The Landscape of Arabic Large Language Models (ALLMs): A New Era for Arabic Language Technology

SHAHAD AL-KHALIFA, is a Researcher at iWAN Research Group, King Saud University, Saudi Arabia
NADIR DURRANI*, is a Senior Scientist at Qatar Computing Research Institute, Qatar
HEND AL-KHALIFA*, is a Professor at King Saud University and Head of iWAN Research Group, Saudi Arabia
FIROJ ALAM, is a Senior Scientist at Qatar Computing Research Institute, Qatar

ACM Reference Format:
Shahad Al-Khalifa, Nadir Durrani*, Hend Al-Khalifa*, and Firoj Alam. 2025. The Landscape of Arabic Large Language Models (ALLMs): A New Era for Arabic Language Technology. 1, 1 (June 2025), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

The emergence of ChatGPT marked a transformative milestone for Artificial Intelligence (AI), showcasing the remarkable potential of Large Language Models (LLMs) to generate human-like text. This wave of innovation has revolutionized how we interact with technology, seamlessly integrating LLMs into everyday tasks such as vacation planning, email drafting, and content creation. While English-speaking users have significantly benefited from these advancements, the Arabic world faces distinct challenges in developing Arabic-specific LLMs. Arabic, one of the languages spoken most widely around the world, serves more than 422 million native speakers in 27 countries and is deeply rooted in a rich linguistic and cultural heritage [13]. Developing Arabic LLMs (ALLMs) presents an unparalleled opportunity to bridge technological gaps and empower communities. The journey of ALLMs has been both fascinating and complex, evolving from rudimentary text processing systems to sophisticated AI-driven models. This article explores the trajectory of ALLMs, from their inception to the present day, highlighting the efforts to evaluate these models through benchmarks and public leaderboards. We also discuss the challenges and opportunities that ALLMs present for the Arab world.

1 Foundations of Arabic NLP

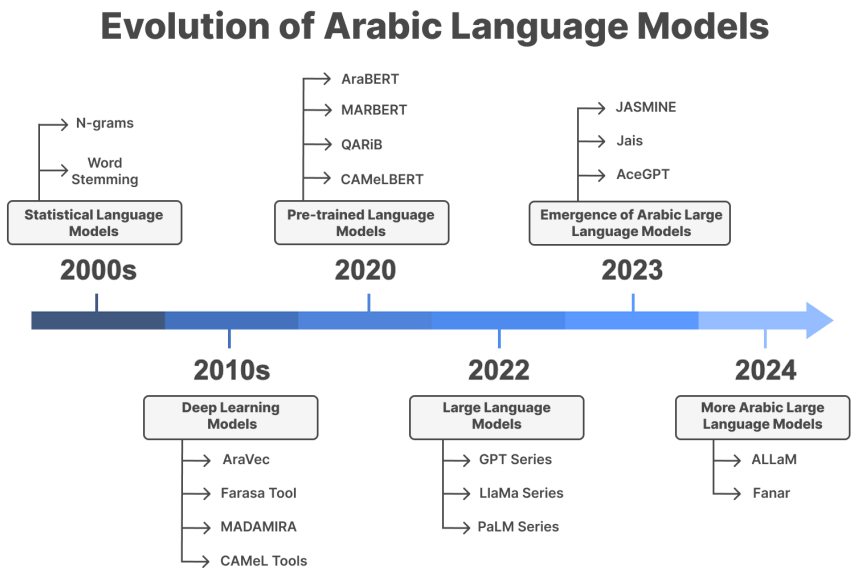
The story of Arabic NLP began in 1985 when pioneers like Sakhr Software[1] tackled the unique challenges posed by Arabic's rich morphology and complex syntax. Early systems, such as morphological analyzers, laid the groundwork for computational tools by addressing tasks like word

*Corresponding authors.
[1]http://www.sakhr.com/index.php/en/

Authors' Contact Information: Shahad Al-Khalifa, shahadalkhalifa90@gmail.com, is a Researcher at iWAN Research Group, King Saud University, Riyadh, Riyadh, Saudi Arabia; Nadir Durrani*, is a Senior Scientist at Qatar Computing Research Institute, Doha, Doha, Qatar, ndurrani@hbku.edu.qa; Hend Al-Khalifa*, is a Professor at King Saud University and Head of iWAN Research Group, Riyadh, Saudi Arabia, hendk@ksu.edu.sa; Firoj Alam, is a Senior Scientist at Qatar Computing Research Institute, Doha, Doha, Qatar, fialam@hbku.edu.qa.

