Evaluate the LLMs

English models:

We evaluated our models on 128 rows of the "*SQuAD-v1.1*" dataset and recorded the following metrics for the generated answers.

Models	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Faithfulness	Relevance	Bias	Toxicity
Gpt-4o-mini	47.32	30.46	47.32	8.80	Gpt4_English_Faithfulness_scores.xlsx	Gpt4_English_Relevance_scores.xlsx	Gpt4_English_Bias_scores.xlsx	Gpt4_English_toxicity_scores.xlsx
google/gemma-2-2b	0.56	0.38	0.59	0.18				
Naseej/noon-7b	75.82	41.70	75.22	2.79	Noon_English_Faithfulness_scores.xlsx	Noon_English_Relevance_scores.xlsx	Noon_English_Bias_scores.xlsx	Noon_English_toxicity_scores.xlsx
HeshamHaroon/Arabic-llama3	50.0	37.0	49.0	18.0				
Omartificial-Intelligence-Space/Arabic-llama3.1-16bit-FT	74.37	31.50	74.37	1.26				

Arabic Models:

We evaluated our models on 100 rows of the "*Generated*" dataset we created and recorded the following metrics for the generated answers.

General access

Models	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Faithfulness	Relevance	Bias	Toxicity
Gpt-4o-mini	39.47	19.25	38.81	55.05				
google/gemma-2-2b (**Arabic prompt**)	15.0	6.0	15.0	11.0				
Naseej/noon-7b	21.54	4.50	21.04	4.58	Noon_Arabic_Faithfulness_scores.xlsx	Noon_Arabic_Relevance_scores.xlsx	Noon_Arabic_Bias_scores.xlsx	Noon_Arabic_toxicity_scores.xlsx
HeshamHaroon/Arabic-llama3 (eng_prompt)	26.0	12.0	25.0	26.0				
Omartificial-Intelligence-Space/Arabic-llama3.1-16bit-FT	19.41	1.35	19.41	1.23	Omar_Arabic_Faithfulness_scores.xlsx	Omar_Arabic_Relevance_scores.xlsx	Omar_Arabic_Bias_scores1.xlsx	Omar_Arabic_toxicity_scores.xlsx