1. Overview: What Is Deployment Documentation?

Deployment documentation explains how to put a trained model into production so users or applications can actually use it.

It’s the bridge between:

“We have a trained model!” and “It’s now running safely and efficiently on real servers.”

🧠 2. Key Concepts Concept Description Inference Running the trained model to make predictions. Serving Hosting the model via an API or web service. Containerisation Packaging model code and dependencies (usually using Docker). CI/CD Continuous integration & deployment pipelines for updates. Monitoring Tracking model health, latency, and performance post-deployment.

Model Deployment Guide: Multilingual Chatbot v2.0

1. Objective

This document describes how to deploy the multilingual chatbot model (v2.0) for production use, including environment setup, model serving, and monitoring.

---

2. Environment Requirements

Component	Version	Notes
Python	3.10	Required for running model scripts
TensorFlow	2.12	Model training and inference
Docker	24.0+	Containerised deployment
FastAPI	0.95	REST API framework
AWS EC2	t3.large	Recommended deployment server

---

3. Deployment Steps

Step 1: Clone the Repository

```
git clone https://github.com/lesliewrites/chatbot-model.git
cd chatbot-model

Step 2: Build Docker Image
docker build -t chatbot-v2 .

Step 3: Run Model Container
docker run -d -p 8080:8080 chatbot-v2

Step 4: Access API Endpoint
```

```
Once deployed, the model can be accessed via:

POST https://api.lesliewrites.tech/v2/predict


Example request:

{
  "language": "de",
  "message": "Wie kann ich mein Passwort zurücksetzen?"
}


Example response:

{
  "intent": "password_reset",
  "confidence": 0.93,
  "response": "Um Ihr Passwort zurückzusetzen, klicken Sie bitte auf 'Passwort vergessen'."
}

4. Monitoring
| Metric  | Description         | Tool       |
| :------ | :------------------ | :--------- |
| Latency | Time per prediction | Prometheus |
| Uptime  | API availability    | Grafana    |
| Logs    | Model activity      | CloudWatch |


## 5. Deployment Steps

| Version | Change Summary       | Deployed |
| :------ | :------------------- | :------- |
| 1.3     | Initial release      | Feb 2025 |
| 2.0     | Multilingual upgrade | Nov 2025 |


6. Troubleshooting

Issue: Docker build fails
Fix: Ensure Docker daemon is running

Issue: API timeout
Fix: Check EC2 instance memory usage

7. Contact

Maintainer: Leslie Amadi
Email: dr.leslieamadi@lesliewrites.tech

Last Updated: 7 November 2025
```

```
---

### 🌍 4. Real-World Examples

| Company | Deployment Type | Example |
|:--|:--|:--|
| **OpenAI** | API Deployment | `api.openai.com/v1/...` (chat completions, embeddings) |
| **Google Cloud AI** | Model Hosting | Vertex AI endpoints |
| **Hugging Face** | Model Serving | Inference API + Spaces |
| **Anthropic** | Safety-first hosting | Claude’s secure inference pipelines |

---

### 🎯 5. Output Summary

Once you save, you’ll have:
✅ `model-deployment.md` → can export to `model-deployment.pdf`
✅ Add the link to your portfolio:
```