A Comprehensive Arabic Large Language Models Survey

Abdelrhman Tarek Waheed[1], Mohamed Marie[2]

[1]Data Science Program, Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

[1]Abdelrahmantarek_pgs@fci.helwan.edu.eg

[2]Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

[2]MohamedMarie@yahoo.com

Abstract— **Arabic Large Language Models (LLMs) have recently witnessed significant progress, making them essential for a wide range of applications in Arabic-speaking areas. Although they perform well on many NLP tasks, problems remain because of the complexities of Arabic's morphology, dialects and script. This survey covers the state-of-the-art Arabic LLMs, including their architectures, capabilities, and performances. It also discusses some challenges, such as limited high-quality datasets and computational resources, and provides potential avenues for future research in order to enhance the robustness of models and the breadth of Arabic-specific resources.**

Keywords— **Deep Learning (DL), Natural Language Processing (NLP), Large Language Models (LLMs), Arabic Large Language Models, Transformers, Neural Networks (NN).**

I. Introduction

Deep learning and natural language processing (NLP) have become the heart of revolutionizing the way machines understand and process human language. This progress has been largely driven by large datasets and the creation of sophisticated neural network models. Leading this charge are large language models (LLMs), which have demonstrated exceptional capabilities in solving challenging NLP tasks. These models have opened the doors for various applications in the field of NLP allowing for improvements in usage when it comes to machine translation, sentiment analysis, and conversational AI. LLMs are evolving further and remain integral to the future of technology across multiple fields [1].

One of the most significant developments in the history of LLM evolution was the release of the transformer architecture. This enables them to rapidly access and prioritize relevant data from large datasets, as opposed to conventional models that do not utilize an attention mechanism. It's this innovation that allows LLMs to scale to unthought-of sizes, giving rise to capabilities and abilities not dreamed of. Since their introduction, transformers have become the underlying architecture of choice for most modern natural language processing (NLP) tasks, enabling revolutionary advances that have reshaped both the targets of NLP and what is possible from models in the research lab and in practice [2].

Although large language models (LLMs) have seen great advances in many languages, Arabic has remained mostly absent from the family of models in recent times. Although LLMs like GPT-3, BERT and T5 have proven to be effective for various aspects of languages such as English; the Arabic language, which is spoken by more than 400 million people around the world, has not yet gained as much attention in the LLM domain. This underrepresentation results in a missed opportunity for language technologies that can significantly improve life for Arabic speakers, stifling access to state-of-the-art natural language processing (NLP) applications. The early focus of most work in NLP on English meant that little attention was given to creating the NLP tools that were needed to contend with the unique linguistic, cultural, and syntactical features of spoken Arabic [3].

A significant problem that Arabic LLMs face is the scarcity of high-quality Arabic datasets for training large models. While English benefits from large corpora of data up to several trillion tokens in size, Arabic datasets are much smaller and less comprehensive. This limited availability of high-quality, diverse, and well-annotated data limits the training of Arabic LLMs with performance comparable to that of their English counterparts. Consequently, Arabic language models struggle to comprehend and produce text with the same level of competence and precision, hindering their usefulness in practical use cases. So, in order to leverage the huge potential of LLMs in Arabic-speaking communities, addressing this data limitation is essential to plugging the gap in Arabic NLP Language Models.

This survey paper begins with an introduction to deep learning, NLP, and the evolution of LLMs, followed by an explanation of the LLMs architectures and applications. Then explores Arabic LLMs, highlighting their challenges, innovations, and a comparison of the leading models in this domain.

II. LARGE LANGUAGE MODELS

Large Language Models (LLMs) can generate human-like text by constructing artificial neural networks. It is trained on lots of text data and can understand, interpret, and generate human languages in various context. Large Language Models (LLMs) are the foundation for many tasks in the domain of Natural Language Processing (NLP), such as translating text, answering questions, and generating content. As compared to traditional models, which achieve prompt performance on a narrow set of tasks but are trained on a task-specific basis, their capacity to generalize across tasks (through approaches that do not rely on task-specific training) is a key distinguishing feature [4].

