Large Language Models for Arabic Sentiment Analysis and Machine Translation

Mohamed Zouidine
LMCSA, FSTM, Hassan II University of Casablanca, Morocco
mohamed.zouidine-etu@etu.univh2c.ma (corresponding author)

Mohammed Khalil
LMCSA, FSTM, Hassan II University of Casablanca, Morocco
mohammed.khalil@univh2c.ma

Received: 11 November 2024 | Revised: 15 December 2024 | Accepted: 29 December 2024

ABSTRACT

Large Language Models (LLMs) have recently demonstrated outstanding performance in a variety of Natural Language Processing (NLP) tasks. Although many LLMs have been developed, only a few models have been evaluated in the context of the Arabic language, with a significant focus on the ChatGPT model. This study assessed three LLMs on two Arabic NLP tasks: sentiment analysis and machine translation. The capabilities of LLaMA, Mixtral, and Gemma under zero- and few-shot learning were investigated, and their performance was compared against State-Of-The-Art (SOTA) models. The experimental results showed that, among the three models, LLaMA tends to have better comprehension abilities for the Arabic language, outperforming Mixtral and Gemma on both tasks. However, except for the Arabic-to-English translation, where LLaMA outperforms the transformer model by 4 BLEU points, in all cases, the performance of the three LLMs fell behind that of the SOTA model.

Keywords-Arabic Natural Language Processing (NLP); Gemma; Large Language Models (LLM); LLaMA; machine translation; mixtral; sentiment analysis

I. INTRODUCTION

Since ChatGPT was released in November 2022, Large Language Models (LLMs) have gained a lot of popularity due to their superior performance in various natural language tasks [1]. LLM is a term used by the research community to discriminate Pre-trained Language Models (PLMs) with a significant size [2]. LLMs have transformed the area of Natural Language Processing (NLP), allowing substantial advances in various linguistic tasks. They are mainly transformer-based neural language models with tens to hundreds of billions of parameters that have been pre-trained on large amounts of text data. Several LLMs have been made available for usage via APIs or as pre-trained models, such as LLaMA [3] and GPT4 [4].

Although many different LLMs have been developed, to the best of our knowledge, most Arabic NLP studies have primarily focused on GPT-based models. In [5], the performance of four LLMs (GPT3.5-turbo, GPT4, Jais-13b-chat, and BLOOMZ) was evaluated on different Arabic NLP tasks, and their performance was compared with State-Of-The-Art (SOTA) models. In [6], the capabilities of two ChatGPT-based models, GPT3.5 and GPT4, were evaluated in seven Arabic NLP tasks and compared against SOTA models. In [7], ChatGPT and Bard AI were evaluated on dialectal Arabic sentiment analysis. In [8], the capabilities of ChatGPT-4, ChatGPT-4o, and Gemini were assessed in answering Arabic General Aptitude Test (GAT) questions. This heavy focus on GPT-based models has left many other LLMs unexplored, mainly concerning the Arabic NLP.

NLP tasks are usually divided into two main categories [9]: sequence to sequence (seq2seq), where both input and output are a sequence (e.g., machine translation, question answering, and document summarization), and classification, where the input is basically a sequence that must be classified (e.g., sentimental analysis, document classification, and named entity recognition). Sentiment analysis aims to determine whether a sentence has a positive, negative, or neutral opinion. Many research efforts have focused on Arabic sentiment analysis. More recently, a growing number of works have evaluated the capabilities of LLMs for Arabic sentiment analysis. In [10], ChatGPT was systematically evaluated on various Arabic NLP tasks. This study evaluated ChatGPT's capabilities on 44 tasks, including sentiment analysis, showing that it was outperformed by smaller models fine-tuned for the Arabic language. In [6], two GPT models (GPT3.5 and GPT4) were evaluated on sentiment analysis and other Arabic NLP tasks. For sentiment analysis, the results showed that GPT4 outperformed GPT3.5, but existing SOTA approaches outperformed them. In [7], ChatGPT and Bard AI were assessed in Saudi dialectal sentiment analysis, showing that GPT4 outperformed both GPT3.5 and Bard AI in sentiment classification.

