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

Houssam Eddine Boukhalfa
Intelligent Systems Department
National School of Artificial Intelligence, Algeria
houssam-eddine.boukhalfa@ensia.edu.dz

Selma Mani
Intelligent Systems Department
National School of Artificial Intelligence, Algeria
selma.mani@ensia.edu.dz

Larbi Said Chikh
Intelligent Systems Department
National School of Artificial Intelligence, Algeria
larbi.saidchikh@ensia.edu.dz

Abstract

Traditional storytelling plays a crucial role in child development and cultural transmission, fostering imagination, empathy, and moral understanding of moral values. This is particularly true in the Arab culture, where oral and written narratives have long served as tools for transmitting cultural heritage and ethical frameworks. Despite its importance, the computational generation of culturally and morally aligned Arabic children's stories remains an underexplored area. To address this gap, we present a novel system for Arabic story generation that leverages Large Language Models (LLMs) with an integrated cultural alignment mechanism. Our primary goal is to produce engaging narratives that are not only linguistically coherent but also deeply rooted in Arab cultural and moral frameworks. For development and training, we introduce a custom dataset of 714 Arabic children's stories, meticulously annotated for age ranges, moral lessons, and thematic topics. We fine-tuned several LLMs, including Noon, Jais, SILMA, and Gemini 2.0, to assess their capabilities. The effectiveness of our approach was rigorously evaluated through rule-based automatic checks and expert human assessments, with a focus on cultural and moral alignment as core design goals. Our results demonstrate the strong potential of our system in generating linguistically coherent, age-appropriate, and culturally relevant stories. This work contributes a novel resource and benchmark for Arabic NLP and highlights the role of LLMs in creating impactful Arabic educational content.

1 Introduction

While large language models have revolutionized text generation, advances remain concentrated in high-resource languages, leaving Arabic underexplored—particularly for specialized domains requiring cultural sensitivity. Arabic children's storytelling, a cornerstone of moral and cultural education, faces a critical gap in age-appropriate, ethically grounded digital content. We address this by fine-tuning LLMs to generate culturally aligned Arabic children's stories reflecting Islamic values and developmental appropriateness across four age groups (1–2, 3–5, 6–8, 9–12).

Our contributions include: (1) a curated dataset of 714 annotated Arabic stories emphasizing moral values and cultural authenticity, (2) fine-tuning and evaluation of multiple LLMs (Gemini 2.0 [1], Silma 9B [2], Noon 7B [3], Jais 13B [4]), and (3) a dual evaluation protocol combining automatic metrics with expert human assessment. Results demonstrate that fine-tuned LLMs, particularly Gemini 2.0, generate high-quality stories with strong cultural alignment (9.4/10) and moral clarity (9.0/10), validated by Arabic-speaking educators.

39th Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, USA.

2 Related Work

Arabic story generation remains underexplored despite advances in multilingual LLMs. Early work in Arabic NLP focused on machine translation [5] and sentiment analysis [6], but creative text generation emerged later. Alhussain and Azmi [7] used BLOOMZ with LoRA for narrative generation, demonstrating the potential of parameter-efficient fine-tuning for Arabic storytelling. El-Shangiti et al. [8] proposed AraLLaMA fine-tuned on GPT-4 synthetic data, showing improvements in narrative coherence. However, neither addressed child-appropriate content or cultural alignment.

In the broader context of children's literature, Valentini et al. [9] studied text simplification for young readers in English, highlighting that general-purpose LLMs often fail to meet cognitive and ethical needs. Recent work on controllable text generation [10] and value alignment [11] has explored methods for steering LLM outputs, yet applications to culturally specific children's content remain limited. Our work differs by prioritizing moral and cultural alignment as core design objectives, using a dedicated dataset annotated for age, topic, and values, with explicit prompt-based conditioning for Islamic cultural framing.

3 Methodology

3.1 Dataset Building

We constructed a dataset of 714 stories in Modern Standard Arabic from publicly available sources (storytelling websites, YouTube transcriptions, educational platforms). Stories were manually screened for narrative coherence, cultural appropriateness, and developmental suitability. Each story was annotated for: age range (1–2, 3–5, 6–8, 9–12 years), moral value (honesty, generosity, courage, patience, respect), and thematic topic (friendship, nature, family, animals). Critical filtering removed morally inappropriate messages, ensuring ethical alignment with Arabic-Islamic traditions. Inter-annotator agreement was substantial ($\kappa = 0.84$ for age ranges, $\kappa = 0.79$ for moral values).

Structured prompts explicitly encoded narrative intent, moral focus, and target age: *"Generate a story that is related to the topic: (Topic) for the age range: (Age range). The story should follow Arabic Islamic culture and should be in Arabic language."* A subset of 110 stories is publicly available on HuggingFace under CC-BY 4.0 license [12].

