Responsible design and use of large language models

ChatGPT and Bard have taken the world by storm! In many circles, even in casual conversations, discussions about the huge positive influences of Large Language Models (LLMs) and the sobering risks they pose to our socio-economic fabric are happening in the same breath. Some of the concerns revolving around the use of LLMs like spread of misinformation and production of unfair, biased, or harmful content are legitimate to a greater or lesser extent depending upon the context. However, the anxiety around the negative impact that such advancing technology may have on the human-race is amplified by the AI apocalypse warnings from prominent figures in the world of AI. Even the developers of such technologies are swiftly addressing the public concerns about the potential risks associated with the use of the LLMs. For instance, OpenAI — the developers of ChatGPT, employs a [Content] 'Moderation API' to assess if the output of ChatGPT contains sexual, violent, hateful or harmful content in accordance with their organization's content policy[1].

Businesses committed to Responsible AI principles need to follow suit. To fully realise the potential of this technology responsibly in specific use-cases, they must recognise the different ways LLMs can be used in applications and establish an appropriate governance layer to evaluate and monitor the context of application from an ethical perspective throughout its lifecycle.

How to use LLMs, the different factors to consider

LLMs, like any other AI models, are essentially socio-technical systems — an inextricable bundle of code, data, subjective parameters and people [2]. Evaluating such systems from an ethical perspective requires examining not only the characteristics of technology but also the context of its use. For businesses considering utilisation of the LLM-technology for a use-case, it is crucial to analyse both the technology as well as the scenario-specific angles.

To begin, businesses must be aware of the various ways LLMs can be employed in context-specific applications. Three different mechanisms enable businesses to leverage this technology:

Using existing LLM Models 'as-is' via APIs

Technology companies have developed cutting-edge LLMs, allowing end users, such as enterprises, consumers, and enthusiasts, to access these models through APIs. This enables them to explore the possibilities offered by these technologies. For example, OpenAI allows anyone with online access to register and experience ChatGPT-3.5. OpenAI achieved an impressive end-user base of 100 million within the first two months of ChatGPT's launch through this mechanism [3].

Building LLMs

Enterprises may find the use of the existing LLMs 'as-is' inadequate for scenario-specific applications due to performance and compliance concerns (e.g., data-security). In such cases, enterprises can develop their own context-specific LLMs in two ways:

a. Finetune existing LLMs:
Many developers of generic LLMs allow fine-tuning of their LLMs (training the front/adaptive layers of the LLM using context-specific data) to cater to the performance and compliance needs of customer-enterprises. HuggingFace, for example, offers various models that enterprises can customize for specific use-cases [4].

b. Building from scratch:
For deep domain specific use-cases, enterprise may consider creating field-specific LLMs from scratch. It is essential to note that LLMs are built from scratch following the architecture outlined in the ground-breaking paper "Attention is all you need," published in 2017[5].

Each approach has strengths and limitations, as listed in Table 1, and its suitability depends on the context of the use-case for which LLMs are being considered.

Table 1: Strengths and limitations of the different approaches of using LLMs

Approach for using LLM	Strengths	Limitations
Using existing LLMs 'as-is' as APIs	• Easy access to collective knowledge across domains • Straight forward consumption of raw output by simple API calls from the existing LLM models • Proven performance on various language tasks	• Potential for biases and ethical concerns in training data • Lack of control over training data and architecture for the enterprise • May not be suitable for scenario specific tasks
Finetune existing LLMs (Retrain the adaptive layers)	• Most effective solution for an enterprise as it allows appropriate retraining of adaptative layers for use in scenario-specific applications • Ready-to-use framework available in market for enablement of finetuned LLM in an enterprise	• Chances of bias incorporation from the frozen layers of LLM • Possible data security and privacy threat
Building LLMs from Scratch	• Maximum control over training data and architecture • More flexibility for customization • No chances of any bias incorporation	• Requirement of huge resources, time, and expertise for model training • Requirement of humongous volume of appropriate data to train model to achieve better results

There are advantages and limitations associated with each of the approaches, as listed in Table 1, and the suitability of an approach depends on the context of the use-case for which LLMs are being considered. To assess the suitability from both performance and ethical AI perspectives, it is essential to consider how the three approaches can help improve application performance and ensure responsible LLM design and use. Businesses should evaluate the approaches from a feasibility-of-implementation angle as well.