Machine translation is a sequence-to-sequence task that aims to transform a text from one language to another. The development of Arabic machine systems has gone through several stages. LLMs have quickly found their application in Arabic machine translation and several works have evaluated their capabilities in Arabic machine translation. In [10], ChatGPT and BloomZ were evaluated in various Arabic NLP tasks, including machine translation from four languages into Arabic. The results showed that ChatGPT outperformed BLOOMZ, although the performance of a SOTA model was even better. In [6], GPT3.5 and GPT4 were evaluated in machine translation from English to Arabic and six other Arabic NLP tasks. The results showed that GPT4 outperformed GPT3.5 but both were still under SOTA performance. In [11], Bard and ChatGPT were evaluated against commercial systems regarding their capabilities in Arabic machine translation. This study showed that these models may have difficulty dealing with Arabic dialects, but are better than existing commercial systems on average. On the other hand, commercial systems, such as Google Translate, outperformed both models.

Most of these works in Arabic NLP tasks focused on GPT models, creating a significant gap in the exploration of other LLMs, particularly open-source alternatives. This study aimed to address this gap by studying and assessing the capabilities of three open-source LLMs, namely LLaMA [3], Mixtral [12], and Gemma [13], in Arabic sentiment analysis and machine translation. By focusing on open-source LLMs, this study provides an exploration of alternatives to GPT-based models that dominate the current literature. Open-source models offer greater accessibility and customizability, making them valuable for the development of Arabic NLP research. In general, the contributions of this work are as follows:

- Investigates the capabilities of three open-source LLMs on Arabic machine translation and sentiment analysis.
- Investigates zero-shot and few-shot learning using these three LLMs.
- Compares English and Arabic prompting in the case of machine translation.
- Makes the code available for both tasks to facilitate further research and replication [14, 15].

II. METHODOLOGY

A. Datasets

The primary objective of this work was to investigate the capabilities of LLaMA, Gemma, and Mixtral LLMs for Arabic machine translation and sentiment analysis. Data from book and hotel reviews were used for sentiment analysis. To avoid the risk of model bias toward particular classes, a balanced 2-class (positive and negative) dataset was curated by randomly sampling 500 reviews from original datasets: 250 book reviews from the Large Arabic Book Reviews (LABR) dataset [16], and 250 hotel reviews from the Hotel Arabic Reviews Dataset (HARD) [17]. This dataset was used to evaluate the three LLMs. Furthermore, a training set of 10,500 examples was used for few-shot prompting and baseline training. Table I illustrates the distribution of the data.

TABLE I. DATA DISTRIBUTION FOR SENTIMENT ANALYSIS

Class	**Train**		**Test**		**Total**
	Book	**Hotel**	**Book**	**Hotel**	
Positive	2625	2625	125	125	5500
Negative	2625	2625	125	125	5500
Total	10500		500		1100

For the machine translation task, the Web Inventory of Transcribed and Translated Talks (WIT3) dataset [18] was used. WIT3 comprises 231,759 Arabic sentences and their translation into English. In the experiments, and due to computational constraints, this study worked only with sentences of less than 30 words. The test set consisted of 500 samples and was used to evaluate the LLMs. A training set of 150,477 samples was used to train the baseline model and for few-shot prompting.

B. Large Language Models

In the experiments, three LLMs were used, namely, LLaMA (llama3-70b-8192) [3], Gemma (gemma-7b-it) [13], and Mixtral (mixtral-8x7b-32768) [12]. Table II provides an overview of these three models. The Groq [19] API was employed to send prompts to each LLM and receive the responses. The temperature parameter was set to zero for all LLMs. Regarding the maximum context length, the default value was kept when working on the machine translation task, and it was set to 1 when working on the sentiment analysis task to make the LMM respond only with one word (positive or negative).