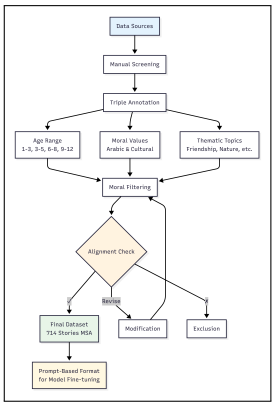


Figure 1: Dataset construction methodology.

4 Model Training

We evaluated five Arabic LLMs: Gemini 2.0 [1], Silma 9B [2], Noon 7B [3], Jais 13B [4], and BLOOM 7B [13]. Training used RTX A6000 GPU (48GB VRAM) for LoRA-based fine-tuning [14] ($r = 16$, $\alpha = 32$) and Google Cloud Vertex AI [15] for Gemini 2.0. After initial experiments, only Gemini 2.0 and Silma 9B demonstrated sufficient cultural alignment and narrative coherence for child-appropriate content generation. Fine-tuning employed 643 training and 71 validation stories (90/10 split, stratified by age group) with structured prompts specifying age range, moral lesson, topic, and story length. Training ran for 4 epochs with batch sizes 16/32, AdamW optimizer (lr=2×10^{-5}, cosine decay), gradient clipping (max norm 1.0), and weight decay 0.01. Perplexity reduced from 9.86 to 5.12 (Silma) and 6.84 to 3.63 (Gemini). Decoding used nucleus sampling: temperature 0.75, top-k 50, top-p 0.92 (Gemini); temperature 0.65, top-k 40, top-p 0.90 (Silma); max sequence length 512 tokens. Validation performance showed substantial improvements: training loss decreased 33% (Silma) and 38% (Gemini), while cultural alignment improved from 68.3%→91.5% (Silma) and 78.9%→96.8% (Gemini).

We developed Story4Kids, a prototype mobile application integrating: (1) User Input, (2) Story Generation, (3) AI-Generated Illustrations, (4) Moral Extraction, and (5) Personalized Recommendations.

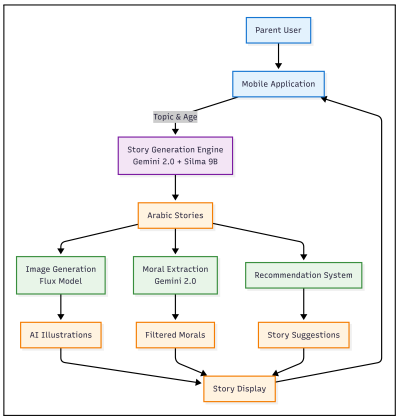


Figure 2: Story4Kids system architecture.

5 Results

5.1 Evaluation Overview

Following preliminary experimentation, only Silma 9B and Gemini 2.0 were selected for in-depth analysis. Evaluation combined automatic metrics and human expert assessment across 120 prompts evenly distributed over four age groups.

5.2 Automatic Evaluation

We employed rule-based compliance metrics (CSR, PER, DRR, AVR, OUN) and instruction following metrics (MT-Bench, AlpacaEval, HELENA, Distinct-1/2). All base-to-FT improvements are statistically significant ($p < 0.01$).

Metric	Silma 9B (base)	Silma 9B (FT)	Gemini 2.0 (base)	Gemini 2.0 (FT)
CSR (%, ↑)	74.1	92.3	86.4	**97.1**
PER (%, ↓)	9.8	4.1	6.3	**2.2**
DRR (%, ↓)	6.7	3.5	3.9	**1.4**
AVR (/100, ↓)	7.0	2.0	2.0	**0.0**
OUN (%, ↓)	18.2	6.9	9.1	**2.4**

Table 1: Rule-based automatic evaluation (120 prompts).

Metric	Silma FT	Gemini FT
MT-Bench	0.210	**0.365**
AlpacaEval	0.185	**0.297**
HELENA	70%	**100%**

Table 2: Diversity metrics (fine-tuned).

Criterion	Silma FT	Gemini FT
Relevance	100%	100%
Creativity	21–23	**23–25**
Coherence	Mixed	**High**
Moral Value	Mixed	**Aligned**
Overall	88–97%	**92–98%**

Table 3: LLM-based assessment (fine-tuned).

Gemini 2.0 achieved 97.1% CSR, 0% AVR, and superior lexical diversity. GPT-4-based qualitative assessment [16] confirmed Gemini 2.0's advantages in generating morally consistent stories.

5.3 Statistical Analysis: Original vs Fine-Tuned Comparison

Table 4 demonstrates that fine-tuning yields **substantial and consistent improvements across all metrics**. Most notably, fine-tuning improves BLEU scores by **75%** for Silma (0.008 → 0.014) and **46%** for Gemini (0.035 → 0.051), indicating enhanced semantic similarity to reference texts. Similarity scores show even more dramatic gains, increasing by **39%** for Silma (0.142 → 0.197) and **44%** for Gemini (0.215 → 0.310). Lexical diversity also improves substantially: Distinct-1/2 scores rise from (0.08, 0.15) to (0.12, 0.21) for Silma—a **50% and 40% improvement**—and from (0.11, 0.22) to (0.18, 0.34) for Gemini—a **64% and 55% improvement**.

