

Large language models-based metric for generative question answering systems

Hazem Abdel Azim, Mohamed Tharwat Waheed, Ammar Mohammed

School of Computing and Digital Technologies, ESLSCA Univeristy, Cairo, Egypt

Article Info

Article history:

Received Mar 20, 2024

Revised Aug 13, 2024

Accepted Aug 30, 2024

Keywords:

Evaluation metrics

Generative question answering

Large language models

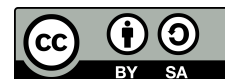
Likert-scale scoring

Zero-shot prompting

ABSTRACT

In the evolving landscape of text generation, which has advanced rapidly in recent years, techniques for evaluating the performance and quality of the generated text lag behind relatively. Traditionally, lexical-based metrics such as bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), metric for evaluation of translation with explicit ordering (METEOR), consensus-based image description evaluation (CIDER), and F1 have been utilized, primarily relying on n-gram similarity for evaluation. In recent years, neural and machine-learning-based metrics, like bidirectional encoder representations from transformers (BERT) score, key phrase question answering (KPQA), and BERT supervised training of learned evaluation metric for reading comprehension (LERC) have shown superior performance over traditional metrics but suffered from a lack of generalization towards different domains and requires massive human-labeled training data. The main contribution of the current research is to investigate the use of train-free large language models (LLMs) as scoring metrics, evaluators, and judges within a question-answering context, encompassing both closed and open-QA scenarios. To validate this idea, we employ a simple zero-shot prompting of Mixtral 8x7 B, a popular and widely used open-source LLM, to score a variety of datasets and domains. The experimental results on ten different benchmark datasets are compared against human judgments, revealing that, on average, simple LLM-based metrics outperformed sophisticated state-of-the-art statistical and neural machine-learning-based metrics by 2-8 points on answer-pairs scoring tasks and up to 15 points on contrastive preferential tasks.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Hazem Abdel Azim

School of Computing and Digital Technologies, ESLSCA Univeristy

Cairo, Egypt

Email: hazem.abdelazim@eslsc.edu.eg

1. INTRODUCTION

Question answering (QA), dating back to the seminal work of Hirschman and Gaizauskas [1], has long aspired to equip computer systems with the ability to furnish accurate and pertinent responses to posed inquiries, leveraging either predefined context or curated knowledge bases. QA systems are typically decomposed into two key components [2]: a retriever and a reader. The retriever's function is to search among an extensive collection of passages and retrieve the most relevant passage given the query. The reader's function is to comprehend the passage and answer the query from the given passage or set of passages retrieved. The current research focuses on the reader component, namely, the reading comprehension (RC) task, and in particular, different metrics are used to measure the performance of the RC task.

Traditional RC-QA systems [3] rely on extractive "span-based" QA, whether in a closed or open domain. Span-based extractive QA means giving a passage and a question, and the task of the AI model is to extract the answer from within the passage in a span [start-end] indices. Accordingly, the metrics used to evaluate those systems were designed to capture lexical-based similarities between the model answers and the ground-truth ideal answers created by human annotators. Recent QA systems are generative [4], sometimes known as "abstractive" QA, cater for generating a "semantically" correct answer from within the passage, and do not necessarily capture a span of answers. More advanced metrics are required to evaluate those generative responses.

Generally, the current landscape of QA metrics can be categorized into three broad categories: lexical-statistical metrics, embedding-based metrics, and neural bidirectional encoder representations from transformers (BERT)-based models [5]. Lexical-statistical metrics are the more conventional metrics used for several years. They rely on token matching, whether exact match (EM) or relaxed (F1-score), with different n-gram variants. These metrics include bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), metric for evaluation of translation with explicit ordering (METEOR), consensus-based image description evaluation (CIDER), EM, and F1-score. BLEU is precision-centric and widely used in evaluating translation tasks [1]; ROUGE is recall-centric and commonly used in summarization tasks [6]. Although these traditional metrics have provided acceptable performance for span-based extractive QA systems, they suffer from critical drawbacks as they do not capture semantic features in the tokens.