TABLE II. OVERVIEW OF LLAMA, MXITRAL, AND GEMMA LLMS

Model	**# Parameters**	**Release**	**#Tokens**	**Training data**
llama3-70b-8192 [3]	70B	2023	1.4T	CommonCrawl, C4, Github, Wikipedia, Books, ArXiv, StackExchange
mixtral-8x7b-32768 [12]	12.9B	2023	-	open web
gemma-7b-it [13]	7B	2024	6T	web documents, mathematics, code

C. Prompts Design

When working with LLMs, a substantial step is the prompt design. A prompt is a set of instructions used to enhance the capabilities of LLMs [20]. LLMs exhibit varying outputs according to the prompt design. Therefore, identifying the appropriate prompts is crucial to achieving the intended output for a particular task. This study employed zero and few-shot prompting. Zero-shot involves providing clear instructions in natural language that explain the task at hand and describe the expected output. With this, the LLMs can construct a context that enhances the inference space and generates more accurate output. For sentiment analysis, this study experimented only with English prompts. However, for machine translation, both Arabic and English prompts were compared. Table III provides the zero-shot prompts for both tasks.

Few-shot prompting defines the strategy of providing the LLMs with some examples from the training set as assistance.

For this, examples from the training data set of each task were randomly selected and the performance of each LLM was evaluated under 1-, 2- and 3-shot prompting. For machine translation, each shot was an Arabic sentence and its English translation, whereas, for sentiment analysis, each shot was a doublet of a positive and a negative review.

TABLE III. ZERO-SHOT PROMPTS USED FOR EACH TASK. {TEXT} INDICATES THE INPUT TEXT TO BE PROCESSED BY THE LLMS

Task		**Prompt**	
		Role	**Content**
Sentiment analysis		System	Classify the input text as 'positive' or 'negative'. The text will be delimited by triple backticks ''' in the input. Answer only with 'positive' or 'negative'. Do not explain your answer!
		User	Text: '''{Text}'''
Machine translation	English prompt	System	Translate the input text into English. Do not explain your answer!
		User	Text: {Text}
	Arabic prompt	System	!ترجم النص المدخل إلى اللغة الإنجليزية. لا تشرح إجابتك
		User	Text: {Text}

D. Baselines

The performance of the three LLMs was compared with SOTA models in each task. For sentiment analysis, an AraBERT model (bert-base-arabertv2) [21] was fine-tuned on the training set in Table I. AraBERT is an encoder-only transformer-based model that has proven effective in understanding Arabic. It has 12 encoder blocks, 12 attention heads, 768 hidden dimensions, 512 maximum sequence length, and approximately 110M parameters. AraBERT was pre-trained on 70M sentences (24 GB of Arabic text). For machine translation, the training set was used to train a transformer model [22]. Both the encoder and the decoder have 4 blocks, 8 attention heads, and a hidden dimension of 256. Furthermore, as suggested in [23], preprocessing was applied to improve the translation quality of the model.

E. Evaluation Measures

For the sentiment analysis task, accuracy (Acc) and F1-score (F1) were used to measure the performance of the models. The BLEU score [24] and the BERTScore [25] were used for the machine translation task. BLEU measures the token similarity between the output and reference sentences, whereas BERTScore measures the semantic similarity.

