Avestia Publishing
Journal of Machine Intelligence and Data Science (JMIDS)
Volume 6, Year 2025
ISSN: 2564-3282
DOI: 10.11159/jmids.2025.003

Advanced Large Language Models Tuning and Applications for Product Analysis in R&D: A Comprehensive Investigation

Elena Barzizza[1], Nicolò Biasetton[1]*, Giorgio Caligiuri[2], Daniele Pennisi[2], Luigi Salmaso[1]
[1]University of Padova, Department of Management Engineering
Stradella San Nicola, 3 - 36100, Vicenza, Italy
elena.barzizza@phd.unipd.it; nicolo.biasetton@unipd.it; luigi.salmaso@unipd.it
[2]University of Padova, Department of Civil Environmental and Architectural Engineering
Via Marzolo, 9 - 35131 Padova, Italy
giorgio.caligiuri@studenti.unipd.it; daniele.pennisi@studenti.unipd.it
*Corresponding author

Abstract *- In today's competitive market, staying abreast of concurrent product characteristics is imperative for companies to maintain a competitive edge. Leveraging open-source Large Language Models (LLMs) presents a promising avenue for efficient and comprehensive analysis. This paper delves into the current landscape of commercial and open-source LLMs, assessing their potential for analyzing product characteristics. Additionally, it explores the feasibility of fine-tuning these models, including the utilization of Retrieval Augmented Generation (RAG), to enhance response accuracy and depth. Through evaluation, Mistral 7B emerges as a suitable open-source model for implementation, balancing performance with computational constraints. Furthermore, it outlines the process of refining LLMs using proprietary data, market intelligence, patent insights, and data gathered from web scraping to develop a comprehensive analytical tool for R&D purposes. This tool enables efficient extraction, analysis, and visualization of pertinent information, empowering decision-makers to steer innovation effectively.*

Keywords: Large Language Models, R&D, Product Characteristics, Smart data

1. Introduction

In the landscape of Research and Development (R&D), the ability to effectively monitor and analyze concurrent product characteristics stands as a cornerstone for sustaining competitiveness and fostering innovation. Traditional methods of market analysis often face significant challenges in providing timely and comprehensive insights, primarily due to the vast and diverse array of available data. However, recent advancements in natural language processing (NLP), especially with the advent of Large Language Models (LLMs), offer a promising solution to this challenge. LLMs, such as the Generative Pre-trained Transformer (GPT) series, exhibit unparalleled capabilities in comprehending and synthesizing human-like text. This renders them exceptionally well-suited for dissecting textual data pertaining to product characteristics (1,2)

By harnessing LLMs, analysts are furnished with a tool that seamlessly mimics human interaction, enabling nuanced exploration of intricate datasets. The present work embarks on a comprehensive exploration of the contemporary landscape of both commercial and open-source LLMs ripe for adoption. Furthermore, it undertakes a preliminary investigation into the feasibility of tuning these models or employing Retrieval Augmented Generation (RAG). Such endeavors are aimed at augmenting the responses to potential queries by adding contextual information through embedding, thereby aiming for heightened precision and comprehensiveness in insights (3).

This multifaceted inquiry not only sheds light on the cutting-edge capabilities of LLMs but also paves the way for their optimization in the domain of product

Date Received: 2024-04-12
Date Revised: 2024-10-08
Date Accepted: 2025-02-18
Date Published: 2025-03-04

characteristic analysis within R&D frameworks. Ultimately, it aspires to furnish decision-makers with unparalleled insights, thus catalysing informed strategies and bolstering innovation trajectories.

The present paper is organized the following way: Section 2 report an overview of most important commercial and open source LLMs as well as a comparison based on literature measures; Section 3 discusses the integration of proprietary data, market data, customer preference data, patent data and web-scraped data for tuning LLMs in product characteristic analysis and tool development. Lastly section 4 reports come conclusion on the present work.

2. LLMs: commercial and open source products evaluation

The adoption of LLMs in R&D necessitates a systematic approach to ensure optimal utilization of these models. Firstly, selecting an appropriate open-source LLM that aligns with the requirements and resources of the company is crucial. Factors such as model architecture, pre-training data, and computational requirements need to be considered during this selection process. Private models made available through API requests allows for an easier management but with less control over the models. The most important are:

- Claude 3, powerful state of the art model released by Anthropic (04/03/2024) in 3 versions: Haiku, Sonnet and Opus (from the least to the most powerful) .
- ChatGPT, market-leading family of models developed by OpenAI and currently available mainly in 3 versions: GPT-4, GPT-4 Turbo and GPT-3.5 Turbo. (4)
- Gemini, developed by Google and currently available as Gemini 1 Pro, with the more powerful Gemini 1 Ultra and the recent Gemini 1.5 Pro (15/02/2024) both available in preview (5,6)
- Command, made by Cohere and available in 3 versions: Command Light, Command and the recent addition Command-R optimized for RAG (11/03/2024, also available freely for research purposes) (7).
- Titan, developed by Amazon as a core model for Bedrock and available in 2 versions: Light and Express. (8)
- Jurassic 2, developed by AI21 labs id offered in 3 versions: Light, mid and Ultra (9,10).
- Mistral, developed by MistralAI and available in 3 versions: Small, Medium and Large.

