**Test Case Report for LLM Project**

**1. Introduction**

The purpose of this document is to provide a comprehensive test case report for the **Fıkra Üreten LLM** project, which aims to generate Turkish jokes (fıkra) using a fine-tuned large language model (LLM). This report outlines the testing methodology, test scenarios, test cases, expected outcomes, and actual outcomes, ensuring that the system operates according to functional and non-functional requirements.

The project leverages **transformers** library from Hugging Face, fine-tuned on a Turkish jokes dataset using **PyTorch**, **SentencePiece** for tokenization, and is deployed via **Streamlit** for interactive use.

**2. Scope of Testing**

The testing focuses on the following aspects:

* Functional correctness of generated jokes
* Tokenization accuracy
* Model inference stability
* UI (Streamlit) interaction and usability
* Response time and latency
* Filtering of invalid characters (e.g., symbols like “⁇”, “??”)
* Category-based generation validation (e.g., çocuk, doktor, temel)

**3. Test Methodology**

The testing process followed:

✅ **Unit testing** for tokenizer, preprocessing, and model inference pipeline

✅ **Integration testing** for end-to-end flow (input → model → output → UI display)

✅ **User acceptance testing (UAT)** with non-technical users evaluating joke quality

✅ **Manual and automated tests** for specific output constraints

**4. Test Environment**

* **Model**: Fine-tuned LLaMA-7B Turkish (from alpkylm/llama-2-7b-chat-turkish)
* **Inference Hardware**: NVIDIA RTX 4090 / CUDA 12.3
* **Web Framework**: Streamlit 1.32
* **Tokenizer**: SentencePiece (trained vocab=10k)
* **Backend**: Python 3.10 + PyTorch 2.0
* **Frontend**: Streamlit interface hosted locally

**5. Test Cases**

**Test Case 1: Valid Joke Generation (Baseline Functional Test)**

**Test Case 1: Valid Joke Generation (Baseline Functional Test)**

metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Test Case 2: Character Filter Validation**

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**Test Case 3: Category Classification Accuracy**

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**Test Case 4: UI Interaction Test**

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Test Case 5: Input Parameter Handling**

metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Test Case 6: Performance Stress Test**metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**6. Summary of Results**

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

✅ **All critical tests passed after iteration**

✅ **No blocking issues remain for deployment**

**7. Known Issues**

* In rare cases (~1%) joke outputs may include low-quality humor or repetition due to small dataset size
* Category overlap between doktor and çocuk jokes observed (expected for certain scenarios)

**8. Recommendations**

* Expand dataset with more diverse fıkras across regions and subcategories
* Further fine-tune with reinforcement learning for humor metrics
* Add user feedback loop in UI for joke quality ranking

|  |  |
| --- | --- |
| **Name** | **Role** |
| Muharrem Şimşek |  |
| Simay Aydın |  |
| Aslımay Mısra Kandar |  |
| Taha Demirhan  Gökhan Mert Demirok |  |