Noise Steering for Controlled Text Generation: Improving Diversity and Reading-Level Fidelity in Arabic Educational Story Generation

Haziq Mohammad Khalid
b00100601@aus.edu

Salsabeel Shapsough
sshapsough@aus.edu

Imran Zualkernan
izualkernan@aus.edu

American University of Sharjah

Abstract

Generating diverse, pedagogically valid stories for Arabic early-grade reading assessments requires balancing tight constraints on vocabulary, reading level, and narrative structure against the need to avoid repetitive plots that undermine assessment validity. We investigate *noise steering*, injecting calibrated Gaussian perturbations into the internal representations of transformer models at inference time, as a training-free diversity method evaluated across five small Arabic-centric language models (7–9B parameters). We compare four injection strategies against high-temperature sampling baselines, measuring diversity, quality, constraint adherence, and reading grade level. Residual stream noise consistently improves narrative diversity with minimal quality or constraint cost and preserves early-grade reading level across all Arabic-centric models. Attention entropy noise injection (AENI) stabilizes the otherwise unreliable attention-logit noise while recovering quality. High-temperature sampling inflates reading grade level and causes catastrophic collapse on several models. We find internal representation-level perturbation to be a more suitable diversity strategy than output-level stochasticity for constrained educational content generation.

1 Introduction

The Early Grade Reading Assessment (EGRA) is a widely used framework for evaluating foundational reading skills in young learners (Dubeck and Gove, 2015). A key component involves short narrative passages that must satisfy strict pedagogical constraints: controlled vocabulary, appropriate reading level, clear narrative structure, and cultural neutrality. Manually producing such passages at scale is labor-intensive, making automatic generation an attractive alternative, particularly in Arabic, where rich morphology, orthographic variation, and diglossia introduce additional linguistic complexity (Habash, 2010). Recent work has shown that large language models can generate fluent Arabic stories aligned with EGRA criteria, but two tensions persist (Abdelghafur et al., 2025; Rai et al., 2024). First, *diversity vs. constraint adherence*: increasing output variety risks violating the tight pedagogical requirements that make EGRA passages valid assessment instruments (Holtzman et al., 2020). Second, practical deployment in under-resourced educational settings favors small language models (SLMs) of 7–9 billion parameters over large frontier models. Yet, these smaller Arabic-centric models are more susceptible to generation instability when diversity is pushed. We address both tensions by investigating *noise steering*: injecting calibrated Gaussian perturbations into the internal representations of transformer models at inference time, without any fine-tuning. We evaluate four injection strategies across five Arabic SLMs and show that residual stream noise and attention entropy noise injection (AENI) consistently improve narrative diversity while preserving reading grade level and constraint adherence, outperforming output-level approaches such as high-temperature sampling that raise diversity at the cost of pedagogical validity.

2 Background

2.1 EGRA Constraints

EGRA-aligned texts are constructed under strict pedagogical and linguistic constraints (Dubeck and Gove, 2015). These include controlled passage length, age-appropriate vocabulary, a clear narrative structure, limited characters, grammatical correctness, and cultural neutrality. These constraints ensure comparability across assessments and support the generation of both literal and inferential comprehension questions. However, manually creating such passages is labor-intensive and requires domain expertise, limiting scalability (Abdelghafur et al., 2025).

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 581–597

2.2 Challenges in Arabic EGRA Generation

Generating EGRA-aligned content in Arabic introduces additional challenges due to the language's rich morphology, orthographic ambiguity, and diglossia between Modern Standard Arabic and dialectal varieties, compounded by the limited availability of high-quality annotated datasets for children's literature (Habash, 2010; Abdelghafur et al., 2025). While recent work has shown that models such as GPT-4 and Jais can generate fluent, EGRA-aligned stories, inconsistencies in narrative structure, deviations from target word counts, and repetitive outputs with minimal structural variation remain persistent issues (Abdelghafur et al., 2025; Rai et al., 2024; El-Shangiti et al., 2024).

2.3 Diversity vs. Constraint Trade-off

A central challenge in EGRA-based generation is balancing diversity with constraint satisfaction. Increasing diversity through stochastic decoding can lead to output degeneration and reduced control over pedagogical requirements (Holtzman et al., 2020), while overly constrained prompting often results in repetitive and structurally similar outputs (Rai et al., 2024).

Recent approaches attempt to address this trade-off. Shankarnarayanan et al. (2024) show that inspiration-based prompting can improve narrative diversity, while Madaan et al. (2023) demonstrate that iterative self-refinement can enhance output quality without model fine-tuning. Despite these advances, achieving consistent diversity while maintaining strict adherence to EGRA constraints remains unresolved, particularly in Arabic.

3 Related Works

3.1 Output-Level Diversity: Stochastic Decoding Strategies

A common approach to diversity in LLM generation is stochastic decoding, which modifies the token probability distribution at each step. Temperature scaling (Ackley et al., 1985) flattens or sharpens this distribution, while top-k sampling (Fan et al., 2018) restricts candidates to the k most probable tokens, and nucleus (top-p) sampling (Holtzman et al., 2020) selects the smallest token set whose cumulative probability exceeds a threshold. These methods trade output quality for diversity as their parameters increase, yet they share a fundamental limitation: all intervention occurs at the final vocabulary logit, leaving the model's internal computations, embeddings, attention patterns, and intermediate representations entirely unchanged. Locally Typical Sampling (Meister et al., 2023) approaches diversity differently by selecting tokens whose information content stays close to the conditional entropy of the model, reducing degenerate repetition while preserving fluency. A complementary line of work intervenes on the output vocabulary itself: Dineen et al. (2026) randomly mask logits during self-play training to stop the model from collapsing onto a narrow set of token sequences. Their mask is applied during RL training, while ours is applied at inference, making the two approaches complementary. We use temperature, top-p, top-k primary decoding baselines. Our internal noise methods are independent to all of them and can be combined with any output-level sampler.

3.2 Internal Noise Injection in Transformers

A smaller but closely related body of work explores injecting perturbations into the internal representations of transformer models rather than into the output distribution. Kang et al. (2024) is the most directly comparable prior work, they apply input-dependent Gaussian noise to the hidden states of early LLM layers during *fine-tuning* to produce diverse paraphrases of training examples for knowledge injection. Kang et al. (2024) demonstrate that latent-level perturbation yields more semantically diverse augmentations than data-level paraphrasing alone. This finding motivates our hypothesis that similar perturbations at *inference time* can promote creative diversity in generation. Unlike Kang et al. (2024)'s method of applying noise during training to improve knowledge retention, our methods apply noise at inference time to diversify outputs without modifying model weights.

The residual stream framework of Elhage et al. (2021) motivates our choice of injection sites. Embeddings, attention logits, and block outputs each represent distinct points at which information is written to the model's shared communication channel, producing qualitatively different perturbation effects.

3.3 Entropy-Aware Generation

Several works use the model's own uncertainty as a signal for adaptive decoding. EDT (Zhang et al., 2024) raises temperature when the output vocabulary entropy is low, encouraging exploration precisely when the model is overconfident. Our atten-

tion entropy-based method shares this motivation but monitors entropy over attention distributions rather than output logits, intervening at the attention level, before the model's repetitive tendencies manifest in the output distribution.

