Zero-Shot Phonics: Evaluating Constraint-Adherent Phonics Story Generation in Large Language Models

Maria Monica Manlises
De La Salle University
Manila, Philippines
maria_monica_manlises@dlsu.edu.ph

Ethel Ong
De La Salle University
Manila, Philippines
ethel.ong@dlsu.edu.ph

Abstract

Phonics stories are essential for early literacy, requiring controlled repetition of grapheme-phoneme (GP) patterns while remaining simple, readable, and developmentally appropriate. Generating such texts poses a challenge for large language models (LLMs), which must handle multiple linguistic and pedagogical constraints. To investigate how these constraints can be balanced, we systematically vary prompt design across 16 configurations and evaluate six LLMs in a zero-shot setting, producing 8,688 outputs (39,096 stories). Outputs are assessed using a multi-dimensional framework covering phonological alignment, developmental lexical appropriateness, readability, and narrative quality. Results show that while LLMs generate highly readable and age-appropriate text, they exhibit variability in phoneme control and narrative coherence. Prompt design and model choice significantly affect performance, revealing trade-offs across constraints. These findings highlight the challenges of controllable educational text generation and the importance of prompt design in balancing instructional objectives. We release our prompts, generated stories, and evaluation framework to support future work in phonics-based story generation.

1 Introduction

Phonics stories are a foundational tool in early literacy instruction, designed to reinforce grapheme-phoneme (GP) correspondences through controlled and repetitive exposure. However, unlike general narrative text, these stories must satisfy multiple instructional constraints, requiring consistent reinforcement of target sound patterns while remaining simple, readable, and developmentally appropriate for early learners (National Reading Panel, 2000; Ehri, 2020).

Phonics instruction plays a critical role in early literacy and speech development, teaching learners to decode written language by associating graphemes—letters or letter combinations—with their corresponding phonemes—sounds or sound combinations (National Reading Panel, 2000; Ehri, 2020). Decodable texts, including phonics stories, are essential components of systematic phonics instruction, as they support the development of phonetic decoding skills through structured GP reinforcement (Ehri, 2020). For example, a phonics story targeting the grapheme ‘s’ and the phoneme /s/ may include sentences such as "***S**am **s**ee**s** the **s**nake in the gra**ss**.*" By reinforcing GP decoding, these stories serve as effective tools for speech and reading development (Cheatham and Allor, 2012; Pennell et al., 2024; Murphy Odo, 2024).

Large language models (LLMs) have demonstrated strong capabilities in generating coherent and contextually appropriate text, including narratives, with minimal prompting (Xie et al., 2023; Tian et al., 2024). Prior work has explored guiding LLMs to produce age-appropriate children’s stories through prompt design and lexical simplification (Valentini et al., 2023). However, existing work does not evaluate whether LLMs can simultaneously satisfy phonological and pedagogical constraints. Unlike general storytelling, this task requires balancing readability and developmental appropriateness with consistent and meaningful reinforcement of target GP mappings.

In this paper, we investigate the effectiveness of zero-shot prompting in generating phonics stories that emphasize specific GP mappings. We evaluate outputs using a multi-dimensional framework spanning phonological alignment, developmental lexical appropriateness, readability, and narrative quality.

To clarify terminology used throughout this work, while the terms “decodable texts” and “phonics stories” are often used interchangeably, we use “phonics stories” to refer specifically to short, structured narratives designed for early readers, whereas “decodable texts” more broadly include instruc-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 914–932

tional materials that emphasize controlled GP exposure without necessarily maintaining narrative structure. In this work, we primarily use the term "phonics stories", while also examining the effect of terminology in prompt design.

Our main contributions are as follows:

1. We use zero-shot prompting for large-scale phonics story generation across 16 prompt configurations and 6 LLMs, producing a dataset[1] of 8,688 outputs (39,096 stories).

2. We define a multi-dimensional evaluation framework comprising 12 metrics that capture phonological alignment, developmental lexical appropriateness, readability, and narrative quality.

3. We provide a detailed analysis of prompt design and model choice, identifying trade-offs between phonological accuracy, linguistic simplicity, and narrative coherence.

2 Related Work

2.1 Automated Children's Story Generation

Automated story generation evolved from early rule-based systems such as TALE-SPIN (Meehan, 1977) and MINSTREL (Turner, 1992) to modern neural approaches (Roemmele et al., 2017). While early systems lacked flexibility and relied heavily on handcrafted rules, subsequent data-driven methods improved narrative coherence through structured knowledge integration and learned representations (Liu and Singh, 2002; Roemmele et al., 2017).

More recently, the emergence of LLMs has significantly advanced the quality of generated stories, achieving improvements in fluency, coherence, and contextual relevance (Brown et al., 2020). However, despite these advances, instructionally constrained storytelling, particularly for educational purposes such as phonics story generation, remains underexplored, as most work prioritizes general narrative quality over pedagogical alignment.