On the other hand, the semantic capturing aspect has been addressed in the second category of embedding-based metrics, which utilize token embeddings to provide a more nuanced similarity score and mitigate the limitations of lexical metrics [7], [8]. While these metrics offer flexibility and improve QA scoring compared to lexical metrics, they nonetheless encounter challenges adapting to specific contexts due to their static nature, failing to consider the contextual nuances of tokens within questions or answers [9]. For instance, a word like "bank" would yield the same static embedding vector in different contexts, such as "depositing a paycheck in the bank" and "crossing the river bank". Those limitations were handled in the third category: Neural BERT-based models, using different variants of BERT architectures [10], to capture contextualized embeddings, which showed superior performance correlating with human judgments compared to other categories. Several models were reported recently like BERTscore [8], which relies either on words or contextualized embeddings and cosine similarity to generate a numeric score. Bilingual evaluation understudy with representations from transformers (BLEURT) [11] is a refined version of BERTscore that empowers augmented synthesized data to train the model. Another advanced version uses BERT to train the model to learn certain critical weights for each token, like in key phrase question answering (KPQA) models [9]. The refinement here is that instead of treating tokens in the model answer and ground truth gold answers equally, they are weighted based on their importance in answering the question. Standard BERT architecture is followed by a softmax classifier layer to generate the weights for each token, and those weights are incorporated into conventional metrics like ROUGE, BLEU, and the BERTscore metric. A BERT-based direct supervised learning approach adopted by [12], which learns the required rating directly using massive training labelled data. The model is called learned evaluation metric for reading comprehension (LERC). This model is based on BERT architecture that has undergone fine-tuning based on human judgment scores. LERC takes as input a passage (context), question, reference, and candidate, and the output score measures the accuracy of the candidate as compared to the ground truth human judgement. The preceding neural-BERT systems have demonstrated significantly superior performance to traditional lexical and static embedding metrics, especially within the domains for which they are trained. However, they are hindered by a complex training procedure, necessitating costly manual annotation of samples due to their reliance on large amounts of human-annotated data for training. Additionally, they exhibit limited generalization across various domains, and there is still more room for improvements on out-of-distribution data, particularly on contrastive pairs [12].

Recently, a fourth category based on using large language models (LLMs) in scoring as a judge shows significant promise compared to the preceding three categories. Utilizing LLM with carefully crafted prompts [13] has demonstrated remarkable success in various tasks, both within academic benchmarks [14] and real-world settings [15]. However, to our knowledge, no published research has yet reported on using LLMs as a scoring agent for RC tasks in a QA context to mimic the human judgments on a Likert scale and the simpler binary tasks for correct/incorrect answers. Thus, this research's primary contribution lies in exploring a fourth category, employing GPT LLMs and zero-shot prompting to assess the correlation between model scores and human judgments compared to other state-of-the-art QA metrics. We conducted experiments using

this proposed approach on 10 state-of-the-art datasets [12], [16]–[23], encompassing diverse domains and QA styles. The first eight datasets feature human scores based on a 5-Likert scale, while the last two consist of binary openQA datasets. In these latter datasets, the task assigned to the LLM is to determine whether the answer is correct compared to a gold (ground truth) answer. Through careful prompting, we directed the LLM to execute the scoring task. This discriminative task is particularly challenging, especially with the 5-Likert scale judgment, as the LLM must distinctly differentiate between closely labelled categories between 1 and 5.

The rest of the paper is organized as follows: in section 2, we discuss the research method, in subsection 2.1., we discuss the proposed LLM mixtral-based scoring model and subsection 2.2. describes the datasets employed in this research, providing a foundation for the empirical analysis. The findings from our empirical analysis across all datasets are presented and discussed in section 3. Finally, section 4 concludes the paper with a summary of key insights and conclusions drawn from the research.

2. METHOD

In this section, we describe the methodology used in our research to develop and evaluate the proposed LLM-based metric, as well as the the datasets used for evaluation.

