

Chatbot Recomendador de Videojuegos

Juan Camilo Niño
1202498

Nicolás Acevedo
1202520

Simón Porras Villalobos
1202603

Resumen—Este trabajo presenta el desarrollo de un chatbot recomendador de videojuegos basado en técnicas de aprendizaje profundo, específicamente utilizando ajuste fino eficiente mediante LoRA. El sistema emplea el modelo TinyLlama-1.1B especializado con información extraída de la plataforma Steam para generar recomendaciones precisas y contextualizadas sobre videojuegos.

I. INTRODUCCIÓN

En los últimos años, los modelos de lenguaje natural basados en IA generativa han transformado la interacción humano-máquina, permitiendo sistemas capaces de producir respuestas coherentes y contextualizadas. Sin embargo, los modelos generativos tradicionales requieren especialización en dominios específicos para ofrecer recomendaciones precisas y fundamentadas en conocimiento relevante del área.

Para abordar la adaptación eficiente de modelos de lenguaje a dominios específicos, el método *Low-Rank Adaptation* (LoRA) ha emergido como un mecanismo efectivo para el ajuste fino de grandes modelos con un número reducido de parámetros entrenables [1]. LoRA permite especializar modelos sin necesidad de recursos computacionales extensivos, siendo ampliamente utilizado en chatbots basados en modelos LLaMA y similares.

En este proyecto se emplea TinyLlama-1.1B, un modelo compacto derivado de la arquitectura LLaMA, diseñado para ofrecer un rendimiento competitivo con bajos requerimientos de hardware. La especialización de TinyLlama mediante ajuste fino con LoRA permite construir un recomendador de videojuegos eficiente, capaz de proporcionar sugerencias informadas basadas en el conocimiento adquirido durante el entrenamiento con datos de Steam.

II. ESTADO DEL ARTE

En los últimos años ha crecido notablemente la investigación y las aplicaciones prácticas que emplean modelos de lenguaje especializados para construir chatbots más fiables y orientados a dominios específicos. A continuación se sintetizan las líneas más relevantes del estado del arte relacionadas con este trabajo.

II-A. Ajuste eficiente de LLMs: LoRA y variantes

Para adaptar modelos de lenguaje a dominios concretos sin entrenar todos los parámetros, las técnicas de *parameter-efficient fine-tuning* (PEFT) como LoRA han ganado amplia aceptación. LoRA inserta adaptaciones de bajo rango en las capas del modelo, reduciendo memoria y tiempo de

entrenamiento al mantener los pesos base congelados; implementaciones como QLoRA combinan cuantización y LoRA para afinar modelos cuantizados con eficiencia significativa [1]. Trabajos posteriores examinan variantes y mejoras en la retención de información durante la cuantización y en la adaptación dinámica del rango para tareas específicas [2].

II-B. Modelos compactos y despliegue práctico

La investigación hacia modelos compactos y optimizados (p. ej. familias tipo TinyLlama y versiones reducidas de LLaMA) facilita despliegues en entornos con recursos limitados sin sacrificar capacidades básicas de comprensión y generación. Estos modelos, combinados con técnicas de PEFT, ofrecen una alternativa viable para prototipos y demos interactivas, permitiendo la integración en plataformas como Hugging Face Spaces y frontends ligeros como Gradio.

II-C. Especialización mediante datos de dominio

Los sistemas de recomendación basados en modelos de lenguaje se benefician significativamente de la especialización con datos del dominio objetivo. La calidad de los datos de entrenamiento, incluyendo descripciones de productos, reseñas de usuarios y metadatos estructurados, afecta directamente la capacidad del modelo para generar recomendaciones relevantes y precisas. Estudios empíricos emplean métricas tanto automáticas (BLEU, ROUGE, METEOR) como evaluaciones humanas centradas en relevancia, utilidad y fluidez para validar chatbots especializados.

