Resumen Ejecutivo del Proyecto de Obtención de Grado

PROTECCIÓN Y PRIVACIDAD DE DATOS DELICADOS EN CONSULTAS A LLMs: PRESERVACIÓN DE PRIVACIDAD

Dr. Luis Miguel Escobar Vega, luis.escobar@iteso.mx

1) Descripción general del proyecto propuesto

El Privacy-Preserving Prompt Tuning [1] es una técnica que permite adaptar los LLMs (Large Language Models) como chatGPT[2], Palm2[3] o Llama[4] a nuevos escenarios de servicio usando datos propios del usuario. No obstante, el uso de datos privados conlleva riesgos asociados. Por ello, es imperativo garantizar su protección, asegurando que estos datos se utilicen únicamente de manera local y no se expongan globalmente. Una de las complejidades radica en que el entrenamiento de los LLMs se realiza directamente con estos datos privados para optimizar su rendimiento. Por ello, hemos visto la necesidad de desarrollar un método que extraiga la esencia de dichos datos sin divulgar información específica, permitiendo que los LLMs operen de manera eficaz y al mismo tiempo respeten la privacidad. Adicionalmente, se publican documentos que podrían contener información sensible o identificable (como registros judiciales, comunicaciones internas e informes periodísticos). Estos se comparten, en muchos casos, por razones de transparencia. La tarea de anonimizar esta información suele ser manual, lo que requiere una inversión de tiempo considerable. La anonimización de información sensible en LLMs es crucial para proteger la privacidad del usuario, cumplir con regulaciones, minimizar riesgos legales y financieros, actuar con ética y mantener la confianza de los usuarios. Es tanto una exigencia legal como una responsabilidad ética para las entidades que implementan LLMs.

2) Objetivo General

Desarrollar un plugin que asista al usuario en la detección y manejo de datos sensibles, tanto en los prompts como en la definición del contexto.

Algunas preguntas que podría considerar al desarrollar el plugin:

- ¿Qué tipos de datos sensibles se deben detectar?
- ¿Qué técnicas de aprendizaje automático se utilizarán para detectar datos sensibles?
- ¿Cómo se proporcionarán sugerencias al usuario sobre cómo eliminar o anonimizar la información sensible?
- ¿Cómo se registrarán los cambios realizados a los datos?

3) Entregabe esperado

Aplicación o especificación del planteamiento, deseable desarrollar plugin con el que se mida los resultaos obtenidos y se pueda comparar con otros modelos del estado del arte.

4) Vinculación o colaboración

Este proyecto se vincula con el programa de investigación de doctorado con el número de 399053 del Conacyt. Además, refuerza la base teórica del alumno, porque pone en práctica su aplicación pues permitirá a los interesados relacionarse con problemáticas reales en la implementación de tecnologías de inteligencia conversacional.

5) Asignaturas de la MSC relacionadas con el desarrollo del proyecto

Análisis y diseño de algoritmos [MSC2229A],
Aprendizaje automático (Machine learning) [MSC1007A] y
Aprendizaje profundo (Deep learning) [MSC2498A].

6) Participación en el proyecto

Este TOG pretende contar con un alumno de la Maestría en Sistemas Computacionales. Se requiere conocimientos generales de frameworks para procesamiento natural del lenguaje, y herramientas de desarrollo de software como Python y conocimiento básico de al menos alguna plataforma de ML como tensorflow, keras o torch.

Bibliografía relacionada

En este apartado se debe incluir una lista de bibliografía (formato IEEE) que se relacione con el desarrollo del TOG para que el alumno pueda revisarla y ahondar más en el proyecto propuesto. A manera de ejemplo:

[1] *Y. Li, Z. Tan, Y. Liu. Privacy-Preserving Prompt Tuning for Large Language Model Services, https://arxiv.org/abs/2305.06212*

[2] *OpenAI.* (s.f.). OpenAI. Source https://openai.com.

[3] *Bard. Google.* 2023. [En línea]. Source https://bard.google.com.

[4] *Meta AI.* (2022). Code Llama: Large Language Model Coding. Retrieved from https://ai.meta.com/blog/code-llama-large-language-model-coding

[5] Alcaraz, A., & González-Rodríguez, S. (2023, May 10). Privacy-Preserving Prompt Tuning for Large Language Model Services. arXiv preprint arXiv:2305.06212.

[6] OpenAI. (2023, February 14). Introducing ChatGPT Enterprise. Retrieved from https://openai.com/blog/introducing-chatgpt-enterprise>

[7] *I. Goodfellow, Y. Bengio, A. Courville,* (2016). Deep Learning. MIT Press.

[8] S. Raschka, V. Mirjalili, (2017). Python Machine Learning. *Packt Publishing; Edición 2nd.*

[9] *Alcaraz, A.* (2023, July 20). Integrating ontologies with large language models for decision-making. Medium. Retrieved from https://medium.com/@alcarazanthony1/integrating-ontologies-with-large-language-models-for-decision-making-bb1c600ce5a3

CV del proponente

El Dr. Luis Miguel Escobar Vega es profesor en el ITESO y especialista en semántica interpretativa, QASs y modelos de lenguaje. Ha contribuido significativamente a la semántica computacional a través de sus publicaciones, enfocadas en optimizar sistemas de recuperación de datos. Además, ha representado activamente su campo en congresos internacionales, discutiendo los beneficios de combinar sistemas estadísticos con análisis semántico.