Explaining Large Language Models: Integrating Explainable AI (XAI) with LoRA

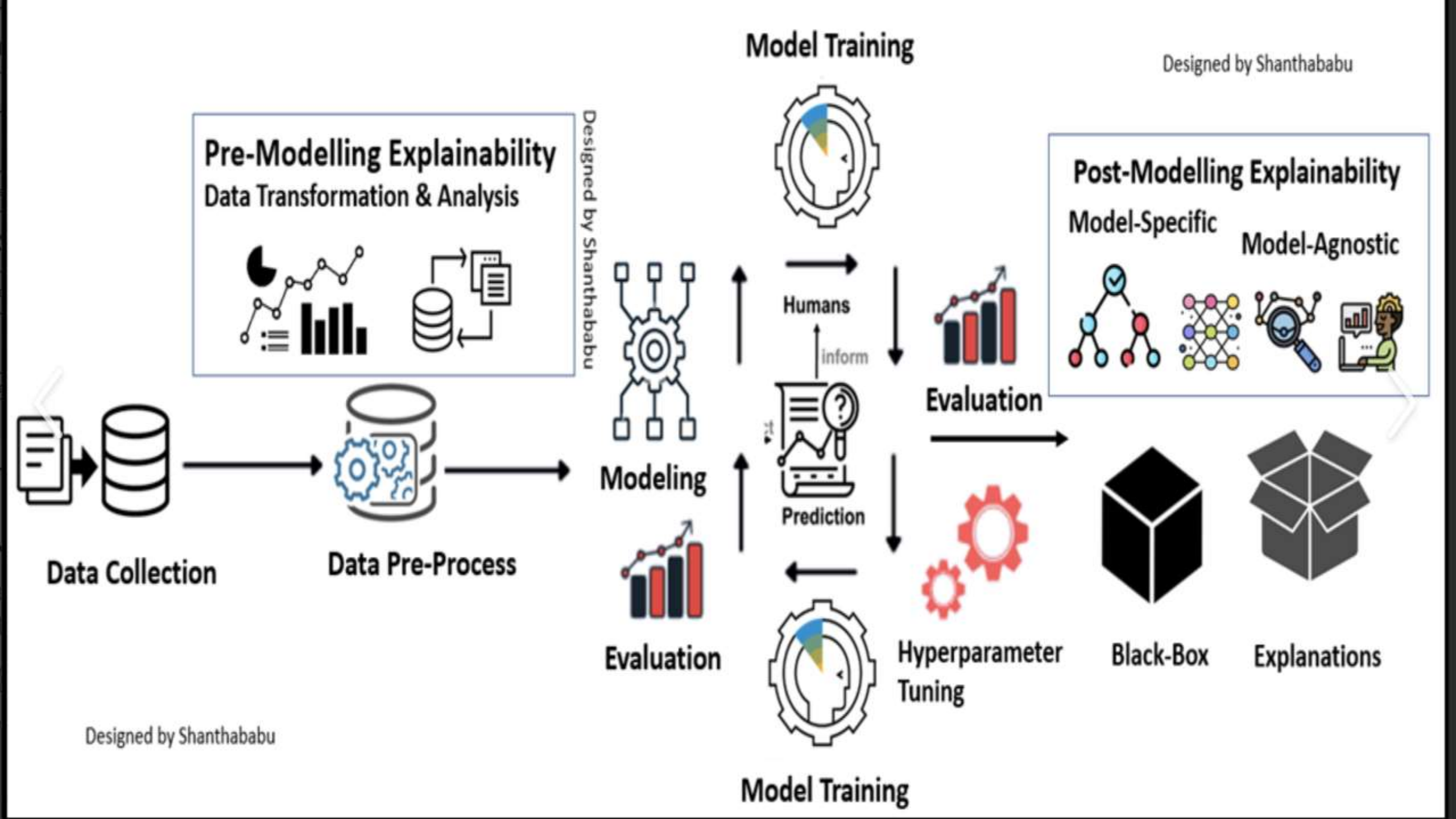
Presented by Alisha Rath

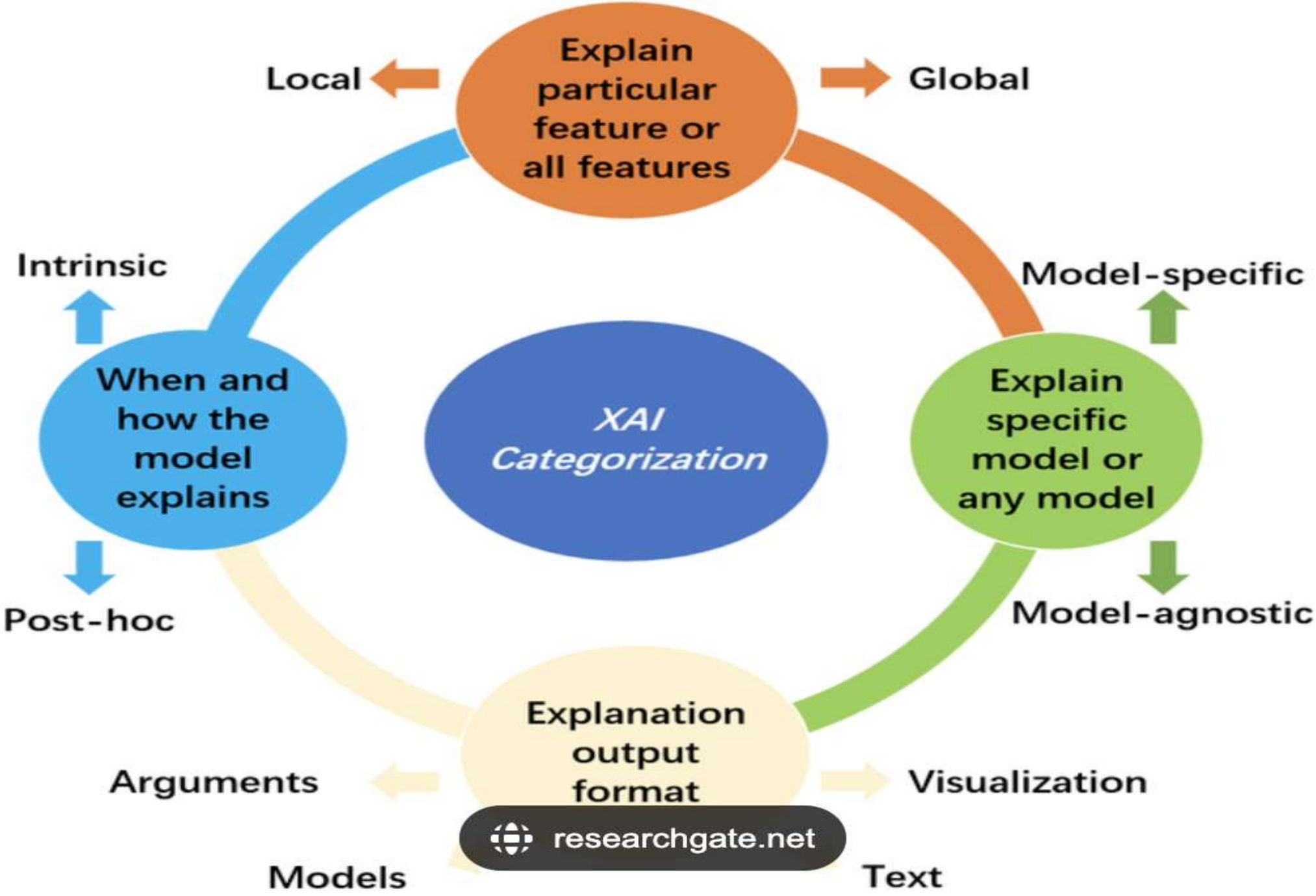
What is Explainable AI (XAI)?

XAI is the process of making AI decisions understandable to humans.It addresses the "black-box" nature of complex AI models, including LLMs.Key benefits:Improves trust in AI models.

Helps identify biases and errors.

Makes regulatory compliance easier in industries like healthcare and law.


Model Training
Designed by Shanthababu
Pre-Modelling Explainability
Data Transformation & Analysis
Designed by Shanthababu
Post-Modelling Explainability
Model-Specific
Model-Agnostic
Humans
inform
Evaluation
Modeling
Prediction
Data Collection
Data Pre-Process
Evaluation
Hyperparameter Tuning
Black-Box
Explanations
Designed by Shanthababu
Model Training


Explain particular feature or all features
Local
Global
Intrinsic
Model-specific
When and how the model explains
XAI Categorization
Explain specific model or any model
Post-hoc
Model-agnostic
Explanation output format
Arguments
Visualization
researchgate.net
Models
Text

How XAI Works

XAI techniques provide insights into how models make decisions by:**Feature Importance**: Highlights which input features (tokens, words) were crucial in predictions.

Model Behavior Tracking: Provides transparency on how internal model layers (e.g., attention heads in LLMs) respond to inputs.

Counterfactual Explanations: Shows how small changes to input would alter the outcome.

Types of XAI Techniques

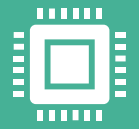
SHAP (SHapley Additive exPlanations):Explains the output by attributing importance to each input feature.

LIME (Local Interpretable Model-agnostic Explanations):Simplifies the model's predictions locally for better understanding.

Attention Maps:Visualizes which parts of the input the model is "attending" to when making predictions.