3.4 Constrained Generation and Arabic Educational LLMs

Constrained decoding methods (Willard and Louf, 2023) enforce formal output constraints by masking invalid tokens at each step, but do so at the cost of forcing the model toward low-probability and often repetitive continuations. This phenomenon is seen clearly in Early-Grade Reading Assessment (EGRA) story generation (RTI International, 2016; Zualkernan and Shapsough, 2024), where tight vocabulary and structural constraints narrow the valid token set and exacerbate the repetitive nature of outputs. Noise injection has been shown to systemically degrade a model's ability to follow constraints and guidelines (Shahani et al., 2025), motivating careful calibration. Our noise decay strategy addresses this directly by reducing perturbation magnitude as generation progresses, allowing the model to explore creative directions early in the story while gradually restoring constraint following behaviour toward the conclusion. We evaluate story generation in this setting using a suite of small Arabic-centric and multilingual instruct models (Sengupta et al., 2023; Bari et al., 2024; Huang et al., 2024; Fanar Team et al., 2025), measuring diversity via Self-BLEU (Zhu et al., 2018) and semantic dispersion.

4 Noise Steering methods

4.1 Embedding Noise

Embedding noise injects Gaussian noise directly into the output of the token embedding lookup table during the decoding phase. This is the earliest possible injection site: the perturbed vector is the model's sole representation of the newly generated token before any transformer computation has occurred. All subsequent hidden-layer inputs are equivalent to the previous block's residual output and therefore have no independent embedding-level analogue.

Formally, let $\mathbf{e}_t \in \mathbb{R}^D$ denote the embedding of the token sampled at decode step t. We perturb it as:

$$\tilde{\mathbf{e}}_t = \mathbf{e}_t + \varepsilon, \quad \varepsilon \sim \mathcal{N}\big(0,\, \sigma(t)^2\mathbf{I}\big) \qquad (1)$$

where $\sigma(t) = \sigma_{\text{emb}} \cdot \delta(t)$ is the decayed standard deviation from Section 4.5. Noise is applied exclusively to decode steps, the prefill pass over the prompt is left unperturbed, ensuring that the model's understanding of the instruction and constraints is not corrupted. Because the perturbation enters at the base of the residual stream, it is amplified and transformed by every subsequent attention and MLP layer, producing broad, high-level variation in the story's trajectory.

4.2 Attention-logit Noise

Attention-logit noise injects Gaussian noise into the output of the self-attention sub-layer at selected transformer blocks, specifically after the output projection W_O but *before* the subsequent MLP sub-layer and before the residual addition. Within each targeted block the computation is:

$$\mathbf{x} = \text{LayerNorm}(\mathbf{h}^{(l-1)}) \qquad (2)$$

$$\mathbf{a} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) VW_O + \varepsilon \qquad (3)$$

$$\mathbf{h}^{(l)} = \mathbf{h}^{(l-1)} + \mathbf{a} + \text{MLP}\Big(\mathbf{a} + \mathbf{h}^{(l-1)}\Big) \qquad (4)$$

$$\varepsilon \sim \mathcal{N}\big(0, \sigma(t)^2\mathbf{I}\big)$$

where $\mathbf{h}^{(l)}$ is the residual stream at layer l and $\sigma(t) = \sigma_{\text{attn}} \cdot \delta(t)$ follows the cosine decay schedule of Section 4.5. The KV cache is left intact: the last token position of the attention output tensor (shape $(B, 1, D)$ during decoding) is perturbed only, leaving all cached key and value states unchanged.

By targeting the attention branch's contribution to the residual stream in isolation, before the MLP has processed the attended context, this method perturbs the information routing step of each selected block without corrupting the feed-forward transformation. This gives it a qualitatively different character from residual stream noise, which acts on the full block output after both sub-layers have run.

4.3 Residual Stream Noise

Residual stream noise injects Gaussian noise into the transformer's residual stream after each selected block has completed its full computation of both the self-attention and MLP sub-layers. The noise at layer l and decode step t is:

$$\tilde{h}_t^{(l)} = h_t^{(l)} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,\, \sigma(t)^2\,)$$

where $h_t^{(l)}$ is the block output at the last token position and $\sigma(t) = \sigma_{\text{res}} \cdot \delta(t)$ is the decayed standard deviation from Section 4.5. Noise is applied only during the decoding phase when new tokens are being generated, leaving the prompt representation unperturbed. Injecting at the full block output means the noise enters the residual stream at the point where the block's complete contribution has been written, and propagates forward through all subsequent layers.

4.4 Attention Entropy Noise Injection

Attention Entropy Noise Injection (AENI) conditions noise magnitude on the peakedness of the model's own attention distribution at each decode step. The motivation, shared with EDT (Zhang et al., 2024), is that a model is most likely to produce repetitive or formulaic output precisely when it is most confident: attention heads that concentrate their weight on a small set of tokens signal over-reliance on a narrow context window, which in constrained story generation often corresponds to copying earlier story elements. AENI responds by injecting more noise when attention is peaked and less when it is diffuse, intervening adaptively rather than applying a fixed perturbation schedule.

At each targeted layer l and decode step t, we read the post-softmax attention weight tensor $\mathbf{W}^{(l)} \in \mathbb{R}^{B\times H\times T_k}$, where H is the number of heads and T_k is the current key-sequence length. We extract the last query row (the current token's attention pattern) $\mathbf{w} = \mathbf{W}^{(l)}_{:,:,-1,:}$ and compute a peakedness score as the mean per-head maximum attention weight:

$$\phi(\mathbf{w}) = \text{mean}_h \max_j w_{h,j} \tag{5}$$

This score lies naturally in $[0, 1]$, is independent of context length, and requires no normalisation: a peaked head drives its maximum toward 1, while a uniform distribution drives it toward $1/T_k$. Gaussian noise is then added to the self-attention output after W_O but before the MLP and residual addition, with effective standard deviation:

$$\sigma_{\text{eff}}(t) = \sigma_{\text{aeni}} \cdot \phi(\mathbf{w}) \tag{6}$$

where σ_{aeni} is the per-model ceiling calibrated by RMS as in Section 5.1. As with attention-logit noise, only the last token position of the attention output (shape $(B, 1, D)$ during decoding) is perturbed and the KV cache is left intact.

4.5 Noise Decay

Applying constant noise throughout generation risks degrading constraint adherence, consistent with findings that activation-level noise systemically harms instruction and constraint following in LLMs (Shahani et al., 2025). We therefore apply a cosine decay schedule that reduces noise magnitude as generation progresses, allowing the model to explore creatively in the early tokens while recovering clean, constraint respecting decoding for the middle and conclusion of its story.

Let t denote the current decode step and T the decay horizon. The noise standard deviation at step t is scaled by:

$$\delta(t) = \frac{1}{2}\left(1 + \cos\left(\frac{\pi \cdot \min(t,\, T)}{T}\right)\right)$$

so that $\delta(0) = 1$ (full noise at the first generated token) and $\delta(T) = 0$ (no noise at and beyond the horizon), with the effective standard deviation given by $\sigma(t) = \sigma_{(\text{t}-1)} \cdot \delta(t)$.

5 Experimental Setup

5.1 Noise Magnitude and Model Calibration

A key challenge when applying noise across multiple models is that the scale of internal representations varies substantially between architectures and models (Dettmers et al., 2022; Sun et al., 2024). A fixed noise standard deviation that is undetectable for one model may be catastrophically large for another. To ensure that noise is applied at a consistent *relative* scale across all models, we calibrate the noise magnitude using the root mean square (RMS) of each model's residual stream activations.