2.2 LLM Capabilities

2.2.1 Story Generation

Recent studies consistently show that LLMs outperform earlier neural models in story generation tasks, particularly in terms of fluency and overall narrative quality (Xie et al., 2023). However, these gains are accompanied by limitations in originality and control, especially when compared to human-authored stories.

In the context of children's storytelling, Valentini et al. (2023) further demonstrate that while LLMs can generate readable and coherent narratives, they often struggle to satisfy developmental constraints such as age-appropriate vocabulary in zero-shot settings. Their findings highlight the importance of lexical simplification, which significantly improves both readability and alignment with target age groups. Complementing this, Alasmari et al. (2025) move beyond surface-level fluency by proposing an automated framework for evaluating narrative cohesion grounded in Gérard Genette's narrative theory. By operationalizing cohesion into computable features and achieving high classification performance, their work provides a foundation for systematically evaluating and improving story structure in both human-authored and AI-generated texts.

Together, these studies suggest that while LLMs are strong generators of fluent narratives, ensuring developmental appropriateness and structural quality requires additional mechanisms for control and evaluation.

2.2.2 Constrained Generation

Following explicit constraints specified in prompts is not just a challenge in general children's storytelling but is also central in phonics-based story generation of LLMs. Prior work shows that even relatively simple constraints, such as age-level requirements, are not consistently satisfied in zero-shot settings (Valentini et al., 2023), raising concerns about the reliability of instruction-following in educational contexts.

More broadly LLMs' controlled generation capabilites have been evaluated across various constraints such as sentiment conditioning, keyword inclusion, prefix constraints, and numerical planning where adherence was found to vary depending on task formulation (Sun et al., 2023). Similarly, Marbun and Shishido (2026) demonstrated how readability targets in prompt formulation can improve lexical appropriateness compared to traditional prompting stratefies.

Collectively, these works emphasize that while LLMs possess strong generative capabilities, achieving reliable constraint adherence requires deliberate prompt design and task-specific strategies.

[1] https://huggingface.co/datasets/monicamanlises/zero-shot-phonics

2.2.3 Phonological Tasks

In order to generate phonics stories, LLMs must exhibit some degree of phonological awareness. While LLMs demonstrate strong theoretical knowledge of phonetics and phonology in controlled evaluations, they often struggle to apply this knowledge in generative settings (Peng et al., 2025), indicating a gap between passive understanding and practical use. This limitation is further supported by Suvarna et al. (2024), who evaluate LLMs on tasks such as grapheme-to-phoneme (G2P) conversion, syllable counting, and rhyme generation. Although advanced models perform relatively well among LLMs, they still fall short of rule-based systems and human performance, particularly in tasks requiring fine-grained phonological precision.

These findings highlight a key challenge for phonics story generation: while LLMs can approximate phonological reasoning, their limitations necessitate careful model selection and evaluation.

3 Task Definition

We define phonics story generation as a constrained phonological text generation task. Given a target GP pair or phonic set, the objective is to generate a short story that simultaneously satisfies three main requirements:

1. **Phonological Alignment:** The story should reinforce the target GP mapping through repeated and meaningful usage.

2. **Developmental Appropriateness:** The vocabulary, sentence structure, and overall readability should align with the intended reading level of early learners, ensuring that texts are both accessible and instructionally effective.

3. **Narrative Quality:** The story should be coherent, cohesive, and easy to follow.

Unlike general story generation, this task requires balancing competing constraints. Increasing phoneme repetition may improve phonological reinforcement but reduce narrative naturalness, while prioritizing readability may weaken phonological control. As a result, phonics story generation serves as a challenging testbed for controllable text generation under educational constraints.

4 Method

This study follows a generation-evaluation pipeline to assess the capability of LLMs in generating phonics stories for early readers. In a zero-shot setting, multiple LLMs were prompted to produce short phonics stories and decodable texts designed to reinforce specific GP correspondences. The generated outputs were subsequently evaluated using a multi-dimensional framework encompassing phonological target alignment, developmental appropriateness, readability, and narrative quality. This methodology enables a systematic investigation of LLM performance in constrained educational text generation.

4.1 Models

We evaluated six LLMs: Claude-3.7-Sonnet (`claude`), GPT-4o (`gpt4o`), GPT-4o-Mini (`gpt4o_mini`), Gemini-2.0 Flash (`gemini`), DeepSeek-V3 (`deepseek`), Mistral-Large-2.1 (`mistral`). These model versions were available as of March 2025.

The models `claude`, `gpt4o`, and `gpt4o_mini` were selected due to their demonstrated strengths in phonological subtasks such as grapheme-to-phoneme (G2P) conversion and syllable counting in prior benchmarks (e.g., PhonologyBench (Suvarna et al., 2024)). Batched API calls were used for these models.

To have a balance of three open-source and three closed-source models, `mistral`, `deepseek`, and `gemini` were also included. These were accessed via standard single-prompt endpoints.

