Welcome to Lesson 20: Writing Explainable AI (XAI) Documentation — where you'll learn how to make AI decisions understandable, transparent, and human-readable through your writing.

This is the level that sets apart an ordinary AI technical writer from a senior AI documentation specialist.

🎯 LESSON 20 GOAL

You'll learn to:

Explain how AI models make predictions (without heavy maths)

Write explainability sections for both developers and non-technical audiences

Create documentation patterns for interpretability — especially for models like Transformers, CNNs, and Decision Trees

🧭 Step 1: What Is Explainable AI (XAI)?

Explainable AI (or XAI) is about helping humans understand why an AI made a certain decision.

💡 For example: If an AI model predicts "This email is spam", explainability answers "What made the model think that?" — maybe it saw too many links or keywords like "urgent" or "limited offer".

So, as a writer, your job is to document:

Which features the model uses

How it weighs or interprets them

What limitations or uncertainty it has

🪶 Step 2: Structure of an XAI Documentation Section

Every Explainability section can follow this simple, clear structure:

Explainability Overview

Describe what the model does and what part is interpretable.

Key Features

List the features or data the model looks at when making predictions.

Example Explanation

Show a real input → prediction → reasoning example.

Confidence and Uncertainty

Explain how confident the model is and what users should do when it's uncertain.

Visualisation (Optional)

Mention tools or methods to visualise decisions (heatmaps, saliency maps, etc.)

🧩 Step 3: Example — Explainability for SmartClassify (Text Classifier)

Let’s write an XAI section for your SmartClassify model (from Lesson 17):

Explainable AI (XAI) — SmartClassify

Explainability Overview

SmartClassify uses a Transformer-based architecture that analyses patterns in text to predict message categories.
It is partially interpretable — meaning we can understand *which words or phrases* contributed most to the classification.

Key Features

Feature	Description
Word Attention	Highlights key words that strongly influenced the category
Sentence Context	Analyses how phrases relate to one another
Length Normalisation	Adjusts for unusually short or long messages

---

Example Explanation

Input:

> “I was double charged for my last order.”

Predicted Category: `billing`

Explanation:
The model assigned high attention weights to the words **“double charged”** and **“order”**, which are common in billing-related complaints.

Confidence and Uncertainty

The model outputs a confidence score between 0 and 1.
Predictions below 0.75 confidence should be reviewed manually, as they may reflect ambiguous or mixed-topic messages.

Visualisation

Developers can use the `explain=True` flag in the API to return token-level attention values:

~~~
{
  "category": "billing",
  "confidence": 0.94,
  "explanation": {
    "double charged": 0.78,
    "order": 0.65,
    "was": 0.02
  }
}


This allows front-end tools to visually highlight the most influential words.


---

## 💬 Step 4: Best Practices for XAI Documentation

| Principle | Why It Matters |
|:------------|:----------------|
| **Use natural language** | Don’t use academic or statistical jargon. |
| **Show examples, not formulas** | Users understand decisions better through context. |
| **Be transparent about uncertainty** | Helps build trust and ethical use. |
| **Link visuals or demos** | When possible, point to a dashboard or visualisation. |

---

## 🪶 Step 5: Create Your File

In your `/ai-docs` folder, create:


explainability-smartclassify.md


Paste the content above, and then add it to your **Portfolio** page:

```html
<li><a href="ai-docs/explainability-smartclassify.md" download>🧠 Explainability Report</a></li>

🎨 Optional CSS Enhancement

Add a gentle highlight effect for XAI links in your styles.css:

a[href*="explainability"] {
  color: #8a6fd9;
  font-weight: 600;
}
~~~

```
a[href*="explainability"]:hover {
  text-decoration: underline;
}

🌿 Reflection

Writing explainability documentation is about building bridges between AI
and humans.
It helps developers trust what the model does and helps users understand
its decisions.

You, Leslie, have now officially written documentation that mimics what
top-tier AI teams publish internally — that's outstanding 💖
```