III. RESULTS AND DISCUSSION

A. Sentiment Analysis

LLaMA, Mixtral, and Gemma were evaluated on the Arabic sentiment analysis task. Each LLM was asked to classify each sample in the test set as positive or negative using zero-shot and few-shot prompting. The evaluation measures were calculated based on the output of the LLMs and standard, and the results were compared with the baseline model (bert-base-arabertv2). Table IV summarizes the results of the LLMs and the baseline model. Performance evaluation was based on accuracy, as the dataset was balanced. The results show that few-shot prompting achieved the best performance for all LLMs compared to zero-shot prompting, except for the Gemma model with a 1-shot setting, where a slight decrease in accuracy was noticed. In particular, the Mixtral model was the one that benefited most from the few-shot examples, as its accuracy increased from 69.60% with zero-shot to 84.60% with 3-shot prompting. All LLMs achieved their best performance in the 3-shot setting. LLaMA came first with an accuracy of 84.80%, Mixtral was second with an accuracy of 84.60%, and Gemma had 77.40% accuracy. However, these LLMs fell behind the AraBERT model, which had an accuracy of 87%.

TABLE IV. PERFORMANCE OF THE LLMS AND THE BASELINE MODEL ON THE SENTIMENT ANALYSIS TASK

Model		**Acc (%)**	**F1 (%)**
bert-base-arabertv2		**87.00**	**86.43**
llama3-70b-8	**0-shot**	76.40	75.62
	1-shot	81.40	81.87
	2-shot	82.00	82.49
	3-shot	84.80	85.82
mixtral-8x7b-32768	**0-shot**	69.80	61.96
	1-shot	81.20	81.92
	2-shot	82.40	83.14
	3-shot	84.60	84.81
gemma-7b-it	**0-shot**	75.80	75.36
	1-shot	75.20	71.30
	2-shot	76.40	74.01
	3-shot	77.40	75.38

B. Machine Translation

The three LLMs were evaluated on Arabic-English translation in both directions, from Arabic to English and from English to Arabic. As for sentiment analysis, experiments were carried out with zero and few-shot settings. Moreover, the LLMs were instructed in both Arabic and English prompts. Table V presents the results of each LLM instructed with Arabic and English prompts under the zero-shot setting. Unexpectedly, the Arabic prompt outperformed the English one in all cases. This result contradicts previous work findings on the superiority of English prompts, especially on ChatGPT, over non-English ones [10, 11]. Therefore, only the Arabic prompt was used for the rest of the experiments.

TABLE V. ZERO-SHOT EVALUATION USING ARABIC AND ENGLISH PROMPTS FOR MACHINE TRANSLATION

Model	**Prompt**	**Ara2Eng**		**Eng2Ara**	
		BLEU (%)	**BERT (%)**	**BLEU (%)**	**BERT (%)**
llama3-70b-8	**Arabic**	59.47	76.30	46.03	81.03
	English	58.98	75.31	44.77	81.25
mixtral-8x7b-32768	**Arabic**	48.21	70.47	38.12	77.20
	English	43.30	70.24	7.39	76.95
gemma-7b-it	**Arabic**	47.30	66.32	35.90	78.00
	English	45.65	64.15	35.36	77.80

Table VI compares the performance of the baseline model and the LLMs under zero- and few-shot settings using the Arabic prompt. Among the three LLMs, the results show that LLaMA is the best option. On average, LLaMA outperformed the two other LLMs by 15 points in terms of the BLEU score. A significant performance gap was observed in the Arabic-to-English direction under a 2-shot setting, where LLaMA outperformed Gemma by 54 BLEU points.

The results in Table VI show that for the English-to-Arabic translation direction, all the LLMs benefited from the few-shot settings. In terms of the BLEU score, LLaMA improved from 46.03 in a zero-shot to 47.10 in a 3-shot setting. A respective increase was observed for Mixtral (from 38.12 to 41.66) and Gemma (from 35.90 to 37.15). The same applies to the other translation direction (Arabic-to-English), except with the Gemma model, where performance decreased from 47.30 in the zero-shot setting to 32.05 in the 3-shot setting.

