# Test Case Report – Fine-tuned GPT-2 on Nasreddin Hoca Jokes

Project Title: LLM Risk Analysis

Model: GPT-2 (fine-tuned on Nasreddin Hoca joke dataset)

Dataset: nasreddin\_hoca\_jokes.json – containing paired input-output fields

Date: May 7, 2025

Prepared by: Team 9

## 1. Objective

The goal of this report is to evaluate the performance, consistency, and reliability of a fine-tuned Turkish GPT-2 model trained on a curated dataset of Nasreddin Hoca jokes. This evaluation focuses on six key aspects:  
- Whether the model generates valid outputs  
- Structural and grammatical correctness  
- Category alignment (i.e., jokes must fit the Nasreddin Hoca theme)  
- Semantic coherence and humor quality  
- Response time  
- Repetition of joke patterns

## 2. Test Case Summary

|  |  |  |
| --- | --- | --- |
| Test Case No | Test Name | Purpose |
| TC-FT-001 | Empty Output Generation | Checks if the model fails to generate any output |
| TC-FT-002 | Sentence Structure Test | Evaluates sentence completeness and grammar |
| TC-FT-003 | Category Inconsistency | Detects generation of non-Nasreddin Hoca content |
| TC-FT-004 | Humor Quality Assessment | Evaluates whether the output is actually a joke and contextually funny |
| TC-FT-005 | Response Time Measurement | Measures the average time taken to generate an output |
| TC-FT-006 | Joke Pattern Repetition | Assesses diversity of jokes generated |

## 3. Detailed Test Cases

### TC-FT-001: Empty Output Generation

Input: "Write a Nasreddin Hoca joke"

Actual Output: "" (blank output)

Status: ❌ Failed

Notes: This issue occurred in approximately 5% of generation attempts, especially under default decoding settings without sampling.

### TC-FT-002: Sentence Structure Test

Input: "Generate a Nasreddin Hoca joke involving the marketplace"

Actual Output: "Hoca went to the market. There tomato. Then smiled."

Status: ❌ Failed

Notes: Incomplete thoughts, broken sentence structure, missing connectors.

### TC-FT-003: Category Inconsistency

Input: "Write a joke about Hoca riding his donkey backward"

Actual Output: "Then Batman showed up and stopped him."

Status: ❌ Failed

Notes: Pop culture references suggest insufficient training or data leakage.

### TC-FT-004: Humor Quality Assessment

Input: "Nasreddin Hoca went to the judge"

Actual Output: "Judge: Why are you riding the donkey backward? Hoca: I’m not going forward, I’m returning!"

Status: ✅ Passed

Notes: Culturally appropriate, humorous, and coherent.

### TC-FT-005: Response Time Measurement

Input: Google Colab (T4 GPU)

Actual Output: ~1.3 seconds per prompt (Max: 2.5 seconds)

Status: ✅ Passed

Notes: Performance is acceptable for interactive applications.

### TC-FT-006: Joke Pattern Repetition

Input: 30 generations analyzed

Actual Output: 18 included the “riding the donkey backward” theme

Status: ⚠️ Warning

Notes: High repetition; improve sampling diversity (top-k, top-p).

## 4. Evaluation Summary

|  |  |  |
| --- | --- | --- |
| Evaluation Aspect | Status | Notes |
| Output Validity | ⚠️ Moderate | Occasional blank generations observed |
| Sentence Completeness | ❌ Poor | Fragmented and grammatically incorrect in some outputs |
| Category Adherence | ❌ Poor | Occasional generation of irrelevant or unrelated content |
| Humor Coherence | ✅ Strong | Successfully generated culturally relevant jokes |
| Response Time | ✅ Efficient | Fast enough for production use |
| Output Diversity | ⚠️ Needs Work | Frequent repetition of certain joke patterns |
| Overall Test Accuracy | ~70% | Moderate success with room for further optimization |

## 5. Recommendations

- Post-processing filters can be applied to remove repetitive phrases.  
- Improve category control by adding classification signals or prompts like "Generate a traditional joke about Hoca" during training.  
- Apply top-k sampling (k=50) or nucleus sampling (p=0.9) to increase diversity.  
- Extend the fine-tuning phase with more diverse examples and better padding strategies to prevent abrupt sentence endings.

# Test Case Report – Custom Sample LLM (Transformer-based)

Model: Custom LLM (trained from scratch using Transformer architecture)

Dataset: Same Nasreddin Hoca dataset (input-output structured)

Date: May 7, 2025

Prepared by: Team 9

## 1. Objective

This report evaluates a custom-built LLM trained from scratch using Transformer architecture on the same Nasreddin Hoca joke dataset. The goal is to compare its output quality against the fine-tuned GPT-2 model and analyze its performance based on clarity, structure, cultural relevance, and humor quality.

## 2. Test Case Summary

|  |  |  |
| --- | --- | --- |
| Test Case No | Test Name | Purpose |
| TC-SM-001 | Complete Sentence Generation | Checks if the model generates full and coherent sentences |
| TC-SM-002 | Category Alignment | Verifies alignment with Nasreddin Hoca theme |
| TC-SM-003 | Cultural Humor Detection | Evaluates the model’s understanding of traditional Turkish humor |
| TC-SM-004 | Response Time Measurement | Measures average output generation time |
| TC-SM-005 | Content Diversity | Checks the variation in joke themes |

## 3. Detailed Test Cases

### TC-SM-001: Complete Sentence Generation

Input: "Write a Nasreddin Hoca joke"

Actual Output: "Nasreddin Hoca once sold his neighbor’s donkey, saying he was only borrowing it for a lesson."

Status: ✅ Passed

Notes: All outputs consisted of full, grammatically correct sentences.

### TC-SM-002: Category Alignment

Input: "Write a traditional Nasreddin Hoca joke"

Actual Output: "When the judge asked why Hoca was late, he said the sun took longer today."

Status: ✅ Passed

Notes: The model produced thematically aligned and culturally relevant jokes.

### TC-SM-003: Cultural Humor Detection

Input: "Tell a joke in the style of Nasreddin Hoca"

Actual Output: "Hoca: I never lie! Unless it’s to avoid chores."

Status: ✅ Passed

Notes: The model captured the playful tone and humor style of Nasreddin Hoca accurately.

### TC-SM-004: Response Time Measurement

Input: Colab with small Transformer (scratch)

Actual Output: ~2.2 seconds per prompt

Status: ✅ Passed

Notes: Slightly slower than GPT-2 but consistent output.

### TC-SM-005: Content Diversity

Input: Generated 30 samples

Actual Output: Only 3 mentioned 'donkey', others varied

Status: ✅ Passed

Notes: Demonstrated high diversity across generated content.

## 4. Result Comparison: Fine-tuned GPT-2 vs Custom Sample LLM

|  |  |  |  |
| --- | --- | --- | --- |
| Test Case | Fine-tuned GPT-2 Result | Custom Sample LLM Result | Better Model |
| Sentence Structure | Fragmented and sometimes incomplete | Grammatically correct and complete | Custom LLM |
| Category Alignment | Occasional off-topic generations | Strong thematic alignment | Custom LLM |
| Humor Quality | Moderate with some cultural fit | High humor accuracy and tone | Custom LLM |
| Response Time | ~1.3s average | ~2.2s average | Fine-tuned GPT-2 |
| Diversity | Repetitive (donkey theme) | High variation in topics | Custom LLM |
| Overall Performance | ~70% test success | ~90% test success | Custom LLM |