Jheronimus Academy of Data Science

Thesis project: format for the Research Proposal

Name - Vishal Sehgal **Student number -** 2109374

Preliminary title of your thesis

Enhancing AI Transparency: Leveraging Large Language Models to Automate the Generation and Discovery of Semantic Model Cards

Main supervisor: I.P.K. Weerasingha Dewage

A brief description of your thesis proposal

Provide an abstract of 500 words maximum

With the rapid advancement of technology, the complexity of AI models has significantly increased. Developers are building generative AI applications which contain one or more models which make it challenging to determine the most suitable model for specific use cases [2]. Due to the infinite numbers of AI applications, people find it hard to access open and closed models. The main reason is Model cards [3][6]. Despite the availability of vast datasets, the creation of comprehensive and effective model cards remains an ongoing issue, limiting the discoverability, transparency, and reproducibility of AI models. According to the recent study *approximately 44.2% of the 74,970 AI models on Hugging Face have model cards, accounting for 90.5% of the platform's total download traffic* [1]. This is due to substantial lack of detailed information, such as environmental impact and model limitations, in these model cards, particularly among the most popular models. This lack of transparency undermines trust in AI models and hinders broader adoption.

When compared to smaller, less popular models, this disparity in model documentation

becomes even more pronounced. Large language models (LLMs) offer a promising avenue to bridge this gap by enhancing model card quality and supporting both small and large models in improving their documentation [5]. Some preliminary solutions, such as using NLP to augment web services for automated discovery and composition, as well as data-to-text generation, have shown potential. However, research in this domain remains limited. This thesis aims to explore how large language models and advanced model documentation techniques can be leveraged to automatically generate, complete, compare, and enhance the discoverability of semantic model cards. By improving model documentation, this research seeks to increase transparency and trust in AI models across the board.

References

1. Chen, L., Tan, C., & Zhang, Z. (2024). Contrastive self-supervised sequential recommendation. arXiv preprint arXiv:2402.05160. https://doi.org/10.48550/arXiv.2402.05160

2. Shumailov, I., Baskin, C., Sezener, E., Bansal, K., & Goldblum, M. (2023). Dodging attackers in neural networks with conflicting gradients. *Nature Machine Intelligence, 5*, 861–870. https://doi.org/10.1038/s42256-023-00692-8

3. GitHub. (2023, October 17). Introducing GitHub models. *GitHub News and Insights*. https://github.blog/news-insights/product-news/introducing-github-models/

4. ChatGPT. Response shared via conversation: https://chatgpt.com/share/6713ee7b-0214-8003-bcbb-62c311034c17 [Personal communication - Text improvement].

5. Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). *Trust in AI: Progress, Challenges, and Future Directions*. arXiv. https://doi.org/10.48550/arXiv.2403.14680

6. Liang, W., Rajani, N., Yang, X., Ozoani, E., Wu, E., Chen, Y., Scott Smith, D., & Zou, J. (2024). *Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI*. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-024-00857-z

Jheronimus Academy of Data Science

I am doing the master thesis based on (please select the appropriate option):

- My own company, namely: _____
- External company or organization, namely: _____
- Research project @ JADS, namely: _____

Problem statement and research question

a) Describe your field of study and the existing body of knowledge. What lacunae still exist in this area? What has been ignored so far? What question exists in scholarly literature, in theory, or in practice that points to the need for meaningful understanding and deliberate investigation?

The field of study for this thesis project focuses on the documentation of models, particularly in the generation, completion, and enhancement of model cards, with a focus on large language models (LLMs). While model cards have gained some attention, academic research in this area remains limited. A recent study indicates that although LLMs like GPT-4 and Claude are vital for performing various tasks, the automation of model card generation still requires substantial development to become a standard practice.[1]

Although some research has been conducted using natural language processing (NLP), the relationship between model cards and LLMs is still an open field of inquiry. Specifically, there is a need for further investigation into how these model cards should be documented to effectively highlight the models' behaviors, biases, and ethical implications [3]. While there is growing interest in creating interactive model cards that allow users to explore the model's attributes in real-time, this area is still underdeveloped.