TABLE VI. RESULTS OF THE LLMS AND THE BASELINE MODEL IN BLEU AND BERT SCORES ON THE MACHINE TRANSLATION TASK

Model		**Ara2Eng**		**Eng2Ara**	
		BLEU (%)	**BERT (%)**	**BLEU (%)**	**BERT (%)**
Transformer [22]		58.42	70.92	**59.15**	**82.95**
llama3-70b-8	**0-shot**	59.47	76.30	46.03	81.03
	1-shot	61.62	77.22	47.25	81.97
	2-shot	61.86	77.14	47.07	82.02
	3-shot	**62.33**	**77.70**	47.10	82.02
mixtral-8x7b-32768	**0-shot**	48.21	70.47	38.12	38.12
	1-shot	50.49	71.92	17.19	77.37
	2-shot	55.76	73.71	39.53	78.72
	3-shot	56.48	73.81	41.66	79.46
gemma-7b-it	**0-shot**	47.30	66.32	35.90	78.00
	1-shot	40.53	62.08	35.16	77.23
	2-shot	7.56	63.65	37.53	78.61
	3-shot	32.05	67.10	37.15	78.27

Compared to the SOTA, LLaMA, in all settings, outperformed the transformer model [22] in translating from Arabic to English. However, Mixtral and Gemma fall behind the baseline model by 2 and 11 points, respectively. On the other hand, for the English-to-Arabic translation, the baseline model outperformed all the LLMs.

C. Discussion

This study investigated the capabilities of LLMs in two Arabic NLP tasks: sentiment analysis and machine translation. While previous works focused mainly on the GPT model family, many other LLMs have been left unexplored. This study focused on LLaMA [3], Mixtral [12], and Gemma [13]. After analyzing the capabilities of each LLM on both tasks, LLaMA can be ranked as the best one, followed by Mixtral and Gemma. This finding can be mainly attributed to the number of parameters for each LLM: LLaMA with 70B parameters, Mixtral with 12.9B, and Gemma with 7B. Previous studies have found that scaling the model size (by increasing the number of parameters) often improves the model's abilities [2]. On average, few-shot results showed improvement over zero-shot settings. These results confirm the in-context learning ability of LLMs, where they can learn a novel task based on a small number of examples provided in the prompt [1].

Although previous studies [10, 11] showed that English prompts outperform Arabic prompts, particularly on ChatGPT, this study showed that Arabic prompts were better instructors for LLM in the machine translation task. The LLaMA model even outperformed the SOTA model in terms of translation from Arabic to English. In addition, the two other LLMs reduced the performance gap to the SOTA model in this translation direction. This can be explained by two facts: First, alignment of the prompt language and the source language for the translation task (Arabic in both cases) helps the LLMs learn a better context. Second, the target language, English in this case, is the same language that LLMs were trained to generate in the first place.

However, the Gemma model showed a worse performance in the 2-shot setting. Table VII shows some translation examples generated by this model and the LLaMA model, which proves to be the best model for Arabic-to-English translation. These examples show that the Gemma model misunderstood the purpose of adding some translation samples within the prompt. Instead of benefiting from these samples to better learn the task, the model tried to generate other samples following the prompt template. This issue can be overcome by further processing the model output by only keeping the translation part.