Counterfactual Analysis:Investigates "what if" scenarios to determine how input changes affect output.

What is LoRA (Low-Rank Adaptation)?



LoRA is a technique used to fine-tune large pre-trained language models by applying low-rank updates.Key advantages:Reduces computational costs.

Enables fine-tuning with fewer parameters.

Helps target specific behaviors in models, making it easier to analyze and explain changes.

How to Integrate XAI with LoRA

Fine-Tuning with LoRA:Use LoRA to selectively fine-tune parts of the model (e.g., attention layers) to capture specific knowledge.

XAI Analysis After LoRA Fine-Tuning:Once fine-tuning is done, use **XAI** tools like **SHAP** to understand which input features (tokens, words) influenced the model's decisions.

Attention Maps can help track the effects of LoRA fine-tuning on attention distribution.

Post-Fine-Tuning Explanation:Analyze how LoRA's low-rank updates change the model's behavior and make the model more interpretable.

Real-World Example

Content:Example: A legal decision support system using LoRA-fine-tuned LLM for predicting court case outcomes.

Process:

Fine-tune specific layers using LoRA.	Use SHAP to explain why the model predicts a certain outcome.	Use **attention maps** to visualize the focus of the model during its decision-making.

Challenges and Future Directions

Challenges:

- Handling multi-modal models (text + images).
- Computational costs of fine-tuning and explainability.

Future Directions:

- Improved integration of XAI techniques for complex AI architectures.
- More interpretable fine-tuning strategies using LoRA in different AI applications.

Conclusion

- Integrating **Explainable AI (XAI)** with **LoRA** enhances transparency in large models like GPT.XAI helps understand why certain decisions are made, and LoRA fine-tuning ensures targeted behavior adjustments, enabling clearer explanations.