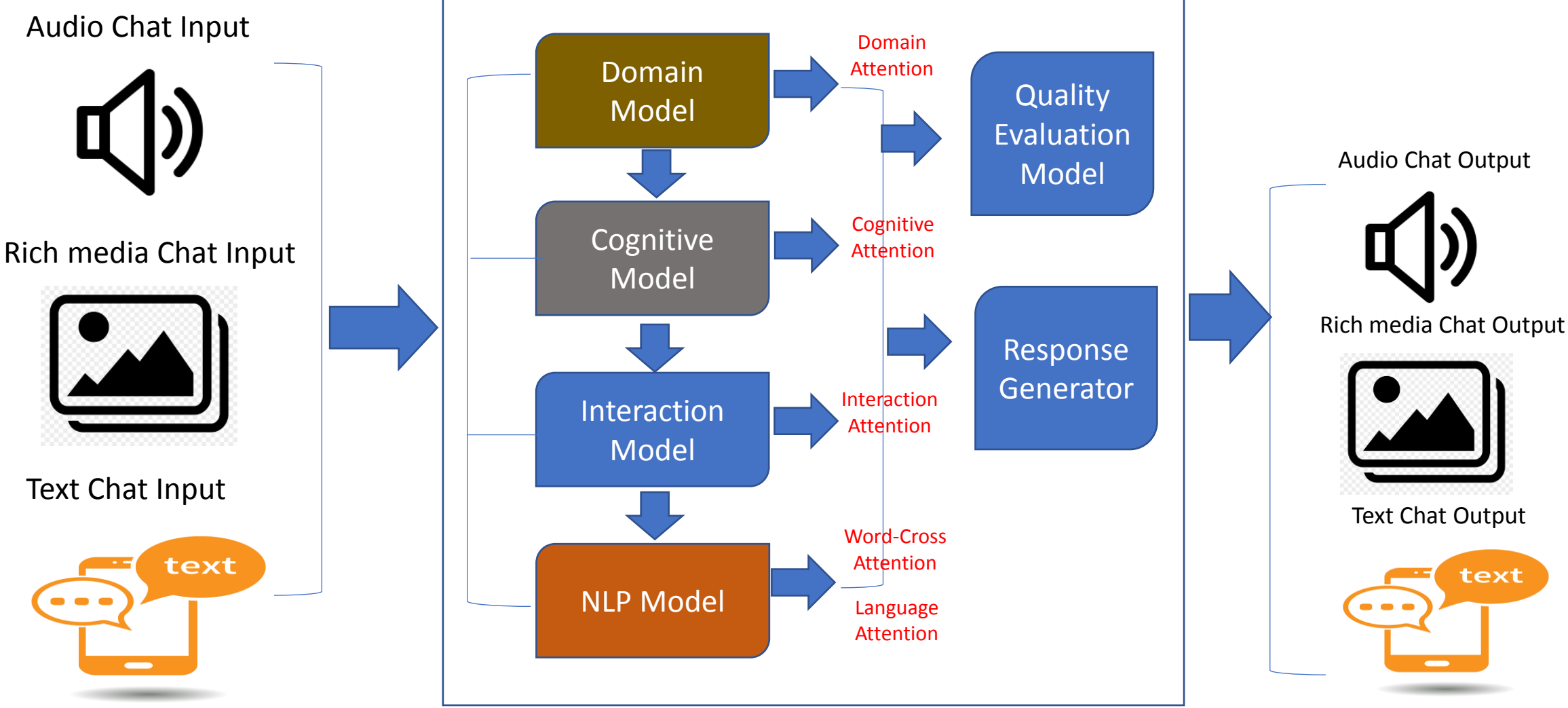
Testable AI Chatbot System

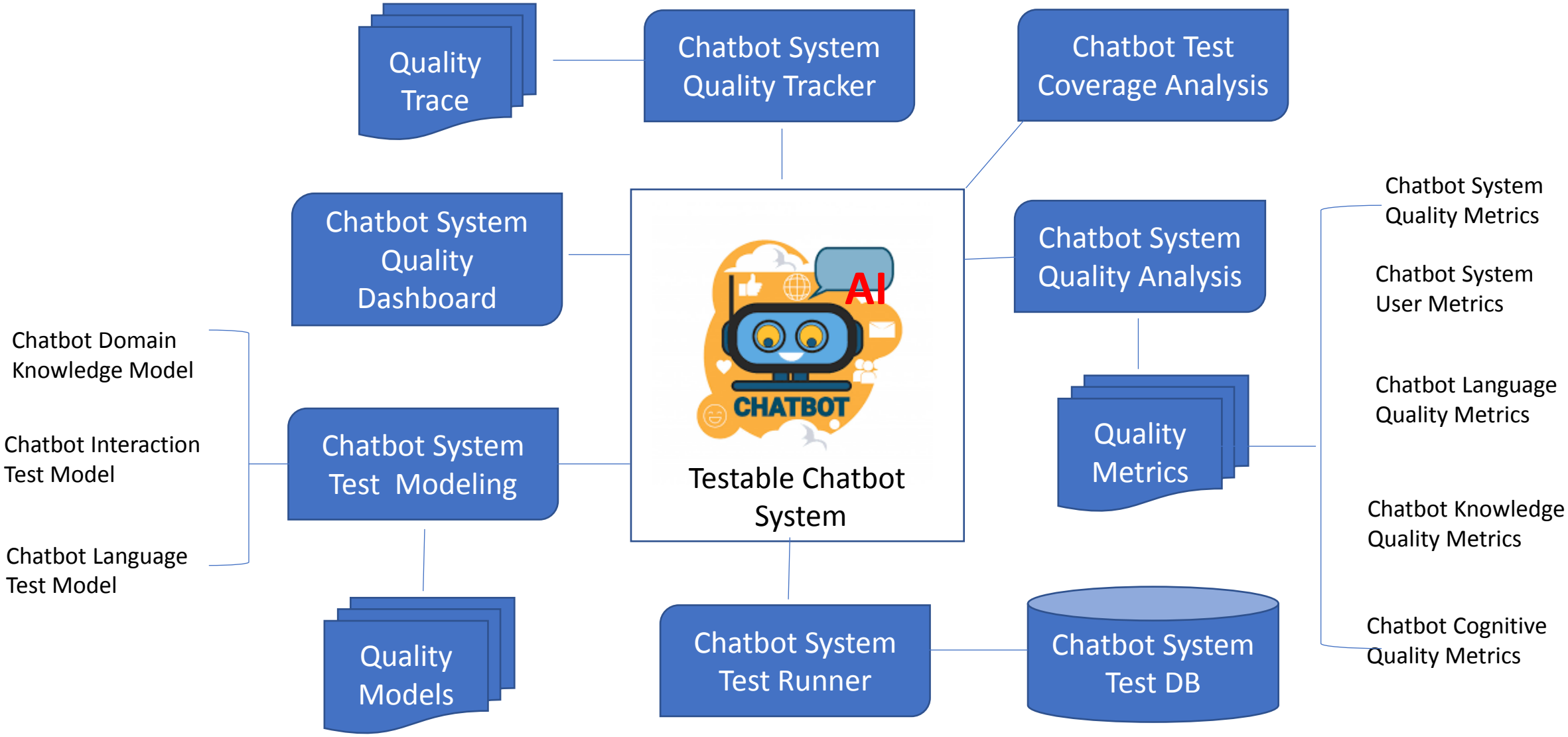
- **Quality Evaluation Platform, Model, Infrastructure, and Evaluation Metrics**