Language models have come a long way from simple statistical models in the last couple of decades to being based on more complex neural networks. Surrounding the period of early models such as n-grams which simply calculated the statistical probabilities of subsequent words in a sequence, with newer models introducing deep learning techniques like Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) networks. The emergence of transformer-based models in 2017, beginning with the Transformer architecture [1], represented a turning point, allowing for increasingly more efficient and scalable models that could deal with long-range dependencies in text more accurately and faster [5].

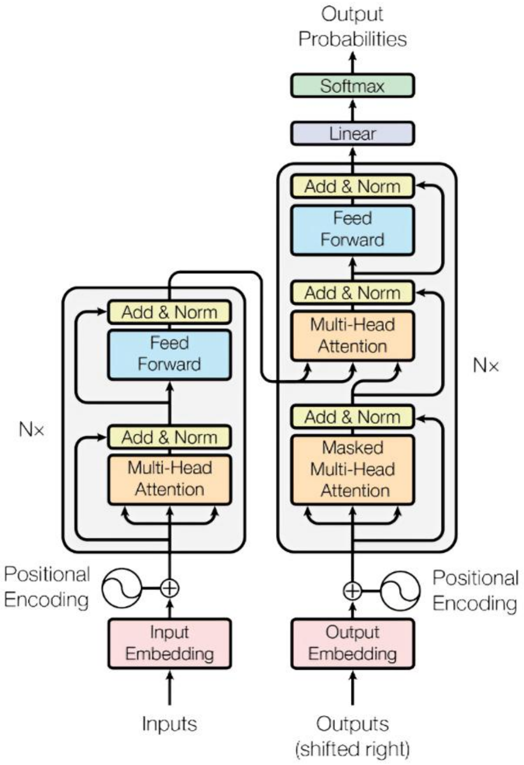

Figure 1- first introduced transformer architecture

Additionally, it is important to remember that the transition from statistical models to neural models to transformer-based models is not an incremental upgrade or simply a feature addition, but rather a significant jump in capabilities (i.e., language modelling). Early models such as n-grams used frequency counts to predict the sequence of words but had limitations in modeling long-range dependencies and understanding context. Previously, neural models like RNNs and LSTMs abstracted this space efficiently but struggled with scalability and training efficiency. Transformers, which were proposed by Vaswani et. al [1], do not have these limitations, as they leverage a self-attention mechanism to enable fully parallelized processing of full sequences during training, effectively capturing context at scale with potentially greater efficiency.

Training LLMs is computationally expensive, and they require extensive processing power and memory. Once trained, these models contain billions of parameters (in the case of GPT-3); CUDA cores on Graphics Processing Units (GPUs) and Matrix cores on Tensor Processing Units (TPUs) process petabytes of data to establish the correlation of human language with the inverse-distance weighting method. This work is usually distributed over cloud-based infrastructures or proprietary hardware. While regarding LLMs development, there are three major processes inherent in the construction of LLMs are **pre-training**, **fine-tuning**, and **instruction-tuning**.

- pre-training process trains a model over large collections of texts to produce representations with a general understanding of language.
- fine-tuning takes this pretrained model and specializes it for specific tasks using a smaller, task-specific dataset.
- Instruction-tuning takes this a level up, allowing the model to follow more complicated instructions, execute several tasks with few or no labeled data.

These processes work hand-in-hand to enable LLMs to demonstrate outstanding flexibility and efficacy over a broad set of NLP tasks [2].

Recently, a few very interesting architectures have taken LLMs further. OpenAI’s GPT-3 and GPT-4 autoregressive models excel at generating coherent, contextually relevant text. Google's PaLM, the Pathways Language Model, is designed to integrate several modalities, while BERT, the Bidirectional Encoder Representations from Transformers model, is designed to comprehend text in both directions for tasks like sentiment analysis and question answering. T5 (Text-to-Text Transfer Transformer) reframes all NLP tasks into text-to-text tasks, allowing the model to be applicable and perform well on a host of tasks. Each of these architectures has made unique contributions to LLM development, enhancing performance in various NLP domains [3].