ACM XXXX-XXXX/2025/6-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Fig. 1. Evolution of Arabic Language Models



segmentation and root extraction—critical for processing a language with intricate grammatical structures. As illustrated in Figure 1, the early 2000s saw the rise of statistical models, with techniques like n-grams and word stemming being widely utilized for various NLP tasks such as text classification, information retrieval, and machine translation. These models offered improvements over rule-based approaches but were constrained by limited data availability and struggled to generalize across Arabic's diverse dialects and linguistic complexities. Despite these challenges, statistical methods provided a stepping stone for future innovations, setting the stage for more advanced approaches. The 2010s marked a paradigm shift with the adoption of deep language models, bringing with them powerful tools like word embeddings and Farasa Arabic word processing tool [1],[2] which significantly enhanced the accuracy and adaptability of NLP systems. Techniques like LSTMs and CNNs enabled breakthroughs in sentiment analysis, machine translation, and dialect identification, allowing for more nuanced understanding of Arabic text. However, the inherent diversity of Arabic, including its numerous dialects and morphological richness, continued to pose challenges, underscoring the need for even more sophisticated and scalable models.

2 The Rise of Transformers and ALLMs

Building on the challenges faced by earlier models in handling Arabic's rich linguistic diversity, the advent of transformer architectures in 2017 marked a turning point for NLP. With the introduction of the self-attention mechanism, transformers offered a more robust framework for understanding the complexities of Arabic text, paving the way for a new era of Arabic-specific models. These architectures allowed models to better understand context and relationships within text, significantly improving performance across a wide range of tasks.

[2]https://farasa.qcri.org/

Among the most influential transformer-based models was BERT (Bidirectional Encoder Representations from Transformers), which revolutionized NLP by setting new benchmarks in understanding language nuances. Building on this success, specialized Arabic models like AraBERT [10] and QARIB [2] were developed, significantly improving performance in tasks such as sentiment analysis, named entity recognition, and dialect identification. Additionally, tools such as CAMeL[3] and Farasa[4] offered support for various Arabic language processing tasks. These models became essential tools across a wide range of applications, showcasing the transformative potential of BERT-inspired architectures in Arabic NLP.

Following the release of ChatGPT in 2022, the Arab world saw significant advancements in Arabic language processing. Models such as JASMINE [12] and Jais [26] set new benchmarks for Arabic language understanding and generation. JASMINE excelled in commonsense reasoning and text classification tasks, while Jais showcased advanced capabilities in instruction-response tasks. Jais-chat, a fine-tuned variant, demonstrated remarkable fluency in conversational contexts, and Atlas-Chat introduced optimizations for handling dialectal Arabic, particularly in casual and everyday use cases.

Newer models like AceGPT [17], ALLaM [11], Fanar [27], Peacock [8], and Dallah [7] have expanded the scope of ALLMs further. AceGPT and ALLaM leverages reinforcement learning from AI feedback to enhance instruction-following and contextual understanding. Fanar specializes in understanding Arabic dialects and generative Arabic tasks while also being a multimodal ALLM, capable of handling both text and image-based tasks. Peacock, another multimodal ALLM, integrates visual and linguistic capabilities, demonstrating success in tasks like visual question answering and image captioning.

These advancements stress the growing diversity in ALLMs, with each model tailored to address specific linguistic or cultural challenges. The taxonomy now spans general-purpose models, conversational agents, domain-specialized systems, and multimodal platforms, each contributing uniquely to the Arabic NLP ecosystem. Despite these advancements, challenges persist, such as the need for better handling of dialectal variations and contextual nuances. However, these models represent significant progress toward fully unlocking the potential of Arabic NLP.