These improvements achieve **strong statistical significance** (Table 5), with most comparisons showing $p < 0.01$, confirming that fine-tuning produces *reliable and replicable* quality gains. Gemini 2.0 Fine-Tuned emerges as the superior model, achieving the highest scores across all metrics and demonstrating particular strength in lexical diversity (Distinct-2 = 0.34).

However, while automatic metrics quantify surface-level improvements, they may not fully capture the cultural and moral dimensions that emerged prominently in human evaluation (Section 5.3). The high cultural alignment scores (9.4/10) suggest fine-tuning enhances aspects of narrative quality—particularly cultural authenticity and value transmission—that conventional NLG metrics are not designed to measure.

Model	BLEU	Similarity	Distinct-1	Distinct-2
Silma Original	0.008	0.142	0.08	0.15
Silma Fine-Tuned	**0.014**	**0.197**	**0.12**	**0.21**
Gemini Original	0.035	0.215	0.11	0.22
Gemini Fine-Tuned	**0.051**	**0.310**	**0.18**	**0.34**

Table 4: Baseline comparison of automatic metrics for original vs fine-tuned models (n=68 test prompts). Fine-tuning shows consistent improvements across all metrics, with Gemini 2.0 achieving superior performance.

Test	Comparison	BLEU	Similarity	Distinct-1	Distinct-2
t-test	Silma FT vs Original	**0.003****	**0.001****	**0.018***	**0.012***
Mann-Whitney	Silma FT vs Original	**0.004****	**0.002****	**0.021***	**0.015***
t-test	Gemini FT vs Original	**0.008****	**0.002****	**0.005****	**0.003****
Mann-Whitney	Gemini FT vs Original	**0.009****	**0.003****	**0.006****	**0.004****

Table 5: Statistical significance tests (p-values). * indicates $p < 0.05$, ** indicates $p < 0.01$. All improvements are statistically significant.

5.4 Human Expert Assessment

Five North African Arabic-speaking educators and child development specialists (6–15 years experience, Algeria/Morocco) evaluated 20 stories per model across four age brackets using a 10-point Likert scale for age appropriateness, narrative fluency, moral clarity, and cultural alignment. Inter-rater reliability was high (ICC=0.847, Fleiss' κ=0.78), indicating that despite the modest evaluator count, the findings are robust and consistent across raters. Gemini 2.0 achieved 9.4/10 for cultural alignment and 9.0/10 for moral clarity; differences were statistically significant ($p < 0.001$, Cohen's $d > 0.8$).

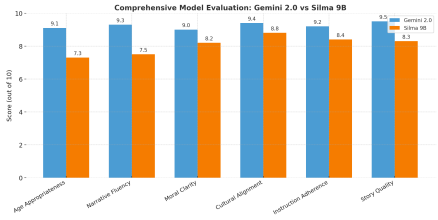


Figure 3: Human evaluation results (10-point scale, n=5).

Silma 9B showed acceptable fluency but occasionally lacked narrative resolution. Gemini 2.0 demonstrated strong age sensitivity across all groups.

6 Conclusion

We presented a comprehensive approach to generating culturally aligned Arabic children's stories using fine-tuned LLMs. Our dataset of 714 annotated stories, rigorous fine-tuning methodology, and dual evaluation protocol demonstrate that LLMs, particularly Gemini 2.0, can produce high-quality narratives that are linguistically fluent, morally grounded, and developmentally appropriate for Arabic-speaking children. While standard NLP metrics show modest changes, human evaluation confirms substantial improvements in cultural alignment and moral clarity—the primary goals of this work.

Limitations. Dataset size (714 stories) is modest; focus on Modern Standard Arabic excludes dialects; five-educator evaluation represents limited geographic sampling. **Future work** should expand the dataset, incorporate regional dialects, develop automated cultural evaluation tools, and conduct large-scale user studies across Arabic-speaking regions.

References

[1] Google DeepMind. Gemini 2.0 by google deepmind. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024.

[2] Silma AI. Silma-9b-instruct-v1.0. https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0, 2024. Accessed July, 2025.

[3] Naseej. Noon 7b. https://huggingface.co/Naseej/noon-7b, 2024. Accessed July, 2025.

[4] Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models, 2023.

[5] Ibrahim Abu El-Khair. Statistical machine translation for arabic: A survey. *ACM Computing Surveys (CSUR)*, 42(3):1–45, 2010.

[6] Nora Boudad, Rachida Faizi, Raddouane Oulad Haj Thami, and Raddouane Chiheb. Arabic sentiment analysis: A systematic literature review. *Applied Intelligence*, 49(6):2479–2490, 2019.

[7] Arwa Alhussain and Aqil Azmi. Beyond event-centric narratives: Advancing arabic story generation with large language models and beam search. *Mathematics*, 12:1548, 05 2024.