On the contrary, Open-source models, result to be much more flexible but less powerful:

- Llama 2, family of models developed by Meta and available in 3 sizes: 7B, 13B and 70B (11).
- Mistral 7B and Mixtral 8x7B, the two open models developed by MistralAI (12,13).
- Gemma, family of light models recently released by Google (21/02/2024) and available in 2 sizes: 2B and 7B (14,15).

Deciding to focus on open source model Table 1 reports some important characteristics (namely number of parameters per model and relative VRAM necessary for usage, pre-training size, context window size, type of license and commercial use availability) of the most common LLMs.

Table 1: Open source LLM characteristics

MODEL	Number of Parameters	GPU/RAM required (fp/bf)	Pre-train size (tokens)	Context Window	License	Free Commercial Use
Gemma 2B	2.51B	8GB	2 trillions	8k	Gemma License	YES
Gemma 7B	8.54B	24GB	6 trillions	8k	Gemma License	YES
Llama 2 7B	7B	16GB	2 trillions	4k	Llama 2 Community License	YES (limit of 700 mil. users)
Llama 2 13B	13B	32GB	2 trillions	4k	Llama 2 Community License	YES (limit of 700 mil. users)
Llama 2 70B	70B	145GB	2 trillions	4k	Llama 2 Community License	YES (limit of 700 mil. users)
Mistral 7B	7.3B	16GB	N/A	32k	Apache 2.0	YES
Mixtral 8x7B	46.7B (12.9B active)	100GB	N/A	32k	Apache 2.0	YES

With the aim to select one open source LLM to implement the tuning and the product analysis, many indices developed to evaluate language models can be adopted. Table 2 report the models' performances, evaluated with the aid of various popular benchmark commonly used, focusing in particular on:

- Knowledge capabilities and knowledge retention, using benchmarks like MMLU (16) ARC-e and GPQA (17).
- Text comprehension and reasoning over text, using benchmarks like DROP (18), HotPotQA, TriviaQA and Kilt.
- Common sense and reasoning, using benchmarks like HellaSwag(19), Big-Bench-Hard (20) and WinoGrande.
- General perception and usability, using benchmark like Chatbot Arena (21) (ranked elo system obtained through anonymous "duels" evaluated by humans).

Note that not all measures are available for all models and that they have been chosen to evaluate also commercial models (not reported here) and the most performing model on each relevant measure are highlighted.

Table 2: Open source LLM comparison

MODEL	**MMLU (%)**	**ARC-e (%)**	**HotPotQA (%)**	**KILT (%)**	**TriviaQA (%)**	**HellaSwag (%)**	**WinoGrande (%)**	**Big-Bench-Hard (%)**	**ChatBOT Arena**
Gemma 2B	42.3	3.2			3.2	1.4	5.4	5.2	85
Gemma 7B	4.3	**1.5**			3.2	1.2	2.3	**5.1**	029
Llama 2 7B	4.4	8.7			6.6	7.2	9.5	2.6	027
Llama 2	5.6	5.2			4,0	0.7	2.9	9.4	043
Llama 2	**9.9**	9.9	2	3.2	**3.0**	**5.4**	0.4	1.2	082
Mistral 7B	2.5	**0.5**			2.5	1.0	4.2	**6.1**	073
Mixtral	**0.6**	**3.6**	6	6.0	**1.5**	**4.4**	7.2		**114**

The selection of the Mistral 7B model was driven by a multi-criteria evaluation that balanced both quantitative performance and practical resource constraints. Specifically, performance metrics such as the ARC-e and Chatbot Arena scores, alongside benchmark indicators like MMLU, highlighted its competitive capabilities. Moreover, Mistral 7B offers a 32k context window and operates within our hardware limitations (16GB GPU memory), making it particularly suitable for our R&D application. These factors, combined with its scalability and efficiency compared to alternatives such as Gemma 7B, led to its preferential selection.

3. Tuning LLMs for Product Characteristic Analysis and tool development:

Once a LLM is chosen, the integration of proprietary data, patent information, and web-scraped data becomes paramount to enhance the model's performance and relevance to the company's domain. Tuning LLMs involves fine-tuning the model parameters and training on domain-specific data to improve its effectiveness in analyzing product characteristics.

To this purpose we decide to adopt both proprietary data, market data and customer preference data. Additionally, incorporating pertinent patent data allows the model to capture insights into technological advancements and intellectual property landscapes relevant to the company's industry and for the product type under investigation. Web-scraped data from various sources further enriches the training dataset, enabling the model to stay updated with the latest developments and competitor offerings.

