# Video Game Recommendation Chatbot

Juan Camilo Niño
1202498

Nicolás Acevedo
1202520

Simón Porras Villalobos
1202603

***Abstract*—This work presents the development of a video game recommendation chatbot based on deep learning techniques, specifically using efficient fine-tuning through LoRA. The system employs the TinyLlama-1.1B model specialized with information extracted from the Steam platform to generate accurate and contextualized video game recommendations.**

## I. Introduction

In recent years, natural language models based on generative AI have transformed human–machine interaction, enabling systems capable of producing coherent and contextualized responses. However, traditional generative models require specialization in specific domains to provide accurate recommendations grounded in relevant domain knowledge.

To address the efficient adaptation of language models to specific domains, the *Low-Rank Adaptation* (LoRA) method has emerged as an effective mechanism for fine-tuning large models with a reduced number of trainable parameters [1]. LoRA enables model specialization without the need for extensive computational resources and is widely used in chatbots based on LLaMA models and similar architectures.

This project employs TinyLlama-1.1B, a compact model derived from the LLaMA architecture, designed to offer competitive performance with low hardware requirements. Fine-tuning TinyLlama with LoRA enables the construction of an efficient video game recommender capable of providing informed suggestions based on knowledge acquired through training on Steam data.

## II. State of the Art

In recent years, research and practical applications involving specialized language models for building more reliable, domain-oriented chatbots have grown substantially. The most relevant areas of the state of the art related to this work are summarized below.

### A. Efficient LLM fine-tuning: LoRA and variants

To adapt language models to specific domains without training all parameters, parameter-efficient fine-tuning (PEFT) techniques such as LoRA have gained widespread acceptance. LoRA inserts low-rank adapters into model layers, reducing memory usage and training time by keeping base weights frozen; implementations such as QLoRA combine quantization and LoRA to fine-tune quantized models with significant efficiency [1]. Later works examine variants and improvements in information retention during quantization and dynamic rank adaptation for specific tasks [2].

### B. Compact models and practical deployment

Research on compact and optimized models (e.g., TinyLlama families and reduced LLaMA versions) facilitates deployment in resource-limited environments without sacrificing core comprehension and generation capabilities. These models, combined with PEFT techniques, provide a viable alternative for prototypes and interactive demos, enabling integration into platforms such as Hugging Face Spaces and lightweight frontends like Gradio.

### C. Domain specialization through domain data

Recommendation systems based on language models benefit significantly from specialization using data from the target domain. The quality of training data—including product descriptions, user reviews, and structured metadata—directly influences the model's ability to generate relevant and accurate recommendations. Empirical studies employ both automatic metrics (BLEU, ROUGE, METEOR) and human evaluations focused on relevance, usefulness, and fluency to validate specialized chatbots.

### D. Synthetic dataset generation

The generation of synthetic question–answer datasets is a common practice to specialize models in concrete domains and improve coverage of expected queries. This technique allows expanding the training set with representative examples of the interactions the chatbot will face in production.

### E. Open challenges

Despite significant advances, several challenges remain: managing bias and toxicity inherited from training data; balancing generalization and specialization to avoid overly generic or overly specific responses; and developing standardized metrics that better correlate with user-perceived utility in specific domains. Recent research proposes additional validation mechanisms and continuous model adaptation to mitigate these issues [1], [2].

*a) Conclusion of the state of the art.:* Efficient fine-tuning using LoRA on compact models is a promising practice for building specialized domain chatbots, enabling efficient adaptation and practical deployment in resource-limited environments when accompanied by good practices in data curation, human evaluation, and post-deployment monitoring [1], [2].

## III. Project Justification

The development of a video game recommendation chatbot is supported by the wide availability of data from the Steam platform and the need to provide specialized and contextually relevant recommendations. LoRA has been validated as a highly efficient method for adapting models to specific domains without requiring high memory or long training times [1]. The combination of LoRA with TinyLlama and specialized Steam data provides a suitable framework for building a lightweight, specialized system useful for users seeking video game recommendations.

## IV. Resources Used

- **Hugging Face Spaces:** a platform that enables rapid creation and deployment of ML demos in minutes.
- **TinyLlama-1.1B:** a generative language model based on the LLaMA architecture, optimized for efficiency and suitable for conditioned generation tasks on limited hardware.
- **Gradio:** an open-source Python framework for rapidly building interactive web interfaces for ML models.
- **Steam datasets:** including game metadata and user reviews collected from Steam (23K games and approximately 31M reviews) [3], [4].