For each model we register forward hooks on nearly all transformer blocks (excluding the first and last layers) and collect block output tensors during *decode* steps only (i.e., when the sequence length is 1 and the KV cache is active), matching the phase at which noise is actually injected. The per-layer RMS is computed over the last generated token as:

$$\text{RMS}(l) = \sqrt{\frac{1}{C}\sum_C x^2} \tag{7}$$

where x is the block output tensor at layer l averaged over the batch and hidden dimensions, and C is the number of decode steps collected. We then aggregate these per-layer values into a single scalar

using the median across layers, which is robust to outlier layers at either extreme of the network.

The residual noise standard deviation for each model is then set as:

$$\sigma_{\text{res}} = \alpha \cdot \text{median}_l\big(\text{RMS}(l)\big) \tag{8}$$

where α is a fixed scaling coefficient set to 0.175 across all experiments. This formulation ensures that the injected noise is always proportional to the model's own activation scale, making the effective perturbation strength comparable across models.

5.2 Generation

We compare each noise-injection method to the following baselines: a *noise-free* baseline ($T = 1.0$), a *High-temperature* ($T = 1.8$) baseline with $topk = 40$, and a *High-temperature* ($T = 1.8$) baseline with $topp = 0.9$. The truncation values follow prior work: Holtzman et al. (2020) report $k = 40$ in their main comparison and note that nucleus values "are usually in $[0.9, 1)$". We pair these with an elevated $T = 1.8$, within the high-temperature regime explored by Nguyen et al. (2025). More aggressive settings (higher T, looser or no truncation) would likely produce still higher collapse rates, and absolute numbers would shift under per-model tuning. However, our principal claims concern the qualitative trade-off between output-level stochasticity and internal-representation perturbation, which we expect to be robust to such tuning.

Each condition generates 50 stories independently. Story i under every condition is generated with the same seed, which makes generation reproducible and aligns the underlying sampler RNG state across conditions where the sampling distribution coincides. However, since different conditions modify the sampling distribution itself (e.g. via temperature, top-k, top-p, or internal noise), the seed alone does not eliminate sampling variability between conditions. We therefore report all aggregate metrics over the full 50-story sample per condition and use the statistical tests in Appendix A to assess whether the observed effects are reliable rather than artefacts of sampling variance. All models are prompted with the same instruction to produce a story adhering to the EGRA constraints described in Section 5.3.3.

5.3 Evaluation

5.3.1 Creativity Scores

Vendi Score: The *Vendi Score* (Friedman and Dieng, 2023) measures the diversity of a set of items by quantifying how evenly distributed they are across predefined categories. For our stories, embeddings were obtained using the BAAI/bge-m3 model, with higher values indicating greater diversity.

Lexical Diversity: We compute Self-BLEU (Zhu et al., 2018), treating each story in turn as a candidate and the remaining 49 as references. The lexical diversity score is $1 - \overline{\text{Self-BLEU}}$, so that higher values indicate less overlap across the generated set.

5.3.2 Readability and Linguistic Quality

We judge readability, grammatical correctness, linguistic quality, and appropriateness of reading level are evaluated using GPT5.3 Chat (OpenAI, 2026) as an LLM judge. Each story is scored independently on these dimensions, providing a complementary quality signal to the diversity metrics above.

5.3.3 Constraint Following

EGRA constraint adherence is evaluated through a combination of rule-based checks and LLM judgment. Constraints that are able to be verified exactly are checked programmatically. These constraints include: Story length not exceeding 60 words, Use of exactly one proper noun, Use of the present tense

Constraints that require semantic and structural understanding are evaluated using GPT5.3 CHAT (OpenAI, 2026) as a judge. These constraints include:

- Narrative structure includes intro, a middle dilemma, and an ending with resolution.
- Gender-balanced: includes both a boy and a girl.
- Avoids gender/religion/other stereotypes.
- Vocabulary suitable for children and the local context.

5.4 Models and Metrics

Models: We evaluate five small ($\leq$9B parameter) Arabic-centric instruct models. **Jais 8B** (Sengupta et al., 2023) is pretrained from scratch on a large Arabic and English corpus. **ALLaM 7B** (Bari et al., 2024) adapts LLaMA via Arabic tokeniser expansion and continued pretraining. **Fa-**

nar 9B (Fanar Team et al., 2025) emphasises morphology-aware Arabic tokenisation and dialectal coverage. **AceGPT** (Huang et al., 2024) is a bilingual (Arabic–English) model built on the LLaMA architecture with continued pretraining on large-scale Arabic data and instruction tuning for dialogue and generation tasks. **Phi-4-mini** is included as a compact multilingual baseline. Together these models span the main architectural and training strategies present in the current Arabic SLM landscape.

Quality (Qual.↑). The quality score is an average of four LLM-judge dimensions assessed per story: overall readability, logical soundness, grammatical correctness, linguistic quality. Each dimension is scored independently out of 10 and the four scores are averaged into a single value per story, which is then averaged across the 50 stories in a condition.

Constraint Violations (Viol.↓). We report the average number of constraints broken per story and the total constraints broken, where a lower score is better and zero indicates full compliance. This combines the programmatic and LLM-judge checks from Section 5.3.3: each of the seven constraints (three programmatic, four LLM-judged) contributes one violation point if failed, giving a maximum possible score of 7 per story.

6 Results

6.1 Total Modal Collapse

Before comparing diversity and quality metrics, we first examine instances of Total modal collapse (TMC): cases where generation degenerates into repetitive or incoherent token sequences, producing stories that are effectively unusable. Table 1 reports the collapse rate for each model and decoding condition. Any condition with a collapse rate at or above 40% is excluded from further evaluation, as the surviving outputs would constitute a non-representative sample. Conditions with collapse rates greater than 20% (and less than 40%) are excluded from Figure 1 but are discussed further in Sec 6.2.

High-temperature baselines. The high temperature conditions (HiTemp-k and HiTemp-p) are the most fragile overall. AceGPT and Phi-4-mini both exceed the 40% threshold under both high-temperature settings, with collapse rates as high as 84% and 66% respectively. This confirms the well-known instability of aggressive output-level sampling: flattening the output distribution beyond a certain point causes the model to assign non-negligible probability to incoherent continuations, from which recovery is rare once the context has been corrupted (Nguyen et al., 2025).

Residual stream noise. residual stream noise (L-Res), records a 0% collapse rate across every model tested. Because this method injects noise after each transformer block has already completed its full computation, the perturbation enters the network at a point where the model's internal representations are well-formed, and the signal propagates forward in a controlled manner. This makes residual noise a reliable location to inject noise without worrying about generation failures.