For instance, businesses with low-to-medium AI maturity levels should avoid building an LLM from scratch, as it is time and resource-intensive, requiring advanced data-science (NLP) skill sets and extensive data and computational power. Table 2 presents an overview of LLMs from these different dimensions.

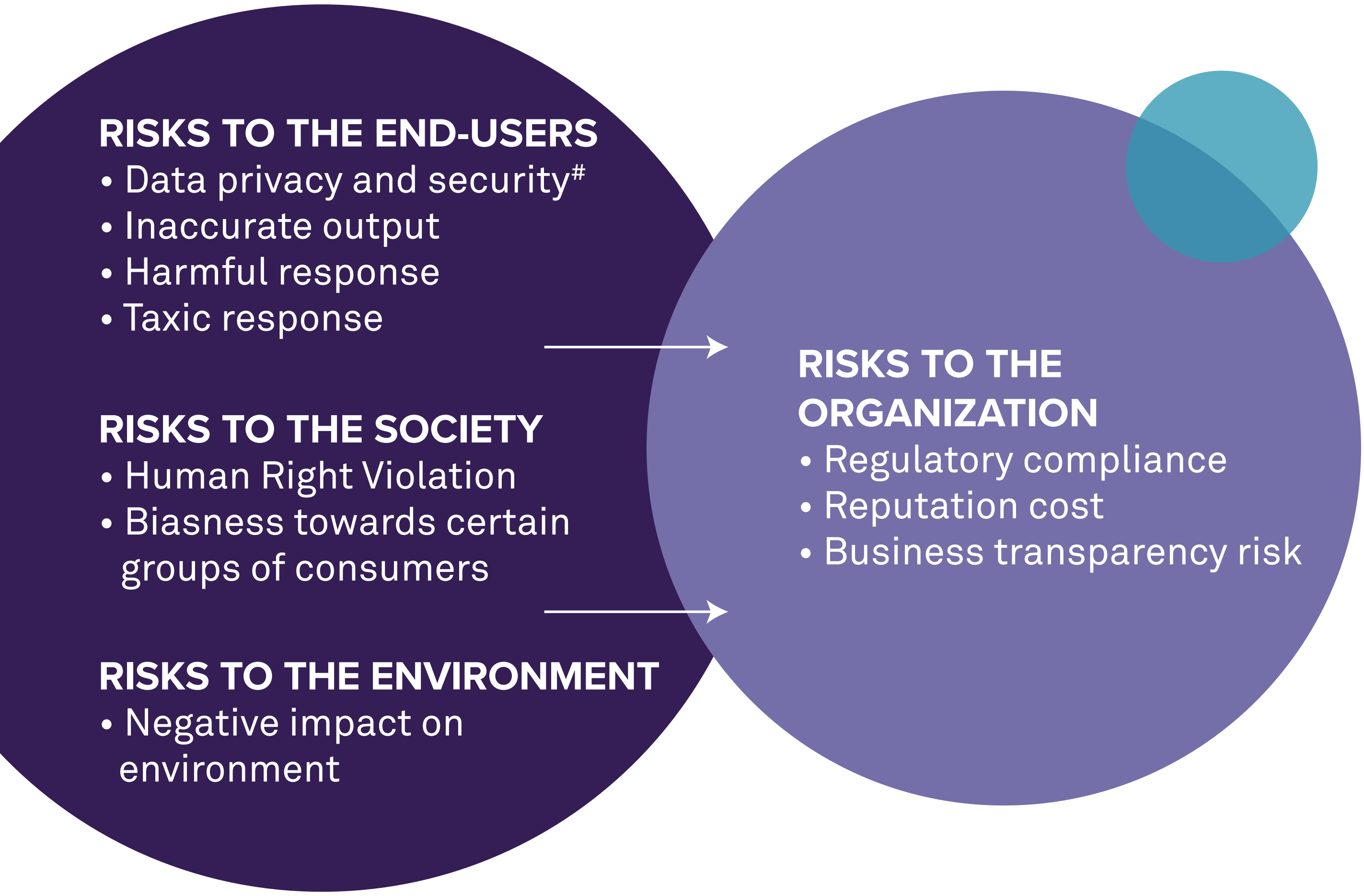
Table 2: Evaluating llms from different decision-making dimensionstable

