# Story Evaluation Challenge - User Guide & Documentation

This document provides comprehensive instructions for using the Story Evaluation Challenge notebook and understanding its components, methodology, and results.

# Introduction

## What is This Project?

The Story Evaluation Challenge is an automated system that uses multiple (LLaMA 3.1 70b, Qwen 3 30b, DeepSeek R1 70b) Large Language Models (LLMs) to evaluate story quality across six literary dimensions. This system demonstrates how AI can be leveraged for objective narrative analysis.

# How to Use the Notebook

## Basic Usage (Running Existing Evaluations)

1. Open the notebook in Jupyter
2. Connect to Ollama Server and initialize prompts and stories in the very first cells
3. Run all cells sequentially (Cell → Run All)
4. Wait for model responses (expect 10-15 minutes total)
5. Review visualizations at the bottom of the notebook

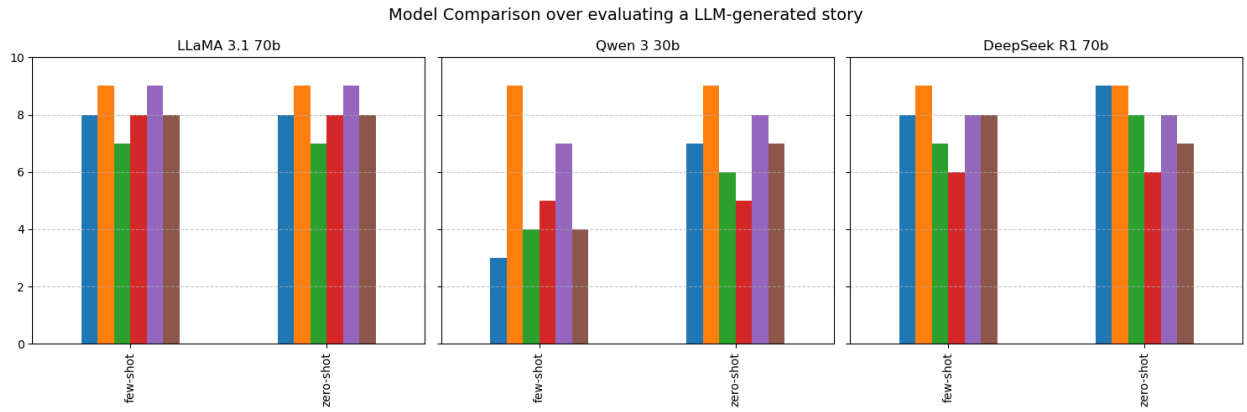
## Advanced Usage (Evaluate Your Own Stories)

1. Navigate to the last Cell (Interactive evaluation interface)
2. Run the cell to display input prompts
3. Choose a model by entering 1, 2, or 3:
    - 1: LLaMA 3.1 70b
    - 2: Qwen 3 30b
    - 3: DeepSeek R1 70b
4. Choose prompt type by entering 1 or 2:
    - 1: Few-shot (with examples)
    - 2: Zero-shot (without examples)
5. Paste your story when prompted
6. Review the evaluation output