4.2 Prompt Design

We designed 16 structured prompts grouped into three prompt sets. These were constructed to evaluate the ability of LLMs to adhere to phonological, developmental, and narrative constraints. The prompts targeted either individual GP pairs or phonic sets and varied in their degree of constraint and instructional framing.

4.2.1 Phoneme Selection

Target GP mappings are derived from Pearson's *Phonics Progression Chart* (Pearson Education, n.d.), which follows a structured synthetic phonics curriculum aligned with established literacy research (Ehri, 2020; National Reading Panel, 2000; Rose, 2006). Graphemes are preserved in orthographic form, while phonemes are represented in ARPABET. Table 3 lists these GP pairs, as well as corresponding target age, phonic phase, and phonic set.

4.2.2 Prompt Variants

All prompts were initialized with a shared system instruction (Figure 1) that defines the model's role and objective. Three prompt sets were then constructed, each with a distinct focus, along with minor variations within each set.

To examine consistency, half of the prompts required the model to generate a single story, while the remaining prompts required the generation of eight stories. Additionally, prompts alternated between the terms "phonics stories" and "decodable texts" to analyze the effect of terminology on generation behavior, despite their conceptual overlap.

The full set of prompt variants is provided in Appendix A.

Prompt Set 1: Single Grapheme-Phoneme Pair and Narrative Focused Generation (Prompts 1-8) This set targeted individual GP pairs. Variants included basic prompts and plot-driven prompts. These prompts evaluated the ability of LLMs to reinforce a single phoneme within a coherent narrative.

Prompt Set 2: Phonic Set Focused Generation (Prompts 9-12) This prompt set introduced constraints based on a given phonic set, emphasizing the GP pairs within the set while permitting supporting GP pairs from earlier sets to improve fluency and phonic progression.

Prompt Set 3: Age-Constrained Generation (Prompts 13-16) This set incorporated developmental constraints. Phases 2-3 were designed for children aged 4–5, while Phase 5 was designed for children aged 5–6. These age ranges were originally aligned with the UK Department for Education's *Letters and Sounds* framework, in which Phases 2 and 3 are introduced during Reception (ages 4–5) and Phase 5 is taught in Year 1 (ages 5–6) (Department for Education and Skills, 2008). Although *Letters and Sounds* was formally withdrawn in 2024 as part of the shift toward updated validated systematic synthetic phonics programs, its phase structure remains widely referenced in both research and instructional materials. This progression is also reflected in Pearson's Phonics Progression Chart and broader systematic synthetic phonics curricula (Pearson Education, n.d.; Rose, 2006).

4.3 Data Cleaning and Post-processing

Generated outputs were standardized for consistency and evaluability. The process included removing extraneous labels, re-prompting malformed outputs, and normalizing formatting.

While the target structure specifies exactly four sentences, minor deviations (3-5 sentences) were retained to avoid introducing bias through aggressive filtering. These cases were included in evaluation to preserve the integrity of model outputs.

4.4 Evaluation Metrics

Generated stories are evaluated using 12 automated metrics spanning four dimensions: phonological alignment, developmental lexical appropriateness, readability, and narrative quality. Full formal definitions are provided in Appendix B. For clarity, metrics are referenced by both name and index (1-12) throughout the paper.

4.4.1 Phonological Target Alignment (Metrics 1-4)

Phonological target alignment measures adherence to target GP mappings. A word is considered aligned if the target grapheme appears in the word and the corresponding phoneme appears in its ARPABET transcription. Metrics include **(M1) Story-level Phoneme Density**, **(M2) Sentence-level Phoneme Density**, **(M3) Mean Phonics Level Deviation**, and **(M4) Maximum Phonics Level Deviation**.

4.4.2 Developmental Lexical Appropriateness (Metrics 5-6)

Lexical appropriateness captures the suitability of vocabulary for early readers. This is quantified using Age of Acquisition (AoA) norms (Kuperman et al., 2012; Valentini et al., 2023). Metrics include **(M5) Mean AoA Deviation** and **(M6) Maximum AoA Deviation** from the target age range.

4.4.3 Readability (Metrics 7-10)

Readability measures the linguistic simplicity of the text and is assessed using standard metrics, including **(M7) Flesch Reading Ease**, **(M8) Flesch-Kincaid Grade Level**, **(M9) Gunning Fog Index**, and **(M10) Automated Readability Index**. Thresholds aligned with early readers (ages 4-6) are used for interpretation (Table 4).

4.4.4 Narrative Quality (Metrics 11-12)

Narrative quality evaluates coherence and cohesion in the generated stories. These metrics

are scored using an LLM-as-a-judge framework (Claude-Haiku-4.5) with rubric-based prompting to improve consistency and reduce variance. Scores are normalized to a 0-1 scale. **(M11) Coherence** measures global narrative organization, including setup, logical progression, causal relationships, and resolution. **(M12) Cohesion** assesses local connectedness across sentences, including referential clarity, lexical consistency, and the absence of contradictions.