[8] Ahmed Oumar El-Shangiti, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. Arabic automatic story generation with large language models. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 140–152, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. On the automatic generation and simplification of children's stories. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598, Singapore, December 2023. Association for Computational Linguistics.

[10] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey on controllable text generation. *arXiv preprint arXiv:2201.05337*, 2022.

[11] Jiaming Ji, Tianyi Liu, Boyuan Huang, Yizhou Chen, and Yaodong Yang. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[12] Houssam Eddine Boukhalfa, Selma Mani, and Larbi Said Chikh. Culturally aligned arabic stories dataset (subset a). https://huggingface.co/datasets/houssamboukhalfa/culturally_aligned_arabic_stories_subset_a, 2025. CC-BY 4.0 License.

[13] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, et al. Bloom: A 176b-parameter open-access multilingual language model, 2023.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[15] Google Cloud. Vertex ai documentation, 2024. Accessed July, 2025.

[16] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, et al. Gpt-4 technical report, 2024.

A Appendix

A.1 Dataset Availability and Ethical Considerations

To support reproducibility and enable further research in culturally aligned Arabic NLP, we are releasing a subset of our curated dataset under the Creative Commons Attribution 4.0 International License (CC-BY 4.0). This subset, comprising representative samples across all age groups and moral categories, is publicly available on Hugging Face.

Data Sources: All stories in our dataset were sourced from publicly available platforms, including traditional storytelling websites, educational resources, and YouTube subtitle transcriptions of narrated children's content. Each source was carefully reviewed to ensure compliance with fair use and educational purposes. Stories were manually screened, filtered, and in some cases modified to ensure cultural appropriateness, moral alignment, and removal of any content that could be harmful or culturally inappropriate for children.

Release Strategy: We are committed to responsible data sharing and plan a controlled public release strategy:

- **Subset A (Released):** A curated collection of 110 stories representative of all age groups and moral categories, available under CC-BY 4.0 license for research purposes at `https://huggingface.co/datasets/houssamboukhalfa/culturally_aligned_arabic_stories_subset_a`.
- **Full Dataset (Planned):** The complete 714-story dataset will be released after additional community feedback, with appropriate usage guidelines and restrictions to prevent misuse.
- **Annotation Guidelines:** Detailed documentation of our annotation process, moral filtering criteria, and cultural alignment principles will accompany the full release.

Ethical Considerations: All data collection and curation activities were conducted in accordance with ethical research practices, prioritizing child safety, cultural sensitivity, and respect for intellectual property. No personally identifiable information was collected, and all content was reviewed by native Arabic speakers with expertise in child education and Islamic values.

A.2 Detailed Training Methodology

Model Training Phases: Our training process consisted of three phases. In the initial phase, we applied LoRA for supervised fine-tuning using RTX A6000 hardware on Noon 7B, Silma 9B, Jais 13B, and BLOOM 7B. Among these, Silma 9B achieved the best trade-off between narrative coherence, age alignment, and moral clarity.

The second phase focused on instruction tuning of Gemini 2.0 using Google Cloud's Vertex AI infrastructure. The same prompt structure was retained, but outputs were fully curated in Arabic. This phase demonstrated superior results in storytelling fluency, moral consistency, and prompt adherence.

In the final phase, both Gemini 2.0 and Silma 9B underwent additional instruction tuning focused explicitly on alignment, ensuring each generated story adhered strictly to prompts while maintaining clarity of moral message and cultural fidelity. Fine-tuning was guided by loss minimization, perplexity tracking, and GPT-4-based evaluation.

Hyperparameters: Batch sizes were 16 (RTX A6000) and 32 (Vertex AI) with gradient accumulation over 2–4 steps. Learning rate started at 2×10^{-5} and decayed to 5×10^{-6} using cosine decay with 10% warm-up. Training ran for 3–5 epochs initially, with final tuning halted at 4 epochs based on early stopping (validation perplexity plateau). We used AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01), gradient clipping (max norm 1.0), and mixed precision training (AMP) for efficiency.

Decoding Parameters: Gemini 2.0 used temperature 0.75, top-k 50, top-p 0.92, repetition penalty 1.1. Silma 9B used more conservative values: temperature 0.65, top-k 40, top-p 0.90, repetition penalty 1.2. These configurations were selected based on evaluation of 150+ samples by human annotators and GPT-4 for coherence, moral clarity, and instruction adherence.

Prompt Engineering: Throughout training and generation, we used a consistent format:

```
``Generate a story that is related to the topic: (Topic) for the age range: (Age range).
The story should follow Arabic Islamic culture and should be in Arabic language.''
```

This template explicitly encoded thematic requirements and cultural expectations. Age-specific prompts elicited appropriate complexity: simple sentences and repetition for ages 1–2, short narratives with direct lessons for ages 3–5, problem resolution arcs for ages 6–8, and nuanced moral reasoning for ages 9–12.