## V. System Operation Description

The system is deployed on Hugging Face Spaces with a Gradio web interface. The TinyLlama-1.1B model specialized through LoRA fine-tuning processes user queries and generates video game recommendations based on knowledge acquired during training with Steam data. The model was trained with game descriptions, user reviews, and a synthetic set of question–answer prompts about video games, enabling it to understand and respond to recommendation queries meaningfully.

To ensure coherence and variety in responses, decoding parameters (temperature, top-k, repetition penalty) are adjusted in the Transformers library, avoiding repetitive or out-of-context responses. The system is designed to maintain natural conversations while providing relevant recommendations based on user preferences.

## VI. Project Development

The project began with the selection and preparation of data from Steam, including game metadata and millions of user reviews. These data were cleaned and structured for use in the model fine-tuning process.

Initially, DistilGPT2 was used, but due to coherence issues in responses, the project migrated to TinyLlama-1.1B, a more suitable model for this type of application. LoRA fine-tuning was then applied to specialize the model in the video game domain, following the technique described by Dettmers et al. [1].

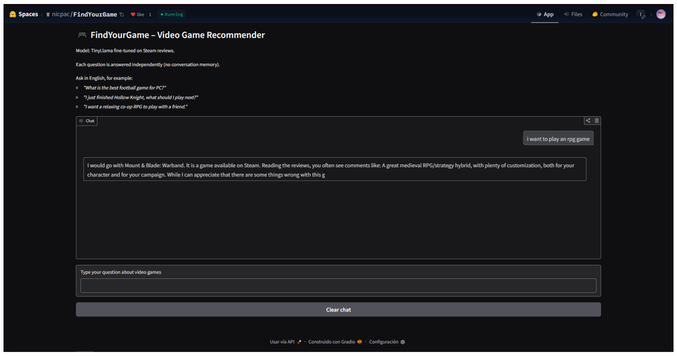
To reinforce thematic specialization, a synthetic dataset of 50,000 question–answer prompts about video games was generated. This synthetic dataset complemented real Steam data, providing the model with broad coverage of expected query types and interaction styles.

The LoRA fine-tuning process enabled the model to internalize knowledge about video games, genres, gameplay mechanics, and user preferences, supporting the generation of contextualized recommendations. Finally, the specialized model was deployed using Gradio on Hugging Face Spaces, allowing direct and intuitive interaction with end users.

## VII. Results

The developed system demonstrated the ability to generate coherent and relevant video game recommendations. Fine-tuning through LoRA effectively specialized the model in the video game domain without requiring excessive computational resources. The trained model showed understanding of genres, mechanics, and game characteristics, providing appropriate suggestions based on user preferences.

The Gradio interface facilitated user interaction, offering an intuitive and accessible experience. Deployment on Hugging Face Spaces enabled public availability of the chatbot for demonstration and real-user testing.



FindYourGame – Video Game Recommender

Model: TinyLlama fine-tuned on Steam reviews.

Each question is answered independently (no conversation memory).

Ask in English, for example:

- *"What is the best football game for PC?"*
- *"I just finished Hollow Knight, what should I play next?"*
- *"I want a relaxing co-op RPG to play with a friend."*

i want to play an rpg game

I would go with Mount & Blade: Warband. It is a game available on Steam. Reading the reviews, you often see comments like: A great medieval RPG/strategy hybrid, with plenty of customization, both for your character and for your campaign. While I can appreciate that there are some things wrong with this g

Type your question about video games

Figure 1. FindYourGame chatbot interface deployed on Hugging Face Spaces.

## VIII. Conclusions

The main challenges included ensuring the chatbot maintained natural dialogue and generated accurate recommendations based on the knowledge acquired during training. Fine-tuning through LoRA proved to be an effective technique for specializing the compact TinyLlama model in the video game domain.

Key achievements of the project include the successful implementation of an efficient fine-tuning pipeline using LoRA, the generation of a specialized synthetic dataset of 50,000 prompts, and the deployment of a functional prototype accessible on Hugging Face Spaces for demonstration.

Overall, the developed system demonstrates the feasibility of adapting compact language models through efficient fine-tuning techniques to create chatbots specialized in specific domains, offering relevant video game recommendations with moderate computational requirements.

# References

[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