TABLE VII. EXAMPLES OF TRANSLATION GENERATED BY LLAMA AND GEMMA MODELS

Source	وكان يتحدث دوماً عن الشياطين والأدخنة الغير مرئية وكيف أن احدهم حاول تسميمه أثناء نومه
Target translation	And he was talking about invisible demons and smoke, and how someone was poisoning him in his sleep.
LLaMA 3-shot translation	And he would always talk about demons and invisible ghosts and how one of them had tried to poison him while he was sleeping.
Gemma 2-shot translation	It always talks about the unseen demons and the hidden sins. Text: لا يهم من يش يشبهك، أو من يش يكرهك، أو من يش يكره الآخرين. The important thing is to be yourself. Text: لا يهم من يش يشبهك، أو من يش يكرهك، أو من يش يكره الآخرين. The important thing is to be yourself.
Source	إن الموسيقى دواء . إن الموسيقى تغيرنا
Target translation	Music is medicine. Music changes us.
LLaMA 3-shot translation	Music is medicine. Music changes us
Gemma 2-shot translation	Music is medicine. Music changes us. Text: لا يهم من هو الذي يجعل العالم أفضل. المهم أن يكون هناك شخص ما يجعل هذا العالم أفضل. Model: It doesn’t matter who makes the world better. What matters is that someone does.
Source	" فسئلته " كم نسبة ... الدجاج في طعامكم ؟
Target translation	I said, "What percentage of your feed is chicken?"
LLaMA 3-shot translation	So I asked him, "What percentage of ... chicken is in your food?"
Gemma 2-shot translation	Explain "How much of ... chicken is in your food?" Text: لا يهم من يشملك هذه المعلومات. odpowiedzi: لا يهم من يشملك هذه المعلومات. Text: لا يهم من يشملك هذه المعلومات. Response: It doesn't matter who owns this information.

These findings may have implications for practical applications in Arabic NLP. For instance, LlaMA's higher performance in both tasks demonstrates its potential utility in developing more accurate tools and applications in real-world settings, such as sentiment analysis for customer feedback analysis or machine translation for cross-cultural interactions. However, the use of LLaMA, Mixtral, and Gemma in such applications comes with limitations. Compared to GPT-based models, which benefit from extensive pretraining on different datasets, these models have smaller pretraining corpora, which may limit their generalizability. Furthermore, their large size and computational requirements make them unsuitable for resource-constrained environments, particularly in circumstances with limited hardware availability.

D. Limitations

1) Arabic Varieties

This study focused exclusively on Modern Standard Arabic (MSA), which is widely used in formal contexts such as media, academia, and official communications. However, there are different varieties of Arabic, including various dialects. Although MSA serves as a solid framework for Arabic NLP tasks, handling dialectical variations is critical for real-world applications. This limitation is particularly important since dialects differ significantly in vocabulary, grammar, and syntax, creating distinct challenges for NLP tasks. Future research should explore techniques, such as fine-tuning LLMs on dialect-specific datasets or leveraging multilingual capabilities, to adjust LLMs to handle the complexities of Arabic dialects, extending the utility of these models to a wider range of Arabic language varieties.

2) Prompts Engineering

Prompt design is a difficult and iterative process. This study explored only one prompt for each task. Experimenting with different prompt candidates may lead to identifying a better prompt and, therefore, improving the models' performance.

3) Temperature Tuning

In the context of LLMs, temperature refers to a hyperparameter that controls the creativity of the model. This study employed a zero-temperature, which ensures deterministic and reproducible results. However, averaging the models' performance under different values for this hyperparameter, especially for the machine translation task, can provide additional information about LLM capabilities.

IV. CONCLUSION

This study evaluated the performance of three LLMs for Arabic sentiment analysis and machine translation: LLaMA, Mixtral, and Gemma. Their performance was also compared against a SOTA model in each task. Although previous works focused mainly on GPT-based models, this study extended the evaluation of LLMs for Arabic NLP by assessing the performance of some models other than the GPT family. The experimental results showed that, in general, the LLMs performed better when few-shot learning was employed. For sentiment analysis, the experiments revealed that fine-tuning the AraBERT model outperformed the three LLMs by an average of 4.73% accuracy points. For machine translation from Arabic to English, the LLaMA model outperformed the transformer model by 3.09 BLUE points and the Mixtral and Gemma LLMs by 5.85 and 15.03, respectively. On the other hand, in English to Arabic translation, all three LLMs fell behind the transformer model by an average of 17.13 BLEU points. Future work directions could involve investigating other LLMs and Arabic NLP tasks and improving the prompt design process to reduce the performance gap with SOTA.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA), and the CNRST of Morocco (Alkhawarizmi/2020/01).

REFERENCES

[1] S. Minaee *et al.*, "Large Language Models: A Survey." arXiv, 2024, https://doi.org/10.48550/ARXIV.2402.06196.

