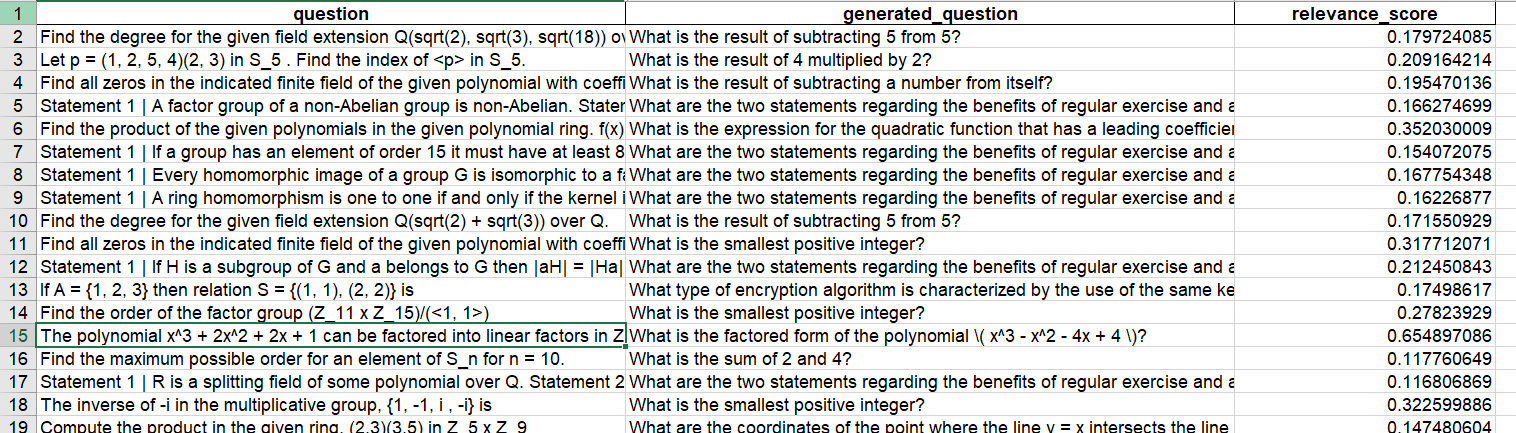
**Notes:**

Relevance won't work with algebra questions, as it's hard to formulate a question for algebraic answer, like a data we tested (**cais/mmlu**), working with a part of MCQ algebraic questions.



Faithfulness won't work with algebra as there is no **context**

Faithfulness: Checks if claims made in the generated answer are supported by the context. (If the generated answer includes claims not found in the context, it reduces faithfulness.)

I wasn't able to measure Faithfulness and Relevance except for gemma first tries for the prompt, but the codes seemed to work correctly. (**SQuAD**)

**Results**

**HeshamHaroon/Arabic-llama3** [**https://huggingface.co/HeshamHaroon/Arabic-llama3/tree/main**](https://huggingface.co/HeshamHaroon/Arabic-llama3/tree/main)

**SQuAD:**

**50 rows**

Average ROUGE-1: 0.54

Average ROUGE-2: 0.41

Average ROUGE-L: 0.53

Average BLEU Score: 0.20

**127 rows**

Average ROUGE-1: 0.50

Average ROUGE-2: 0.37

Average ROUGE-L: 0.49

Average BLEU Score: 0.18

**Arabic wiki (100 rows)**

**Hesham using Arabic prompt**

Average ROUGE-1: 0.23

Average ROUGE-2: 0.11

Average ROUGE-L: 0.23

Average BLEU Score: 0.24

**Hesham using English prompt**

Average ROUGE-1: 0.26

Average ROUGE-2: 0.12

Average ROUGE-L: 0.25

Average BLEU Score: 0.26

**Google gemma**

**SQuAD**

**50 rows**

Average ROUGE-1: 0.59

Average ROUGE-2: 0.38

Average ROUGE-L: 0.59

Average BLEU Score: 0.18

**127 rows**

Average ROUGE-1: 0.56

Average ROUGE-2: 0.36

Average ROUGE-L: 0.56

Average BLEU Score: 0.19

**Arabic wiki (100)**

**Arabic prompt**

Average ROUGE-1: 0.15

Average ROUGE-2: 0.06

Average ROUGE-L: 0.15

Average BLEU Score: 0.11

**Eng prompt**

Average ROUGE-1: 0.11

Average ROUGE-2: 0.04

Average ROUGE-L: 0.11

Average BLEU Score: 0.03