against.

3.2 Training Process

The training workflow consisted of four iterative steps:

1) Initialize Prompting: Provide the LLM with the rubric, the policy to be scored, and instructions to score each of the five categories with justifications and relevant statements.
2) Correct the LLM: After the LLM produces its initial scores, provide the corresponding librarian scores, justifications, and supporting statements.

3) Analyze, Compare, and Learn: Instruct the LLM to compare its scores against the correct answers, identify where it went wrong, and explain why the librarian's scoring better aligns with the rubric.
4) Score a New Policy: Ask the LLM to apply what it has learned by scoring a previously unseen policy, then repeat the cycle.

The LLM was expected to produce three outputs for each category: a score within the defined range, a brief justification for that score, and the relevant policy statements that drove its decision.

3.3 Student Scoring Exercise

To better understand the difficulty of the scoring task, and to validate whether the rubric and instructions were clear enough for a non-expert to apply, the student team attempted to score two policies themselves using the same materials given to the LLM. The exercise confirmed that broad-level agreement (e.g., identifying a policy as "Neutral" overall) was achievable, but precise score placement within a range was difficult without clearer guidelines. This mirrored the LLM's own struggles and reinforced the team's hypothesis that the rubric needed to be made more systematic and objective.

3.4 Prompt Engineering

Rather than interacting with the LLM conversationally, the team developed a structured prompting workflow intended to minimize variability and align LLM outputs with librarian scoring standards. Throughout the semester, the team explored several different prompting strategies, each designed to address a specific source of inconsistency identified in testing.

3.5 Correction and Training Prompts

The baseline prompting workflow involved three sequential messages. The first message introduced the rubric and asked the LLM to score a policy across all five categories. The second message provided the human-generated scores and asked the LLM to compare them against its own, explaining any discrepancies. The third message asked the LLM to score a new policy using what it had learned.

Rather than asking for a single integer, the team shifted to allowing the LLM to provide a score range (e.g., 21–31), which better reflected the inherent ambiguity in the rubric and reduced the impact of small misinterpretations.

3.6 Policy Length, Accuracy, and Segmented Prompting

A consistent finding across the semester was that LLM accuracy degrades as policy length increases. Shorter policies tended to produce more accurate scores across all five categories. Longer, more complex policies led to significantly more errors. This finding aligns with broader research on LLM performance, where increased input size correlates with reduced modeling accuracy. For example, the team tried a segmented approach where each category was scored in a separate prompt to reduce length. When scoring the UChicago policy using this segmented prompting, the Collection score shifted from 85 to 45, moving it closer to the expert range of 65–75.

3.7 Intermediary Thinking Step

Another improvement came from inserting an intermediary reasoning step before score assignment. Rather than asking the LLM to immediately produce a score, the team first prompted it to make observations about the policy without committing to any number. Once it made observations, the LLM was asked to score the policy. This approach produced better results, particularly on shorter policies. For example, when tested on harvard.txt, this approach produced noticeably better alignment with librarian scores: the Collection score shifted from 65 down to 30–40, compared to the librarian range of 42–52, and the Processing score moved from 60 to 25–35, against a librarian range of 21–31. The intuition is that the model was previously skimming policies and converging too quickly on a score, and the forced observation step encouraged deeper reading. Results were strongest on shorter policies, though the approach showed promise across the board.

3.8 Rubric Revision

Near the end of the semester, the team also began experimenting with a revised rubric that replaced subjective descriptive criteria with more objective, checkable guidelines. The original rubric required the model (and human scorers) to make judgment calls when scoring policies. The revised rubric moved toward binary-style checkpoints that could be applied more consistently. When tested across seven policies, the revised rubric produced more accurate scores compared to the original, with particular improvement in the Collection and Processing categories.

4. Challenges and Next Steps

Throughout this project, our team encountered several core challenges that shaped both our process and our thinking. Without a stable rubric baseline, it was difficult to determine whether the LLM was genuinely improving or simply producing different results across runs. Inconsistent output formatting made comparisons between LLM and librarian scores harder than expected, and longer, more ambiguous policies consistently pushed the model toward less reliable outputs.

Moving forward, the most promising directions include converting qualitative rubric criteria into binary checkpoints to reduce subjectivity, studying how model performance degrades as policy length and complexity increase, and prompting the model to reason through its scoring before committing to a number, an approach that showed early signs of improving consistency.

Ultimately, the goal of this work is not to replace librarian expertise, but to scale it. A reliable LLM scoring tool could give Cornell a reusable framework for evaluating not just vendor privacy policies, but a wide range of institutional documents, from accessibility compliance to data governance agreements. For us as students, this project was also a hands-on introduction to the real challenges of applied AI, from prompt engineering to model evaluation, and those lessons will carry forward well beyond this class.