To fine-tune the LLM effectively, we implemented a robust methodology for integrating heterogeneous data sources, including proprietary datasets, market intelligence, and web-scraped data. The process involves: (1) Data pre-processing: This includes cleaning, normalization, deduplication, and anonymization to ensure consistency. (2) Quality assurance: Automated filtering algorithms assess data relevance and quality, which is further supplemented by expert review to validate proprietary and market data. (3) Data alignment: Diverse data streams are harmonized through a common schema to ensure compatibility during fine-tuning. (4) Iterative validation: The tuning process incorporates cross-validation and periodic performance assessments to maintain high accuracy and domain relevance. This structured approach ensures that the fine-tuning dataset is both comprehensive and reliable, thus enhancing the model's overall performance in product characteristic analysis.

The tuned LLM serves as the foundation for developing a comprehensive tool tailored to the needs of R&D departments. This tool enables analysts to extract, analyze, and visualize relevant information pertaining to concurrent product characteristics efficiently. By leveraging the power of LLMs, the tool offers capabilities such as trend analysis, competitor benchmarking, and predictive modeling, empowering R&D teams to make informed decisions and drive innovation effectively.

Scalability is a critical aspect of our solution. The tuning pipeline has been designed to accommodate increasing data sizes and complexity inherent to R&D applications. By leveraging the 32k context window of Mistral 7B, along with parallel processing, optimized data batching, (considering also the possibility of enhancing with cloud-based GPU clusters), we effectively mitigate computational constraints such as memory limitations and extended training times. Preliminary experiments confirm that the system maintains efficiency under larger data volumes. Future work will focus on further enhancing scalability through distributed training paradigms and dynamic resource allocation strategies, ensuring sustained performance as data demands escalate.

4. Conclusions

The adoption and tuning of LLMs for product characteristic analysis offer several benefits, including enhanced competitiveness, improved decision-making, and accelerated innovation cycles. However, challenges such as data privacy concerns, model interpretability, and computational resources pose potential hurdles that need to be addressed.

Although our study is centered on R&D applications, the underlying methodology—comprising data integration, fine-tuning, and performance evaluation—is inherently adaptable. We posit that with domain-specific adjustments, the approach can be generalized to other industries such as healthcare, finance, and manufacturing. Nonetheless, it is important to note that differences in data availability, contextual nuances, and operational constraints may limit the direct applicability of our tuned LLMs outside the R&D context. Future work will explore these limitations and investigate necessary adaptations for broader domain transferability.

In conclusion, the adoption and tuning of LLMs present a promising approach for analyzing concurrent product characteristics in R&D environments. By integrating proprietary data, patent information, and web-scraped data, companies can leverage the power of LLMs to gain actionable insights and maintain a competitive edge in their respective industries.

5. Acknowledgements

Nicoló Biasetton and Luigi Salmaso thanks the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004).

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [citato 11 aprile 2024]. p. 1877–901. Available on: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; 2019 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/1810.04805
3. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training.
4. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report [Internet]. arXiv; 2024 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2303.08774
5. Gemini Team, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: A Family of Highly Capable Multimodal Models [Internet]. arXiv; 2024 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2312.11805
6. Reid M, Savinov N, Teplyashin D, Lepikhin D, Lillicrap T, Alayrac J baptiste, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context [Internet]. arXiv; 2024 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2403.05530

7. Context by Cohere [Internet]. 2024 [citato 11 aprile 2024]. Command R: RAG at Production Scale. Available on: https://txt.cohere.com/Command-R

8. Amazon Web Services, Inc. [Internet]. [citato 11 aprile 2024]. Modelli di fondazione per l'IA generativa – Amazon Titan – AWS. Available on: https://aws.amazon.com/it/bedrock/titan/

9. Announcing Jurassic-2 and Task-Specific APIs [Internet]. [citato 11 aprile 2024]. Available on: https://www.ai21.com/blog/introducing-j2

10. Simplifying Our Jurassic-2 Offering [Internet]. [citato 11 aprile 2024]. Available on: https://www.ai21.com/blog/simplifying-our-jurassic-2-offering

11. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models [Internet]. arXiv; 2023 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2307.09288

12. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. Mistral 7B [Internet]. arXiv; 2023 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2310.06825

13. mistralai/Mistral-7B-Instruct-v0.2 · Hugging Face [Internet]. [citato 11 aprile 2024]. Available on: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

14. Gemma Team, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, et al. Gemma: Open Models Based on Gemini Research and Technology [Internet]. arXiv; 2024 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2403.08295

15. google/gemma-7b · Hugging Face [Internet]. 2024 [citato 11 aprile 2024]. Available on: https://huggingface.co/google/gemma-7b

16. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring Massive Multitask Language Understanding [Internet]. arXiv; 2021 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2009.03300

17. Rein D, Hou BL, Stickland AC, Petty J, Pang RY, Dirani J, et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark [Internet]. arXiv; 2023 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2311.12022

18. Dua D, Wang Y, Dasigi P, Stanovsky G, Singh S, Gardner M. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs [Internet]. arXiv; 2019 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/1903.00161

19. Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y. HellaSwag: Can a Machine Really Finish Your Sentence? [Internet]. arXiv; 2019 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/1905.07830

20. Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models [Internet]. arXiv; 2023 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2206.04615

21. Chiang WL, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference [Internet]. arXiv; 2024 [citato 11 aprile 2024]. Available on: http://arxiv.org/abs/2403.04132