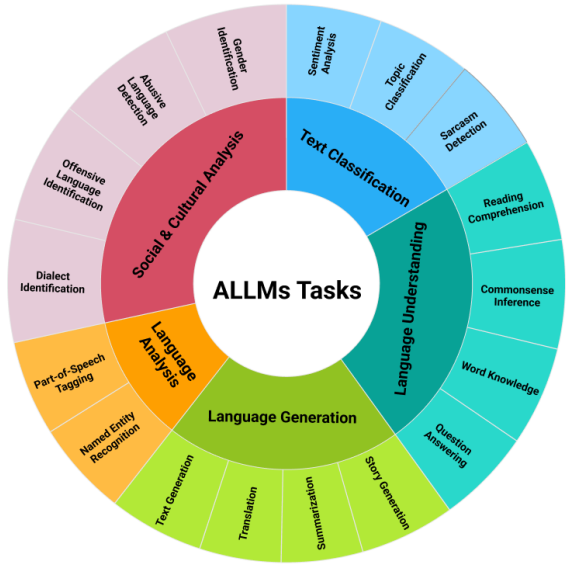
3 Datasets and Benchmarks for ALLMs

Data serves as the cornerstone for building LLMs, serving as their linguistic and knowledge-based foundation. Various forms of datasets, such as those used for pretraining, supervised fine-tuning (SFT), and benchmarking, serve as the foundation for developing LLMs. As illustrated in Figure 2, ALLMs tackle a wide spectrum of downstream tasks, including language generation, understanding, classification, and social-cultural analysis. For ALLMs, a common trend has been adapting datasets originally created for English LLMs through translation. It is mainly due to the scarcity of digital Arabic content needed to train LLMs of substantial size (e.g., several billion parameters). Another notable trend in ALLM development is the combined use of English and Arabic datasets. Additionally, some efforts incorporate code datasets to enhance the model's reasoning capabilities.

Pre-training. For pre-training, the datasets include web content (e.g., Common Crawl), Wikipedia, books, news, and code, covering a wide range of disciplines [20]. Every ALLM development initiative curates, filters, and processes these datasets within their custom pipelines. A common practice across these initiatives is data de-duplication and various types of filtering (e.g., Jais employs rule-based filtering, whereas Fanar uses syntactic, semantic, and model-based filtering). For machine translation, Fanar places greater emphasis on in-house systems for translating English to Modern

[3]https://github.com/CAMeL-Lab/camel_tools
[4]https://farasa.qcri.org/

Fig. 2. Overview of the various capabilities and downstream tasks tackled by ALLMs.



Standard Arabic (MSA) and MSA to dialects.[5] Additionally, efforts across different ALLMs have focused on including dialectal datasets to enhance their capabilities in handling Arabic dialects. Although translating data can introduce Western cultural biases, developing ALLMs with billions of parameters would not have been possible without translated data. Moreover, models trained without translated data tend to exhibit higher training loss [11].

SFT. Instruction tuning is essential for enabling an LLM to engage in dialogue-style interactions with users. Across all ALLM initiatives, the curation of SFT datasets often began with publicly available English datasets (e.g., Super-NaturalInstructions,[6] Natural Questions,[7] P3,[8] xP3[9]), which were subsequently translated into Arabic [26, 27]. Some initiatives also developed their own in-house datasets. For instance, Jais created *NativeQA*, a set of question–answer pairs focused on the UAE and the surrounding region, as well as *SafetyQA* and *DoNotAnswer*, to ensure the model avoids engaging in unsafe conversations, including discussions on self-harm, sexual violence, or identity attacks. For cultural alignment, relevant efforts include the development of CIDER to ensure alignment with Arabic norms [9].

Benchmarking. Benchmarks play a critical role in evaluating language model performance across various tasks. For Arabic, benchmarks have evolved significantly over time, reflecting the increasing sophistication of models. Early benchmarks like AraBench [24] focused on specific tasks such as machine translation, while later benchmarks like ALUE [25], ARLUE [3], and ARGEN [21] offered broader evaluation scopes across multiple tasks. With the rise of ALLMs, specialized benchmarks emerged to assess advanced capabilities: ORCA for text classification [14], Dolphin for

[5]http://mt.qcri.org/
[6]https://huggingface.co/allenai/open-instruct-sni-13b
[7]https://huggingface.co/datasets/google-research-datasets/natural_questions
[8]https://huggingface.co/datasets/bigscience/P3
[9]https://huggingface.co/datasets/bigscience/xP3

natural language generation [22], and benchmarks like AlGhafa [6] and Qiyas [4] for multiple-choice evaluation.