Broad Dimensions	Factors	Existing LLM (Use 'As-Is' Via APIs)	Finetune Existing LLMs (Retrain Adaptive Layers)	Building LLM From Scratch
Feasibility of implementation of performant LLMs for use-cases	Data requirements for building an LLM	No model-building data required as the developers of these existing LLMs have already trained the model with vast amount of data from a wide variety of text corpora.	Enterprises would need use-case specific data to finetune existing LLMs. Volume of data required for finetuning is less than the volume of data required to build LLMs from scratch, in general.	Building a robust and performant LLM from scratch requires huge volume of data.
	Architecture	Architectural considerations not required in this context.	Architecture considered by the LLM-provider cannot be changed. Enterprises need to consider the adaptive layers required for finetuning.	Enterprises must decide on the choice of architecture as it has a significant impact on the model's performance.
	Hyperparameter-tuning	Pre-built LLMs provide no scope of hyperparameter tuning	Finetuning LLMs provides on option to tune hyperparameter from the adaptive layers (specific to scenarios) as hyperparameters of frozen layers of the architecture are inaccessible to customization.	A proper strategy around the selection of hyperparameters and its optimization based on the training data are required.
	Time and resources	Time and resource consideration required only to put in place mechanisms to feed inputs (prompts) into an existing LLM and consume its output	As it needs some retraining hence consumption of time and resources are required. However, there are lot of available frameworks in market which easily enables this capability.	A colossal time-consuming and highly skilled data-science resource-intensive process as it requires developing and testing of complex advanced architecture.

Broad Dimensions	Factors	Existing LLM (Use 'As-Is' Via APIs)	Finetune Existing LLMs (Retrain Adaptive Layers)	Building LLM From Scratch
Feasibility of use from a Responsible AI perspective	Bias and Fairness	Prebuilt LLMs can perpetuate and amplify harmful biases present in the training data. There is a chance that these models may not generate equitable outcomes in specific applications.	Retraining the adaptive layers may not eradicate all the application-specific unwanted biases that may have creeped into it from the frozen section of the model.	The enterprise has ultimate control in terms of bias and fairness as it owns the data and the governance oversight for the choice of architecture and the process of development and deployment.
	Privacy & Security	Uploading data, in the form of inputs/prompts, to an existing LLM may pose a data privacy and security threat.	Uploading data into finetuned LLMs may also pose a data privacy and security threat. Some of the leading LLM-providers are coming up with architectural designs to address these concerns.	The enterprise has complete control to mitigate data privacy and security risks.
	Governance and regulation	Adherence to organisational level governance and regulation may prove to be difficult in this context.	Adherence to organisational level governance and regulation may be limited in this context.	The enterprise can incorporate all the necessary governance frameworks and regulatory norms to make the developed LLM compliant.
	Auditing and testing	Prebuilt LLMs can be audited and tested in the form of 'Black Box Testing' to identify potential issues pertaining to its use in a use-case.	LLMs built with transfer learning should be tested from both development and usage perspectives.	Built from scratch LLMs should require thorough development testing as well as usage-based assessment.
	Document ation	There are extensive documentations and user guides from a usage perspective. However, LLM-providers may not reveal the details of the data used to build the LLM or the specifics of the architecture.	Finetuning LLMs providers offer an extensive documentation on the usage and on the modifications of adaptive layers. However, LLM-providers may not reveal the details of the data used to build the LLM or the specifics of the architecture.	Enterprise needs to produce all the documentation pertaining to the LLMs that they built from scratch along with user support.

