Explainability Methods Comparison

SHAP
(SHapley Additive exPlanations)

Key Features:

- Game-theoretic approach
- Feature importance scores
- Global & local explanations
- Model-agnostic

Strengths:

- Theoretically grounded
- Consistent
- Additive

Best Use Case:

Understanding feature contributions to predictions

Visualization Example
(Heat map / Bar chart / Attention map)

LIME
(Local Interpretable Model-agnostic Explanations)

Key Features:

- Local surrogate models
- Perturbation-based
- Instance-specific
- Text highlighting

Strengths:

- Intuitive
- Local fidelity
- Fast

Best Use Case:

Explaining individual prediction instances

Visualization Example
(Heat map / Bar chart / Attention map)

Attention Weights
(Transformer Self-Attention)

Key Features:

- Built into model
- Token-level importance
- Multi-head analysis
- No extra computation

Strengths:

- Native to model
- Efficient
- Multi-layer

Best Use Case:

Understanding what the model focuses on

Visualization Example
(Heat map / Bar chart / Attention map)