Attention noise and AENI. Raw attention-logit noise (Attn) shows high collapse rates, reaching 28% and 34% on ALLaM and Fanar respectively. However, our Attention Entropy Noise Injection method (AENI) eliminates collapse entirely across all five models. Rather than applying a fixed perturbation at every step, AENI scales the noise magnitude according to how peaked the model's own attention distribution is at that moment. When the model attends broadly and diffusely, little or no noise is injected; perturbation is reserved for steps where attention has collapsed onto a narrow set of tokens, which is precisely when formulaic or repetitive output is most likely to follow. This adaptive

Table 1: Modal collapse rate (% of 50 generated stories) per model and decoding strategy. **Italic**: method excluded from further evaluation due to ($\geq 40\%$ collapse rate);

Model	Baseline	HiTemp-k	HiTemp-p	Embed	Attn	AENI	L-Res
ALLaM 7B	0%	0%	4%	22%	28%	0%	0%
AceGPT 8B	2%	*84%*	*80%*	0%	0%	0%	0%
Fanar 9B	0%	2%	0%	22%	34%	2%	0%
Jais 8B	0%	0%	0%	2%	4%	0%	0%
Phi-4-mini	0%	*50%*	*66%*	*100%*	*44%*	0%	0%

gating prevents the runaway instability that fixed attention noise produces.

Embedding noise. Embedding noise (Embed) causes complete collapse on Phi-4-mini (100%), and elevated rates on ALLaM (22%) and Fanar (22%). The Phi-4-mini model in particular is very sensitive to embedding noise, even at noise levels $\alpha < 0.175$ virtually every generation collapses. As a compact multilingual model, Phi-4-mini has had comparatively less Arabic training data, meaning its token embedding space for Arabic is more sparsely and less robustly organized. Because embedding noise acts at the very first step of the residual stream, before any transformer computation has occurred, any perturbation there propagates through every subsequent layer.

For a model whose Arabic representations are already fragile, this earliest injection site offers no opportunity for the model to compensate.

6.2 Diversity, Quality, and Constraint Adherence

Figure 1 plots each model-method condition in the space of Vendi Score (diversity, x-axis) against LLM-judged quality (y-axis), with marker size encoding the average number of constraint violations per story. Beyond these core metrics, we also track estimated reading grade level per condition, as diversity methods that push stories toward richer vocabulary or more complex sentence structures may improve narrative variety and quality scores while silently violating the early-grade readability that EGRA passages require. Full numerical results are in Appendix Tables 6 and 4.

Residual Stream Noise. L-Res is the most consistently beneficial intervention. It achieves a 0% collapse rate for every model and improves Vendi Score in all five cases: +1.28 for ALLaM (7.14Ä8.42), +0.28 for AceGPT (9.95Ä0.23), +0.14 for Fanar (8.47Ä8.61), +2.38 for Jais (4.71Ä7.09), and +1.19 for Phi-4 (10.45Ä1.64). Constraint adherence is largely undisturbed, with violations changing by at most 0.78 per story across all models. Reading grade level is preserved throughout: ALLaM and Fanar remain at Grade 3, Jais at Grade 2, and AceGPT at Grade 4, all consistent with their respective baselines. These results indicate that

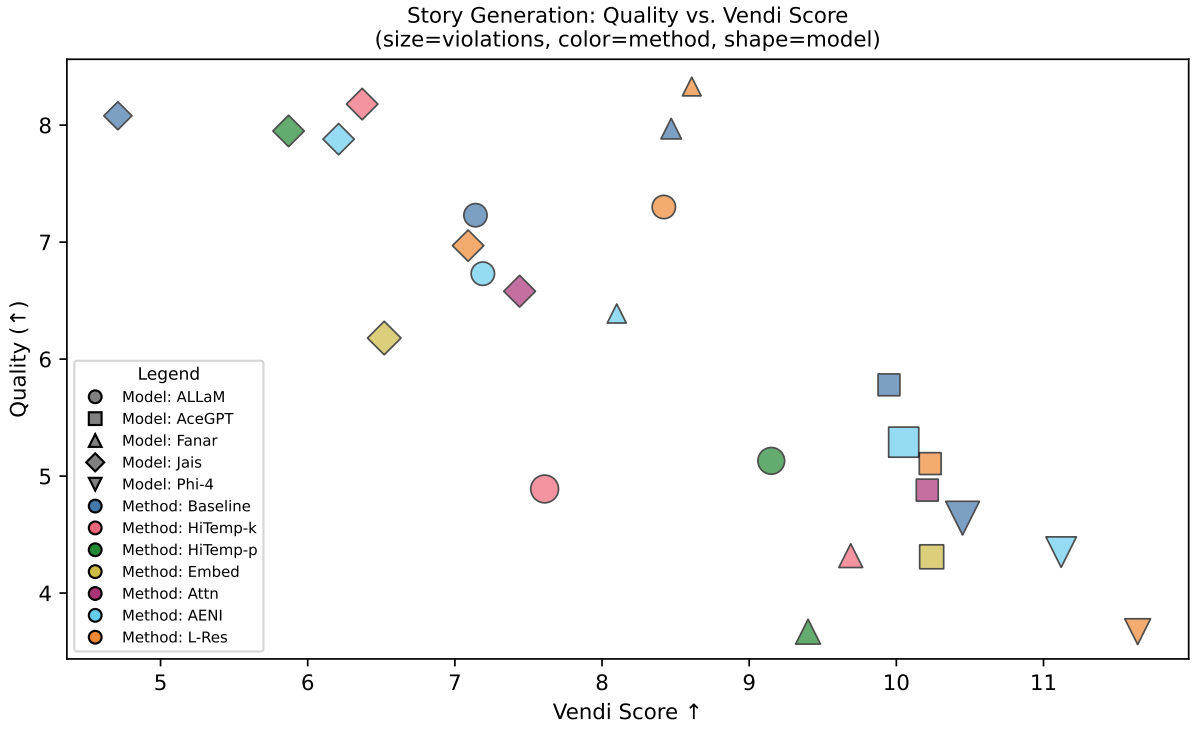

Figure 1: Story quality versus output diversity (Vendi Score) for all evaluated conditions with ≤ 10 total modal collapses. Each point represents one model–method combination: marker **shape** encodes the model and marker **colour** encodes the decoding method; marker **area** is proportional to the total number of EGRA constraint violations across the 50 generated stories (larger = more violations). Methods that cluster toward the lower-left or produce large markers trade quality or diversity for higher violation rates. Conditions with > 10 modal collapses are excluded to avoid diversity scores inflated by repetitive collapse outputs

perturbing the model's internal representations allows it to diversify narrative choices, such as character, setting, and plot structure, without reaching for more complex language to do so. The quality impact depends on the baseline. For Fanar and ALLaM, which already produce strong stories (quality scores of 7.97 and 7.23 respectively), L-Res maintains or improves quality outright: Fanar reaches 8.33 while gaining diversity, placing it in a genuinely better region of the quality-diversity space. For weaker models the diversity gains come at a greater quality cost: AceGPT drops from 5.78 to 5.11, Jais from 8.08 to 6.97, and Phi-4 from 4.64 to 3.67. In Jais's case the large Vendi Score gain suggests the model was simply not exploring the narrative space at baseline. In Phi-4's case, even gentle perturbation strains its more limited Arabic generation capability, and grade level drifts to Grade 5, the one exception to L-Res's otherwise consistent grade-level preservation.