Assessing the impact of LLMs in use-cases

Businesses, with an understanding of the decision-making dimensions of the three approaches to use LLMs, should then evaluate the potential impact of an appropriate scenario-specific LLM application on its end-users, society, and environment and the organization itself (as depicted in Figure 1).



RISKS TO THE END-USERS

- Data privacy and security#
- Inaccurate output
- Harmful response
- Taxic response

RISKS TO THE SOCIETY

- Human Right Violation
- Biasness towards certain groups of consumers

RISKS TO THE ENVIRONMENT

- Negative impact on environment

RISKS TO THE ORGANIZATION

- Regulatory compliance
- Reputation cost
- Business transparency risk

Figure 1: Risk of LLMs upon its end-users, society, environment, and organization

To gauge the impact of a scenario specific LLM system on its end-users, businesses must first determine the system's reach or coverage. Let's take the example of a multi-national electro-mechanical component manufacturer for automotive applications. They plan to deploy an LLM-powered chatbot for field-engineers across all operating geographies. Field-engineers, with domain knowledge, can likely distinguish sensible responses from hallucinations (non-sensical outputs) [6]). However, the company must be cautious, as a potentially harmful response slipping past human filters could lead to adverse outcomes. Additionally, consistency of correct responses in different languages must be ensured.

For engineering-help queries, the company can opt for finetuned LLMs or build their own use-case specific LLMs. Finetuned LLMs are preferable to maintain feasibility and environmental commitments [7] [8]. However, implementing guardrails such as deflection-logic and prompt & response filtering for responsible use is vital. It is essential for the business to conduct a thorough evaluation of the LLM system including data security, privacy, contextual correctness and toxicity along with proper documentation.

Businesses should be mindful of risks posed by use-case specific LLMs, such as personal data leakage and unfair outcomes, which can translate to societal risks and regulatory non-compliance. Lack of transparency in LLM operations when using 'as-is' or finetuned models adds to the risks. Table 3 presents mechanisms to assess LLM systems comprehensively, including their ability to match human-level common-sense knowledge.

Table 3: Evaluating LLMs from impact perspective

LLM Usage	Accuracy	Robustness	Fairness	Toxicity [7]	Common sense Knowledge [8]
As-is’ via APIs	Reference from Holistic Evaluation of Language Models (HELM)* framework [9]	Reference from HELM* framework	Equalised Odds, Individual/Group Fairness	Hugging Face based Toxicity Score [7] for detecting hate speech, Test Cases designed with LLM BLEU framework [10] to identify the differences w.r.t. to a reference	HellaSwag* – evaluate physical, grounded, and temporal common sense WinoGrande* - examines physical and social common sense. Social IQA* - evaluates social common sense PIQA* - covers the physical aspect of common sense
Finetuned LLMs	Perplexity, Entropy	Adversarial Accuracy, Distributional Shift	Demographic Parity, Equalized Odds, Treatment Equality, Individual /Group Fairness		
Built from scratch	Perplexity, Entropy, BPC	Adversarial Accuracy, Distributional Shift, Diversity, Bias Mitigation	Demographic Parity, Equalized Odds, Treatment Equality, Individual/Group Fairness		

** Small descriptions on HELM, HellaSwag, WinoGrande, Social IQA, PIQA, etc. are provided in the notes*

Businesses considering the direct integration of LLM systems into their product or service offerings should also assess the risk of these offerings from a usage perspective, aligning with the forthcoming AI regulations in various geographies (for example, EU’s AI Act). More information about this can be found in Ethical AI: Looking beyond accuracy to realize business value [11].

