

Stage 4: Production Deployment and Inference

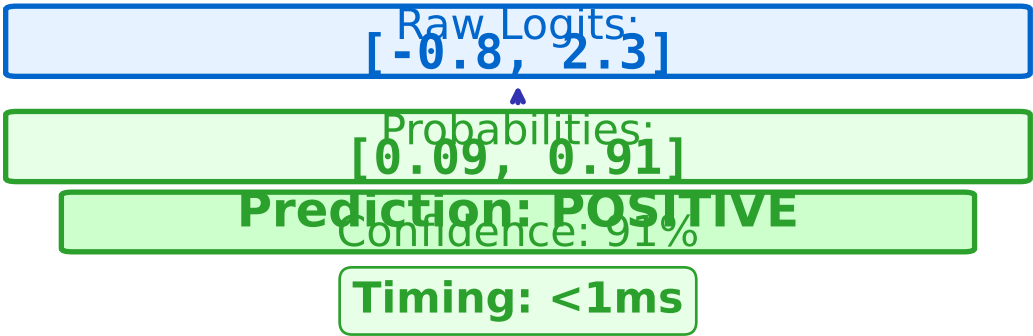
Step 1: Input Processing



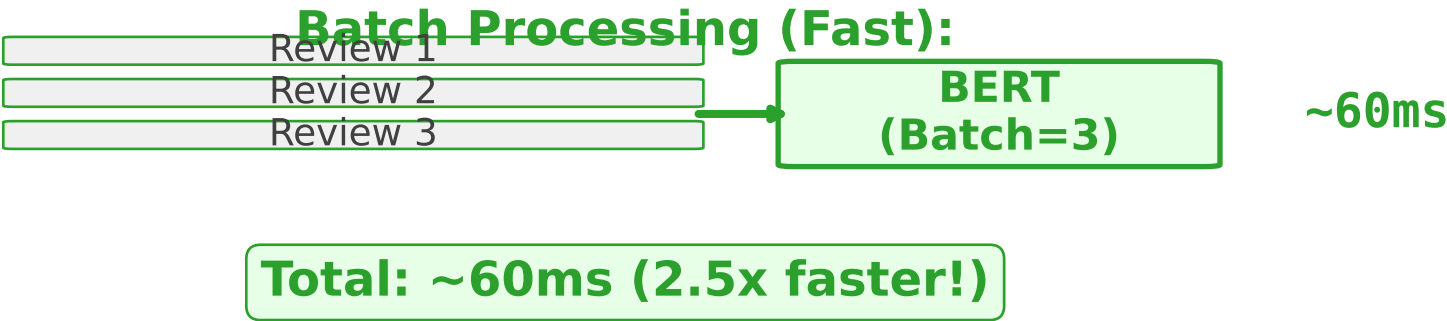
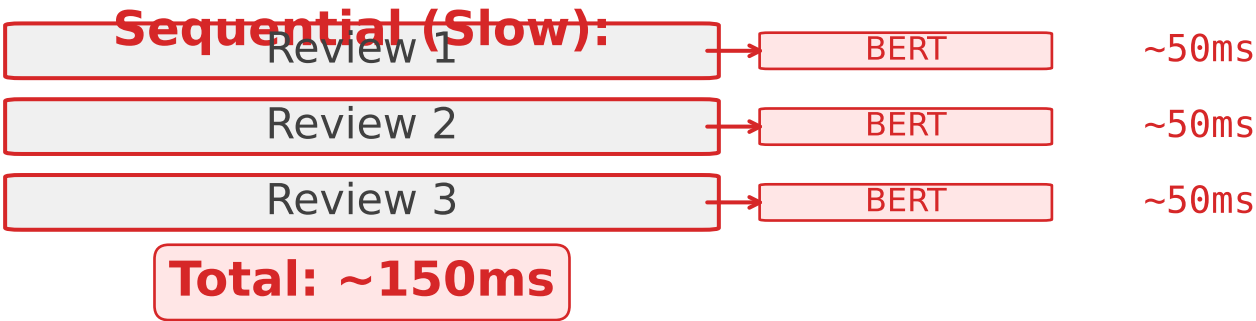
Step 2: Model Inference



Step 3: Output Processing



Production Optimization: Batch Processing



Production Code Example (Hugging Face Transformers)

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch

# Load fine-tuned model
tokenizer = AutoTokenizer.from_pretrained("my-sentiment-model")
model = AutoModelForSequenceClassification.from_pretrained("my-sentiment-model")
model.eval() # Set to inference mode
model.to("cuda") # Move to GPU

# Batch inference
reviews = ["Great movie!", "Terrible acting", "Loved it!"]
inputs = tokenizer(reviews, padding=True, truncation=True, return_tensors="pt")
inputs = {k: v.to("cuda") for k, v in inputs.items()} # Move to GPU

with torch.no_grad(): # Disable gradient computation
    outputs = model(**inputs)
    predictions = torch.softmax(outputs.logits, dim=1)
    labels = torch.argmax(predictions, dim=1)

# Results: labels = [1, 0, 1] (POSITIVE, NEGATIVE, POSITIVE)
```