Further, LLMs have several fundamental architectures that contribute to their high efficiency. Tokenization splits text into smaller, more manageable pieces (tokens) to feed the model as shown in fig2:

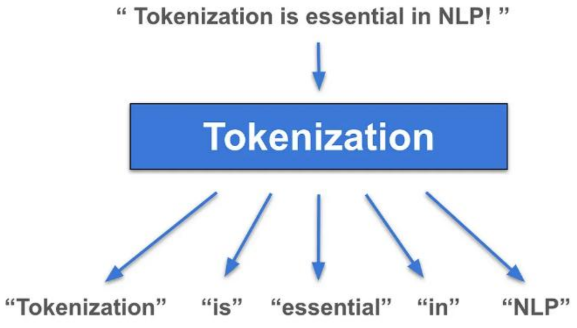

Figure 2 – Tokenization in action

This is where the attention mechanism helps – it enables the model to pay attention to the relevant portions of the input sequence and assign different importance to different tokens according to their context. It is key to understanding complex relationships in language. This is where techniques like layer normalization come into play – normalizing the inputs to each layer helps stabilize the training process and leads to efficient convergence of the model while ensuring consistent performance during the training process [6].

LLMs are used in various domains, enabling new solutions, enhancing existing ones, changing the shape of the industry, and fundamentally advancing what AI systems can do. In the field of healthcare, they help with such things as medical diagnosis, drug discovery, and clinical data analysis [7]. In the context of education, LLMs assist in personalized learning, tutoring and content generation. LLMs are employed by social media platforms to provide better interactivity with users through chatbots, or content moderation. In business, LLMs act as customer service assistants, automate clerical tasks, and produce market analyses. While LLMs have their strengths, there are challenges and limitations to their use. Biased training data can result in biased outputs, since these models are trained on massive amounts of data, possibly even sensitive data. There are ethical issues also in how LLMs are deployed and how their capabilities can be misused. Second, the massive computational and energy requirements for training and deploying these models pose questions about their environmental impact and sustainability [8].

III. Arabic Large Language Models

Arabic is culturally, historically, and intellectually significant as one of the world's most spoken languages. Considering the fact that Arabic has over 400 million speakers worldwide and boasts large dialectal diversity as in terms of Modern Standard Arabic (MSA) and dialects, the need for powerful Arabic natural language processing (NLP) tools is imperative. But, despite the rapid strides in NLP and the advent of large language models (LLMs), the Arabic language continues to be underrepresented in the greater AI space.

The challenges of producing Arabic LLMs comes from how complex Arabic morphology, syntax, diacritics and the availability of high-quality diverse datasets are. Solving these problems need specialized research and building systems that can use models from the various Arabic dialects to Modern Standard Arabic [10].

This chapter presents the state-of-the-art Arabic large language models, including their architectures, training methods, and applications. This analysis provides an insight into how these models tackle the linguistic and technical challenges that are often faced in Arabic NLP, focusing on the most prominent Arabic LLMs. The upcoming sections give a detailed review of important Arabic LLMs, and a further comparative analysis as show in table 1 to show their strengths, weaknesses, and future work horizons.

Jais, developed by G42's Inception, is hailed as the world's most advanced Arabic language model, open-sourced for broad use. Jais is designed to handle a wide range of NLP tasks tailored to the Arabic language. Its architecture is deeply optimized for Arabic dialects and its rich morphological structure. The model's training dataset was specifically curated to improve performance on Arabic-specific linguistic challenges such as diacritization and root extraction. Jais stands out by being one of the first large-scale Arabic LLMs that incorporates both Modern Standard Arabic (MSA) and various Arabic dialects, making it highly versatile for real-world applications in various sectors, from government to business. Jais offers state-of-the-art accuracy in text generation, translation, and sentiment analysis, representing a milestone in Arabic NLP [11].

