Production Best Practices

Deploying LLM Applications at Scale

Security & Authentication

- **API Key Management:** Use environment variables or dedicated secret management services (e.g., AWS Secrets Manager, HashiCorp Vault).
- **Access Control:** Implement robust rate limiting, user authentication, and strict input validation to prevent abuse and injection attacks.
- **Data Privacy:** Never include PII or sensitive data in prompts. Ensure compliance with regulations like GDPR and CCPA.

Error Handling & Reliability

```python
import time
import random

def robust_completion(messages, max_retries=3):
    for attempt in range(max_retries):
        try:
            response = litellm.completion(
                model="gpt-3.5-turbo",
                messages=messages,
                timeout=30
            )
            return response.choices[0].message.content
        except Exception as e:
            if attempt == max_retries - 1:
                raise e
            time.sleep(random.uniform(1, 3))  # Exponential backoff
    return None
```

30-35 minutes | Lab Step 16/17 | Enterprise Ready