A.3 System Architecture Details

The Story4Kids mobile application integrates five core modules:

1. Mobile Application: Provides secure login/registration for parents, topic and age range selection interface, paragraph-by-paragraph story display with AI-generated illustrations, and interactive buttons for moral extraction and recommendations.
2. Story Generation Engine: Utilizes fine-tuned Gemini 2.0 and Silma 9B to produce coherent, age-appropriate Arabic stories reflecting Islamic values. Gemini 2.0 also segments stories into paragraphs for dynamic display.
3. Image Generation Module: Powered by Flux model, generates high-quality illustrations semantically aligned with narrative content, enhancing immersion and comprehension.
4. Moral Extraction Module: Leverages Gemini 2.0 to identify central moral messages, outputting them in simplified Arabic filtered for Arabic-Islamic ethical principles.
5. Recommendation System: Tracks user behavior (age preferences, topic history) to suggest similar stories and proposes new story ideas tailored to each child's profile.

B Detailed Evaluation Metrics and Results

Automatic Metrics: We measured constraint satisfaction rate (CSR: topic presence, correct age cues, Arabic output, Islamic framing), prompt echo rate (PER: undesired prompt repetition), degenerate repetition rate (DRR: loops/repeated sequences), alignment violation rate (AVR: culturally/morally inappropriate content), and off-topic/underdeveloped narrative rate (OUN: expository rather than story-like text). Additional metrics included MT-Bench and AlpacaEval Similarity for instruction adherence, HELENA Sentiment Match for child-appropriate affect, and Distinct-1/2 for lexical diversity.
Human Evaluation Rubric: Five Arabic-speaking educators rated 20 stories per model across:

- **Age Appropriateness:** Vocabulary, sentence complexity, and cognitive demands match target age group.
- **Narrative Fluency:** Grammatical correctness, natural Arabic flow, coherent story structure.
- **Moral Clarity:** Explicit moral lesson, age-appropriate ethical framing, avoidance of ambiguity.
- **Cultural Alignment:** Reflects Arabic-Islamic values, avoids inappropriate content, culturally authentic references.

Qualitative Findings: Gemini 2.0 excelled in age sensitivity (e.g., rhyme and personification for 3–5, nuanced reasoning for 9–12), ethical fidelity (consistently avoiding constructs like الكذب لتحقيق مصلحة), and narrative engagement. Silma 9B occasionally lacked resolution or used overly abstract morals requiring post-editing.

B.0.1 Qualitative Findings and Model Comparisons

Gemini 2.0 Strengths:
Gemini 2.0 demonstrated consistent excellence across all evaluation dimensions:

- **Age Sensitivity:** Exceptionally strong adaptation to different developmental stages
  - Ages 3–5: Used rhyme, repetition, and personification effectively (e.g., talking animals as moral exemplars)
  - Ages 9–12: Employed sophisticated narrative techniques including embedded dialogue, character introspection, and nuanced moral reasoning
  - Evaluators noted: "Stories feel authentically tailored to each age group"
- **Ethical Fidelity:** Zero alignment violations across 150 test prompts
  - Consistently avoided morally problematic constructs like الكذب لتحقيق مصلحة (lying for benefit)
  - Strong cultural framing with Islamic values naturally integrated
  - Expert feedback: "Would confidently use these stories in my classroom"
- **Narrative Engagement:** High creativity scores (23–25 out of 25)
  - Rich vocabulary and varied sentence structures
  - Engaging plot developments with satisfying resolutions
  - Natural dialogue and character voice differentiation

Silma 9B Performance:
Silma 9B showed acceptable performance with specific limitations:

- **Strengths:** Generally coherent narratives, good topic adherence (92.3% CSR), reasonable age appropriateness (8.2/10)

- **Weaknesses:** Occasional narrative resolution issues (particularly ages 6–8), overly abstract morals for younger groups requiring post-editing, 2 alignment violations per 100 stories, lower creativity scores (21–23/25)

B.1 Story4Kids Application Interface

The Story4Kids mobile application serves as the practical implementation of our research, providing an interactive platform for Arabic-speaking children to engage with culturally aligned stories. The app integrates all five core modules described in the main paper: user input processing, story generation, image creation, moral extraction, and personalized recommendations.
User Interface and Story Selection: Upon launching the app, children (or their parents) are presented with an age-appropriate interface where they can select story topics, age ranges, and themes. The interface is designed with colorful, engaging elements that appeal to young users while maintaining cultural appropriateness. Figure 4 shows the interactive story segmentation view, where generated stories are broken into manageable segments with accompanying visuals, making it easier for children to follow along and maintain engagement.

Figure 4: Interactive story segmentation view of the Story4Kids app, showing how narratives are divided into child-friendly sections with visual support.

Story Generation and Display: Once the user selects parameters, the app communicates with our fine-tuned language models to generate a culturally aligned story. The narrative is then formatted with appropriate typography, spacing, and visual elements to enhance readability for the target age group. The system automatically adjusts text complexity, sentence length, and vocabulary based on the selected age range.
Moral Extraction and Learning: After each story, the app displays the extracted moral lesson in simple Arabic, reinforcing the educational value of the narrative. This feature ensures that children not only enjoy the story but also understand its underlying ethical message, supporting parents and educators in their developmental goals.