# There are different ways of implementing LLMs responsibly from a data security and privacy point-of-view using federated learning, differential privacy, etc.; these methods are elaborated in[6]

Conclusion

The advent of advanced Large Language Models (LLMs) brings both excitement and trepidation. Businesses must tread carefully in this landscape, establishing safeguards to harness the potential of LLMs with responsibility and ethics. Incorporating control elements into LLM approaches is essential to align with business objectives and ethical standards. Scrutiny of scenario-based LLM applications is imperative to identify potential adverse effects such as biases and misinformation perpetuation. Mitigating risks and ensuring equitable outcomes through protective measures becomes paramount.

The fast-shifting LLM landscape calls for human-in-the-loop applications, where humans retain ultimate responsibility and decision-making, with LLMs serving as support tools. This approach provides greater control, accountability, and risk reduction. By treading thoughtfully, businesses can embrace the possibilities of LLMs while mitigating their associated challenges and ensuring responsible utilization.

Notes

HELM Framework

Holistic Evaluation of Language Models (HELM) is an assessment approach that provides a comprehensive analysis of the performance, limitations, and capabilities of language models. It considers various factors, including linguistic quality, factual accuracy, bias detection, ethical considerations, and robustness, to provide a holistic understanding of the model's strengths and weaknesses. HELM aims to ensure a well-rounded evaluation that considers the broader implications and challenges associated with deploying language models in real-world applications [12].

HelaSwag

HellaSwag is an evaluation benchmark designed to assess the physical, grounded, and temporal common sense understanding of Large Language Models (LLMs). It focuses on evaluating the models' ability to reason about real-world situations, events, and context, going beyond syntactic and semantic understanding. HellaSwag aims to measure the LLM's capability to generate plausible and contextually appropriate responses, considering nuanced aspects of human-like reasoning, thereby providing insights into the model's comprehension of common-sense knowledge in a broader context [13].

WinoGrande

WinoGrande is an evaluation dataset specifically designed to examine the physical and social common sense understanding of language models. It consists of a set of multiple-choice questions that require reasoning about real-world scenarios, incorporating both physical and social contexts. WinoGrande focuses on challenging the models' ability to comprehend nuanced aspects of common-sense knowledge, such as causality, intention, and social dynamics. By assessing the performance of language models on WinoGrande, researchers gain insights into their capabilities and limitations in understanding and reasoning about common sense in a variety of contexts [14].

Social IQA

Social IQA is an evaluation benchmark that measures the social common sense understanding of language models, assessing their ability to comprehend and reason about social interactions, emotions, intentions, and cultural context [15].

PIQA

PIQA (Physical Interaction: Question Answering) is an evaluation benchmark that focuses on assessing the physical common sense understanding of language models by challenging them with questions related to physical interactions and dynamics in the real world [16].

References

1 T. Markov, C. Zhang, S. Agarwal, T. Eloundou, T. Lee, S. Adler, A. Jiang and L. Weng, "New and improved content moderation tooling," 10 August 2022. [Online]. Available: https://openai.com/blog/new-and-improved-content-moderation-tooling.

2 I. Bartoletti, "Another warning about the AI apocalypse? I don't buy it," 3 May 2023. [Online]. Available: https://www.theguardian.com/commentisfree/2023/may/03/ai-chatgpt-bard-artificial-intelligence-apocalypse-global-rules.

3 K. Hu, "ChatGPT sets record for fastest-growing user base - analyst note," 2 February 2023. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

4 "Fine-tune a pretrained model - Transformers," Hugging Face, [Online]. Available: https://huggingface.co/docs/transformers/training.

5 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

6 Mugunthan, V.; Maximizing the ROI of Large Language Models for the large enterprise., 29 March, 2023. [Online] https://www.dynamofl.com/blogs/maximizing-the-roi-of-large-language-models-for-the-large-enterprise

7 Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, p. 1–38, 2023.

8 "Toxicity," Hugging Face, [Online]. Available: https://huggingface.co/spaces/evaluate-measurement/toxicity.

