**Conversation Analysis Technical Report**

**Application Features:**

1. Profanity Detection Web Application
2. Privacy and Compliance Violation Web Application
3. Upload JSON file(s) to analyze it in real-time
4. Silence and overtalking analysis

**Implementation Recommendation:**

**Question 1: Profanity Detection**

**Scenario 1: Regex**

A pure regex-based profanity filter is a straightforward approach to detecting offensive language. It relies on a predefined list of profane words and patterns, which can be tailored to suit cultural, regional, or domain-specific needs. This method offers flexibility, as organizations can customize the list to include terms relevant to their context.

However, maintaining the profanity list is both a strength and a limitation. While it enables control over what is flagged, it also requires continuous manual updates to remain effective, especially as language evolves. Regex patterns may miss variations in spelling, slang, or obfuscated terms and cannot interpret context, tone, or intent.

Furthermore, regex-based detection cannot account for nuances such as sarcasm, irony, or sentiment. These limitations may lead to false positives or missed detections, impacting the accuracy of the system. As such, while regex offers a simple and cost-effective starting point, it may not be sufficient for more complex or large-scale applications where contextual understanding is important.

**Scenario 2: LLM**

An LLM (Large Language Model)-based profanity filter uses advanced AI and natural language processing to understand context, tone, and intent, offering more accurate and nuanced detection of offensive language compared to regex. It can identify variations, slang, and obfuscated terms, and distinguish between offensive and non-offensive uses of words based on context.

The main advantage of LLMs is their ability to detect profanities in a dynamic and evolving linguistic environment. They can learn and adapt to new slang and trends, reducing the need for manual updates. Additionally, they can interpret nuances like sarcasm and irony, minimizing false positives.

However, LLMs require more computational resources than regex systems, leading to higher costs and potentially slower performance. While they are good at understanding context, they may still struggle with highly subtle or complex language, which can result in false negatives, where offensive language is missed.

LLM-based filter is more accurate and adaptable but comes with higher computational demands and complexity. It’s ideal for large-scale applications where context and evolving language are critical, but may not be necessary for simpler use cases.

**Question 2: Privacy and Compliance Violation**

**Scenario 1: Regex**

Regex is a simple and effective tool for detecting specific privacy and compliance violations, as many predefined patterns are readily available to identify sensitive data like credit card numbers, social security numbers, or email addresses. Its main strength lies in its speed and cost-effectiveness, especially when violations are well-defined.

However, the challenge arises in understanding context. For instance, distinguishing between a date of birth and a regular date requires contextual awareness, which regex cannot provide. Additionally, if sensitive information like account numbers or balances is written in words rather than numbers, regex may fail to detect them. While regex works well for known, specific violations, it becomes less reliable on a larger scale where nuances and evolving data formats are involved. Thus, regex is ideal for targeted detection but lacks the flexibility and adaptability needed for more complex or dynamic compliance environments.

**Scenario 2: LLM**

An LLM-based approach to detecting privacy and compliance violations leverages advanced natural language processing to understand the context in which sensitive information appears. Unlike regex, an LLM can interpret nuances, such as identifying whether a date refers to a date of birth or a general event. It can also recognize when sensitive data like account numbers or balances are written in words, which regex might miss.

The strength of an LLM lies in its ability to analyze data more holistically, understanding complex patterns and evolving formats. It can adapt to new types of violations as language changes, reducing the need for constant manual updates. Furthermore, an LLM can assess the context of the data, minimizing false positives and detecting violations that are more obfuscated or indirect.

LLMs require significant computational resources and are more complex to implement and maintain. While they offer greater accuracy and flexibility, they are best suited for large-scale, dynamic environments where context and adaptation are crucial.

**Question 3: Silence and Overtalk Distribution**

1. **Overtalk (Simultaneous Speaking) Percentage per Call:**

Overtalk refers to instances when both parties speak at the same time during a call. The overtalk percentage is calculated by measuring the duration of simultaneous speech in relation to the total call time.

* 1. **Significance:**
     1. Monitoring overtalk is critical in ensuring clear, effective communication during calls.
     2. A high overtalk percentage may signal issues with call flow, such as participants talking over each other, which could hinder the effectiveness of the conversation.
     3. Reducing overtalk can improve call clarity, enhance customer experience, and lead to better outcomes in support or sales calls.
  2. **Observation:**

**A graph with blue bars and white text

AI-generated content may be incorrect.**

|  |  |  |
| --- | --- | --- |
| **Based on 251 Conversations** | **Value** | **Comment** |
| **Average Silence** | 4.36 % | Good |
| **Average Silence Duration** | 0.68 s | Good |

1. **Silence Percentage per Call:**

Silence refers to the periods when no one is speaking on the call. It can be measured by calculating the duration of silence against the total call time

* 1. **Significance:**
     1. Analyzing silence is important because prolonged periods of silence can lead to disengagement or confusion.
     2. It may suggest that one party is unsure of how to respond, is unprepared, or that the conversation is not moving forward.
     3. Monitoring and reducing unnecessary silence can improve call productivity, maintain engagement, and enhance overall communication quality.
  2. **Observation:**

**A graph with orange bars

AI-generated content may be incorrect.**

|  |  |  |
| --- | --- | --- |
| **Based on 251 Conversations** | **Value** | **Comment** |
| **Average Overtalk** | 10.33% | The conversation duration is low, maybe that is why it is high. Need to check the same metric for longer duration calls. |
| **Average Overtalk Duration** | 0.95 s | Good |