The full evaluation prompt, inspired by the coherence and cohesion frameworks of Alasmari et al. (2025) and Hargood et al. (2011), is provided in Appendix G.

The evaluator models also generated textual justifications alongside their rubric-based scores. These outputs were retained separately for qualitative inspection but were omitted from the main dataset and analyses to maintain focus on aggregate quantitative comparisons.

4.5 Statistical Analysis

Evaluation results were compared across metric, prompt design, and model levels using aggregate analysis, omnibus tests, and post-hoc pairwise comparisons.

Prior to significance testing, a Shapiro-Wilk normality test was applied to each group within a comparison. The test evaluates the null hypothesis that a sample is drawn from a normal distribution; a p-value ≤ 0.05 indicates a significant deviation from normality.

As most metric distributions violated normality assumptions, non-parametric tests were used throughout. For comparisons involving more than two groups, the Kruskal-Wallis (KW) test was applied as the omnibus test. For pairwise comparisons between two groups, the Mann-Whitney U (MWU) test was used.

For comparisons with more than two groups, post-hoc pairwise tests were conducted only when the omnibus test was significant ($\alpha = 0.05$). Bonferroni correction was applied to adjust p-values and control the family-wise error rate.

Omnibus tests determine whether significant differences exist across groups, while post-hoc pairwise tests identify which specific group pairs differ significantly.

Metric	Mean	Std
1. Story-Level Phoneme Density (↑)	0.360	0.222
2. Sentence-Level Phoneme Density (↑)	0.373	0.219
3. Mean Phonics Level Deviation (↓)	1.440	0.997
4. Max Phonics Level Deviation (↓)	1.053	0.728
5. Mean AoA Deviation (↓)	0.206	0.275
6. Max AoA Deviation (↓)	3.378	1.637
7. Flesch Reading Ease (↑)	104.206	11.721
8. Flesch-Kincaid Grade Level (↓)	0.196	1.681
9. Gunning Fog Index (↓)	2.848	1.217
10. Automated Readability Index (↓)	-0.386	2.320
11. Coherence (↑)	0.665	0.231
12. Cohesion (↑)	0.704	0.167

Table 1: Summary statistics across all generated stories. Arrows indicate whether higher (↑) or lower (↓) values are generally preferable.

5 Results and Discussion

Through zero-shot prompting, a dataset of 8,688 outputs comprising 39,096 stories was generated, cleaned, and evaluated across four metric groups. Overall, models consistently produce highly readable and age-appropriate text, but show greater variability in phonological alignment and narrative quality. Both prompt design and model choice significantly affect performance across all metrics.

5.1 Metric-Level Comparison

Table 1 summarizes aggregate performance across all models and prompt configurations. Results indicate consistently high readability and appropriate vocabulary usage, alongside moderate adherence to phonological constraints and more variable narrative quality (see Appendix F).

5.1.1 Phonological Target Alignment

Phoneme density is moderate but acceptable (0.360), indicating emphasis of target phonemes in at least one per three words. Sentence-level density (0.373) is slightly higher, suggesting uneven distribution across sentences. However, phonics level deviation further highlights weak control. The mean deviation (1.44) indicates frequent use of words outside the target level, while a high maximum deviation reflects occasional inclusion of a substantially different phonic phase. The raw values show that around 75% of deviation is negative. Along with a lower maximum phonics level deviation, this indicates the use of phonemes in easier phonic phases.

5.1.2 Developmental Lexical Appropriateness

Lexical appropriateness is generally strong, with low mean AoA deviation (0.206). However, a high

average maximum deviation (3.378) indicates occasional out-of-scope words, even reaching a deviation of 10.920. This reflects stable average performance but inconsistent worst-case behavior, suggesting limited constraint enforcement at the token level.

5.1.3 Readability

Readability is consistently high, with Flesch Reading Ease above 100 and grade-level metrics near or below early elementary levels. This indicates strong control over linguistic simplicity, possibly due to the maximum syllable requirement in the basic prompt in Figure 1.

Compared to other metrics, readability aligns closely with general language modeling objectives, making it the most stable dimension. However, this strength does not guarantee alignment with phonics progression, as linguistically simple text may still fail to systematically reinforce target grapheme-phoneme correspondences.

5.1.4 Narrative Quality

Narrative quality is moderate, with coherence (0.665) lower than cohesion (0.704). This indicates stronger local sentence connections than global narrative structure. Performance varies across models, with stronger models producing more stable narrative flow and weaker models showing greater inconsistency.

5.2 Prompt Design-Level Comparison

Prompt design significantly affects performance across all metric groups, influencing trade-offs between phonological alignment, readability, and narrative quality. Different aspects of prompt design, such as prompt set, terminology, and story number, were evaluated for their performance. Figures 5 and 6 summarize performance across prompt sets and prompts, respectively.

5.2.1 Comparison Across Prompt Sets

All prompt sets achieve high readability ratings, with Prompt Set 2 producing the easiest texts.

