Universidad
Tecnológica
de Pereira

MASTER'S DEGREE IN SYSTEMS AND COMPUTER ENGINEERING

MASTER'S THESIS: NAVIGATING THE LINUX LANDSCAPE: A FEASIBILITY STUDY ON EMPLOYING LLAMA V2 FOR CRAFTING A CONTEXT-AWARE VIRTUAL ASSISTANT TAILORED FOR ARCH LINUX USERS

AUTHOR:
ENG. ANDRÉS MANUEL PRIETO ÁLVAREZ

PEREIRA-2024

NAVIGATING THE LINUX LANDSCAPE: A FEASIBILITY STUDY ON EMPLOYING LLAMA V2 FOR CRAFTING A CONTEXT-AWARE VIRTUAL ASSISTANT TAILORED FOR ARCH LINUX USERS

ENG. ANDRÉS MANUEL PRIETO ÁLVAREZ

Degree project presented as a requirement to opt for the Master's degree in Systems and Computing Engineering

Director: Ph.D José Jaramillo Villegas
Adviser: MsC. Lina Elizabeth Porras Santana

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
MASTER'S PROGRAM IN SYSTEMS AND COMPUTER ENGINEERING
PEREIRA
2024

Year 2024

ACCEPTANCE NOTE

____

____

____

____

Jury

____

Jury

Pereira (March 23, 2024)

DEDICATION

I dedicate this work to the free software community and the developer communities whose collaboration and generosity have enriched my professional and personal development. This work is a testimony of my gratitude to you.

Eng. Andrés Manuel Prieto Álvarez.

ACKNOWLEDGMENTS

I want to express my sincere gratitude to my thesis director, José, and my co-director, Lina, for their invaluable guidance, support and dedication throughout this research process. Her experience and advice were fundamental for the development of this work.

Extending my gratitude to the free software community, who from the beginning welcomed me with open arms and gave me a warm welcome. Their continued support and help were essential to achieving the objectives of this study.

I deeply thank my family for their unconditional support throughout all these years. Your love, understanding, and encouragement have been my greatest strength and motivation on this academic journey.

Finally, I would like to acknowledge all the people who, in one way or another, contributed to this project. Their encouragement and assistance were invaluable on this path toward the completion of this thesis.

Eng. Andrés Manuel Prieto Álvarez.

Abstract

The increasing complexity of Linux systems, coupled with the growing demand for automation in software development and system management, necessitates innovative solutions Debian (2023); Linux (2024). This research explores the feasibility of employing a Large Language Model Devlin et al. (2019); Sheng et al. (2023), specifically LLaMa v2, to create a cost-effective Virtual Assistant Devlin et al. (2019); Hu et al. (2021); Liu et al. (2023); Marukatat (2021); Sheng et al. (2023) tailored for Arch Linux users. The aim is to enhance user experience by providing context-specific suggestions and addressing general queries related to the installed software.

As the complexity of Linux systems continues to increase, driven by the demand for automation in software development and system management, the need for innovative solutions becomes paramount Debian (2023); Linux (2024). This research endeavors to explore the feasibility of utilizing a Large Language Model (LLM), specifically LLaMa v2, to develop a cost-effective Virtual Assistant tailored for users of Arch Linux. The primary objective is to enhance the user experience by providing context-specific suggestions and addressing general queries related to the installed software.

Keywords: GNU/Linux, Large Language Models, Virtual Assistant, Automatization, Natural Language Process

Índice general

Índice de figuras

Índice de tablas

Glossary

Throughout this study, technical and specialized terms are used that may not be familiar to all readers. In order to facilitate understanding of the text and improve accessibility for a wide audience, this glossary has been included:

- **Arch Linux:** A lightweight and flexible Linux® distribution that tries to Keep It Simple.
- **GNU/Linux:** Is an operating system, a set of programs that allow you to interact with your computer and run other programs.
- **Large Language Model or "LLM":** Large language models (LLMs) are very large deep learning models that are pretrained on large amounts of data.
- **Virtual Assistant:** A virtual assistant is a software agent that helps users of computer systems, automating and performing tasks with minimal human-machine interaction.
- **SourceForge:** Is a collaboration website for software projects.
- **Wiki:** A wiki is a web-based collaborative platform that enables users to store, create and modify content in an organized manner.
- **Man Pages:** A man page (Short for "manual page") is a form of software documentation usually found on a Unix or Unix-like operating system.

