Stage 4: Production Deployment and Inference

Step 1: Input Processing

Raw Review Text:
"*Great movie!*"

Tokenization:
[CLS] Great movie ! [SEP]
IDs: [101, 2624, 3185, 999, 102]

Timing: ~1ms

Step 2: Model Inference

Fine-tuned BERT
+ Classifier Head
Embeddings
Layers 1-12
Pooling
Linear

GPU Accelerated

Timing: ~10-50ms

Step 3: Output Processing

Raw Logits:
[-0.8, 2.3]

Probabilities:
[0.09, 0.91]

Prediction: POSITIVE
Confidence: 91%

Timing: <1ms

Production Optimization: Batch Processing

Sequential (Slow):

Review 1 → BERT ~50ms
Review 2 → BERT ~50ms
Review 3 → BERT ~50ms

Total: ~150ms

Batch Processing (Fast):

Review 1
Review 2
Review 3
→ **BERT (Batch=3)** **~60ms**

Total: ~60ms (2.5x faster!)

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)
```