

# Response to Reviewer Feedback

## Crafting Culturally Aligned Narratives: Large Language Models for Arabic Children’s Story Generation

We thank the reviewers for their constructive and thorough feedback. Below, we detail how we have addressed each concern in the revised manuscript.

### 1 Summary of Revisions

**5 Concerns Fully Addressed:** Incomplete visuals, dataset accessibility, statistical testing, fine-tuning evidence, writing quality

**3 Concerns Acknowledged as Limitations:** Small evaluator sample (mitigated), no GPT-4 baseline, dialectal coverage

### 2 Detailed Response to Reviewer Concerns

#### 2.1 [Fully Addressed] Incomplete Visuals

**Reviewer Concern:** “Several figures and diagrams are missing from the manuscript, which weakens presentation and clarity.”

**Our Response:**

We have added all missing figures to both the main paper and appendix:

**Main Paper (3 figures):**

- **Figure 1:** Dataset construction methodology diagram (Section 3)
- **Figure 2:** Story4Kids system architecture diagram (Section 4)
- **Figure 3:** Human evaluation results chart (Section 5.3)

**Appendix (5 additional figures):**

- App screenshots showing interactive story segmentation (Appendix Section 4)
- Training dataset structure examples (Appendix Section 5)
- Generated story samples for ages 3–5, 6–8, and 9–12 (Appendix Section 5)

All figures are now properly labeled, captioned, and referenced in the text for clarity.

#### 2.2 [Fully Addressed] Dataset Accessibility

**Reviewer Concern:** “Licensing details and source permissions are unclear; the dataset is not yet publicly available.”

**Our Response:**

We have released a 110-story subset on Hugging Face under the Creative Commons Attribution 4.0 International License (CC-BY 4.0):

`https://huggingface.co/datasets/houssamboukhalifa/culturally_aligned_arabic_stories_subset_a`

We added a detailed **Appendix Section 1: Dataset Availability and Ethical Considerations** covering:

- **Data sources:** Public platforms, YouTube subtitles, educational resources
- **Ethical sourcing:** Fair use compliance, manual screening for cultural appropriateness
- **Release strategy:** 110-story subset released; full 714-story dataset planned
- **Licensing:** CC-BY 4.0 for research purposes with proper attribution

This ensures reproducibility and enables further research in culturally aligned Arabic NLP.

## 2.3 [Fully Addressed] Statistical Significance Testing

**Reviewer Concern:** “Statistical significance testing is absent, and evaluation lacks quantitative validation.”

**Our Response:**

We have added comprehensive statistical analysis throughout the paper:

**Automatic Evaluation (Section 5.2):**

- **Table 4:** Statistical significance tests showing p-values for all metrics
- Paired t-tests comparing Original vs Fine-Tuned models
- Mann-Whitney U tests for non-parametric validation
- All improvements achieve strong significance:  $p < 0.01$  for most metrics

**Human Evaluation (Section 5.3):**

- Inter-rater reliability: ICC=0.847, Fleiss’  $\kappa$ =0.78
- Effect sizes: Cohen’s  $d > 0.8$
- Significance tests:  $p < 0.001$  for cultural alignment improvements

We also added emphasis on robustness: *“Despite the modest evaluator count, high inter-rater reliability (ICC=0.847) indicates findings are robust and consistent across raters.”*

## 2.4 [Acknowledged] Limited Human Evaluation Sample

**Reviewer Concern:** “Human evaluation is limited to a small sample of five educators.”

**Our Response:**

While we acknowledge this limitation, we have taken steps to demonstrate the robustness of our findings:

**Statistical Robustness:**

- ICC=0.847 indicates **excellent agreement** among evaluators
- Fleiss’  $\kappa$ =0.78 shows **substantial inter-rater consistency**
- These metrics indicate that results are reliable despite the modest sample size

**Acknowledgement:**

- Added to Limitations section: “Five-educator evaluation represents limited geographic sampling”
- Future work explicitly mentions: “Conduct large-scale user studies across Arabic-speaking regions”

**Evaluator Expertise:**

- All evaluators: 6–15 years experience in child education
- Native Arabic speakers from North Africa (Algeria/Morocco)
- Specialists in child development and Islamic cultural education

We believe the high inter-rater reliability provides confidence in the findings. Expanding to more regions is planned for future work but requires additional resources.

## 2.5 [Fully Addressed] Models Are Fine-Tuned, Not Just Prompted

**Reviewer Concern:** “The work is primarily demonstrative rather than scientific. No models are fine-tuned only prompted comparisons are made.”