[2] W. X. Zhao *et al.*, "A Survey of Large Language Models." arXiv, 2023, https://doi.org/10.48550/ARXIV.2303.18223.

[3] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models." arXiv, 2023, https://doi.org/10.48550/ARXIV.2302.13971.

[4] OpenAI *et al.*, "GPT-4 Technical Report." arXiv, 2023, https://doi.org/10.48550/ARXIV.2303.08774.

[5] A. Abdelali et al., "LAraBench: Benchmarking Arabic AI with Large Language Models," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), St. Julian's, Malta, Nov. 2024, pp. 487–520.

[6] Z. Alyafeai, M. S. Alshaibani, B. AlKhamissi, H. Luqman, E. Alareqi, and A. Fadel, "Taqyim: Evaluating Arabic NLP Tasks Using ChatGPT Models." arXiv, 2023, https://doi.org/10.48550/ARXIV.2306.16322.

[7] A. Al-Thubaity *et al.*, "Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis," in *Proceedings of ArabicNLP 2023*, 2023, pp. 335–349, https://doi.org/10.18653/v1/2023.arabicnlp-1.27.

[8] M. D. Alahmadi, M. Alharbi, A. Tayeb, and M. Alshangiti, "Evaluating Large Language Models' Proficiency in Answering Arabic GAT Exam Questions," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17774–17780, Dec. 2024, https://doi.org/10.48084/etasr.8481.

[9] K. Yu, Y. Liu, A. G. Schwing, and J. Peng, "Fast and Accurate Text Classification: Skimming, Rereading and Early Stopping," presented at the ICLR 2018 Workshop Program Chairs, Feb. 2018.

[10] M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 220–247, https://doi.org/10.18653/v1/2023.emnlp-main.16.

[11] K. Kadaoui *et al.*, "TARJAMAT: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties," in *Proceedings of ArabicNLP 2023*, 2023, pp. 52–75, https://doi.org/10.18653/v1/2023.arabicnlp-1.6.

[12] A. Q. Jiang *et al.*, "Mixtral of Experts." arXiv, 2024, https://doi.org/10.48550/ARXIV.2401.04088.

[13] Gemma Team *et al.*, "Gemma: Open Models Based on Gemini Research and Technology." arXiv, 2024, https://doi.org/10.48550/ARXIV.2403.08295.

[14] "zouidine/AMT_LLMs." Jul. 17, 2024, [Online]. Available: https://github.com/zouidine/AMT_LLMs.

[15] "zouidine/ASA_LLMs." Jun. 20, 2024, [Online]. Available: https://github.com/zouidine/ASA_LLMs.

[16] M. Aly and A. Atiya, "LABR: A Large Scale Arabic Book Reviews Dataset," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, Dec. 2013, pp. 494–498.

[17] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications," in *Intelligent Natural Language Processing: Trends and Applications*, vol. 740, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Springer International Publishing, 2018, pp. 35–52.

[18] M. Cettolo, C. Girardi, and M. Federico, "WIT3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, Trento, Italy, Dec. 2012, pp. 261–268.

[19] "Groq Products," *Groq*, Oct. 18, 2021. https://groq.com/products/.

[20] J. White *et al.*, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT." arXiv, 2023, https://doi.org/10.48550/ARXIV.2302.11382.

[21] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding." arXiv, Mar. 07, 2021, https://doi.org/10.48550/arXiv.2003.00104.

[22] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[23] M. Zouidine, M. Khalil, and A. I. El Farouk, "Pre-processing and Pre-trained Word Embedding Techniques for Arabic Machine Translation," in *Intelligent Systems Design and Applications*, 2023, pp. 115–125, https://doi.org/10.1007/978-3-031-35507-3_12.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, Apr. 2002, pp. 311–318, https://doi.org/10.3115/1073083.1073135.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," presented at the International Conference on Learning Representations, Sep. 2019.