Prompt Set 1, which was designed to focus on both specific phonemes and plot, achieves desired outputs. It yields the strongest phonemic alignment but with at least 1 level of deviation from the desired phonic phase. It also achieves a moderate narrative quality, only a close second to Prompt Set 3. With regards to lexical appropriateness, it achieves the lowest mean deviation but also contains more words that deviate greatly from the target age of acquisition.

Meanwhile, Prompt Set 2, which focuses on phonic-set-based generation, has the largest deviation from the desired phonics phase, which may be because it is more lenient on what specific phoneme to use. The texts produced by this prompt set also had the worst coherence and cohesion.

Prompt Set 3, which focuses on age-appropriateness, produces texts with more consistent ages of acquisition rates and produces acceptable phonological alignment, although it is the lowest in terms of phoneme density.

In Figure 10, KW tests revealed that there were significant differences between the prompt sets for all metrics except for metric 3 (Maximum Phonics Level Deviation). Bonferroni-corrected MWU pairwise tests show that all prompt sets are significantly different from each other with regard to readability despite producing extremely readable texts.

Similarly, there are significant differences across all specific prompts on all metrics except for phonics level deviation. Narrative coherence had the most pairs with significant differences (91 out of $\binom{16}{2} = 120$).

5.2.2 "Phonics-Based Story" vs "Decodable Text"

The prompts alternated between using “phonics-based stories" and “decodable texts." No significant differences are observed in phoneme density, indicating stable phonological alignment across the terminology used. However, using "phonics-based story" achieves higher coherence and cohesion, while "decodable text" prompts produce easier texts. This difference was found to be significant using MWU tests (Figure 10).

This suggests that LLMs may understand that stories require richer narratives, while texts simply adhere to the target of being decodable. Thus, even a slight change in the terminology for the target output influences the trade-off between narrative structure and readability, although the effect is less pronounced than other design factors.

5.2.3 Single-story vs Multi-story Prompts

Using MWU tests, we compared each single-story prompt to its equivalent multi-story prompt (e.g., Prompt 1 vs.3), totaling 8 pairs. This showed that all single-story prompts achieve significantly higher phoneme density, while multi-story prompts yield better coherence, cohesion, and readability.

This trade-off is consistent across prompt sets, indicating that output format strongly affects constraint adherence. There was no significant difference found for phonics level deviation metrics.

5.3 Model-Level Comparison

Model choice is the primary determinant of overall performance, showing significant differences across all metrics using KW tests and post-hoc Bonferroni-corrected MWU tests. Particularly, performance differences in terms of phoneme density are most evident and are present in all $(\binom{6}{2} = 15)$ pairs. Further analysis showed that some models showed clear trade-offs with phonological alignment for narrative quality while others show moderate yet acceptable performance across all metrics.

Figure 8 shows `claude` and `gemini` have the strongest phonological alignment, where the target phoneme or phonic set is present in around one per two words. However, they also perform the worst in terms of narrative quality. In contrast, `gpt4o_mini` and `mistral` rank highest in terms of narrative quality but lowest in terms of phonological alignment. Keeping a better balance of both these metric categories are `deepseek` and `gpt4o`, with `deepseek` showing the best set of balanced scores as seen in Figure 7

In samples 1 to 6 in Table 6, we can see stories generated by the models for the GP mapping between grapheme 'le' and phonemes ['L', 'AH L']. This is one case that highlights the trade-off between phonological alignment and narrative quality, where we see that the stories generated by `claude`, `gemini`, and `deepseek` have the target GP pair appearing in more than 1 per 2 words while having low narrative quality. In contrast, `mistral` and `gpt4o_mini` gained perfect scores for coherence and cohesion, but the target GP pair appeared in fewer than 1 in 3 words in the story.

5.3.1 Selected Model Performance Comparison

Figure 9 presents a focused comparison between `claude`—because of its strong alignment—and `deepseek`—because of its strong balanced performance. They are compared across four representative metric categories: phonological alignment (M1), lexical difficulty control (M5), readability (M7), and narrative quality (M11-M12), evaluated across all prompt sets (PS1-PS3).

Phonological Alignment (M1) Across all prompt sets, `claude` consistently achieves higher phoneme coverage, with median values centered around 0.45–0.55, compared to `deepseek`'s lower range of approximately 0.30–0.45. This confirms its stronger adherence to target GP constraints. However, `claude` also exhibits slightly wider interquartile ranges, indicating less consistent enforcement. While weaker in overall coverage, `deepseek`, shows more stable distributions across prompt configurations.

Lexical Difficulty Control (M5) Both models maintain relatively low AoA deviation, suggesting that generated vocabulary generally aligns with the intended difficulty level. Stories by `claude` show tighter distributions closer to zero, indicating more consistent lexical control, while `deepseek`, particularly under PS2, exhibits slightly higher medians and greater spread, suggesting occasional introduction of more difficult words.