Attention Entropy Noise Injection vs. Raw Attention Noise. Raw attention-logit noise (Attn) can achieve high diversity, with Vendi Scores of 9.51 for ALLaM and 7.44 for Jais, but at severe cost: collapse rates of 28% and 34% on ALLaM and Fanar respectively, constraint violations spiking to 5.36, and ALLaM quality collapsing to 2.55, the lowest value in the entire evaluation result in the surviving scores being a biased subset of an otherwise unreliable condition. AENI resolves this by conditioning noise magnitude on the model's own attention entropy, injecting more perturbation when attention is sharply peaked and backing off when it is already diffuse. Collapse rates fall to 0% on ALLaM, AceGPT, and Jais, and just 2% on Fanar, with quality recovering substantially: ALLaM to 6.73, Fanar to 6.39, and Jais to 7.88. Crucially, AENI also preserves reading grade level across all Arabic-centric models and for Phi-4 actually improves it to Grade 3 from the baseline Grade 4, a stronger result than L-Res on that model. The cost of this stability is a moderate Vendi Score reduction relative to raw Attn, but this trade-off is clearly worthwhile for an educational generation system. On AceGPT, already a high-diversity model, the two methods produce near-identical Vendi Scores (10.21 vs. 10.05), confirming that AENI's adaptive mechanism does not sacrifice diversity unnecessarily when it is not needed.

High-temperature Baselines High-temperature sampling is effective only for Jais, whose baseline Vendi Score of 4.71 is the lowest in the evaluation. Both HiTemp-k and HiTemp-p raise diversity meaningfully (to 6.37 and 5.87) while quality stays high at 8.18 and 7.95, suggesting Jais's cautious baseline was due to overfitting to the EGRA task format, producing structurally valid but narratively repetitive stories, and that output-level stochasticity was sufficient to break this pattern without destabilising the underlying generation quality. For every other model the picture is far worse: ALLaM's quality falls to 4.89 and 5.13, Fanar's to 4.32 and 3.67, and both AceGPT and Phi-4 exceed the collapse threshold entirely. The grade-level results are equally damaging: ALLaM's Grade 3 baseline rises to Grade 6 and Grade 5 under the two conditions, and Fanar's Grade 3 baseline drifts to Grade 5 and Grade 6 respectively. By flattening the output probability distribution, high-temperature sampling makes models more likely to reach for less common, more sophisticated vocabulary, precisely the kind of variation that is undesirable in an early-grade assessment context. It treats all forms of output variation as equivalent and has no mechanism for distinguishing a more creative plot from a more adult vocabulary, a distinction that is central to the EGRA task.

Appendix A contains significance tests that reaffirm these findings: L-RES and AENI are the only methods that consistently preserve quality and constraint adherence at a level statistically indistinguishable from the Baseline, while high-temperature sampling, embedding noise and raw attention noise injections are significantly worse on both metrics across the majority of models.

7 Conclusion

We investigated noise steering as a lightweight, fine-tuning-free approach to improving narrative diversity in Arabic EGRA-style story generation across five small language models. L-RES and AENI are the only methods that consistently improve diversity without the substantial degradation in quality, constraint adherence, or early-grade readability that the alternatives produce, the last of which proves a critical differentiator from high-temperature sampling, which inflates reading level and causes catastrophic collapse on several models. Residual stream noise is the most reliable all-around method and yields the greatest gains on models that already generate high-quality stories at baseline. AENI trades a small diversity mar-

gin for substantially better stability, making it the stronger choice when attention-level intervention is preferred. A consistent finding across all methods is that a model's baseline generation quality is a strong predictor of how gracefully it absorbs noise: models with weaker Arabic representation degrade faster regardless of the intervention applied, suggesting that noise steering is best understood as a tool for unlocking latent diversity rather than compensating for limited language capability. Future work should explore layer-specific targeting strategies and extension to other low-resource languages and assessment frameworks.

Limitations

LLM-as-a-Judge Evaluation. Our quality and parts of our constraint adherence scores rely on GPT as an LLM judge, which introduces well-documented limitations. LLM judges are known to exhibit systematic biases including verbosity bias, where longer outputs receive inflated scores, and self-enhancement bias, where a model tends to favor outputs stylistically similar to its own generations (Chen et al., 2024). A particular concern in comparative evaluations is relative scoring bias, where a story may receive an inflated score simply by virtue of being better than the others it is presented alongside rather than on absolute merit. We took two steps to mitigate this: stories were submitted to the judge in fixed-size batches rather than all at once, and stories from different experimental conditions were shuffled together before evaluation, preventing the judge from implicitly comparing outputs within a single run. Quality dimensions such as narrative clarity and vocabulary suitability ideally require human annotators with relevant expertise. The programmatic constraint checks mitigate this concern for verifiable constraints such as word count and tense, but the LLM-judged dimensions: narrative structure, cultural neutrality, and reading appropriateness, should be interpreted with caution. Future work should incorporate human evaluation by Arabic literacy educators to validate the automated scores reported here.

Decoding-Hyperparameter Sensitivity. Our high-temperature baselines use single fixed values of $T = 1.8$, $k = 40$, and $p = 0.9$, chosen as widely used high-diversity defaults in prior work but not separately tuned per model. Since different models exhibit different baseline confidence profiles, the absolute diversity, quality, and collapse numbers for these baselines could shift under per-model hyperparameter sweeps. A similar caveat applies to the single value of $\alpha = 0.175$ used to calibrate noise magnitude across all noise-injection methods. We expect the qualitative ordering between output-level stochastic decoding and internal-representation perturbation to be robust, but a more comprehensive sensitivity study across decoding hyperparameters and per-model α values is left to future work.

Acknowledgments

This work was funded by the American University of Sharjah through FRG24-E-E87. The views and claims expressed in this work are those of the authors and do not necessarily reflect the official position of the University.

References

Y. Abdelghafur, Y. Kaddoura, S. Shapsough, I. Zualkernan, and E. Kochmar. 2025. Tales from the algorithm: Enhancing reading comprehension assessments with ai-generated arabic stories. In *Proceedings of the 10th International Congress on Information and Communication Technology (ICICT 2025)*, volume 1443 of *Lecture Notes in Networks and Systems*, pages 277–288.

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, and Abdulmohsen Al-Thubaity. 2024. ALLAM: Large language models for Arabic and English. *arXiv preprint arXiv:2407.15390*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Process-*

ing, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35.

Jacob Dineen, Aswin RRV, Zhikun Xu, and Ben Zhou. 2026. Vocabulary dropout for curriculum diversity in llm co-evolution. *Preprint*, arXiv:2604.03472.

Margaret M. Dubeck and Amber Gove. 2015. The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development*, 40:315–322.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN — a novel Arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255, Portorož, Slovenia. European Language Resources Association (ELRA).

Ahmed O. El-Shangiti and 1 others. 2024. Arabic automatic story generation with large language models. In *Proceedings of the First Workshop on Arabic Natural Language Processing*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, and Masoomali Fatehkia. 2025. Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Preprint*, arXiv:2210.02410.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Minki Kang, Sung Ju Hwang, Gibbeum Lee, and Jaewoong Cho. 2024. Latent paraphrasing: Perturbation on layers improves knowledge injection in language models. In *Advances in Neural Information Processing Systems*, volume 37.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *Preprint*, arXiv:2407.01082.

OpenAI. 2026. GPT-5.3 Instant ChatGPT model. Accessed: 2026-03-06.

Samarth Rai, Salsabeel Shapsough, and Imran Zualkernan. 2024. Measuring fluency, coherency and logicality of GPT-4 generated EGRA comprehension stories. In *Proceedings of the 2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 201–203.

RTI International. 2016. Early grade reading assessment toolkit. Technical report, RTI International / USAID.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, and 3 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Prithviraj Singh Shahani and 1 others. 2025. Noise injection systemically degrades large language model safety guardrails. *arXiv preprint arXiv:2505.13500*.