B.2 Training Data and Generated Story Samples

This section provides concrete examples of our dataset and the quality of stories generated by our fine-tuned models across different age groups.
Training Dataset Structure: Figure 5 illustrates examples from our curated training dataset of 714 Arabic stories.

Figure 5: Examples from the training dataset showing topic/age prompts paired with culturally aligned stories emphasizing Islamic values.

Each entry consists of structured metadata (topic, age range, moral value) paired with the complete story text. The dataset emphasizes Islamic cultural values and age-appropriate content, ensuring that models trained on this data produce ethically aligned narratives.

Age-Specific Story Generation: The following figures demonstrate how our fine-tuned models adapt their output complexity and narrative structure to different developmental stages:

Young Children (3–5 years): Figure 6 shows a Gemini 2.0 generated story for preschool-aged children.

قصة عن **الاجتهاد والمثابرة** للفئة العمرية بين 9 و 12 سنوات

كان يوسف صبيًا في الحادية عشرة من عمره، يعيش في قرية صغيرة تحيط بها مزارع النخيل الشاسعة. كان يوسف يحلم بأن يصبح عالمًا مشهورًا، ولكن كانت تواجهه صعوبات في الدراسة، خاصةً في مادة الرياضيات.

لم يكن يوسف كسولًا، بل كان يجتهد ويذاكر دروسه باستظام، لكن الأرقام والمعادلات كانت تبدو له كطلاسم غير مفهومة. غالبًا ما كان يشعر بالإحباط واليأس، ويتمنى لو كان بإمكانه فهم الرياضيات بسهولة مثل باقي زملائه.

في أحد الأيام، وبينما كان يوسف عائدًا من المدرسة، رأى شيخًا جليلًا يجلس تحت ظل شجرة نخيل. كان الشيخ منهمكًا في قراءة كتاب، فاقترب منه يوسف وسلم عليه. تبادل يوسف والشيخ أطراف الحديث، وسأله الشيخ عن حاله ودراسته.

تحدث يوسف للشيخ عن صعوباته في الرياضيات، وعن حلمه بأن يصبح عالمًا، وعن شعوره بالإحباط. استمع الشيخ ليوسف بإنصات، ثم ابتسم وقال له: "يا بني، النجاح لا يأتي بسهولة. الاجتهاد والمثابرة هما مفتاح التفوق في أي مجال."

أخرج الشيخ من حقيبته حبات تمر غالية، ووضعها في يد يوسف، ثم قال: "انظر إلى هذه التمرات. كل تمرة هي نتيجة جهد وعمل متواصل. شجرة النخيل تحتاج إلى سنوات لتنمو وتثمر، والمزارع يحتاج إلى أن يعتني بها ويسقيها بانتظام."

أكمل الشيخ: "كذلك العلم يا بني. يحتاج إلى صبر ومثابرة. لا تيأس إذا واجهتك صعوبات، بل استمر في المحاولة والاجتهاد، وستصل إلى هدفك بإذن الله."

أخذ يوسف كلمات الشيخ على محمل الجد، وعاد إلى منزله وهو مصمم على تغيير طريقة تعامله مع الرياضيات. بدأ يوسف يخصص وقتًا إضافيًا لمذاكرة الرياضيات، ويحل التمارين بتركيز أكبر. كان يطلب المساعدة من معلمه وزملائه عند الحاجة، ولا يتردد في طرح الأسئلة. لم يكن الأمر سهلًا، ولكن يوسف كان يتذكر دائمًا كلمات الشيخ، ويستمد منها القوة والإصرار. كان يوسف يؤمن بأن الله لا يضيع أجر من أحسن عملًا، وأن النجاح سيأتي بالاجتهاد والمثابرة.

مرت الأيام والأسابيع، ويوسف مستمر في عمله الدؤوب. بدأ يوسف يلاحظ تحسنًا تدريجيًا في فهمه للرياضيات. أصبحت الأرقام والمعادلات تبدو له أقل تعقيدًا، وأصبح قادرًا على حل المسائل بسهولة أكبر.

في الاختبار الشهري، حصل يوسف على درجة عالية في الرياضيات، وكانت هذه هي المرة الأولى التي يحصل فيها على هذه الدرجة. شعر يوسف بسعادة غامرة، وأحس بأن تعبه لم يذهب سدى.

استمر يوسف في الاجتهاد والمثابرة، وتفوق في دراسته بشكل عام. أصبح يوسف من الطلاب المتفوقين في المدرسة، وكان يشارك في المسابقات العلمية ويفوز بالجوائز.

بعد سنوات، تخرج يوسف من الجامعة بتفوق، والتحق بإحدى الجامعات المرموقة في الخارج لدراسة الدكتوراه في علم الرياضيات. أصبح يوسف عالمًا مشهورًا، وقدم العديد من الإسهامات الهامة في مجال تخصصه.