2.1. Large language model-based metric

The proposed metric in our research is based on capitalizing on open-source Mixtral 8x7B LLM reasoning capabilities as a scoring machine, which we will call LLM-Mixtral. This research answers a hypothesis about whether simple prompt-based zero-shot open-source LLM can outperform all state-of-the-art existing metrics we have covered in the previous sections and correlate better with human judgements. We formally design a prompt containing a question, q, a gold (reference) answer and a model generated answer AI-generated answer as (1):

$$\hat{y} = M_{\text{LLM}}(\text{prompt}) \quad (1)$$

The predicted score \hat{y} is then compared to the corresponding human judgements example of prompt that can be applied as an input to (12): "Here is a question, a set of golden answers (split with /), an AI-generated answer. Can you judge whether the AI-generated answer is correct according to the question and golden answers, answer Yes or No."

We used several prompts depending on the task and the dataset used. An example is shown in Figure 1, to instruct the LLM-Mixtral to generate a human-like judgement on how well the hypothesis candidate answer is aligned semantically with the ground truth reference answer. The predicted judgment \hat{y} could be on a Likert-scale from 1-5 for the first eight datasets or binary judgement (correct/incorrect) for the last two datasets, as will be explained in the next section.

LLM-prompt
<p>Please score the hypothesis about the given question and reference statement. Assign a Likert score from 1 to 5, where one indicates that the hypothesis is completely irrelevant or incorrect compared to the question and reference, and five indicates that the hypothesis is highly relevant and accurate in the context of the question and reference.</p> <p>Question: {question} Reference: {reference} Hypothesis: {answer}</p> <p>Likert Scale: 1 - Completely irrelevant or incorrect 2 - Mostly irrelevant or incorrect 3 - Ambiguous 4 - Mostly relevant and accurate 5 - Highly relevant and accurate</p> <p>Consider carefully how well the hypothesis addresses the question and aligns with the reference, and assign the most appropriate score between 1 and 5 accordingly.</p>

Figure 1. Zero-shot prompt applied to LLM-Mixtral model

2.2. Datasets used in question answering evaluation

Numerous benchmark datasets are available in the literature for evaluating QA. We selected datasets that were deployed in the same research setting, by comparing the metric with human judgments mostly on a Likert scale from 1 to 5 where 5 is the most relevant model answer compared to the gold - ground truth answers. Datasets utilized in our experiments is summarized in Table 1.

Table 1. Summary of datasets

Dataset	Description	References
NarrativeQA	Benchmark for GenQA metrics, with short answers averaging 4.7 words.	[17], [24]
SemEval	Used for GenQA metrics, with very short answers averaging 2.5 words.	[16], [17]
MS-MARCO	Contains human judgments for model-generated answers, known for longer responses.	[17]
AVSD	Collected human judgments on model responses, with longer and complex answers.	[17]
MCScript	Evaluates reasoning within stories for children, assessing comprehension skills.	[16]
CosmosQA	Focuses on commonsense reasoning through everyday blogs, assessing real-world reasoning.	[18]
SocialIQA	Evaluates social reasoning from knowledge-base passages, focusing on social interactions.	[19]
Quoref	Assesses coreferential reasoning within Wikipedia articles for language comprehension.	[20]
Contrastive pairs	Consists of contrastive answer pairs for evaluating models against human judgments.	[12]
EVOUNA (NQ, TQ)	Aggregates outcomes from various Open-QA models on NQ and TQ datasets.	[21]–[23]

3. EXPERIMENTAL RESULTS

We tested our proposed LLM—Mixtral metric on ten different datasets and compared it with all the methods covered in section 3. We chose Mixtral 7 B because very little research has tackled this problem using open-source models, and most of the related research in this area used closed GPT models (OpenAI and Claudera), which are paid services. The second reason is that Mixtral 8x7 B is one of the top performing open source models [25] on general tasks with relatively fewer parameters than many open source LLMs. Mixtral notably exhibits superior performance, matching or surpassing Llama 2 70B and GPT-3.5 on public tasks, with remarkable results in mathematics, code generation, and multilingual tasks. So, we investigate the model's performance in this challenging closed specific task of Likert-scale scoring of QA-generated answers versus human judgments. The third reason is that open source models, for privacy reasons, are more appealing for some government and private sector enterprises where the criticality of data privacy is very high, and they prefer to have their data on-premises, which is achievable using open source models.