ALLaM, a recent addition to the family of Arabic LLMs, brings innovations in the integration of multiple modalities, including text and image understanding. Trained on a diverse set of multimodal data, ALLaM is designed to tackle tasks that require both visual and textual comprehension, such as image captioning, text-based image retrieval, and video analysis. By combining the power of LLMs with vision models, ALLaM enhances Arabic AI's ability to understand and generate content that bridges the gap between textual and visual domains. This makes it an exciting development for Arabic NLP applications in industries like e-commerce, media, and education [12].

AceGPT, is a conversational model based on the GPT architecture and fine-tuned for Arabic dialogue generation. Designed for natural, context-aware conversations, AceGPT handles various Arabic dialects with remarkable fluency. It can be applied in a wide range of scenarios, from customer service chatbots to personal assistants, offering a robust solution for

Table 1 – Most Prominent Arabic Large Language Models Comparative Analysis

References	Model	Year Released	Architecture	Parameters (Size)	Dataset Size	Dataset Diversity	Specialization	Applications	Strengths	Weaknesses
N. Sengupta et al. [11]	Jais	2023	Transformer (GPT-based)	13B	395B tokens	Multilingual, 30% Arabic	General Purpose	NLP tasks, Code generation	Open-source, multilingual training, fine-tuning capabilities	High computational cost
M. S. Bari et al. [12]	ALLaM	2024	GPT-based	15B	~ 400B tokens	High-quality diverse Arabic dataset	General Purpose	Conversational AI, Text completion	Focused on Arabic excellence, competitive performance	Resource-intensive training
H. Huang et al. [13]	AceGPT	2023	GPT-based	6B	~300M sentences	Arabic and English corpus	Bilingual capabilities	Translation, Sentiment Analysis	Bilingual approach increases versatility	Limited parameter count restricts handling of complex contexts
E. Almazrouei et al. [14]	Falcon 40B	2023	Decoder-only Transformer	40B	1T tokens	Multilingual with Arabic focus	General Purpose	Chatbots, Content creation	High parameter count, trained on multilingual and high-quality datasets	Energy-intensive training
A. Koubaa et al. [15]	ArabianGPT	2024	GPT-based	~20B	~200B tokens	Arabic-first corpus	General Purpose	Text generation, QA, Summarization	Optimized for Arabic, dialect support	Limited comparative evaluation in diverse NLP tasks
F. Qarah. [16]	SaudiBERT	2024	BERT-based	330M	~20M sentences	Arabic news, Wikipedia, social media	MSA and Saudi Dialects	Named Entity Recognition (NER), Sentiment Analysis	Focused on specific Arabic subdomains	Smaller model size limits contextual understanding
M. Abdul-Mageed et al. [17]	ARBERT	2021	BERT-based	110M	~5B tokens	MSA corpus	General Purpose	Named Entity Recognition, QA	Lightweight and efficient, suited for a range of NLP tasks	Limited coverage of dialectal Arabic
M. Abdul-Mageed et al. [17]	MARBERT	2021	BERT-based	162M	~10B tokens	Diverse Arabic corpus	Dialects and MSA	Social media analysis, Text classification	Robust performance in dialectal Arabic	Outdated architecture compared to newer models

building intelligent Arabic dialogue systems. What makes AceGPT unique is its deep understanding of colloquial Arabic, which enables it to generate human-like responses in everyday conversational settings. It is a breakthrough in the development of Arabic conversational AI, offering promising applications in interactive voice response (IVR) systems and virtual assistants [13].

Falcon 40B, is a significant model in the development of Arabic LLMs, designed by a collaborative effort to handle both English and Arabic tasks effectively. With a massive parameter size of 40 billion, Falcon 40B exemplifies the trend toward larger and more capable models in the Arabic NLP space. The model is particularly noteworthy for its ability to handle complex code-switching between Arabic and English, an increasingly common challenge in real-world Arabic NLP tasks. Additionally, Falcon 40B has been trained with a diversified Arabic dataset, improving its ability to capture nuances in the language. This model's ability to generate coherent, context-aware text across different domains—including legal, technical, and literary texts—positions it as one of the key developments in Arabic LLM research [14].

