Performance vs Size Tradeoffs

Performance-Size Tradeoffs

Finding the balance between model size, accuracy, and speed according to compression levels

Pareto Frontier

Large Model

Size: 500MB
Accuracy: 95%
Speed: 500ms

Medium Model

Size: 100MB
Accuracy: 93%
Speed: 100ms

Small Model

Size: 20MB
Accuracy: 90%
Speed: 20ms

Tiny Model

Size: 5MB
Accuracy: 85%
Speed: 5ms

Model Selection Criteria

Cloud Deployment

Large/Medium Model

Accuracy prioritized

Mobile App

Medium/Small Model

Balanced performance

Wearable

Small/Tiny Model

Size minimization

Decision-Making Guide:

1. Determine minimum acceptable accuracy
2. Identify target device constraints
3. Select the optimal point on the Pareto curve