Following this evolution, recent ALLM benchmarks increasingly focus on evaluating reasoning and domain-specific competencies. These benchmarks assess a wide range of capabilities, including *World Knowledge* (OpenAI MMLU,[10] ArabicMMLU[11]), *Common Sense Reasoning* (AraSWAG) [12], MQA-KEAL [5] *Reading Comprehension* (ARCD)[12], *Misinformation* (AraTruthfulQA) [11], and *Cultural Alignment* capabilities (ACVA) [17]. Domain-specific benchmarks like ArabLegalEval [16] and multimodal frameworks like CAMEL-Bench [15] and Peacock [8] have further expanded evaluation possibilities. Manual human evaluation has gained significant attention, with approaches ranging from open-ended interactions to comparative assessments. For instance, Fanar's benchmarking involved over 300 testers from various Arab countries providing feedback, while ALLaM employed comparative evaluation between model responses. While standard NLP dataset evaluation has received less attention, community efforts through resources like AlGhafa, LAraBench and LLMeBench[13] are addressing this gap. Focusing on the dialectal evaluation of various capabilities of ALLMs, benchmarks like AraDICE,[14] leverages a post-edited machine translation approach to curate data for MSA, Egyptian, and Gulf Arabic. Additionally, a cultural benchmark has been introduced alongside AraDICE, further enriching the evaluation.

Despite these advancements, challenges persist, including limited dialectal representation and reliance on machine translations, highlighting the need for more authentic and diverse evaluation frameworks. Figure 3 visualizes the pipeline of training and evaluating ALLMs. It begins with the collection and preparation of pretraining datasets, followed by instruction tuning using SFT datasets. These steps result in an ALLM capable of various tasks, which are then assessed using a range of benchmarks. The figure highlights how pretraining and SFT are sequentially fed into the model, producing outputs that are later evaluated through task-specific benchmarks, ensuring a comprehensive assessment of linguistic capabilities, reasoning, and cultural alignment.

4 Challenges and Opportunities in Building Arabic LLMs

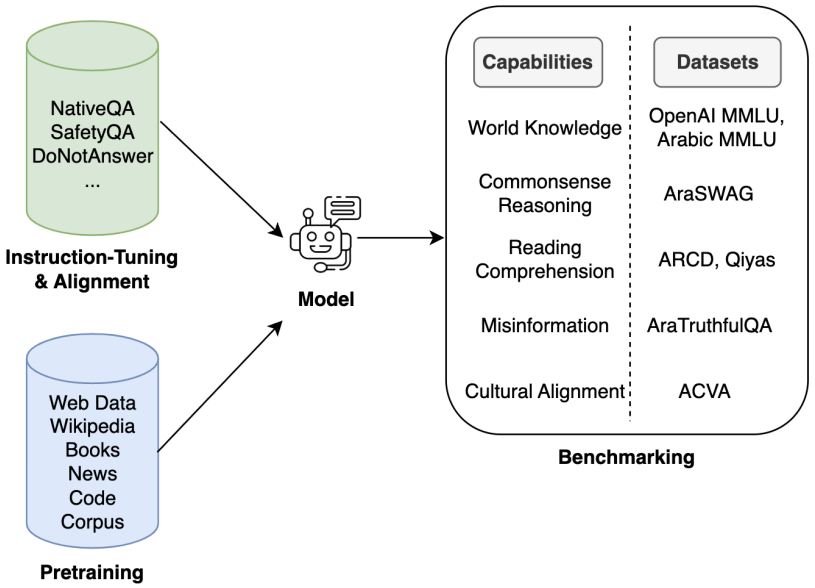
The development of ALLMs faces several interconnected challenges, which must be addressed to unlock their full potential. These challenges include data scarcity, dialectal variation, tokenizer inefficiencies, technical limitations, cultural and safety alignment, and human evaluation constraints. However, despite these obstacles, ALLM development presents transformative opportunities to bridge language technology gaps, foster regional collaboration, and create models that serve the diverse needs of Arabic-speaking communities. Below, we outline key areas where innovation and strategic action can drive progress.