One critical aspect often overlooked is that, even among the most popular models, the environmental impact is poorly documented. Mak et al. emphasize the need for transparent model cards to better communicate these impacts. Yet, model cards are inconsistently updated, and few efforts focus on automating their creation or systematically improving their accuracy. Semantic models for model cards provide a

structured framework for documenting a model's characteristics.[5] However, there is a lack of work on automating the conveyance of critical information on social, environmental, and ethical aspects, which can hinder the potential of model cards as tools for transparent documentation.

The following questions remain open in scholarly literature and point to areas in need of further research:

1. How can we automate the creation and updating of model cards effectively?
2. How can model cards better capture communication biases and ethical concerns?
3. How can we accurately document energy consumption and environmental impact in model cards?
4. What methods can make model cards more interactive, especially for non-technical stakeholders?

References

1.Makridakis S, Petropoulos F, Kang Y. Large Language Models: Their Success and Impact. Forecasting. 2023;5(3):536-549. doi:10.3390/forecast5030030.

2.Mak V, Alomari M, Tan Z, Zeiger J, Zhang M. Evaluating the sustainability of large AI models: A framework for analyzing the environmental impact of GPT-4 and other LLMs. J Am Med Inform Assoc. 2024;31(6):1436-1450. doi:10.1093/jamia/ocad235.

3.Slack, D., Krishna, S., Lakkaraju, H. *et al.* Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nat Mach Intell* ***5***, 873–883 (2023). https://doi.org/10.1038/s42256-023-00692-8

4.ChatGPT. Conversation shared by user. https://chatgpt.com/share/671537d3-738c-8003-9bee-3e89c698644b [Personal communication - Text improvement.

5. Amith, M.T., Cui, L., Zhi, D. et al. Toward a standard formal semantic representation of the model card report. BMC Bioinformatics 23 (Suppl 6), 281 (2022).

https://doi.org/10.1186/s12859-022-04797-6

Please state your target scientific discipline that you intend to contribute to (e.g., Data modeling and analysis methods, Deep learning, Data engineering, HTI and recommender systems, Entrepreneurship, Strategy, Legal/ethics, Social networks, Regulations and institutions, etc.):

Machine Learning, Deep Learning and Data Engineering

Do you intend to pursue a publication based on your thesis and write up the thesis report in the form of a scientific article consistent with standards of a reputable journal/ISI-rated conference from this discipline? (we expect you to agree on this with your main supervisor and (later) 2nd assessor explicitly before answering this question).

Yes/No.

In case Yes: what is the target journal/ ISI-rated conference? ___IEEE Transactions on Knowledge and Data Engineering__

b) What core question needs to be answered to contribute to achieving your research (i.e., resolving the issue identified in the problem statement)? What sub-questions need to be addressed to answer the core question?

Main research question: *To what extent can large language models (LLMs) and model documentation be utilized to automatically generate, complete, compare, and discover semantic model cards?*

Sub Research question

1. How can we use LLMs to automate the generation and completion of semantic model cards?
2. How can LLMs be used to capture communication biases and ethical concerns within models?

3. What strategies can we use to document energy usage and impact on the environment?
4. How can we make model cards more accessible for the non-technical users?
5. How can we automatically update the model card with the update in model?

The innovative character of the proposed project

What is the scientific and practical significance of your thesis? Does it contain an original contribution to the field of existing knowledge, and what is this contribution exactly? (consider using the patterns in scientific contributions per discipline described in the Study Guide and the slides of the Information Session)

The thesis holds both practical and scientific significance.
From a scientific perspective, this thesis aims to advance current development methods by automating the generation and updating of model cards through large language models (LLMs). Additionally, it proposes a structured approach to documenting ethical considerations, biases, and environmental impacts in a consistent manner—an area that has not been thoroughly addressed in existing literature. The thesis also provides methods for automatically identifying aspects such as behavior and performance biases, facilitating further research and promoting transparency and accountability.
On the practical side, the thesis is beneficial for industry practices and policy-making. It aims to enhance the scalability and consistency of model documentation, simplifying adherence to regulations for developers and companies. Additionally, the approach improves transparency for stakeholders, making information accessible to both technical and non-technical users. By addressing environmental impacts, the thesis also raises awareness of the societal effects associated with these technologies.