Readability (M7) In terms of readability, `deepseek` clearly outperforms `claude`. Across all prompt sets, `deepseek` achieves higher Flesch Reading Ease scores, with medians consistently exceeding those of `claude`. Furthermore, `deepseek`'s distributions are more compact, indicating more stable control over sentence complexity. More variability is seen in `claude`'s stories, showing more low-end outliers and reflecting occasional drops in readability.

Narrative Quality (M11–M12) Stories generated by `deepseek` demonstrate stronger narrative performance overall, with higher median scores in both coherence and cohesion across all prompt sets. The gap is especially pronounced under PS2, where `claude` exhibits a notable decline and increased variance. While `claude` improves under PS3, it remains slightly below `deepseek` in both metrics. Tighter distributions present in `deepseek`'s stories indicate more consistent narrative structure.

Trade-off Analysis These results highlight a clear trade-off between constraint satisfaction and overall story quality. While `claude` excels in enforcing phonological constraints and maintaining lexical simplicity, it comes at the cost of reduced readability and weaker narrative structure. In contrast, `deepseek` provides a more balanced performance, achieving higher readability and narrative quality while maintaining moderate phonological alignment. Sample stories 7 to 10 in Table 6 also reflect these trade-offs.

Prompt Effects Across both models, PS2 consistently underperforms relative to PS1 and PS3. This is reflected in lower phoneme coverage, higher lexical deviation, and reduced narrative quality. In contrast, PS1 and PS3 produce more stable and higher-quality outputs, with PS3 often achieving the best balance between constraint adherence and fluency.

6 Conclusion

This work investigates zero-shot phonics story generation as a constrained text generation task requiring simultaneous adherence to phonological alignment, developmental appropriateness, readability, and narrative quality. Through a large-scale evaluation across multiple models and prompt configurations, our findings reveal a consistent pattern: while LLMs produce highly readable and age-appropriate text, they exhibit moderate yet acceptable phonological alignment and variable performance in narrative coherence.

Our analysis highlights systematic trade-offs between constraint categories. Models that achieve stronger phonological alignment often do so at the expense of narrative quality and vice versa. Prompt design further influences these trade-offs, shaping how constraints are prioritized, while model choice determines the overall balance and stability of performance across metrics.

Importantly, results show that no single model consistently satisfies all instructional requirements, underscoring the difficulty of multi-constraint generation in educational contexts. While some models demonstrate more balanced behavior, phonological alignment remains the most challenging dimension, particularly in maintaining consistent and instructionally meaningful reinforcement of target grapheme-phoneme correspondences.

Overall, this work establishes phonics story generation as a stringent testbed for controllable text generation and provides a multi-dimensional evaluation framework for assessing instructional alignment. These findings highlight the need for methods beyond zero-shot prompting, including constrained decoding, structured generation, or fine-tuning approaches, to better align LLM outputs with pedagogical objectives.

By systematically analyzing model behavior across constraints, this study contributes both empirical insights and evaluation tools for future research in educational NLP and controllable text generation.

Limitations

This study has several limitations that suggest directions for future work. First, the evaluated models reflect capabilities available as of early 2025; newer systems may exhibit improved performance. Consequently, greater emphasis should be placed on the generalizability of the proposed generation, evaluation, and analysis framework.

Story generation used default sampling parameters rather than deterministic decoding (e.g., temperature = 0), reflecting real-world usage but limiting reproducibility. In addition, all stories were constrained to four sentences, restricting the evaluation of longer narrative structures and potentially underestimating model capabilities in maintaining coherence over extended contexts.

Phonological alignment was measured using GP matching via G2P conversion. While scalable, this approach does not capture pronunciation variability or contextual phoneme realization. Readability was assessed using formula-based metrics that capture surface-level linguistic complexity but may not fully reflect comprehension difficulty for early learners. Narrative quality was evaluated using a single LLM-based evaluator. While rubric-based prompting improves consistency, reliance on one evaluator introduces potential bias. Incorporating multiple LLM judges, ensemble methods, or human annotation would improve robustness. Although evaluator-generated textual rationales and subcriterion-level assessments were retained during scoring, these qualitative traces were not systematically analyzed in the present work. Future research could examine these explanations to better understand evaluator behavior, improve interpretability, and identify recurring narrative strengths or failure modes across generated stories.

Beyond automated evaluation, the study does not include comparisons against professionally developed or expert-authored phonics stories. As a result, while the proposed metrics enable relative comparison across prompts and models, they do not yet establish normative or pedagogically optimal ranges for high-quality instructional materials. Future work could benchmark generated stories against curated human-authored phonics corpora to better contextualize metric distributions, such as phoneme density and readability, and to identify target ranges associated with effective instructional

design.

The study also does not include expert evaluation or in-field testing and therefore does not measure actual learning outcomes such as phoneme acquisition, reading fluency, or comprehension. Future work should incorporate educator assessment and classroom-based studies to evaluate pedagogical effectiveness.