**Our Response:**

**This concern is factually incorrect.** We conducted full fine-tuning with extensive methodology:

**Section 4: Model Training** clearly describes:

- **LoRA-based parameter-efficient fine-tuning** with rank  $r = 16$ ,  $\alpha = 32$
- **4 epochs of training** on 643 stories, 71 validation stories
- **Perplexity reductions:** 48% (Silma), 47% (Gemini)
- **Validation loss improvements:** 33% (Silma), 38% (Gemini)
- **Cultural alignment improvements:** 68.3%→91.5% (Silma), 78.9%→96.8% (Gemini)

**Table 3 (Baseline Comparison)** explicitly shows:

- “Silma Original” vs “Silma Fine-Tuned”
- “Gemini Original” vs “Gemini Fine-Tuned”
- Quantitative improvements across all metrics (BLEU: +75% for Silma, +46% for Gemini)

**Appendix Section 2** provides extensive fine-tuning details:

- Three-phase training process
- Hyperparameter specifications (AdamW, learning rate schedules, gradient clipping)
- Decoding parameter optimization
- Prompt engineering for cultural alignment

**This is NOT a prompting-only study.** We conducted rigorous fine-tuning with LoRA and demonstrated significant improvements through both automatic and human evaluation, with statistical significance testing.

## 2.6 [Acknowledged] Lack of GPT-4/Claude Baseline

**Reviewer Concern:** “No performance benchmark against strong general LLMs (e.g., GPT-4 or Claude) for Arabic story generation.”

**Our Response:**

We acknowledge this limitation. Our current baseline comparison focuses on Original vs Fine-Tuned versions of Silma and Gemini models.

**Evidence of Fine-Tuning Benefits:**

- Table 3 shows clear improvements from fine-tuning:
  - Silma: BLEU +75%, Similarity +39%, Distinct-1 +50%, Distinct-2 +40%

- Gemini: BLEU +46%, Similarity +44%, Distinct-1 +64%, Distinct-2 +55%
- All improvements are statistically significant ( $p < 0.01$ )

#### **Challenges with GPT-4/Claude Comparison:**

1. API access for Arabic fine-tuning not publicly available for GPT-4
2. Cultural alignment evaluation: GPT-4 may lack deep Islamic cultural knowledge
3. Cost constraints for generating 150+ stories across multiple models
4. Different architectural paradigms make direct comparison challenging

**Contribution:** We have demonstrated that fine-tuning significantly improves performance over original models for culturally aligned Arabic storytelling. Future work could include comparison with other Arabic-capable models as they become available.

## **2.7 [Fully Addressed] Writing and Formatting Issues**

**Reviewer Concern:** “Occasional long sentences, inconsistent model name formatting, and redundant references affect readability.”

#### **Our Response:**

We have revised the manuscript for improved clarity and consistency:

- Standardized model names throughout: Gemini 2.0, Silma 9B, Noon 7B, Jais 13B, BLOOM 7B
- Shortened long sentences and improved paragraph flow
- Cleaned up bibliography and removed redundant citations
- Improved table formatting for consistency (aligned columns, uniform decimal places)
- Added clear section headings and transitions

## **2.8 [Acknowledged] Dialectal Coverage**

**Reviewer Concern:** “Focuses solely on Modern Standard Arabic, without addressing dialectal or regional variation.”

#### **Our Response:**

We acknowledge this limitation and have added it to the paper:

#### **Rationale for Modern Standard Arabic (MSA):**

- MSA is the **formal language used in education** across all Arab countries
- Most appropriate for **children’s educational content**
- Ensures **consistency and comprehension** across diverse Arabic-speaking regions
- Standard for written children’s literature in the Arab world

#### **Acknowledgement in Limitations:**

- Explicitly stated: “Focus on Modern Standard Arabic excludes dialects”
- Added to Future Work: “Incorporate regional dialects for broader inclusivity”

While dialectal variation is important for future extensions, MSA provides the most suitable foundation for educational children’s stories that can be understood universally across the Arab world.

### 3 Summary of Key Improvements

The revised manuscript now provides:

1. **Complete visual documentation:** 8 figures across main paper and appendix
2. **Public dataset release:** 150 stories on Hugging Face with CC-BY 4.0 license
3. **Rigorous statistical validation:** t-tests, Mann-Whitney U, ICC, Cohen’s *d*, effect sizes
4. **Clear fine-tuning evidence:** LoRA methodology, perplexity improvements, validation metrics
5. **Robust evaluation:** High inter-rater reliability (ICC=0.847) validates findings
6. **Transparent limitations:** Acknowledged evaluator sample size, dialectal coverage, baseline scope
7. **Improved presentation:** Consistent formatting, clear writing, proper citations

### 4 Conclusion

We believe these revisions substantially strengthen the scientific contribution and address the core concerns raised by reviewers. The paper now demonstrates:

- **Reproducible methodology** with public dataset and detailed training procedures
- **Statistical rigor** with comprehensive significance testing
- **Clear evidence** that fine-tuning (not prompting) drives improvements
- **Robust findings** despite modest evaluator count (high ICC validates consistency)
- **Honest discussion** of limitations and well-defined future work

We appreciate the reviewers’ feedback, which has significantly improved the quality and clarity of our work.