Proposition, hypotheses, and concepts

What is the central proposition? What are the core assumptions and working hypotheses? (in case you have any of the latter). What are the main concepts you intend to use, and how do you plan to operationalize them?

The central proposition focuses on using large language models (LLMs) to automate the creation, completion, and discovery of semantic model cards. This approach aims to enhance transparency and address ethical and environmental concerns. The core assumption is that current model documentation lacks consistency; thus, automation through LLMs can improve accuracy and scalability, making documentation more accessible for both technical and non-technical users.
The main concepts I intend to explore include LLMs, interactive model cards, biases, ethical and environmental impacts, and automation. Together, these concepts will comprehensively address various aspects of model behavior, including biases and ethical and environmental considerations.

Research design

Describe the research design you plan to use. Make sure it fits the project goal and research question(s). Describe your variables/features and indicate the methods you plan to use and analyses you plan to conduct. What data do you intend to gather, how are you planning to do this, and do you have access to these data? How do you plan to safeguard the reliability and validity of your research?

In this thesis, I will use the Design Science Research (DSR) framework developed by Peffers et al. (2007) [3]. Their approach, designed for information systems, is well-suited for building and refining an LLM-based framework that automates, generates, compares, and enhances model cards [1]. The DSR framework will provide guidance at each step to make model documentation more transparent and accessible. Subsequently, I will focus on designing and developing a tool that utilizes LLMs to automatically create model cards. This tool will iteratively improve by comparing generated model cards with manually created ones, aiming to enhance the transparency, discoverability, and accountability of models by including information on biases, ethical considerations, and environmental impacts [2].

Research Question :
To what extent can LLMs and model documentation be utilized to automatically generate, complete, compare and discover semantic model cards.

Data Gathering

Public repositories(eg. Hugging Face, Github etc) will provide the necessary input for the LLM system. The repo contains pre-existing documentation for various models, which will be processed by the LLM to generate model cards.

The research will use open source LLM like GPT-4,Bert etc. These models will be fine-tuned for the specific task of generating semantic model cards.

Development Phase:

LLM will ingest the document of various models and generate a comprehensive semantic model cards. The process will be iterative with the system undergoing refinements based on experimental outcome. Experiment is going to be conducted to test accuracy, consistency and completeness of the generated model cards by comparing it with manually generated model cards by hugging face etc.

The research will employ triangulation , a combination of accuracy, consistency and reliability. Then finally the system will undergo the iterative refinements based on experimental results , to continuously improve the performance and usability of the automated model cards generation.

References:

1. Tăbușcă A, Mihai T, Iftene A. Tracing the Influence of Large Language Models across the Most Impactful Scientific Works. *Electronics*. 2023;12(24):4957. doi:10.3390/electronics12244957.
2. Zhang Y, Chen X, Jin B, Wang S, Ji S, Wang W, Han J. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery. *arXiv preprint*. 2024. doi:10.48550/arXiv.2406.10833.
3. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.
4. ChatGPT. Conversation shared by user. https://chatgpt.com/share/671d7232-ae44-8003-93a3-7780d152b7e3 [Personal

communication - Text improvement.

Entrepreneurship

Please describe the business and/or societal impact your thesis project will have. Which of the three business impact maturity levels do you target? (1. Novel solution for decreasing costs in the existing business, 2. Novel solution for increasing revenues in the existing business, 3. New revenue generation/new business development/new product/new service).

How do you intend to do so? Which experiments/tests do you want to run to substantiate your impact? Which customers, companies, or organizations will you focus on in this validation? (could also be outside the organization in which you do the master thesis)

This thesis targets maturity level 3—focusing on new revenue generation, business development, and new product or service creation—within the three business impact maturity levels outlined above. The first proposed solution is to integrate large language models (LLMs) to automatically generate and structure model cards based on existing literature and documentation. This project aims to benefit developers working on platforms like Hugging Face and GitHub. Additionally, these platforms could leverage this solution to better align with European regulations and legal requirements. Furthermore, industries such as healthcare and finance would benefit from clearer documentation, enabling them to use and understand these models more effectively.

Name of the arranged 2nd assessor and in case help needed in finding one, suggestions for the expertise of the 2nd assessor:

Damian Tamburri