4.1 Data Scarcity

One of the most pressing challenges in developing ALLMs is data scarcity. Arabic lacks abundant, well-annotated resources, particularly for regional dialects and informal language use. A significant portion of Arabic knowledge remains undigitized, making it inaccessible for training large-scale models. Data limitations arise across three critical phases: during pretraining, the limited availability of diverse, digitized text reduces the model's foundational capabilities; during instruction tuning, the scarcity of high-quality, task-specific annotated data hinders adaptability; and during alignment, the lack of culturally nuanced datasets makes it difficult to ensure ethical and safe AI behavior.

[10] https://huggingface.co/datasets/openai/MMMLU
[11] https://huggingface.co/datasets/MBZUAI/ArabicMMLU
[12] https://github.com/husseinmozannar/SOQAL
[13] http://llmebench.qcri.org
[14] https://huggingface.co/datasets/QCRI/AraDiCE

Fig. 3. Pipeline for Training and Evaluation of ALLMs



To address data scarcity, it is crucial to invest in comprehensive data curation initiatives that encompass MSA, Classical Arabic, and regional dialects to ensure a balanced and diverse dataset composition. Expanding the data pool requires digitizing undigitized Arabic knowledge, including manuscripts, oral traditions, and cultural archives. Additionally, developing targeted data pipelines for pretraining, instruction tuning, and alignment can help mitigate these challenges at each stage. Collaborative efforts between academia, industry, and government institutions can further enhance the availability of high-quality Arabic corpora, supporting the development of more robust ALLMs.

4.2 Handling Dialects

Arabic consists of numerous regional dialects with distinct linguistic features. Since current models are primarily trained on MSA, they struggle to understand or generate colloquial or dialectal inputs. This limitation restricts the applicability of ALLMs for real-world use, as users frequently interact using dialects rather than MSA.

Addressing dialectal variations requires the inclusion of diverse dialectal datasets during both pretraining and fine-tuning stages. Leveraging dialect identification models and synthetic data generation techniques can enhance dialectal coverage. Additionally, the development of multidialectal benchmarks would enable more effective evaluation of dialect-handling capabilities in ALLMs. Models fine-tuned on specific dialects, or equipped with zero-shot dialect adaptation techniques, could further enhance robustness in dialectal Arabic understanding and generation.

4.3 Cultural and Safety Alignment

The Arabic language is deeply intertwined with cultural and religious contexts. Models trained on Western-centric or multilingual data often fail to capture these nuances. While English data provides broader knowledge, it also introduces cultural biases that can misalign ALLMs with the values and expectations of Arabic-speaking communities. This misalignment can lead to inappropriate

model outputs, misunderstandings, or even content that contradicts social and ethical norms in Arabic-speaking regions.

Improving cultural and safety alignment requires the adoption of advanced debiasing techniques to mitigate Western-centric biases introduced by English training data. Additionally, ensuring that ALLMs accurately reflect and respect Arabic cultural and religious values demands a stronger representation of culturally relevant content in training datasets. This can be achieved by incorporating region-specific guidelines, enhancing Arabic-specific content filtering, and involving native speakers in evaluation and alignment processes to refine model behavior.

4.4 Multimodality in Arabic

The development of multimodal ALLMs remains relatively underexplored, facing challenges related to data scarcity, dialectal diversity, and cultural misalignment. Most existing multimodal datasets are western-centric, limiting models' ability to interpret Arabic-specific visual and textual content, such as traditional symbols, calligraphy, and region-specific attire. Additionally, dialectal complexity poses difficulties, as models trained on MSA struggle with spoken dialects in videos, advertisements, and social media. Further, cultural biases in training data result in models that fail to align with Arabic social and ethical norms, leading to potential misinterpretations or inappropriate outputs.

Addressing these challenges requires curating high-quality Arabic multimodal datasets that incorporate linguistic and cultural diversity. Initiatives like Peacock [8] and Dallah [7] represent early efforts but require expansion. Dialect-aware multimodal adaptation and culturally informed model alignment can enhance Arabic LLMs' contextual understanding.