Prepared by Jerry Gao, Ph.D. Professor

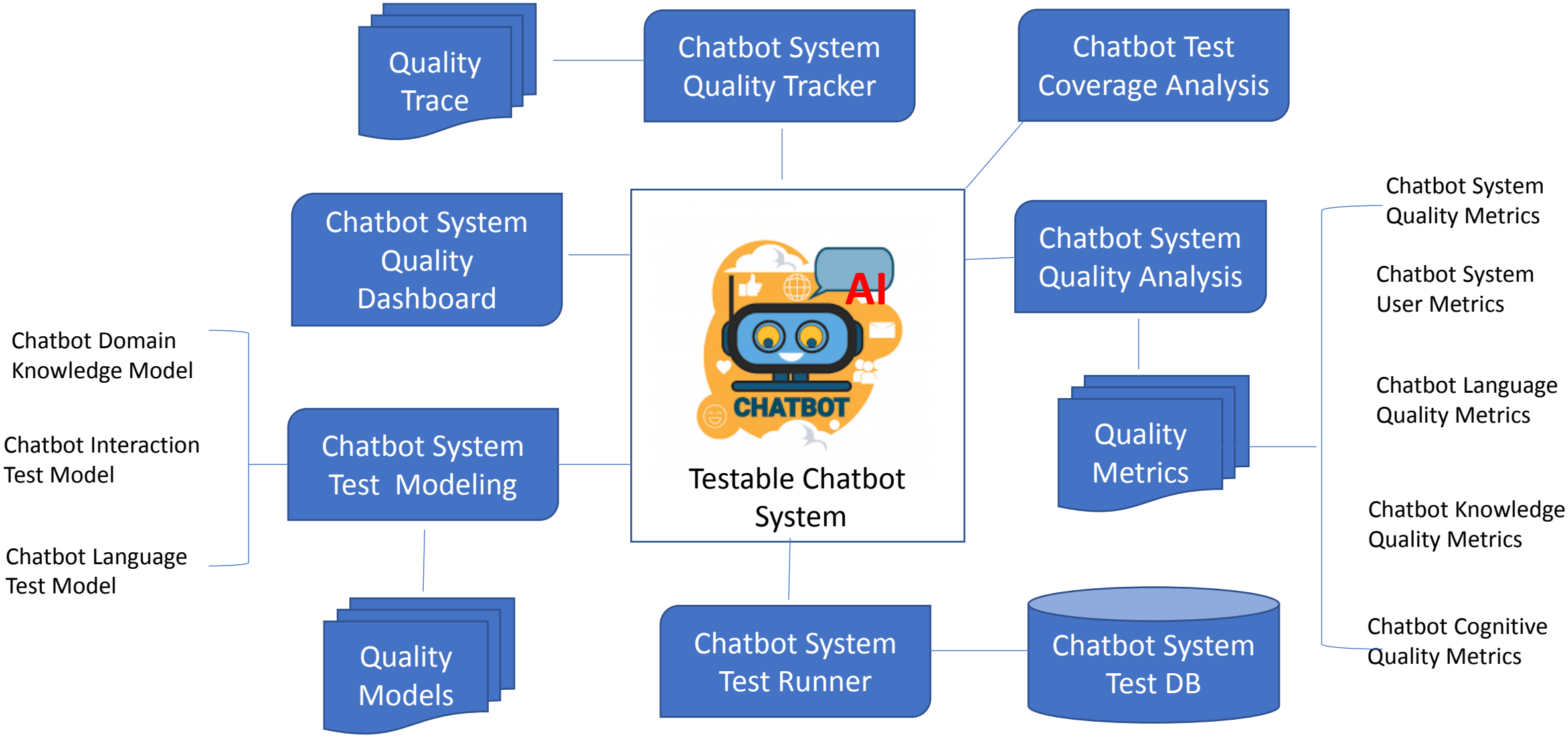
Testable AI Chatbot System Model and Infrastructure