II-D. Generación de datasets sintéticos

La generación de datasets sintéticos en formato pregunta-respuesta es una práctica común para especializar modelos en dominios concretos y mejorar la cobertura de consultas esperadas. Esta técnica permite ampliar el conjunto de entrenamiento con ejemplos representativos del tipo de interacciones que el chatbot enfrentará en producción.

II-E. Retos abiertos

A pesar de los avances, persisten desafíos relevantes: manejo de sesgos y toxicidad heredados en los datos de entrenamiento; balanceo entre generalización y especialización para evitar respuestas demasiado genéricas o excesivamente específicas; y el desarrollo de métricas estándar que correlacionen mejor con la utilidad percibida por usuarios en dominios específicos. Las líneas recientes de investigación proponen mecanismos de validación adicional y adaptación continua del modelo para mitigar estos problemas [1], [2].

II-E0a. Conclusión del estado del arte.: El ajuste fino eficiente mediante LoRA sobre modelos compactos constituye una práctica prometedora para construir chatbots de dominio especializados, permitiendo adaptación eficiente y despliegues prácticos en entornos con recursos limitados, siempre que se acompañe de buenas prácticas en curación de datos, evaluación humana y monitoreo post-despliegue [1], [2].

III. JUSTIFICACIÓN DEL PROYECTO

El desarrollo de un chatbot recomendador de videojuegos se sustenta en la amplia disponibilidad de datos provenientes de la plataforma Steam y en la necesidad de proporcionar recomendaciones especializadas y contextualmente relevantes. LoRA ha sido validado como un método altamente eficiente para adaptar modelos a dominios particulares sin requerir alto consumo de memoria o tiempo de entrenamiento [1]. La combinación de LoRA con TinyLlama y datos especializados de Steam provee un marco adecuado para desarrollar un sistema ligero, especializado y útil para usuarios que buscan recomendaciones de videojuegos.

IV. RECURSOS USADOS

- **Hugging Face Spaces:** plataforma que facilita la creación y despliegue rápido de demostraciones de ML en minutos.
- **TinyLlama-1.1B:** modelo de lenguaje generativo basado en la arquitectura LLaMA, optimizado para eficiencia y adecuado para tareas de generación condicionada en hardware limitado.
- **Gradio:** framework de Python de código abierto para construir rápidamente interfaces web interactivas de modelos de ML.
- **Datasets de Steam:** incluye metadatos de juegos y reseñas de usuarios recopiladas de Steam (23K juegos y aproximadamente 31M reseñas) [3], [4].

V. EXPLICACIÓN DEL FUNCIONAMIENTO DEL SISTEMA

El sistema se despliega en Hugging Face Spaces con una interfaz web de Gradio. El modelo TinyLlama-1.1B especializado mediante ajuste fino con LoRA procesa las consultas de los usuarios y genera recomendaciones de videojuegos basándose en el conocimiento adquirido durante el entrenamiento con datos de Steam. El modelo ha sido entrenado con descripciones de juegos, reseñas de usuarios y un conjunto sintético de preguntas y respuestas sobre videojuegos, lo que le permite comprender y responder a consultas sobre recomendaciones de manera informada.

Para garantizar coherencia y variedad en las respuestas, se ajustan parámetros de decodificación (temperatura, top-k, penalización por repetición) en la librería Transformers, evitando respuestas repetitivas o fuera de contexto. El sistema está diseñado para mantener conversaciones naturales mientras proporciona recomendaciones relevantes basadas en las preferencias expresadas por el usuario.

VI. DESARROLLO DEL PROYECTO

El proyecto comenzó con la selección y preparación de datos provenientes de Steam, incluyendo metadatos de juegos y millones de reseñas de usuarios. Estos datos fueron limpiados y estructurados para su uso en el proceso de ajuste fino del modelo.

Inicialmente se empleó DistilGPT2, pero debido a problemas de coherencia en las respuestas, se migró a TinyLlama-1.1B, un modelo más adecuado para este tipo de aplicaciones. Posteriormente se aplicó ajuste fino mediante LoRA para especializar el modelo en el dominio de videojuegos, siguiendo la técnica descrita por Dettmers et al. [1].