5 Discussion

5.1 Comparing Advancements in Arabic and English LLMs

While ALLMs have made notable strides in language understanding and generation, they still lag behind their English counterparts in planning, reasoning, and agentic frameworks. Advanced English LLMs, such as GPT-4o and DeepSeek-V2, exhibit strong multi-step reasoning and problem-solving capabilities, enabling them to tackle complex mathematical, logical, and decision-making tasks. OpenAI's recent o1 model, for example, integrates deliberation mechanisms to enhance reasoning, making it highly effective in structured problem-solving.

In contrast, ALLMs are still developing their commonsense reasoning abilities, as highlighted by efforts like ArabicSense, MQA-KEAL and AraDICE, which introduce benchmarks to improve reasoning performance. Similarly, planning capabilities—essential for breaking down and executing multi-step tasks—remain an area where ALLMs fall short. While navigation-based models like NavGPT have demonstrated some progress in structured task execution with Arabic instructions, they still struggle with complex planning and reasoning-intensive applications.

Finally, agentic frameworks, which enable AI models to autonomously plan and execute actions with minimal human intervention, are still largely unexplored in Arabic NLP. A recent work explored planing and navigation tasks with instructions in both English and Arabic and demonstrate that some multilingual models struggles in reasoning and planning in the Arabic language due to limitations in their reasoning capabilities, poor performance, and parsing issues [18].

Frameworks such as LangChain and AutoGPT, which have facilitated the development of AI-powered agents in English, do not yet have Arabic-adapted equivalents. Addressing these gaps will require dedicated research, dataset expansion, and tailored model training to ensure that ALLMs can match the sophistication of their English counterparts in these critical AI advancements.

5.2 Lessons from Multilingual LLM Initiatives

Efforts to develop LLMs for languages beyond English offer valuable insights that can inform ALLM development. Notable projects include SeaLLM (for Southeast Asian languages) [23] and EuroLLM (for European languages) [19]. These models focus on addressing linguistic bias, enhancing cultural alignment, and optimizing resource efficiency for underrepresented languages.

SeaLLM, for example, extends vocabulary coverage and applies specialized instruction tuning to improve its understanding of Southeast Asian languages while respecting local norms. EuroLLM focuses on multilingual tokenization, balancing language representation, and improving translation across diverse European languages. ALLMs can benefit from these initiatives in several ways. First, extended vocabulary and specialized fine-tuning could enhance Arabic dialect representation. Second, comprehensive data collection and multilingual tokenization would help Arabic LLMs better capture linguistic nuances. Lastly, cultural and legal alignment techniques from these models could improve ethical considerations in Arabic AI systems.

5.3 The Societal Impact of Arabic LLMs

Despite these challenges, ALLMs hold significant potential across multiple sectors, offering transformative applications in education, governance, healthcare, and cultural preservation. In education, they can enhance language learning, bridge literacy gaps, and democratize knowledge accessibility. In governance, they can improve public service delivery, streamline communication, and support e-governance initiatives. In healthcare, they can enable language-specific solutions such as medical transcription and effective patient communication in Arabic. Moreover, in cultural preservation, these models can digitize and preserve endangered dialects and oral traditions, contributing to heritage conservation.

However, realizing this potential requires overcoming several barriers, including the availability of high-quality domain-specific datasets, ethical considerations in AI-driven governance, and ensuring the accuracy and reliability of medical applications. Additionally, the cultural sensitivity of ALLMs remains a key concern, particularly in educational and governmental use cases where misinformation or bias could have significant consequences. Addressing these challenges will require ongoing research, interdisciplinary collaboration, and region-specific adaptation of AI policies to ensure that ALLMs can serve these domains effectively and responsibly.

5.4 Building a Sustainable Arabic AI Ecosystem

Achieving this societal impact requires a strong and sustainable Arabic AI ecosystem. Despite recent progress, limited regional collaboration and infrastructure across Arabic-speaking countries continue to hinder large-scale development. While resource-rich nations have invested in AI research, the absence of a cohesive research network and weak industry-academia integration prevent widespread progress. Additionally, the challenge of attracting and retaining AI talent in the Middle East further limits local expertise, leading to a reliance on external research initiatives rather than homegrown advancements.