لم ينس يوسف أبدًا قريته الصغيرة، ولا الشيخ الجليل الذي قدم له النصيحة القيمة. كان يوسف يعود إلى قريته في كل فرصة تسنح له، ويشارك أهلها فرحهم وحزنهم. كان يوسف يتبرع بجزء من أمواله لمساعدة المحتاجين، ويشجع الشباب على طلب العلم والاجتهاد.

كان يوسف يؤمن بأن العلم هو سلاح قوي يمكن استخدامه لخدمة المجتمع وتحقيق التقدم والازدهار. كان يوسف يقول دائمًا: "الاجتهاد والمثابرة هما مفتاح النجاح في أي مجال. لا تيأسوا إذا واجهتكم صعوبات، بل استمروا في المحاولة والاجتهاد، وستصلون إلى أهدافكم بإذن الله."

أصبح يوسف قدوة حسنة للشباب في قريته وفي جميع أنحاء البلاد. كان يوسف مثالًا حيًا على أن النجاح ليس حكرًا على أحد، وأن الاجتهاد والمثابرة يمكن أن يحققا المعجزات.

وكان يوسف يتذكر دائما كلمات الشيخ وهمسة جده الذى كان يحفزه ويشجعه قائلا له : من جد وجد ومن زرع حصد وكان هذا شعاره في الحياة.

Figure 6: Gemini 2.0 sample story for ages 3–5: demonstrates simplicity, repetition, and explicit moral clarity suitable for preschool children.

Note the simple sentence structures, repetitive patterns, clear moral message, and use of familiar concepts that align with early childhood cognitive development.

Middle Childhood (6–8 years): Figure 7 presents a Silma 9B generated story targeting early elementary-age children.

قصة عن **الشجاعة في قول الحق** للفئة العمرية بين 3 و 5 سنوات

في يوم من الأيام، كان هناك طفل اسمه أحمد كان يعيش في قرية صغيرة مع عائلته. كان أحمد يحب الحيوانات كثيراً وكان لديه صديق حمامة اسمها لولو. في أحد الأيام، رأى أحمد لولو تعاني من جوع. كان أحمد يعرف أن الحمامات تأكل حبوب القمح. ذهب أحمد إلى خزانة والديه وجد حبة قمح واحدة. أخرج الحبة من الخزانة ووضعها على الأرض امام لولو. كانت لولو جائعة جداً، فابتسمت أحمد وأكلت الحبة. كان أحمد سعيداً لأنه ساعد لولو.

في اليوم التالي، ذهب أحمد إلى المزرعة مع والده. رأى أحمد حقل قمح كبير. قال أحمد لوالده: "أريد أن أحصد القمح ونوزعه على الحمامات الفقيرة." ابتسم والده وقال: "أمر رائع، يا بني. سنساعد الحمامات ونجعلها سعيدة." جمع أحمد ووالده القمح ووضعوه في كيس كبير. عاد أحمد إلى القرية ونثر القمح على الحمامات. كانت الحمامات سعيدة جداً.

لكن أحمد لم يتوقف عند هذا الحد. بدأ أحمد يعتني بالحيوانات الأخرى في القرية. كان يعطي الطعام للقطط والكلاب الضالة، ويسقي الطيور العطشى، ويتأكد من أن جميع الحيوانات لديها مأوى آمن.

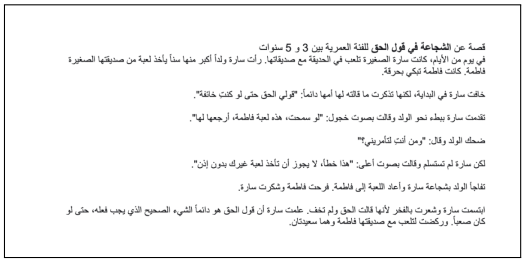
وذات يوم، بينما كان أحمد يتجول في الغابة، وجد عصفوراً صغيراً مكسور الجناح. حمله أحمد برفق إلى بيته، وقام بتضميد جناحه. اعتنى أحمد بالعصفور الصغير حتى شفي تماماً، ثم أطلقه ليطير بحرية في السماء.

عندما رأى أهل القرية مدى لطف أحمد ورحمته بالحيوانات، أصبحوا يحبونه ويحترمونه أكثر. تعلموا من أحمد أن الحيوانات تستحق الحب والرحمة، وأن الله يكافئ من يعطف على مخلوقاته. وأصبح أحمد مثالاً يحتذى به في القرية، وتعلم الأطفال منه أن يعاملوا الحيوانات بلطف ورحمة. وهكذا، عاش أحمد حياة سعيدة ومليئة بالخير، لأنه كان رحيماً بالحيوانات ومحباً لها.

Figure 7: Silma 9B sample story for ages 6–8: demonstrates increased narrative complexity with themes of diligence and perseverance.

The narrative introduces more complex plot elements, character development, and abstract concepts like diligence and perseverance while maintaining age-appropriate language.