Aadhith Shankarnarayanan, Taufiq Syed, Salsabeel Y. Shapsough, and Imran A. Zualkernan. 2024. Once upon a GPT-4: Enhancing diversity in automated reading comprehension story generation with classic tales. In *Proceedings of the 2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 196–200.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Shimao Zhang, Yu Bao, and Shujian Huang. 2024. EDT: Improving large language models' generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1097–1100.

Imran Zualkernan and Salsabeel Shapsough. 2024. Towards using large language models to automatically generate reading comprehension assessments for early grade reading assessment. pages 3772–3782.

A Statistical Significance: L-Res and AENI vs. Alternatives

Test choice. We compare each method's per-story quality and constraint-violation distributions to the Baseline using the Kruskal-Wallis omnibus test followed by Dunn's pairwise post-hoc with Holm correction. This single non-parametric procedure is the appropriate choice for our data: Levene's test rejects equality of variances ($p < 0.05$) in 7 of 10 (model, metric) groups, ruling out ANOVA + Tukey HSD; the metrics are also bounded and non-normal (quality $\in [0, 10]$ with TMC zero-mass; violations integer counts $\in [0, 7]$). Effect size is the matching non-parametric statistic, the rank-biserial correlation $r \in [-1, +1]$, with the convention that $r > 0$ indicates the method is *better than the Baseline* on that metric (quality higher, violations lower); $|r| < 0.1$ is negligible, ≥ 0.3 medium, ≥ 0.5 large, $|r| = 1$ is the maximum the metric supports.

Every alternative method significantly degrades performance whenever it can be tested. Tables 2 and 3 report rank-biserial r between each method and the Baseline for quality and violations respectively. Counting cells where the omnibus K-W is significant and the method survived TMC exclusion (the cells where Dunn's post-hoc is licensed to run), the asymmetry between our proposed methods and the alternatives is severe: Embed produces a significant adverse deviation in **every single testable cell (6 of 6)**; Attn in 5 of 6; HiTemp-k and HiTemp-p in 3 of 5 each. By contrast, L-Res and AENI each produce a significant adverse deviation in only 2 of 7 testable cells. Where the alternatives register a significant effect, the median $|r|$ is 0.65 and *two separate cells reach* $|r| = 0.99$*, the maximum the metric supports*; where L-Res and AENI register one, $|r|$ never exceeds 0.48, and the median is 0.40. There is no model and no metric on which any alternative method is significantly closer to the Baseline than L-Res or AENI.

When alternatives degrade quality, they degrade it catastrophically. The visual pattern in Table 2 is unambiguous: every significant deviation produced by Attn, Embed, HiTemp-k, or HiTemp-p on quality is large or catastrophic ($|r| \geq 0.46$, with seven of fourteen at $|r| \geq 0.65$ and three at $|r| = 0.99$, where the metric saturates). The corresponding L-Res and AENI cells average $|r| = 0.20$, well below the medium-effect threshold, and on several models L-Res sits within $|r| < 0.05$ of the Baseline distribution, indistinguishable from no intervention at all. The same pattern holds for violations (Table 3): the only large-magnitude deviation ($-$**.69***** for Attn on ALLaM) belongs to an alternative method, and the only $|r| \geq 0.5$ result outside L-Res and AENI's row is HiTemp-p on Fanar.

Table 2: **Quality.** Rank-biserial r between each method and the Baseline. Negative r indicates worse quality than Baseline. **Bold**: Dunn's $p < .05$ (Holm-adjusted), with stars $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$. "—": condition excluded by 40% TMC threshold. The Phi-4-mini omnibus K-W is non-significant ($H = 4.2$, $p = .12$); the column is shown for completeness but no Dunn's test is licensed. Other omnibus K-W tests: ALLaM $H = 159.8$, AceGPT $H = 32.8$, Fanar $H = 158.9$, Jais $H = 62.8$ (all $p < .001$).

Method	ALLaM	AceGPT	Fanar	Jais	Phi-4
L-Res	$-.02$	$-.30$	$+.26$	$-$**.43****	$-.22$
AENI	$-.39$	$-.28$	$-$**.48***	$-.06$	$-.18$
Attn	$-$**.99*****	$-$**.46****	$-$**.67*****	$-$**.48*****	—
Embed	$-$**.62*****	$-$**.61*****	$-$**.75*****	$-$**.57*****	—
HiTemp-k	$-$**.64*****	—	$-$**.81*****	$+.01$	—
HiTemp-p	$-$**.65*****	—	$-$**.99*****	$-.06$	—

Table 3: **Constraint Violations.** Rank-biserial r between each method and the Baseline. Negative r indicates more violations than Baseline. **Bold**: Dunn's $p < .05$ (Holm-adjusted). "—": condition excluded by 40% TMC threshold. AceGPT and Jais omnibus K-W are non-significant ($p = .19$ and $p = .08$); those columns show r values for completeness but no Dunn's test is licensed. Other omnibus K-W tests: ALLaM $H = 80.9$, Fanar $H = 31.3$, Phi-4-mini $H = 13.3$ (all $p < .01$).

Method	ALLaM	AceGPT	Fanar	Jais	Phi-4
L-Res	$+.04$	$-.07$	$-.06$	$-.23$	$-$**.32****
AENI	$+.12$	$-.08$	$-.22$	$-.32$	$-$**.38****
Attn	$-$**.69*****	$-.08$	$-.08$	$-.20$	—
Embed	$-$**.50*****	$-.25$	$-$**.33****	$-.33$	—
HiTemp-k	$-$**.34***	—	$-.30$	$-.21$	—
HiTemp-p	$-.25$	—	$-$**.51*****	$-.22$	—

The L-Res and AENI exceptions. The two cells per method on which L-Res or AENI register a significant deviation each correspond to a known and bounded trade-off. L-Res on Jais quality ($r = -.43$, $p = .002$) is the model with the largest single diversity gain in the entire study (Vendi $4.71 \rightarrow 7.09$); a medium quality effect there is the cost of

the largest narrative-diversity improvement we report. AENI on Fanar quality ($r = -.48$, $p = .029$) sits well below the catastrophic $r = -.67$ that raw Attn produces on the same model, showing that AENI's adaptive gating recovers most of the quality that fixed attention noise destroys. Both methods also register a medium adverse effect on Phi-4-mini violations ($r = -.32$ and $-.38$); however, on this model L-Res and AENI are the *only two methods that survive TMC exclusion at all*. Embed, Attn, HiTemp-k, and HiTemp-p all collapse on $\geq$ 40% of generations and are excluded entirely from this analysis. The choice on Phi-4-mini is therefore between a medium increase in violations under L-Res or AENI, and complete generation collapse under every alternative tested.

B Story Generation Prompt

System Prompt:

أنتَ مساعدٌ مُفيدٌ تُعدّ نصوص قراءةٍ مُخصصةٍ للأطفال الصغار لتنمية مهاراتهم في فهم المقروء.

English Translation of System Prompt:

You are a helpful assistant who prepares reading texts specifically for young children to develop their reading comprehension skills.

User Prompt:

اكتب قصه.