3.1. Experiment I: comparison with key phrase question answering metric

We benchmarked LLM-Mixtral against the datasets used in [9]. Based on the LLM - prompt in Figure 1, the resulting output is parsed to get the Likert scale judgment from 1 to 5. The question, candidate answer, and ground truth reference are grabbed and applied to the LLM model for each dataset. The Pearson correlation coefficient is computed for the LLM-Mixtral and human judgments.

As depicted in the results in Table 2, the simple proposed model LLM-Mixtral outperforms all metrics on average and for 3 out of four datasets. Test sets are used in the comparative study. The lexical metrics are far behind in terms of correlation with human judgments. The KPQA provides a relatively good correlation but, on average, is somewhat less than the simple LLM-Mixtral metric.

Table 2. Benchmarking LLM-Mixtral against Lexical and KPQA metrics [18]

Metric	MS-MARCO	AVSD	Narrative-QA	Sem-Eval	Average
BLEU-1	0.349	0.58	0.634	0.359	0.4805
BLEU-4	0.193	0.499	0.258	-0.035	0.22875
ROUGE-L	0.309	0.585	0.707	0.566	0.54175
METEOR	0.423	0.578	0.735	0.543	0.56975
CIDER	0.275	0.567	0.648	0.429	0.47975
BERTScore	0.463	0.658	0.785	0.63	0.634
BLEU-1-KPQA	0.675	0.719	0.716	0.362	0.618
ROUGE-L-KPQA	0.698	0.712	0.774	0.742	0.7315
BERTScore-KPQA	0.673	0.729	0.782	0.741	0.73125
LLM-Mixtral	0.691	0.749	0.818	0.777	0.75875

3.2. Experiment II: comparison with learned evaluation for reading comprehension metric

We benchmarked LLM-Mixtral on different datasets and different models presented by the authors in [12]. The results are shown in Table 3. BERT semantic textual similarity benchmark (STS-B) is a BERT-based model fine-tuned on the sentence similarity task, STS-B [26]. LERC, as described in section 3, is a LERC based on supervised fine-tuning and a 40k+ dataset. The proposed LLM-Mixtral metric outperformed BERT STS-B and LERC on average, except for the Quoref dataset. LERC, the second performer after LLM-Mixtral, was competitive in two datasets, CosmosQA and Quoref. In general, it produced an average correlation of 0.744, while our proposed LLM-Mixtral achieved a moderate correlation of 0.882, a remarkable increase of 14 points.

Table 3. Benchmarking LLM-Mixtral against Lexical and LERC metrics [12]

	Narrative QA	MCScript	CosmosQA	Quoref	Average
BLEU_1	0.472	0.260	0.670	0.578	0.460
METEOR	0.615	0.502	0.711	0.716	0.611
ROUGE-L	0.495	0.297	0.701	0.604	0.490
BERTScore	0.534	0.194	0.779	0.286	0.447
BERT STS-B	0.686	0.449	0.789	0.750	0.638
LERC	0.738	0.694	0.824	0.741	0.744
LLM-Mixtral	0.884	0.795	0.824	0.735	0.822

3.3. Experiment III: comparing LLM-Mixtral with LERC on out-of-distribution datasets

Although LERC achieved good performance on some of the datasets, investigating the training and test datasets used in LERC showed that the data is statistically biased, which provides doubt on the generalization capabilities of LERC. As shown in Figure 2, the distribution of the training and test data has similar biases. To verify that we applied LERC on totally unseen out-of-distribution data from dataset 1, namely Microsoft machine reading comprehension (MSMARCO) and audio-visual scene understanding (AVSD). The correlation results in Table 4 showed a lower performance as expected compared to LLM-Mixtral, which is one of the critical advantages of using LLM-Mixtral based metric as it's dataset and domain agnostic, and is not influenced by a training distribution biases.