Testable AI Chatbot System – Machine Learning Model Architecture

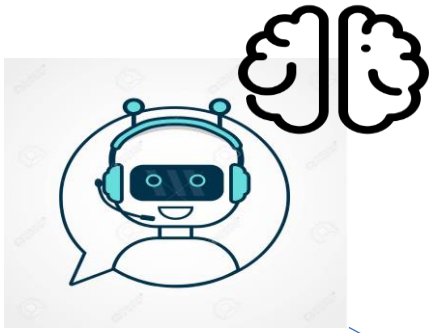


AI Chatbot System Test Platform

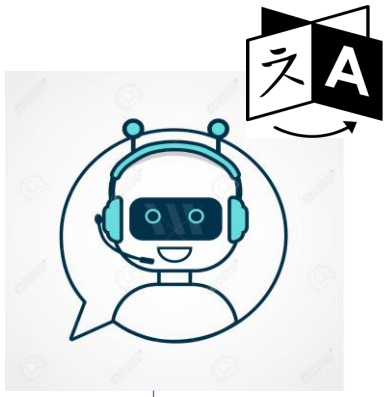


AI Chatbot System Evaluation Metrics and Parameters

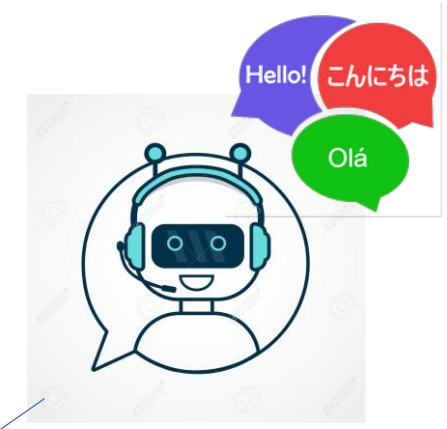
Chatbot Cognitive Quality Metrics



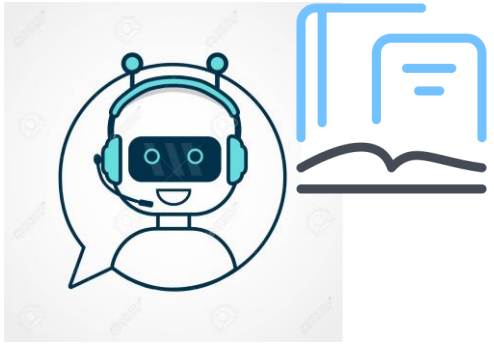
Chatbot Language Quality Metrics



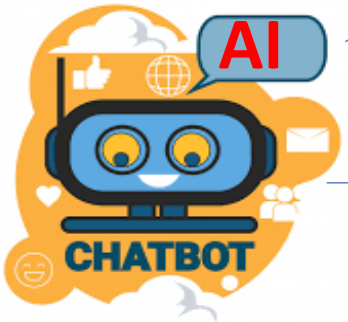
Chatbot System Quality Metrics



Chatbot Knowledge Quality Metrics

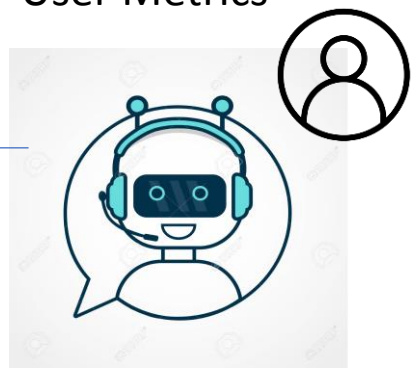


Test



Testable Chatbot

Chatbot System User Metrics



AI Chatbot System Evaluation Metrics and Parameters

System Quality Metrics	System Quality Evaluation Metrics	System Cognitive Evaluation Metrics	Domain Knowledge Evaluation Metrics	System Content Evaluation Metrics	Language Evaluation Quality Metrics
Accuracy Metrics	System accuracy in chatting and conversations	System attention accuracy	Domain knowledge accuracy	System content accuracy metrics	Syntax and wording accuracy metrics
Performance Metrics	System chat performance metrics	System cognitive performance metrics	System domain knowledge engine performance metrics	Chat contention generation performance metrics	System language generation performance metrics
Correctness Metrics	System chat correctness metrics	Cognitive correctness evaluation metrics	Domain knowledge correctness metrics	System content correctness metrics	Chat language correctness metrics
Consistency Metrics	System chat consistency metrics	Cognitive consistent evaluation metrics	Domain knowledge consistency metrics	System content consistency metrics	Chat language consistency metrics
Relevancy Metrics	System chat relevancy metrics	Cognitive relevancy evaluation metrics	Domain knowledge relevancy metrics	System content relevancy metrics	Chat language relevancy metrics