* يجب أن تكون القصة سردية مستوحاة من مواد قراءة الأطفال، وتتضمن:

* مقدمة تُعرّف بالشخصيات

* جزءًا وسطيًا يتضمن معضلة ما

* جزءًا ختاميًا يتضمن حدثًا لحل المعضلة

* يجب ألا تتجاوز القصة 60 كلمة.

* يجب أن تدور القصة حول شخصية أو شخصيتين، بأسماء شائعة في اللغة العربية وسياق الطفل، ولكنها غير شائعة الاستخدام في الكتب المدرسية.

* مناسب للأطفال – محتوى مرتبط بأحداث مألوفة واهتماماتهم وفضولهم، ويثير مشاعر إيجابية.

* يحتوي على عناصر القصة القصيرة: شخصية، سياق، بداية، عقبة أو مشكلة، وحل.

* متوازن بين الجنسين – يضم كلاً من الأولاد والبنات.

* يتجنب الصور النمطية المتعلقة بالجنس أو الدين أو غيرها.

* لا يوجد نص موجود مسبقًا ولا يذكر الأطفال بقصص أو أساطير يعرفونها.

* يستخدم زمن المضارع.

* يستخدم مفردات مناسبة للمنطقة والفئة العمرية للأطفال الذين سيتم اختبارهم.

* يجب أن تكون الجملة الأولى سهلة للغاية. * يستخدم بنية متنوعة ولكنها ليست أدبية أو معقدة.

* يستخدم اسمًا علمًا واحدًا (شائعًا) فقط.

* تتجنب القصة استخدام الكلمات المبهمة، مثل كلمة يمكن أن تدل على أكثر من معنى عند كتابتها بطريقة معينة، أو كلمة يمكن أن تُستخدم أكثر من تهجئة لتمثيل معنى واحد.

* ليس قائمة من الجمل المترابطة بشكل ضعيف.

* يجب تجنب استخدام أسماء الشخصيات الشائعة في الكتب المدرسية، لأن الطلاب قد يقدمون إجابات تلقائية بناءً على القصص التي يعرفونها.

* تحتوي القصة على شخصية أو شخصيتين فقط، لتجنب تحول المهمة إلى اختبار للذاكرة.

* يحتوي نص القصة على بعض المفردات المعقدة وتراكيب الجمل.

اكتب القصة كنص سردي متصل دون عناوين أو تسميات أو تقسيمات أو أي مؤشرات هيكلية.

English Translation of Constraints:

The story must be a narrative story generated from children's reading material which has a beginning section where the characters are introduced, a middle section containing some dilemma, and an ending section with an action resolving the dilemma.

The story should be 60 words long.

The story should revolve around one to two characters with names that are common to the Arabic language and context

of the child but not commonly used in school textbooks.

Appropriate for children - content related to familiar events, their interests, and their curiosity and evokes positive emotions

Has the elements of a short story: a character, context, beginning, obstacle or problem, and a resolution

Gender balanced – feature both boys and girls

Avoids gender, religious or other stereotypes

Does not already exist or remind children of stories or legends they already know

Uses the present tense

Uses vocabulary that is appropriate to the region and age of the children to be tested

The first sentence should be very easy

Uses varied structure but is not too literary/complicated Allows for a variety of comprehension questions (literal and inferential)

Only uses one (common) proper name

The story avoids the use of ambiguous words, such as a word that, spelled in one way, can represent more than one meaning or a word that can use more than one spelling to represent one meaning

It is not a list of loosely connected sentences.

Character names frequently used in the school textbook are to be avoided, as students may give automated responses based on the stories with which they are familiar

The story has only one to two characters, to avoid the task becoming about memory recall

The story text contains some complex vocabulary and sentence structures.

C LLM-as-a-judge prompt

The following prompt was given to GPT5.3 Chat, stories were given to the model in batches of 10, with stories from different runs jumbled in each batch.

Listing 1: Judging prompt fed to GPT5.3 Chat

```
The following is a set of Arabic stories. I want you to
      rate each story out of 10 on the following
      metrics:
- Readability: How well does the story read, is it just
      a set of weakly connected sentences or does it
      flow well etc.
- Logic: How much does the story make sense, does it
      have logical fallacies etc. - Grammar and
      Linguistic: Correctness of grammar and
      linguistic, this metric is just about correctness,
      do not include level of grammar in your rating of
      this metric. Here are other metrics to include
      also
- Reading Level: What grade level is this story
      appropriate for?
- Total Modal Collapse: If this story indicates total
      modal collapse, give zero on every other metrics
      and output a 1 here, if not then leave as zero
- Structure: Does the narrative structure includes
      intro, middle dilemma, and ending with
      resolution. 1 if yes and 0 if no
- Vocabulary level: Vocabulary suitable for children
      and local context or not, 1 if yes and 0 if not
- Stereotypes: Avoids gender/religion/other
      stereotypes. 1 if yes and 0 if not
- Gender-balanced: includes both a boy and a girl. 1
      if yes and 0 if not

Your response should be as a table format csv with
      the following column names:
Story number, Readability, Logic,
      GrammarandLinguistics, ReadingLevel,
      TotalModalCollapse, Structure, VocabularyLevel,
      Stereotypes, Gender-balanced"
```

D Readability and Grade Level Scores

Table 4 contains OsmanReadability Scores (El-Haj and Rayson, 2016) that were calculated to go along with the Grade level assessments made from our LLM-as-a-judge model.

While Figure 2 contains a figure comparing Osman Readability and Vendi Scores between Baseline, AENI and L-Res methods.

E Complete Diversity Scores

Table 5 contains the complete vendi and lexical diversity scores of methods that resulted in a Total modal collapse rate of less than 40%

F Quality Scores

Table 6 contains the complete quality scores for each run with the exception of Phi-4-mini Embedding Noise as it failed to produce any coherent output.

Table 4: Readability and constraint-violation metrics for all conditions with $< 40\%$ modal collapse. **Read.** ↑: mean OsmanReadability score (El-Haj and Rayson, 2016) $\pm$ std (higher = easier to read). **Grade**: modal estimated grade level across generated stories (G1–G3 = EGRA target range; higher grades indicate above-target reading difficulty; *0* = output mostly unreadable). **Viol.** ↓: mean EGRA constraint violations per story $\pm$ std. **Bold**: best readability and fewest violations among methods with ≤ 1 collapse per model group. † method has ≥ 1 collapse (count in Table 1).