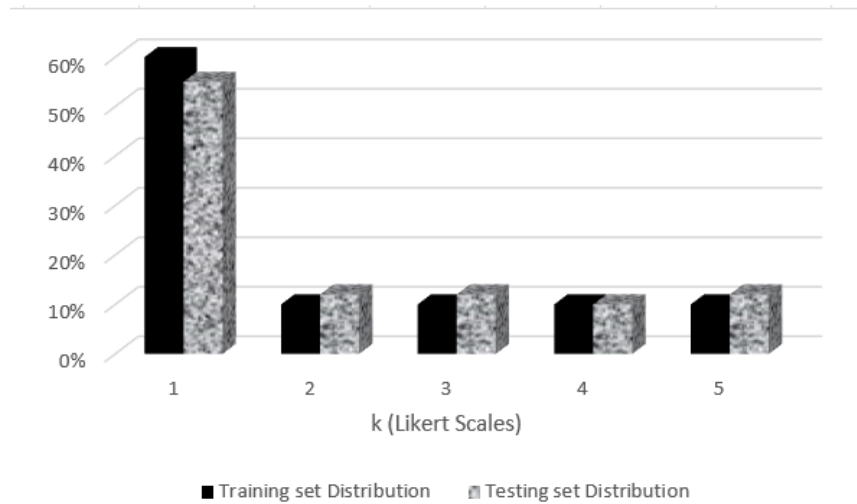


Figure 2. Biases in the training and test sets used in LERC

Table 4. Comparison of models LERC and LLM-Mixtral on MS-MARCO and AVSD datasets [17]

Model	MS-MARCO	AVSD
LLM-Mixtral	0.691	0.749
LERC	0.601	0.621

3.4. Experiment IV: contrastive scoring task

The experiment was conducted on the contrastive pairs dataset [12]. This dataset assesses the preference between two possible answers. The results are summarized in Table 5, with the accuracy of the results reported. Lexical-based metrics performed poorly, as expected since the contrastive pairs were designed to have similar token overlap with the reference. On the other hand, the sentence similarity model STS-B outperformed others, likely because it generalizes beyond token overlap. The LERC model, presented in this research setting, achieved the best results, with an average accuracy of 80%. Our proposed LLM-Mixtral metric, earned an impressive average accuracy of 95%. This result supports our hypothesis that LLM-based models outperform conventional and state-of-the-art models in this scoring task.

Table 5. Results of contrastive pairs experiment on datasets [12]

Metric	NarrativeQA	MCScript	CosmosQA	SocialQA	Avg.
BLEU-1	53	54	52	55	53.5
ROUGE-L	53	57	53	53	61.2
METEOR	60	62	57	53	54
BERTScore	70	58	74	62	66
BERT STS-B	70.6	70	59.3	66.6	66.6
LERC	80	87.3	72.6	81.3	80.3
LLM-Mixtral	96	94	96	94	95

3.5. Experiment V: open question answering datasets

The previous experiments were conducted using closed QA, where the answer text is provided within a given context passage. In the current experiment, we aim to evaluate the metric on a more challenging task on commonly used OpenQA datasets, namely natural questions (NQ), Trivia question answering (TQ), and event and opinion understanding in natural language (EVOUNA) datasets. LLM-Mixtral outperformed BERTScore applied on the same dataset as shown in Table 6, which summarizes the relative performance of LLM-Mixtral over other state-of-the-art models. The best-performing neural-BERT model is chosen for each subset of the ten datasets used in the experimentation. The incremental difference between the proposed LLM-Mixtral model and the best-performing neural-BERT model ranges from 2.7 points to 8.4 points on the answer-pairs scoring task and 14.7 points on the contrastive answer-pairs task. is selected on each subset of the ten datasets used in the experimentation. The incremental difference between the proposed LLM-Mixtral model and the best-performing neural-BERT model ranges from 2.7 points to 8.4 points on the answer-pairs scoring task and 14.7 points on the contrastive answer-pairs task.

Table 6. Comparative analysis of best performing neural BERT models with LLM-Mixtral

Datasets	Model	Avg. performance	LLM-Mixtral	Difference
MS-MARCO, AVSD, NarrativeQA, SemEval	ROUGE-L-KPQA	73.15	75.87	2.72
CosmosQA, MCScript, NarrativeQA, Quoref	LERC	74.44	82.2	7.76
NaturalQuestions	BERTScore	80.84	88.2	7.36
TriviaQA	BERTScore	85.28	93.68	8.4
Contrastive pairs Datasets (CosmosQA, MCScript, NarrativeQA, SocialQA)				
	LERC	80.3	95	14.7