1 Introduction.

As the landscape of Linux systems continues to evolve, the intricate nature of these platforms poses significant challenges for users, particularly in terms of system management and software development. The versatility and continuous evolution of GNU/Linux, coupled with its open-source philosophy, contribute to a dynamic environment marked by diverse distributions and frequent updates Debian (2023); Enterprise (2021); Linux (2024). In this context, the demand for effective automation tools becomes essential to streamline tasks and improve user productivity Enterprise (2021, 2022).

However, despite the growing popularity of virtual assistants in various domains such as shopping, travel, and weather querying, GNU/Linux users have been left without native support from mainstream products like Amazon Alexa, Google Assistant, or Siri Hoy (2018). While projects like Deeping Assistant (currently under development) or KDE Connect provide some similar features, there is a notable gap in research and development for virtual assistants tailored specifically for GNU/Linux systems, particularly those running Rolling Release distributions like Arch Linux. The continuous updating process inherent in Rolling Release systems presents unique challenges that traditional virtual assistants are not adequately equipped to handle. Therefore, the exploration of additional features, such as those provided by autonomous agents, becomes crucial to achieving a continuous loop of context for the virtual assistant.

In light of these challenges, this research aims to explore innovative solutions to address the specific needs of Arch Linux users Linux (2024). By focusing on system management and software-related queries, the study seeks to leverage advanced technologies, particularly a Large Language Model (LLM) known as LLaMa v2 Devlin et al. (2019); Sheng et al. (2023), to develop an intelligent and efficient virtual assistant. LLMs, characterized by their natural language processing capabilities, offer a promising avenue for creating cost-effective virtual assistants tailored to the unique requirements of GNU/Linux users Devlin et al. (2019); Hu et al. (2021); Liu et al. (2023); Marukatat (2021); Sheng et al. (2023).

1.1 Problem context.

The GNU/Linux ecosystem, known for its diversity and flexibility, presents unique challenges for users, especially those using Rolling Release systems like Arch Linux Linux (2024). The versatility and constant evolution of these systems offer a wide range of options and functionality, but also create complexity in terms of system management and software-related queries. GNU/Linux users often face difficulties in performing basic system administration tasks, such as installing and updating packages, due to the lack of efficient and user-oriented tools.

Furthermore, the absence of specific virtual assistants for GNU/Linux, as well as the lack of native support for major products such as Amazon Alexa, Google Assistant or Siri, leaves GNU/Linux users without access to automation tools that could improve significantly your productivity and experience Hoy (2018). Although there are projects like Deeping Assistant and KDE Connect that offer some similar functionality, there is still a notable gap in the research and development of intelligent and efficient virtual assistants for GNU/Linux, especially for Rolling Release systems like Arch Linux.

The dynamic nature of Rolling Release systems, with their frequent updates and continuous changes, poses additional challenges for users, as they may experience difficulties in keeping up to date with the latest software versions and managing system dependencies effectively. These challenges negatively impact the user experience and can affect the productivity and efficiency of the overall system.