9 X. L. Li, A. Kuncoro, J. Hoffmann, C. de Masson d'Autume, P. Blunsom and A. Nematzadeh, "A Systematic Investigation of Commonsense Knowledge in Large Language Models," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 2022.

10 Liang et. al., "Holistic Evaluation of Language Models," arXiv, 2022.

11 K. Doshi, "Foundations of NLP Explained - Bleu Score and WER Metrics," Medium, 9 May 2021. [Online]. Available: https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b.

12 B. K. Mitra and M. Smith, "Ethical AI: Looking beyond accuracy to realize business value," Wipro Limited, March 2022. [Online]. Available: https://www.wipro.com/blogs/bhargav-kumar-mitra/ethical-ai-looking-beyond-accuracy-to-realize-business-value/.

[13] Bommasani, Rishi, P. Liang and T. Lee, "HELM," Center for Research on Foundation Models, 19 March 2023. [Online]. Available: https://crfm.stanford.edu/helm/latest/.

[14] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi and Y. Choi, "HellaSwag: Can a Machine Really Finish Your Sentence?," arxiv, 2019.

[15] K. Sakaguchi, R. Le Bras, C. Bhagavatula and Y. Choi, "WinoGrande: An Adversarial Winograd Schema Challenge at Scale," arXiv, 2019.

[16] M. Sap, H. Rashkin, D. Chen, R. Le Bras and Y. Choi, "SocialIQA: Commonsense Reasoning about Social Interactions," Allen Institute for Artificial Intelligence, Seattle, 2019.

[17] Y. Bisk, R. Zellers, R. Le Bras, J. Gao and Y. Choi, "PIQA: Reasoning about Physical Commonsense in Natural Language," Proceedings of 34th AAAI Conference on Artificial Intelligence, vol. 34, pp. 7432-7439, 202 0.

About the Authors

SOUMYA TALUKDER

is currently working as a Consultant at Wipro Limited. He has 9+ years of experience as a data scientist, where he has worked majorly on Retail and Telecom domains. He has good experience in Statistical, Machine Learning and AI model development.

DIPOJJWAL GHOSH

is currently a Principal Consultant at Wipro Limited, India. He received his M. Tech. in Quality, Reliability and Operations Research from Indian Statistical Institute, Kolkata. He has 16+ years of research and analytical experience in various domains including retail, manufacturing and energy & utilities. Dipojjwal has published multiple research and popular technology articles up to date.

SILADITYA SEN

is a Data Scientist at Wipro Limited. He has received his M. Sc. in Statistics from Presidency University, Kolkata. He has close to 8 years of experience in the field of data science in Retail, Telecom and Utility domains. He is quite proficient in building classical statistical, Machine Learning and AI models.

BHARGAV MITRA

is a Data Science Expert and MLOps Consultant with an entrepreneurial mindset. He is working with Wipro as the AI & Automation Practice Partner for Europe and leading the practice's global initiatives on Responsible AI. Bhargav has over 18 years of 'hands-on' experience in scoping, designing, implementing, and delivering business-intelligence driven Machine/Deep Learning. He holds a DPhil in Computer Vision from the University of Sussex and an MBA from Warwick Business School.

ANINDITO DE

is CTO of the AI Practice at Wipro Limited. His primary responsibilities are building capabilities across different areas of AI and ML and bringing to life AI driven intelligent solutions for customers. With over two decades of experience, he has been a part of many large technology implementations across sectors and authored multiple technology publications and patents.

Ambitions Realized.

Wipro Limited
Doddakannelli
Sarjapur Road
Bengaluru – 560 035
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading technology services and consulting company focused on building innovative solutions that address clients' most complex digital transformation needs. Leveraging our holistic portfolio of capabilities in consulting, design, engineering, and operations, we help clients realize their boldest ambitions and build future-ready, sustainable businesses. With 250,000 employees and business partners across more than 60 countries, we deliver on the promise of helping our clients, colleagues, and communities thrive in an ever-changing world.

For more information,
please write to us at **info@wipro.com**