4. CONCLUSION

This study explored applying LLMs as an evaluation metric for QA tasks. Our inquiry has resulted in a more profound comprehension of the capabilities of LLMs in assessing, adjudicating, and appraising the performance of QA systems in both closed and open domains. We conducted extensive experiments on ten datasets, comparing our proposed LLM-Mixtral metric with existing methods on QA tasks. The results indicated the superiority of LLM-Mixtral in providing accurate evaluations of answer quality. It outperformed traditional lexical metrics, neural BERT-based models, and KPQA approaches. Mixtral 8x7 B, a simple LLM-based metric, showcased higher correlations with human judgments compared to more sophisticated state-of-the-art statistical and neural machine-learning-based metrics. It reached an impressive Pearson correlation of over 80%. Human judgments in evaluating answer pairs achieved accuracy rates exceeding 95% in contrastive scoring. This superior performance across a diverse range of datasets and models underscores the potential of LLMs in QA evaluation. Our adopted metric exhibited versatility in open-domain QA experiments, specifically

on NQ and TQ datasets. It achieved results closer to human judgments and outperformed over-relaxed lexical matching metrics, bridging the gap between automated scoring and human assessment. The correlation with human judgments on these datasets reinforced the effectiveness of LLM-Mixtral, positioning it on par with GPT-3.5 and outperforming state-of-the-art neural BERT-based models like BERTScore. Our findings open new horizons for applying LLMs in QA evaluation, offering a complementary approach to traditional and neural-based metrics. This research marks a crucial step in pursuing more accurate and effective QA evaluation methods. Some key benefits of using LLM-based metrics over state-of-the-art metrics include customizability, multifaceted evaluation, and train-free capabilities. These features enable us to create a metric that can flexibly perform the judgment task across various datasets without requiring a learning process while still achieving competitive performance. LLM-based metrics are more domain agnostic than most machine learning BERT-based techniques, which showed a distribution domain - bias when correlating with human judgments.




REFERENCES

- [1] L. Hirschman and R. Gaizauskas, "Natural language question answering: The view from here," *Natural Language Engineering*, vol. 7, no. 4, pp. 275–300, 2001, doi: 10.1017/S1351324901002807.
- [2] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, vol. 2, no. 3, 2012.
- [3] M. Rotaru and D. J. Litman, "Improving question answering for reading comprehension tests by combining multiple systems," in *Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains*, 2005, pp. 46–50.
- [4] Y. Liu, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Generative question refinement with deep reinforcement learning in retrieval-based QA system," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1643–1652, doi: 10.1145/3357384.3358046.
- [5] D. Deutsch, T. B. -Weiss, and D. Roth, "Towards question-answering as an automatic metric for evaluating the content quality of a summary," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 774–789, 2021, doi: 10.1162/tacl_a.00397.
- [6] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.
- [7] E. Clark, A. Celikyilmaz, and N. A. Smith, "Sentence mover's similarity: automatic evaluation for multi-sentence texts," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2748–2760, doi: 10.18653/v1/P19-1264.
- [8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTscore: evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020*, 2020, pp. 1–43.
- [9] H. Lee et al., "KPQA: A metric for generative question answering using keyphrase weights," in *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2021, pp. 2105–2115, doi: 10.18653/v1/2021.naacl-main.170.
- [10] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186.
- [11] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7881–7892, doi: 10.18653/v1/2020.acl-main.704.
- [12] A. Chen, G. Stanovsky, S. Singh, and M. Gardner, "MOCHA: A dataset for training and evaluating generative reading comprehension metrics," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 6521–6532, doi: 10.18653/v1/2020.emnlp-main.528.
- [13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv-Computer Science*, pp. 1–46, 2021, doi: 10.48550/arXiv.2107.13586.
- [14] V. Sanh et al., "Multitask prompted training enables zero-shot task generalization," *arXiv-Computer Science*, 2021, doi: 10.48550/arXiv.2110.08207.
- [15] L. Ouyang et al., "Training language models to follow instructions with human feedback," *arXiv-Computer Science*, pp. 1–68, 2022, doi: 10.48550/arXiv.2203.02155.
- [16] S. Ostermann, M. Roth, A. Modi, S. Thater, and M. Pinkal, "SemEval-2018 Task 11: Machine comprehension using commonsense knowledge," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 747–757, doi: 10.18653/v1/S18-1119.
- [17] B. Bi, C. Wu, M. Yan, W. Wang, J. Xia, and C. Li, "Incorporating external knowledge into machine reading for generative question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2521–2530, doi: 10.18653/v1/D19-1255.
- [18] L. Huang, R. L. Bras, C. Bhagavatula, and Y. Choi, "COSMOS QA: Machine reading comprehension with contextual commonsense reasoning," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 2391–2401, doi: 10.18653/v1/d19-1243.
- [19] M. Sap, H. Rashkin, D. Chen, R. L. Bras, and Y. Choi, "Social IQA: Commonsense reasoning about social interactions," in *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 2019, pp. 4463–4473, doi: 10.18653/v1/d19-1454.
- [20] P. Dasigi, N. F. Liu, A. Marasović, N. A. Smith, and M. Gardner, "Quoref: A reading comprehension dataset with questions requiring coreferential reasoning," in *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 2019, pp. 5925–5932, doi: 10.18653/v1/d19-1606.
- [21] C. Wang et al., "Evaluating open-QA evaluation," in *37th International Conference on Neural Information Processing Systems*, 2023, pp. 77013–77042.