In summary, there is a clear need to develop an intelligent and efficient virtual assistant for GNU/Linux, especially for Rolling Release systems like Arch Linux, which can provide accurate answers about the system status and offer suggestions for installing packages, thus improving the experience and productivity of GNU/Linux users as a whole. In the context of the continuous development and evolution of the GNU/Linux ecosystem de Sousa et al. (2009); Linux (2024), users of systems like Arch Linux face significant challenges in terms of system management and software-related queries. Although the demand for efficient virtual assistants to improve productivity and user experience is evident, the lack of specific and suitable solutions for GNU/Linux, especially for Rolling Release systems such as Arch Linux, represents a significant gap in research and development. Additionally, the dynamic nature of Rolling Release systems, with their frequent updates and continuous changes, poses unique challenges that traditional virtual assistants cannot effectively address Hoy (2018). Therefore, there is a need to research and develop an intelligent and efficient virtual assistant, based on a pre-trained Large Language Model, such as LlaMa v2, that can provide accurate answers about the system status and offer suggestions for the installation of packages, thus improving the experience of GNU/Linux users, especially those who use Rolling Release systems such as Arch Linux.

1.1.1 Research question.

How can it be developed an efficient virtual assistant for the GNU/Linux operating system, using a pre-trained LlaMa v2 model, guaranteeing its quality and effectiveness through standard and specific virtual assistant metrics, with the aim of providing accurate answers about the state of the system and offer suggestions for installing packages, thus improving the experience of GNU/Linux users?

1.2 Objectives.

1.2.1 General Objective.

The main objective of this research is to develop an efficient virtual assistant for the GNU/Linux operating system, capable of providing accurate answers about the state of the system and offering suggestions for installing packages, based on comparisons of functionalities or needs. For this, the following objectives are proposed

1.2.2 Specific Objectives.

- Use Llama version 2 to generate a pre-trained large language model (LLM) for the development of the virtual assistant, adapting it specifically to the context and needs of the GNU/Linux operating system.
- Evaluate the performance of the LLM using standard metrics, such as Perplexity, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and response quality metrics.
- Design and implement a virtual assistant that is at the forefront of technology, incorporating advanced features that place it on the frontier of next-generation virtual assistant models.
- Evaluate the virtual assistant in terms such as Task Completion Rate, User Satisfaction, User Retention and Error Rate.
- Research and develop natural language processing (NLP) algorithms and techniques specific to the GNU/Linux environment, in order to improve the understanding and responsiveness of the virtual assistant to queries related to the operating system and package installation .
- Evaluate the performance and effectiveness of the developed virtual assistant through exhaustive tests and comparative analyzes with other existing systems, with the aim of guaranteeing its reliability and usefulness in real use environments.
- Provide complete and accessible documentation that details the operation, implementation and capabilities of the virtual assistant, in order to facilitate its adoption and use by the GNU/Linux user community.

1.3 Justification.

Entering and participating in the GNU/Linux ecosystem presents numerous challenges that can discourage new users and make it difficult for those already familiar with the operating system to participate. The constantly evolving nature of GNU/Linux

Gonzalez-Barahona et al. (2009), with multiple releases and options available, can create significant barriers for those who want to learn and collaborate in the environment. The GNU/Linux community, although rich in knowledge and resources, can often be intimidating to non-developer users, due to the steep learning curve and complexity of interactions between community members. Occasional toxicity in certain community spaces can drive away new users and make it difficult to integrate into the ecosystem Carillo and Marsan (2016).

The open source philosophy and collaborative culture in GNU/Linux encourages diversity and innovation, but it can also create confusion and difficulties for users, both new and experienced. The wide range of distributions, packages, and configurations available de Sousa et al. (2009), can overwhelm users and make decision-making difficult. The paradox of choice paradigm, where an abundance of options can lead to indecision and uncertainty Schwartz (2004), is especially relevant in the context of GNU/Linux. While documentation and resources are available, often scattered across wikis and forums, finding clear and concise solutions can require extensive searching and be time-consuming, which can be daunting for new and experienced users alike.