ArabianGPT, a transformer-based model tailored for Arabic, builds upon the architecture of GPT-2 and GPT-3 but with significant adjustments to better serve the Arabic language's unique syntactical structure. Unlike other GPT variants, ArabianGPT was trained on a vast corpus of Arabic data, encompassing various genres and contexts, such as news articles, literature, and social media. ArabianGPT stands out for its ability to generate high-quality, fluent Arabic text, handling both formal and colloquial forms effectively. The model's design allows it to perform well in diverse NLP tasks, including text generation, summarization, and question answering. ArabianGPT has set a new benchmark for Arabic GPT-based models, demonstrating the power of pre-trained transformers in non-English languages [15].

SaudiBERT, is an Arabic adaptation of the popular BERT architecture, designed to excel in tasks such as named entity recognition (NER) and part-of-speech tagging, which are critical for understanding the intricacies of Arabic grammar and syntax. SaudiBERT was trained on a large corpus of Arabic text, enabling it to understand both Modern Standard Arabic (MSA) and colloquial dialects with high accuracy. Its pre-training approach includes a masked language model objective, allowing it to learn bidirectional context and produce state-of-the-art results in sentence-level understanding. SaudiBERT has been utilized in a wide array of applications, from sentiment analysis to medical text processing, showing its robustness across various domains [16].

ARBERT, is another BERT-based Arabic language model designed to provide enhanced contextual understanding for Arabic NLP tasks. Unlike previous models, ARBERT leverages a variety of Arabic-language resources to train on a more diverse set of corpora, which aids in handling the morphological richness of Arabic. The model's success is attributed to its ability to capture context at both the word and sentence level, facilitating its use in tasks such as machine translation and semantic text similarity. ARBERT's design includes multiple pre-training strategies that improve its understanding of Arabic nuances, such as the handling of prefixes, suffixes, and root words, making it highly effective for syntactically complex languages like Arabic [17].

MARBERT, is a multilingual variant of BERT that includes training on both Arabic and other languages, making it a suitable model for cross-lingual tasks. MARBERT is specifically designed to address the challenges faced by Arabic NLP systems in multilingual environments, where code-switching and mixed-language text are common. The model is pre-trained on a corpus of Arabic and multilingual data, making it particularly effective for tasks such as cross-lingual document classification, sentiment analysis, and entity recognition. Its cross-lingual capabilities set it apart, offering a solution for the growing need to process multilingual content, particularly in regions where Arabic speakers frequently interact with content in other languages like English and French [17].

The Arabic LLM landscape is rapidly evolving, with a diverse array of models catering to different needs and applications. While **Jais** leads with its comprehensive multilingual support and large-scale, open-sourced capabilities, **Falcon 40B** shines with its ability to handle code-switching between Arabic and English, making it a strong contender in mixed-language scenarios. **ArabianGPT** stands out for its tailored approach to Arabic text generation, whereas **SaudiBERT** and **ARBERT** excel in specialized tasks such as NER and text understanding, with SaudiBERT specifically optimized for formal Arabic and ARBERT emphasizing contextual understanding. **MARBERT** brings a cross-lingual advantage, excelling in multilingual environments, while **AceGPT** focuses on generating human-like conversational dialogue in Arabic. Finally, **ALLaM** represents the future of multimodal Arabic models, integrating both textual and visual understanding.

In terms of applications, **SaudiBERT** and **ARBERT** are particularly effective in specialized tasks like sentiment analysis, medical text processing, and machine translation. **Falcon 40B** and **Jais** are more general-purpose, while **AceGPT** targets conversational AI and **ALLaM** pushes the boundaries into multimodal domains.

Despite the impressive progress, challenges such as the scarcity of high-quality Arabic datasets, complex linguistic structures, and the need for more dialect-specific models remain. The future of Arabic LLMs will likely involve further innovations in data augmentation, model efficiency, and cross-lingual capabilities, ensuring that these models better serve the vast and diverse Arabic-speaking world.