Para reforzar la especialización temática del modelo, se generó un conjunto sintético de 50 000 *prompts* en formato pregunta–respuesta sobre videojuegos. Este dataset sintético complementó los datos reales de Steam, proporcionando al modelo una amplia cobertura de tipos de consultas y estilos de interacción esperados.

El proceso de ajuste fino con LoRA permitió que el modelo internalizara conocimiento sobre videojuegos, géneros, mecánicas de juego y preferencias de usuarios, facilitando la generación de recomendaciones contextualizadas. Finalmente, el modelo especializado se desplegó mediante Gradio en Hugging Face Spaces, permitiendo interacción directa e intuitiva con usuarios finales.

VII. RESULTADOS

El sistema desarrollado demostró capacidad para generar recomendaciones de videojuegos coherentes y relevantes. El ajuste fino mediante LoRA permitió especializar efectivamente el modelo en el dominio de videojuegos sin requerir recursos computacionales excesivos. El modelo entrenado mostró comprensión de géneros, mecánicas y características de juegos, proporcionando sugerencias apropiadas según las preferencias expresadas por los usuarios.

La interfaz de Gradio facilitó la interacción con usuarios finales, proporcionando una experiencia de usuario intuitiva y accesible. El despliegue en Hugging Face Spaces permitió que el chatbot estuviera disponible públicamente para demostración y pruebas con usuarios reales.

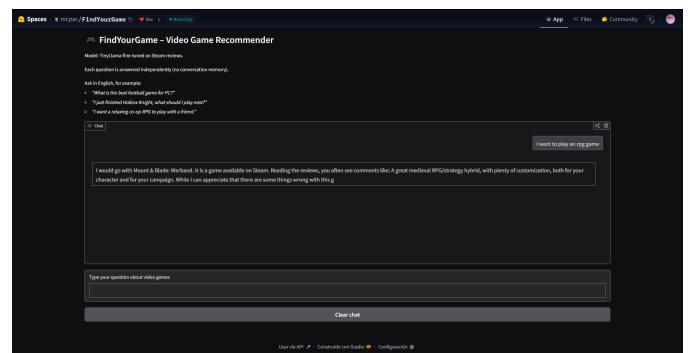


Figura 1. Interfaz del chatbot FindYourGame desplegado en Hugging Face Spaces.

VIII. CONCLUSIONES

Los desafíos principales incluyeron garantizar que el chatbot mantuviera naturalidad en el diálogo y generara recomendaciones precisas basadas en el conocimiento adquirido durante el entrenamiento. El ajuste fino mediante LoRA demostró ser una técnica efectiva para especializar el modelo compacto TinyLlama en el dominio de videojuegos.

Como logros del proyecto cabe destacar la implementación exitosa de un pipeline de ajuste fino eficiente usando LoRA, la generación de un dataset sintético especializado de 50 000 prompts, y el despliegue de un prototipo funcional accesible en Hugging Face Spaces para demostración.

En conjunto, el sistema desarrollado muestra la viabilidad de adaptar modelos de lenguaje compactos mediante técnicas de ajuste fino eficiente para crear chatbots especializados en dominios específicos, ofreciendo recomendaciones de videojuegos relevantes con requerimientos computacionales moderados.

REFERENCIAS

- [1] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient finetuning of quantized LLMs,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: <https://arxiv.org/abs/2305.14314>
- [2] H. Qin *et al.*, “Accurate LoRA-finetuning quantization of LLMs via information retention,” *arXiv preprint*, 2024. [Online]. Available: <https://arxiv.org/abs/2402.05445>
- [3] A. M. Van der Merwe, “Steam Reviews Dataset,” Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/andrewmvd/steam-reviews/data>
- [4] FronKongames, “Steam Games Dataset,” Kaggle, 2024. [Online]. Available: <https://www.kaggle.com/datasets/fronkongames/steam-games-dataset>