The viability of this proposal lies in its innovative approach to address the existing challenges in the GNU/Linux ecosystem. Although the information necessary to learn and use GNU/Linux is already available, its organization and accessibility can be significantly improved. By organizing and structuring this information in a more accessible and understandable way, this proposal aims to enable a next-generation virtual assistant Mauny et al. (2021) to use a machine learning-based language model (LLM) to generate useful responses for users. This would not only drastically reduce search and learning times, but would also make the work of expert users easier by providing them with a more efficient and effective tool for solving problems and finding solutions within the GNU/Linux environment. In summary, by improving the organization and accessibility of existing information Zheng and Fischer (2023), this proposal has the potential to make learning and using GNU/Linux easier and more accessible to a wide range of users, which would contribute to promoting a development more equitable and sustainable technology.

The GNU/Linux ecosystem, although rich in innovation and collaboration Williams (2011), also faces challenges that impact the Sustainable Development Goals (SDGs). The steep learning curve and complexity of community interaction can discourage digital inclusion (Goal 4) and hinder effective collaboration among community members (Goal 17). Furthermore, the abundance of options and solutions available in the GNU/Linux environment can negatively impact work efficiency and performance, making progress towards the goal of decent work and economic growth (Goal 8) difficult. By addressing these challenges and working to improve accessibility, inclusion and collaboration within the GNU/Linux ecosystem, this research seeks to contribute significantly to the achievement of several Sustainable Development Goals, thereby promoting more equitable and sustainable technological development.

Bibliografía

Carillo, K. D. A. and Marsan, J. (2016). "the dose makes the poison"-exploring the toxicity phenomenon in online communities. *Archive.org*.

de Sousa, O. F., de Menezes, M., and Penna, T. J. (2009). Analysis of the package dependency on debian gnu/linux. *Journal of Computational Interdisciplinary Sciences*, 1(2):127–133.

Debian (2023). Qué es gnu/linux.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. Ref: https://arxiv.org/abs/1810.04805.

Enterprise, R. H. (2021). *Annual Linux Market Study.* -edHat. Inc.". Ref: https://www.redhat.com/rhdc/managed-files/li-linux-market-study-ebook-f31489wg-202212-en.pdf.

Enterprise, R. H. (2022). *State of Linux in Public Cloud - Annual Review.* RedHat. Inc.". Ref: https://www.redhat.com/rhdc/managed-files/li-state-of-linux-public-cloud-solutions-ebook-f31743-202208-en_1.pdf.

Gonzalez-Barahona, J. M., Robles, G., Michlmayr, M., Amor, J. J., and German, D. M. (2009). Macro-level software evolution: a case study of a large software compilation. *Empirical Software Engineering*, 14:262–285.

Hoy, M. B. (2018). Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. Technical report, Cornell University. https://arxiv.org/abs/2106.09685.

Linux, A. (2024). Arch linux.

Liu, T., Xiong, Q., and Zhang, S. (2023). When to use large language model: Upper bound analysis of bm25 algorithms in reading comprehension task. In *2023 5th International Conference on Natural Language Processing (ICNLP)*, pages 1–4, Guangzhou, China. IEEE.

Marukatat, S. (2021). Text generation by probabilistic suffix tree language model. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–4, Ayutthaya, Thailand. IEEE.

Mauny, H., Panchal, D., Bhavsar, M., and Shah, N. (2021). A prototype of smart virtual assistant integrated with automation. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 952–957, Coimbatore, India. IEEE.

Schwartz, B. (2004). *The Paradox of Choice: Why More Is Less.* Harper Perennial.

Sheng, Y., Cao, S., Li, D., Hooper, C., Lee, N., Yang, S., Chou, C., Zhu, B., Zheng, L., Keutzer, K., Gonzalez, J. E., and Stoica, I. (2023). S-lora: Serving thousands of concurrent lora adapters. Technical report, Cornell University. https://arxiv.org/abs/2311.03285.

Williams, S. (2011). *Free as in Freedom: Richard Stallman's Crusade for Free Software.* .O'Reilly Media, Inc.".

Zheng, J. and Fischer, M. (2023). Dynamic prompt-based virtual assistant framework for bim information search. *Automation in Construction*, 155:105067.