Chatbot System Cognitive Evolution Metrics and Parameters

Parameters		Descriptions
Attention evaluation metrics		This refers to a set of evaluation metrics focusing of chat attention in different perspectives.
	Language attention	Bilingual evaluation understudying (BLEU{1-4})
	Conversation chat attention	This evaluates chat attention in accuracy/correctness/consistency.
	Domain knowledge attention	This evaluates domain knowledge attention of the chat system in terms of domain and technology terminology, functions, and services in terms of what, why, where, when, who, and how.
	Objective attention	The evaluates chat objective attention in terms of XXX
Understandability metrics		This refers to a set of evaluation metrics focusing of chat understandability.
	Understandability of received questions from clients	This metrics is used to evaluate how well the given chat system is able to understand the received questions from clients.
	Client understandability of generated responses	This metrics is used to evaluate how well the clients are able to understand the generated responses from the chat system.
	Domain understandability	This refers to the metrics that is used to evaluate how well the clients are able to understand the generated responses from the chat system.

Chatbot System User Evaluation Metrics and Parameters

Parameters	Auto Collection and Tracking	Descriptions
Chat Duration Metrics	Yes	Chat duration metrics refer to max/min/average of chat duration time.
Client Review Score	Yes	Chat review score from clients (score: 1-5)
Chat Interaction Rate	Yes	This refers to max/min/average number of chat interactions per chat session.
Chat Abandonment Rate	Yes	This refers to the ratio between no. of abandoned sessions and total no. of chat sessions
User Engagement Rate	Yes	The refers to the ratio between no. of user engaged sessions and total no. of chat sessions.
Goal Completion Rate	Yes	This refers to the ratio between no. of goal completed chats and total chat sessions.
Chatbot Performance Metrics	Yes	This refers to system-and-user performance for chat response time (average/max/min)
Chatbot Consistency Metrics	Yes	These will be used to measure the consistency of chatbot in terms of responses and answers as well as conversation tyles and patterns. (response consistency, conversation pattern consistency, conversation flow consistency)
Chatbot Relevancy Metrics	Yes	These will be used to measure the relevancy of chatbot in terms of domain relevancy, topic/subject relevancy, intention relevancy, as well as content relevancy.
Chatbot Accuracy Metrics	Yes	These will be used to measure the accuracy of chatbot in terms of domain accuracy, topic/subject accuracy, intention accuracy, as well as content accuracy.
Chatbot Correctness Metrics	Yes	These will be used to measure the correctness of chatbot in terms of domain correctness, topic/subject correctness, intention correctness, as well as language correctness.

Chatbot Language-Based Automatic Evaluation Parameters

Parameters	Descriptions
Lexical diversity (distinct-n)	An automatic metric that refers to the number of unique n-grams in the model's responses divided by the total number of generated to- kens (Li et al., 2016).
Average cosine-similarity	An automatic metric that refers to the average cosine-similarity between the mean of the word embeddings of a generated response and ground-truth response (Liu et al., 2016).
Sentence average BLEU-2 score (Liu et al., 2016).	A sentence-oriented evaluation metric with evaluation scores.
Response perplexity	A metric that measures the responses using the likelihood that the model predicts the correct response (Zhang et al. 2018).
Perplexity and SSA	An automatic metric that correlates with human judgment, in contrast to recent findings on other automatic metrics mentioned above.
Perplexity (automatic metric)	An automatic metric that measures how well the model predicts the test set data; in other words, how accurately it anticipates what people will say next.
Language-based Evaluation Metrics	Average/min/max language correctness/accuracy/consistency in chat responses for all chat sessions