A key step in overcoming these barriers is the establishment of a collaborative AI ecosystem. A pan-Arab data consortium could facilitate the sharing of resources, datasets, and infrastructure, allowing researchers across the region to contribute to and benefit from large-scale ALLM development. Strengthening partnerships between governments, academic institutions, and industry leaders can help bridge infrastructure gaps and accelerate innovation.

Equally important is the development and retention of AI talent. To ensure a steady pipeline of skilled researchers and engineers, the region must invest in regional AI education and training programs. Establishing AI research centers, offering competitive funding, and creating career

pathways for AI professionals will encourage local talent to remain and contribute to the Arabic NLP field. Furthermore, fostering international collaborations while ensuring local expertise is developed will be key to advancing ALLM capabilities in the long term.

5.5 Regional Collaboration in AI Development

Regional collaboration has played a crucial role in advancing LLMs for non-English languages, enabling resource-sharing and joint development across multiple nations. The SeaLLM initiative in Southeast Asia and EuroLLM in Europe serve as strong examples of how cross-border cooperation can enhance multilingual AI systems.

SeaLLM was developed through collaboration between Southeast Asian nations, pooling resources, sharing datasets, and co-funding research to create a multilingual model tailored to the region's linguistic diversity. This collective effort ensured high-quality representation for languages with limited AI infrastructure, allowing for more inclusive language technology. Similarly, EuroLLM focuses on supporting the official languages of the European Union, optimizing multilingual tokenization and balancing language representation to create a model that effectively serves diverse linguistic communities.

Arabic AI research has seen significant advancements, with multiple institutions and organizations contributing to the growth of ALLMs. However, unlike coordinated initiatives such as SeaLLM and EuroLLM, most efforts in the MENA region are being developed independently, leading to opportunities for stronger regional collaboration. Given the shared linguistic and cultural heritage across Arabic-speaking countries, a pan-Arab AI consortium could further enhance cooperation by centralizing dataset curation, optimizing computational resources, and aligning research priorities.

By fostering data-sharing agreements, joint model development, and regional funding opportunities, the MENA region can build on its existing progress to drive large-scale ALLM advancements. Strengthening cross-border research networks and leveraging shared linguistic resources would significantly enhance Arabic AI's scalability and impact, ensuring models that better serve the diverse needs of Arabic-speaking communities.

5.6 Ensuring Responsible Development and Evaluation of Arabic LLMs

As ALLMs continue to evolve, their development must be guided by responsible and culturally aware evaluation frameworks. Establishing culturally and linguistically appropriate benchmarks is essential for assessing model performance and ensuring that ALLMs align with the linguistic diversity of the region. To complement this, scalable frameworks for human evaluation should be developed, incorporating feedback from native speakers and domain experts to refine model outputs.

Additionally, technological advancement must be balanced with cultural sensitivity to create inclusive models that effectively serve Arabic-speaking communities. Given the socio-cultural diversity of the Arab world, regional cooperation is necessary to align AI advancements with the ethical, linguistic, and societal needs of different populations. By prioritizing both innovation and cultural awareness, ALLMs can be developed in a way that maximizes their societal benefits while minimizing potential risks.

5.7 Bridging the Research-to-Market Gap in ALLM Deployment

The transition from ALLM research to industrial applications faces significant challenges despite research advances. Most deployments remain limited to narrow tasks rather than comprehensive commercial applications, with models like ALLaM and Jais being exceptions rather than the norm. This gap stems from high computational demands exceeding regional business resources, integration

challenges with existing systems, and the absence of standardized evaluation frameworks for real-world performance metrics.

This limited adoption creates a cyclical problem where sectors that could benefit substantially, such as healthcare, education, and customer service, continue using legacy systems instead of advanced language models. Bridging this gap requires developing deployment-optimized model variants, creating industry-specific benchmarks reflecting real-world requirements, and fostering collaborative ecosystems to transform promising research into commercially viable products serving Arabic-speaking communities.