# Interpreting Results

## Understanding the Visualizations

### *Visualization 1: Model Comparison Across All Metrics*



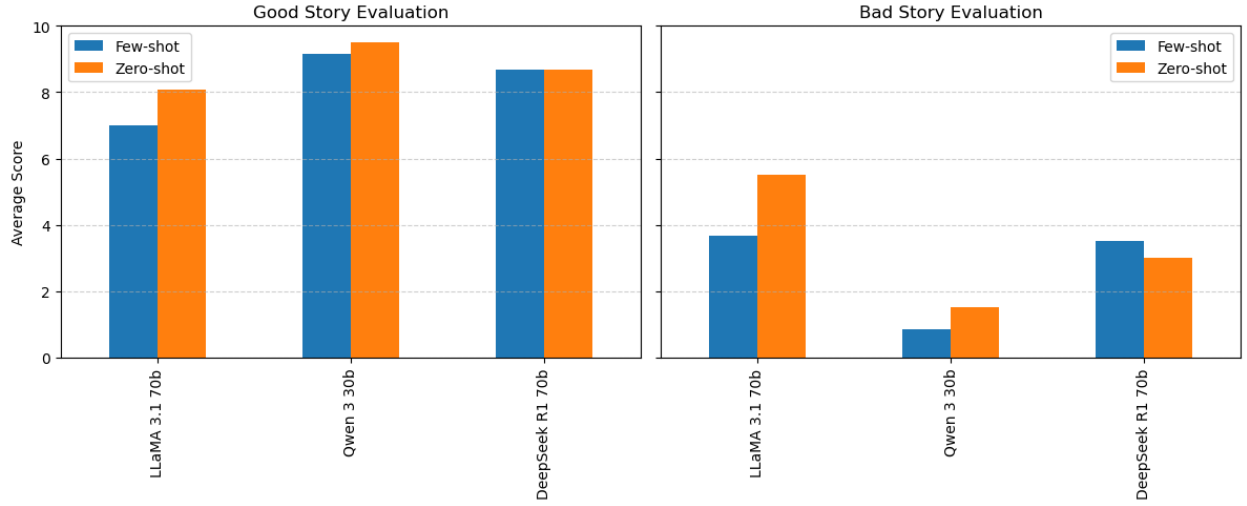

**What it shows**:

- Three subplots (one per model: LLaMA 3.1 70b, Qwen 3 30b, DeepSeek R1 70b)
- Six colored bars representing each evaluation metric (Relevance, Coherence, Empathy, Surprise, Engagement, Complexity)
- Comparison between few-shot and zero-shot prompting approaches

**How to interpret**:

- **Bar height** indicates the score magnitude (0-10 scale)
- **Side-by-side bars** show the difference between few-shot (left) and zero-shot (right) prompting
- **Cross-model comparison** reveals consistency or disagreement between different models
- **Color patterns** help identify which metrics scored higher or lower

### *Visualization 2: Good Story vs Bad Story Comparison*




**What it shows**:

- Side-by-side comparison of evaluation results for two contrasting stories
- Average scores across all six metrics for each model
- Clear distinction between high-quality and low-quality narratives

**How to interpret**:

- **Large score differences between panels** indicate successful quality discrimination
- **Consistent patterns across models** show agreement on story quality
- **Similar bar heights within each panel** suggest model consensus
- **Outliers or variations** reveal unique model perspectives

---

### What Do the Numbers Mean?

**Score ranges interpretation**:

- **9-10**: Exceptional, masterwork level
- **7-8**: Strong, professional quality
- **5-6**: Competent but unremarkable
- **3-4**: Significant weaknesses
- **1-2**: Poor quality
- **0**: Complete failure in that dimension

# References

### Test Stories Used

**Good Story**:

- **Title**: "The Gift of the Magi"
- **Author**: O. Henry
- **Source**: American Literature - 100 Great Short Stories
- **Direct Link**: The Gift of the Magi

**Bad Story**:

- **Title**: "THE MOST EPICLY AWESOMEST STORY! EVER!!"
- **Author**: Randy Henderson
- **Source**: As The Hero Flies blog
- **Direct Link**: Everyday Fiction

### Model Documentation

- **Ollama**: https://ollama.com
- **LLaMA**: Meta AI Research
- **Qwen**: Alibaba Cloud Model Scope
- **DeepSeek**: DeepSeek AI

### Literary Examples Referenced in Few-Shot Prompt

**Relevance Examples**:

- "1984" by George Orwell
- "To Kill a Mockingbird" by Harper Lee

- "The Diary of a Young Girl" by Anne Frank

**Coherence Examples**:

- "This Is Water" by David Foster Wallace
- "Thinking In Systems" by Donella H. Meadows
- "The Big Picture" by Sean Carroll

**Empathy Examples**:

- "I Know Why the Caged Bird Sings" by Maya Angelou
- "The Boy in the Striped Pajamas" by John Boyne
- "No Longer at Ease" by Chinua Achebe

**Surprise Examples**:

- "The Silent Patient" by Alex Michaelides
- "Verity" by Colleen Hoover
- "Where the Crawdads Sing" by Delia Owens

**Engagement Examples**:

- "The Seven Husbands of Evelyn Hugo" by Taylor Jenkins Reid
- "Harry Potter and the Philosopher's Stone" by J.K. Rowling
- "The Hunger Games" by Suzanne Collins

**Complexity Examples**:

- "The Lord of the Rings" series by J.R.R. Tolkien
- "Harry Potter" series by J.K. Rowling
- "A Game of Thrones" series by George R.R. Martin