- [22] T. Kwiatkowski *et al.*, “Natural questions: a benchmark for question answering research,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019, doi: 10.1162/tacl.a.00276.
- [23] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, “TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension,” in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2017, vol. 1, pp. 1601–1611, doi: 10.18653/v1/P17-1147.
- [24] T. Kočický *et al.*, “The narrativeQA reading comprehension challenge,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 317–328, 2018, doi: 10.1162/tacl.a.00023.
- [25] A. Q. Jiang *et al.*, “Mixtral of experts,” *arXiv-Computer Science*, pp. 1–13, 2024, doi: 10.48550/arXiv.2401.04088.
- [26] D. Cer, M. Diab, E. Agirre, I. L. -Gazpio, and L. Specia, “SemEval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 1–14, doi: 10.18653/v1/s17-2001.

BIOGRAPHIES OF AUTHORS






Hazem Abdelazim    is currently a Professor of AI and ML and Dean at ESLSCA University's School of Computing and Digital Technology. He has been locally and internationally recognized for his achievements. He was awarded an 'Invention Achievement Award' from IBM in 1991, the First Scientific Innovation Prize for Arab Scientists (1993), State Excellence and encouragement Award (1995), and MBA Director's Cup (2003) from MSM, Netherlands. His journey included academic positions at Cairo University, AUC, and UAE University, and professional positions as an IBM Research Scientist, and Director of research at Microsoft. His research interests are generative artificial intelligence (AI), LLM, information retrieval, and NLP. He has 35+ publications. He can be contacted at email: hazem.abdelazim@eslsc.edu.eg.



Mohamed Tharwat Waheed    graduated from the Department of Electronics and Communication, Faculty of Engineering, Cairo University in 2006. He received the M.Sc. degree in using reinforcement learning in mobile communication in 2017. He completed his Ph.D. with a focus on the applications of AI/ML in the Telecom industry at Cairo University. In addition to his industry role as a Subject Matter Expert in the technology domain at Vodafone, Egypt. He is a research and teaching doctor at ESLSCA University School of Computing and Digital Technologies. He is also an IEEE Senior Member. His research interests span a diverse spectrum, including IoT in smart cities, 5G, autonomous driving, AI/ML in mobile communication, and the implementation of generative AI in domain-specific tasks. He can be contacted at email: mohamed.mohamed-waheed@vodafone.com.



Ammar Mohammad    earned his bachelor's and master's degrees in computer science from Cairo University, Egypt, and obtained his Ph.D. in computer science from the University of Koblenz-Landau, Germany, in 2010. He has previously served as a researcher and research fellow with the AI Research Group at the University of Koblenz-Landau. Currently, he holds the position of a professor of computer science at both Cairo University and MSA University in Egypt. His research interests encompass machine and deep learning techniques, methods, algorithms, and applications across various domains. He can be contacted at email: ammar@cu.edu.eg.