Finally, this study is limited to English phonics story generation. Extending the framework to other languages and literacy contexts remains an important direction for future research.

Ethical Considerations

This work explores the use of LLMs for generating phonics stories intended for early literacy instruction. While such systems have the potential to support scalable educational content creation, they also raise several ethical considerations.

The generated phonics stories are designed to support early readers but they are not a substitute for professionally developed instructional materials. LLM-generated content may contain inaccuracies in phonological alignment, inappropriate vocabulary, or inconsistencies in narrative structure. They should therefore be used with educator oversight.

This study relies on automated evaluation metrics, including LLM-based assessment for narrative quality. While this enables large-scale analysis, it does not fully capture pedagogical effectiveness. The absence of expert and in-field evaluation means that conclusions about instructional quality should be interpreted with caution.

LLMs are trained on large-scale web data and may reflect biases present in their training corpora. Although phonics stories are typically simple, generated content may still include cultural, social, or linguistic biases that are not appropriate for young learners. Future work should include bias analysis and content filtering to ensure inclusivity and appropriateness.

References

Jawharah Alasmari, Mohammed Alzyoudi, Masheal Alshehri, Rana Alshammari, and Reyouf Aldakan. 2025. An automated predictive model for evaluating narrative cohesion in children's stories: a computational linguistic approach considering gérard genette's narrative structure theory. *International Journal of Adolescence and Youth*, 30(1):2500527.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jennifer P. Cheatham and Jill H. Allor. 2012. The influence of decodability in early reading text on reading achievement: a review of the evidence. *Reading and Writing*, 25(9):2223–2246.

Department for Education and Skills. 2008. Letters and sounds: Principles and practice of high quality phonics. Technical report, UK Department for Education and Skills, Nottingham, UK.

Linnea C. Ehri. 2020. The science of learning to read words: A case for systematic phonics instruction. *Reading Research Quarterly*, 55(S1):S45–S60.

Charlie Hargood, David Millard, and Mark Weal. 2011. Measuring narrative cohesion: A five variables approach. In *Narrative and Hypertext Workshop at Hypertext 11 (06/06/11)*.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Hugo Liu and Push Singh. 2002. Makebelieve: using commonsense knowledge to generate stories. In *Eighteenth National Conference on Artificial Intelligence*, page 957–958, USA. American Association for Artificial Intelligence.

Ronald William Marbun and Makoto Shishido. 2026. A readability-driven prompting framework for accurate grade-specific efl narrative creation. *International Journal of Advanced Computer Science and Applications*, 17(1).

James R. Meehan. 1977. Tale-spin, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'77, pages 91–98, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dennis Murphy Odo. 2024. The use of decodable texts in the teaching of reading in children without reading disabilities: a meta-analysis. *Literacy*, 58(3):267–277.

National Reading Panel. 2000. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups. Technical Report NIH Publication No. 00-4754, National Institute of Child Health and Human Development, Washington, DC.

Pearson Education. n.d. Phonics progression chart. Accessed 2025.

Linkai Peng, Baorian Nuchged, and Yingming Gao. 2025. Spoken language intelligence of large language models for language learning. *Preprint*, arXiv:2308.14536.

Ashley E. Pennell, Rebecca Lee Payne Jordan, Kindel Turner Nash, Kerry Elson, and Woodrow Trathen. 2024. A healthy diet for beginning readers: Decodable texts as part of a comprehensive literacy program. *The Reading Teacher*, 77(5):673–684.

Melissa Roemmele, Andrew S. Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses.

Jim Rose. 2006. Independent review of the teaching of early reading. Technical report, UK Department for Education.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

S.R. Turner. 1992. *MINSTREL, a Computer Model of Creativity and Storytelling*. UCLA Computer Science Department.

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

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

A Prompt Design

All prompt sets incorporated the basic prompt structure in Figure 1 and had respective additional requirements listed in Figures 2, 3, and 4.

System Prompt:
You are a children's story writer and children's educator specializing in speech and language acquisition. You write decodable texts and phonics stories, which are engaging stories for early readers that reinforce grapheme-phoneme associations through repeated exposure to target sounds.

Story Requirements:
- (Prompt-specific additional requirements)
- The story must contain exactly four sentences.
- Each sentence must contain fewer than eleven syllables.
- The story must reinforce the target phoneme(s).

Output Format:
Title
Sentence 1
Sentence 2
Sentence 3
Sentence 4

Figure 1: Base prompt structure for phonics story generation.

Prompt Set 1: Single GP and Narrative Focus

Prompt: Generate a short phonics-based story focusing on the grapheme '{grapheme}', which corresponds to the phoneme(s) '{phoneme(s)}'.

Additional Requirements:
- Each sentence must include at least 2 words containing the grapheme '{grapheme}'.