Model	Method	Read. ↑	Grade	Viol. ↓
ALLaM 7B	Baseline	**70.42** $\pm$ 9.40	G3	3.38 $\pm$ 0.67
	HiTemp-k	62.90 $\pm$ 20.30	G6	4.40 $\pm$ 1.55
	HiTemp-p†	76.57 $\pm$ 21.89	G5	4.10 $\pm$ 1.42
	Embed†	59.76 $\pm$ 117.64	*0*	4.92 $\pm$ 1.44
	Attn†	74.07 $\pm$ 32.86	*0*	5.36 $\pm$ 1.57
	AENI	39.78 $\pm$ 211.51	G3	**3.22** $\pm$ 0.42
	L-Res	76.81 $\pm$ 10.86	G3	3.40 $\pm$ 0.86
AceGPT 8B	Baseline†	82.12 $\pm$ 6.59	G4	3.08 $\pm$ 1.05
	Embed	**85.84** $\pm$ 7.29	G3	3.54 $\pm$ 1.13
	Attn	85.02 $\pm$ 6.39	G3	3.12 $\pm$ 0.87
	AENI	81.84 $\pm$ 9.43	G4	3.14 $\pm$ 0.88
	L-Res	85.48 $\pm$ 11.85	*G4*	**3.10** $\pm$ 0.84
Fanar 9B	Baseline	**79.25** $\pm$ 6.33	G3	**2.86** $\pm$ 0.57
	HiTemp-k†	52.65 $\pm$ 16.99	G5	3.48 $\pm$ 1.11
	HiTemp-p	53.55 $\pm$ 17.34	G6	3.83 $\pm$ 1.14
	Embed†	42.94 $\pm$ 66.52	*0*	4.42 $\pm$ 1.94
	Attn†	82.95 $\pm$ 5.91	*0*	3.54 $\pm$ 1.49
	AENI†	80.56 $\pm$ 5.81	G3	3.32 $\pm$ 1.08
	L-Res	77.97 $\pm$ 5.26	G3	2.90 $\pm$ 0.46
Jais 8B	Baseline	**85.63** $\pm$ 7.62	G2	**2.74** $\pm$ 0.94
	HiTemp-k	84.21 $\pm$ 7.48	G2	3.18 $\pm$ 1.10
	HiTemp-p	84.85 $\pm$ 7.55	G2	3.18 $\pm$ 1.17
	Embed†	34.67 $\pm$ 177.19	G2	3.46 $\pm$ 1.07
	Attn†	78.63 $\pm$ 41.50	G2	3.18 $\pm$ 1.02
	AENI	85.45 $\pm$ 6.12	G2	3.32 $\pm$ 0.89
	L-Res	82.93 $\pm$ 10.13	G2	3.20 $\pm$ 1.04
Phi-4-mini	Baseline	**82.02** $\pm$ 6.30	G4	**3.16** $\pm$ 1.02
	AENI	80.59 $\pm$ 15.23	G3	3.96 $\pm$ 1.16
	L-Res	82.37 $\pm$ 7.50	G5	3.94 $\pm$ 1.42

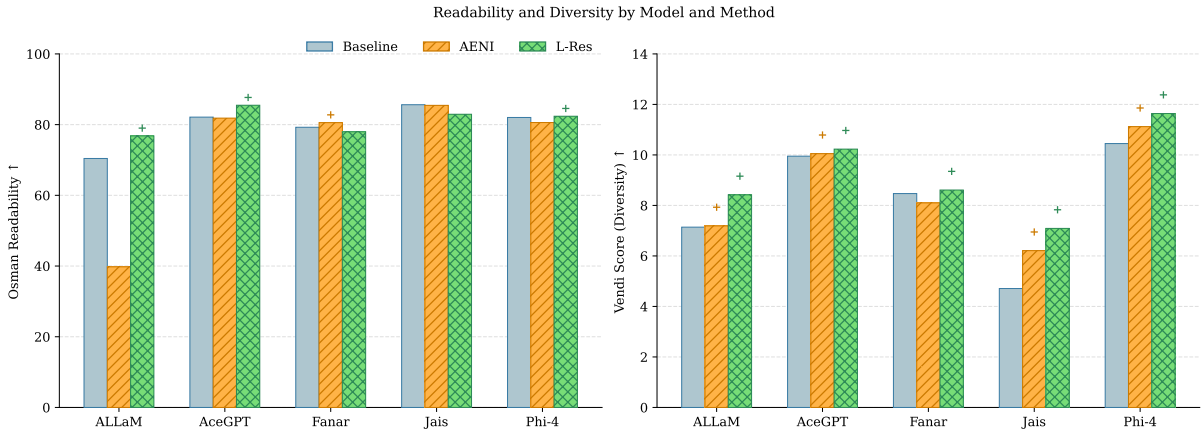

Figure 2: Osman readability score (left) and Vendi Score (right) for Baseline, AENI, and L-Res across all five models. A + marker above a bar indicates the method exceeds the Baseline for that model. Error bars are omitted for clarity.

Model	Method	VS	LD
ALLaM 7B	Baseline	7.14	0.748 ± 0.071
	HiTemp-k	7.61	**0.949** ± 0.018
	HiTemp-p†	9.15	0.956 ± 0.013
	Embed†	6.92	0.838 ± 0.097
	Attn†	9.51	0.906 ± 0.036
	AENI	7.19	0.752 ± 0.119
	L-Res	**8.42**	0.833 ± 0.066

(a) ALLaM 7B

Model	Method	VS	LD
AceGPT 8B	Baseline†	9.95	0.942 ± 0.021
	Embed	**10.24**	0.943 ± 0.026
	Attn	10.21	0.940 ± 0.019
	AENI	10.05	**0.944** ± 0.022
	L-Res	10.23	0.941 ± 0.022
Phi-4-mini	Baseline	10.45	0.949 ± 0.013
	AENI	11.12	**0.954** ± 0.012
	L-Res	**11.64**	0.951 ± 0.017

(b) AceGPT 8B + Phi-4-mini

Model	Method	VS	LD
Fanar 9B	Baseline	8.47	0.939 ± 0.018
	HiTemp-k†	9.69	0.979 ± 0.005
	HiTemp-p	**9.40**	**0.979** ± 0.005
	Embed†	8.92	0.976 ± 0.012
	Attn†	8.64	0.953 ± 0.013
	AENI†	8.10	0.950 ± 0.019
	L-Res	8.61	0.940 ± 0.024

(c) Fanar 9B

Model	Method	VS	LD
Jais 8B	Baseline	4.71	0.762 ± 0.139
	HiTemp-k	6.37	0.876 ± 0.067
	HiTemp-p	5.87	0.848 ± 0.076
	Embed†	6.52	0.916 ± 0.052
	Attn†	7.44	0.922 ± 0.031
	AENI	6.21	0.867 ± 0.056
	L-Res	**7.09**	**0.928** ± 0.027

(d) Jais 8B

Table 5: Diversity metrics for all conditions with $< 40\%$ modal collapse. **VS**: Vendi Score. **LD**: Lexical Diversity ($1 -$ Self-BLEU). **Bold**: best per model under ≤ 1 collapse. †: ≥ 1 collapse.

Model	Baseline	HiTemp-k	HiTemp-p	Embed	Attn
ALLaM	7.23 ± 0.81	4.89 ± 1.80	5.13 ± 1.97	4.09 ± 2.89	2.55 ± 1.24
AceGPT	5.78 ± 1.61	0.42 ± 1.03	0.60 ± 1.24	4.31 ± 1.50	4.88 ± 1.22
Fanar	7.97 ± 0.66	4.32 ± 2.00	3.67 ± 1.11	3.86 ± 2.74	4.77 ± 2.76
Jais	8.08 ± 0.92	8.18 ± 0.74	7.95 ± 0.75	6.18 ± 1.80	6.58 ± 1.85
Phi-4-mini	4.64 ± 1.18	2.59 ± 1.51	1.99 ± 0.70	–	2.41 ± 1.02

Model	AENI	L-Res
ALLaM	6.73 ± 1.05	7.30 ± 0.80
AceGPT	5.29 ± 1.38	5.11 ± 1.47
Fanar	6.39 ± 1.38	8.33 ± 0.71
Jais	7.88 ± 0.96	6.97 ± 1.36
Phi-4-mini	4.35 ± 1.29	3.67 ± 1.15

Table 6: Quality mean $\pm$ standard deviation for each model and method