Answers and metrics generated using Llama and Mistral

	question	Reference	llama_answer	bleu_llama	rouge_llama	bert_llama	mistral_answer	bleu_mistral	rouge_mistral	bert_mistral
0	What is the primary cause of Type 2 Diabetes?	Insulin resistance, where the body's cells do ...	Genetics	0	0	0.885362	The primary cause of Type 2 Diabetes is the bo...	0.276941	0.6	0.924879
1	Name two risk factors for developing Type 2 Di...	Obesity and physical inactivity.	1. Obesity 2. Family history of diabetes	0.033032	0.181818	0.861097	Two risk factors for developing Type 2 diabete...	0.018477	0.125	0.840773
2	List two common symptoms of Type 2 Diabetes.	Frequent urination and excessive thirst.	1. Blurred vision 2. Thirst These symptoms ar...	0.006415	0.05	0.829863	Two common symptoms of Type 2 Diabetes are fre...	0.088892	0.444444	0.892396
3	Which blood test is commonly used to diagnose ...	The A1C test.	The A1C test is commonly used to diagnose Type...	0.090588	0.428571	0.883461	The A1C test is commonly used to diagnose Type...	0.090588	0.428571	0.882688
4	What are two possible complications of poorly ...	Heart disease and kidney damage.	1. Kidney damage 2. Nerve damage	0.086334	0.363636	0.88096	Two possible complications of poorly managed T...	0.153102	0.3	0.895092
5	What is the first-line medication commonly pre...	Metformin.	Metformin is the first-line medication commonl...	0.004361	0.040816	0.872956	The first-line medication commonly prescribed ...	0.036021	0.153846	0.880275
6	How can individuals reduce their risk of devel...	Through weight loss, a healthy diet, and regul...	1. Lifestyle changes, including: a. Avoiding h...	0.006535	0.086957	0.852473	Individuals can reduce their risk of developin...	0.055636	0.1875	0.884282
7	What is essential for effective long-term mana...	Patient education and self-monitoring.	Lifestyle changes, especially in individuals a...	0.021105	0	0.874828	A combination of lifestyle changes, including ...	0.008641	0.068966	0.855595

Comparison Results

```
BLEU Mean Scores:
  LLaMA:   0.0310
  Mistral: 0.0910
------------------------------
ROUGE Mean Scores:
  LLaMA:   0.1440
  Mistral: 0.2885
------------------------------
BERT Mean Scores:
  LLaMA:   0.8676
  Mistral: 0.8820
```

Inference

- **Mistral-7B consistently outperforms TinyLlama-1.1B** across all evaluation metrics.
- Although BLEU scores are low, BERT score shows **high semantic similarity,** especially for Mistral.
- **ROUGE-L** confirms that Mistral preserves more structure from the reference answers.

Conclusion

Mistral is better suited for high-accuracy, context-aware QA in this pipeline.