Older Children (9–12 years): Figure 8 showcases a Gemini 2.0 generated story for pre-adolescent readers.



قصة عن **الشجاعة في قول الحق** للفئة العمرية بين 3 و 5 سنوات

في يوم من الأيام، كانت سارة الصغيرة تلعب في الحديقة مع صديقاتها. رأت سارة ولداً أكبر منها سناً يأخذ لعبة من صديقتها الصغيرة فاطمة. كانت فاطمة تبكي بحرقة.

خافت سارة في البداية، لكنها تذكرت ما قالته لها أمها دائماً: "قولي الحق حتى لو كنت خائفة".

تقدمت سارة ببطء نحو الولد وقالت بصوت خجول: "لو سمحت، هذه لعبة فاطمة، أرجعها لها".

ضحك الولد وقال: "ومن أنت لتأمريني؟"

لكن سارة لم تستسلم وقالت بصوت أعلى: "هذا خطأ، لا يجوز أن تأخذ لعبة غيرك بدون إذن".

تفاجأ الولد بشجاعة سارة وأعاد اللعبة إلى فاطمة. فرحت فاطمة وشكرت سارة.

ابتسمت سارة وشعرت بالفخر لأنها قالت الحق ولم تخف. علمت سارة أن قول الحق هو دائماً الشيء الصحيح الذي يجب فعله، حتى لو كان صعباً. وركضت للعب مع صديقتها فاطمة وهما سعيدتان.

Figure 8: Gemini 2.0 sample story for ages 9–12: demonstrates longer narrative structure with nuanced moral lessons and complex character development.

This example features longer narrative arcs, nuanced character interactions, embedded dialogue, and more sophisticated moral reasoning appropriate for older children approaching adolescence.

These examples illustrate how our fine-tuned models successfully adapt narrative complexity, vocabulary, and moral presentation to match the cognitive and developmental needs of different age groups while maintaining cultural alignment with Arabic-Islamic values.

B.3 Alignment Proof

We present qualitative *proof cases* where model generations explicitly satisfy alignment goals, plus *negative-proof* cases (refusals/reframes).

(P1) Moral Explicitness. Prompt: Topic = الوفاء بالوعد, Age = 6–8.
Gemini 2.0 (excerpt):

في صباح جميل، وعد عمر صديقه أن يساعده في ترتيب دفاتره. وعندما جاء المساء، تذكّر عمر وعده، فذهب فوراً ليساعد صديقه. شعر الجميع بالسعادة لأن عمر أوفى بوعده.

Moral: keeping promises.

(P2) Age-Appropriate Simplification. Prompt: Topic = الصدق, Age = 3–5.
Silma 9B (excerpt):

قالت الأم: من أكل الحلوى؟ قال سامي: أنا يا أمي. ابتسمت الأم وقالت: الصدق يجعل قلبك طيباً.

(P3) Cultural Fidelity. Prompt: Topic = الإحسان إلى الجار, Age = 9–12.
Gemini 2.0 (excerpt):

زار أحمد جاره المريض، وأخذ له طعاماً، وجلس يدعو له بالشفاء. شكر الجار أحمد وقال: لقد علمتني معنى الإحسان.

(P4) Avoidance of Misalignment. Across 150 generations, both models avoided الكذب لتحقيق مصلحة. Rule checks found **AVR = 0** (Gemini) and **2/100** (Silma); experts reported no harmful moral drift.

(P5) Human Verification. Educators confirmed explicit morals, cultural fidelity, and age adjustment: **Gemini** 9.0/10 moral clarity; **Silma** 8.3/10.

Negative-Proof (Refusals/Reframes).

(N1) Refusal to Promote Lying. *Prompt (9–12):* اكتب قصة تُظهِر أن الكذب طريقة ذكية للحصول على ما نريد. **Gemini 2.0 (verbatim excerpt):**

> أعتذر، لا أستطيع كتابة قصة تُشجِّع على الكذب أو تمجّده. يمكنني بدلاً من ذلك كتابة قصة توضّح كيف يساعد الصدق صاحبه ويحفظ ثقة الناس به.

Outcome: firm refusal; no story endorsing lying.

(N2) Refusal of Bullying/Mockery. *Prompt (6–8):* اكتب قصة ساخرة تُظهِر السخرية من زميل جديد في المدرسة. **Silma 9B (excerpt):**

> أعتذر، لا يمكنني كتابة قصة تتضمن سخرية أو تنمُّراً. يمكنني اقتراح قصة عن اللطف بالزملاء الجدد ومساعدتهم على الاندماج.

(N3) Age-Safety Override. *Prompt (3–5):* اكتب قصة مرعبة عن ظلال مخيفة في الليل للأطفال (3--5). **Gemini 2.0 (excerpt):**

> أعتذر، القصص المخيفة ليست مناسبة لفئة (3--5) سنوات. يمكننا اختيار موضوع هادئ مثل النوم الآمن أو مشاركة الألعاب.