Final Model Comparisons and Results

Models	Comparison		Model 1 Avg. Weighted Score	Model 2 Avg. Weighted Score	T-Statistic	P-Value	Significant	Improvement with RAG	Best Model
	Model 1	Model 2							
BASE MODELS	Base_Llama	Base_Mixtral	0.4372	0.4201	2.0569	0.0424	Yes	-	Base_Llama
	Base_Llama	Base_Gemma	0.4372	0.3671	8.5123	0.0000	Yes	-	Base_Llama
	Base_Mixtral	Base_Gemma	0.4201	0.3671	5.1972	0.0000	Yes	-	Base_Mixtral
RAG MODELS	RAG_Llama	RAG_Mixtral	0.3384	0.3909	-5.5521	0.0000	Yes	-	RAG_Mixtral
	RAG_Llama	RAG_Gemma	0.3384	0.3689	-3.6347	0.0005	Yes	-	RAG_Gemma
	RAG_Mixtral	RAG_Gemma	0.3909	0.3689	3.0485	0.0030	Yes	-	RAG_Mixtral
BASE VS RAG	Base_Llama	RAG_Llama	0.4372	0.3384	9.7615	0.0000	Yes	No	Base_Llama
	Base_Mixtral	RAG_Mixtral	0.4201	0.3909	3.5793	0.0005	Yes	No	Base_Mixtral
	Base_Gemma	RAG_Gemma	0.3671	0.3689	-0.2673	0.7898	No	Yes	RAG_Gemma

Final Conclusion

- **Llama Base Model** consistently outperforms in both base and mixed comparisons, showing significant results in its favor.
- **Mixtral RAG Model** demonstrates strong performance among the RAG configurations, particularly when compared to other RAG models.

Best Fit for Ananda Chatbot: Considering both the base and RAG results, **Llama Base Model** is the best choice for the chatbot due to its strong performance in weighted average scores and significant statistical outcomes across multiple comparisons.