IV. Conclusion

Arabic LLMs are a key breakthrough in the field of natural language processing, with transformative implications for Arabic-speaking communities. Despite their increasing relevance and ability to process various NLP tasks, overcoming the many challenges including the language's rich morphology, dialectal proliferation and script variability remain a barrier to their true potential. This survey presented an overview of state-of-the-art Arabic LLMs, explored their architectures, strengths, and limitations, and focused on significant challenges, such as the scarcity of high-quality datasets and computational constraints. The future development of large and broad Arabic datasets, model architectures, and effective training methods must be prioritized in order to further advance Arabic LLMs. Addressing these challenges will not only improve the strength and diversity of Arabic LLMs but will also expand their use in areas such as education, healthcare, and business. Additionally, boosting co-operative research and resource-sharing among the AI community will be crucial for making significant strides. In short, these initiatives will make Arabic LLMs more inclusive, accessible, and impactful, enabling Arabic-speaking communities to leverage cutting-edge AI solutions and participate in shaping the future of AI globally.

ACKNOWLEDGMENT

We would like to extend our heartfelt gratitude to the Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, for providing the necessary resources and fostering an academic environment that greatly contributed to the completion of this paper. Our sincere thanks go to Dr. Mohamed Marie for his exceptional guidance, insightful feedback, and constant support throughout this journey. His expertise and dedication have played a crucial role in the success of this work.

REFERENCES

[1] Vaswani et al., "Attention is all you need," arXiv, Jun. 2017. [Online]. Available: https://arxiv.org/abs/1706.03762 .

[2] Radford *et al.*, "Learning transferable visual models from natural language supervision," *arXiv*, May 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[3] Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.04805.

[4] Radford *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, Jun. 2018. [Online]. Available: https://openai.com/index/language-unsupervised/.

[5] J. Y. L. Bougie *et al.*, "Understanding deep learning requires rethinking generalization," *NeurIPS*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html.

[6] Y. Xiong *et al.*, "Causality in reinforcement learning," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 10695-10704, 2020. [Online]. Available: https://proceedings.mlr.press/v119/xiong20b.

[7] Rajpurkar, P., Chen, E., Banerjee, O., & et al. (2022). AI for Healthcare: Applications of Deep Learning in Medicine. Journal of the American Medical Association, 327(5), 391–399.

[8] M. U. Hadi *et al.*, "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage," *TechRxiv*, DOI: 10.36227/techrxiv.23589741.v1, Year. [Online]. Available: https://www.techrxiv.org/doi/full/10.36227/techrxiv.23589741.v1.

[9] A. Gupta, "Tokenization and its application," *Medium*, [Online]. Available: https://medium.com/@arnavgupta16092004/tokenization-and-its-application-69ce6d90ed13.

[10] M. Mashaabi, S. Al-Khalifa, and H. Al-Khalifa, "A Survey of Large Language Models for Arabic Language and its Dialects," *arXiv*, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.20238.

[11] N. Sengupta *et al.*, "Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models," *arXiv*, Sep. 2023. [Online]. Available: https://arxiv.org/pdf/2308.16149v2.

[12] M. S. Bari et al., "ALLaM: Large Language Models for Arabic and English," arXiv, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.15390.

[13] H. Huang et al., "AceGPT, Localizing Large Language Models in Arabic," arXiv, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.12053.

[14] E. Almazrouei et al., "The Falcon Series of Open Language Models," arXiv, Nov. 2023. [Online]. Available: https://arxiv.org/abs/2311.16867.

[15] A. Koubaa et al., "ArabianGPT: Native Arabic GPT-based Large Language Model," arXiv, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.15313.

[16] F. Qarah, "SaudiBERT: A Large Language Model Pretrained on Saudi Dialect Corpora," arXiv, May 2024. [Online]. Available: https://arxiv.org/abs/2405.06239.

[17] M. Abdul-Mageed et al., "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," arXiv, Jan. 2021. [Online]. Available: https://arxiv.org/abs/2101.01785.