References

[1] Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A Fast and Furious Segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, John DeNero, Mark Finlayson, and Sravana Reddy (Eds.). Association for Computational Linguistics, San Diego, California, 11–16. https://doi.org/10.18653/v1/N16-3003

[2] Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-Training BERT on Arabic Tweets: Practical Considerations. (2021). arXiv:2102.10684 [cs.CL]

[3] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7088–7105. https://doi.org/10.18653/v1/2021.acl-long.551

[4] Shahad Al-Khalifa and Hend Al-Khalifa. 2024. The Qiyas Benchmark: Measuring ChatGPT Mathematical and Language Understanding in Arabic. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, Mourad Abbas and Abed Alhakim Freihat (Eds.). Association for Computational Linguistics, Trento, 343–351. https://aclanthology.org/2024.icnlsp-1.35/

[5] Muhammad Asif Ali, Nawal Daftardar, Mutayyba Waheed, Jianbin Qin, and Di Wang. 2025. MQA-KEAL: Multi-hop Question Answering under Knowledge Editing for Arabic Language. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 5629–5644. https://aclanthology.org/2025.coling-main.377/

[6] Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa Evaluation Benchmark for Arabic Language Models. In *Proceedings of ArabicNLP 2023*, Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham (Eds.). Association for Computational Linguistics, Singapore (Hybrid), 244–275. https://doi.org/10.18653/v1/2023.arabicnlp-1.21

[7] Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A Dialect-Aware Multimodal Large Language Model for Arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 320–336. https://doi.org/10.18653/v1/2024.arabicnlp-1.27

[8] Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A Family of Arabic Multimodal Large Language Models and Benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12753–12776. https://doi.org/10.18653/v1/2024.acl-long.689

[9] Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged Al-shaibani. 2024. CIDAR: Culturally Relevant Instruction Dataset For Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12878–12901. https://doi.org/10.18653/v1/2024.findings-acl.764

[10] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak

(Eds.). European Language Resource Association, Marseille, France, 9–15. https://aclanthology.org/2020.osact-1.2/

[11] M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLaM: Large Language Models for Arabic and English. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=MscdsFVZrN

[12] El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT Models for Few-Shot Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 16721–16744. https://doi.org/10.18653/v1/2023.emnlp-main.1040

[13] Naaima Boudad, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb. 2018. Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal* 9, 4 (2018), 2479–2490.

[14] AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A Challenging Benchmark for Arabic Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9559–9586. https://doi.org/10.18653/v1/2023.findings-acl.609

[15] Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad S. Khan, Salman Khan, and Rao M. Anwer. 2024. CAMEL-Bench: A Comprehensive Arabic LMM Benchmark. *arXiv preprint arXiv:2410.18976* (2024). arXiv:2410.18976 [cs.CV] https://arxiv.org/abs/2410.18976

[16] Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. 2024. ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 225–249. https://doi.org/10.18653/v1/2024.arabicnlp-1.20

[17] Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, Localizing Large Language Models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 8139–8163. https://doi.org/10.18653/v1/2024.naacl-long.450

[18] Malak Mansour, Ahmed Aly, Bahey Tharwat, Sarim Hashmi, Dong An, and Ian Reid. 2025. Language and Planning in Robotic Navigation: A Multilingual Evaluation of State-of-the-Art Models. *arXiv preprint arXiv:2501.05478* (2025).

[19] Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual Language Models for Europe. arXiv:2409.16235 [cs.CL] https://arxiv.org/abs/2409.16235

[20] Malak Mashaabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2024. A Survey of Large Language Models for Arabic Language and its Dialects. *arXiv preprint arXiv:2410.20238* (2024).

[21] El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. AraT5: Text-to-text transformers for Arabic language generation. *arXiv preprint arXiv:2109.12068* (2021).

[22] El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A Challenging and Diverse Benchmark for Arabic NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1404–1422. https://doi.org/10.18653/v1/2023.findings-emnlp.98

[23] Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs – Large Language Models for Southeast Asia. arXiv:2312.00738 [cs.CL] https://arxiv.org/abs/2312.00738

[24] Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking Dialectal Arabic-English Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5094–5107. https://doi.org/10.18653/v1/2020.coling-main.447

[25] Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic Language Understanding Evaluation.

In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb (Eds.). Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 173–184. https://aclanthology.org/2021.wanlp-1.18/

[26] Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149* (2023).

[27] Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. (2025). arXiv:2501.13944 [cs.CL] https://arxiv.org/abs/2501.13944