Figure 2: Prompt Set 1: Single GP and Narrative Focus

Across all prompt sets, systematic variations, as seen in Table 2 were applied. The phrase "phonics story" was replaced with "decodable text" in selected prompts. Prompts requesting a single story were modified to generate "8 different" stories. Additional narrative constraints (e.g., requiring a plot

Prompt Set 2: Phonic Set Focus

Prompt: Generate a short phonics story that gives primary emphasis to the target graphemes in the list {grapheme_set}, which correspond to the phonemes {phoneme_set}, respectively. The majority of the key vocabulary should highlight these target grapheme-to-phoneme correspondences. You may also include graphemes from {lower_grapheme_set_list} to create complete and readable words, but they should play a supporting role.

Additional Requirement:
- The story should be designed to reinforce recognition and pronunciation of the target grapheme-to-phoneme correspondences.

Figure 3: Prompt Set 2: Phonic Set Focus

Prompt Set 3: Age-Constrained

Prompt: Generate a short phonics story for children aged {age} years old. The phonics story must focus on the grapheme '{grapheme}', which corresponds to the phoneme(s) '{phoneme(s)}'.

Additional Requirements:
- The story must be age-appropriate and engaging.
- Each sentence must include at least 2 words containing the grapheme '{grapheme}'.

Figure 4: Prompt Set 3: Age-Constrained

with consistent characters) were included in selected prompts.

B Evaluation Metrics

Notation.

- Each story $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$ consists of sentences.
- Let $W_i = \{w_{i1}, w_{i2}, \ldots, w_{iN_i}\}$ be the words in sentence s_i.
- Let $W = \bigcup_{i=1}^{4} W_i$ be all words in the story, with total size $|W| = \sum_{i=1}^{4} N_i$.
- $\mathbf{1}(\cdot)$ is the indicator function.
- $L(w)$ denotes the set of phonics phases associated with word w.
- Ph_t denotes the target phoneme.
- PS_t denotes the target phonics set.
- PP_t denotes the target phonics phase.
- $A(w)$ denotes the Age of Acquisition (AoA) of word w.
- $\mathcal{A} \subseteq W$ is the set of words with valid AoA values.
- $[a_{\min}, a_{\max}]$ is the target age range.

B.1 Metric 1: Phoneme Density (Story-Level)

Measures the proportion of words in the story that contain the target phoneme or a phoneme from the target set (for Prompt Set 2). This captures how strongly the story reinforces the target at the overall story level. Higher values indicate better phonological alignment.

$$\mathrm{M}_1 = \frac{1}{|W|} \sum_{w \in W} \mathbf{1}(w \text{ contains phoneme}) \quad (1)$$

B.2 Metric 2: Phoneme Density (Sentence-Level)

Measures the average proportion of words containing the target phoneme or a phoneme from the target set (for Prompt Set 2) computed at the sentence level. This ensures that phoneme usage is distributed consistently across sentences rather than concentrated in only one part of the story.

$$\mathrm{M}_2 = \frac{1}{4} \sum_{i=1}^{4} \left(\frac{1}{|W_i|} \sum_{w \in W_i} \mathbf{1}(w \text{ contains phoneme}) \right) \quad (2)$$

B.3 Metric 3: Mean Phonics Level Deviation

Measures the deviation of the average phonics phases associated with each word from the target phonics phase. Note that tricky words listed in 3 were counted towards that phonics phase. Values closer to 0 indicate better alignment.

$$\mathrm{M}_3 = \left| \frac{1}{|W|} \sum_{w \in W} \left(\frac{1}{|L(w)|} \sum_{\ell \in L(w)} \ell \right) - PP_t \right| \quad (3)$$

Variation	Prompt IDs
"phonics-based story" → "decodable text"	2, 4, 6, 8, 10, 12, 14, 16
Single story → "8 different" stories	3, 4, 7, 8, 11, 12, 15, 16
Plot requirement added	5, 6, 7, 8

Table 2: Systematic prompt variations applied across prompt sets.

Target Age	Phase	Set	Grapheme-Phoneme Mappings	Tricky Words
4-5	2	1	s→['S'], a→['AA','AE','AH'], t→['T'], p→['P']	the, to, I, no, go, into, and
		2	i→['IH'], n→['N'], m→['M'], d→['D']	
		3	g→['G'], o→['AA'], c→['K'], k→['K']	
		4	ck→['K'], e→['EH'], u→['AH'], r→['R']	
		5	h→['HH'], b→['B'], f→['F'], ff→['F'], l→['L'], ll→['L'], ss→['S']	
	3	6	j→['JH'], v→['V'], w→['W'], x→['K','S']	he, she, we, me be, was, you, they, all, are, my, her
		7	y→['Y'], z→['Z'], zz→['Z'], qu→['K','W']	
		8	ch→['CH'], sh→['SH'], th→['TH','DH'], ng→['NG']	
		9	ai→['EY'], ee→['IY'], igh→['AY'], oa→['OW'], oo→['UH']	
		10	ar→['AA','R'], or→['AO','R'], ur→['ER'], ow→['AW'], oi→['OY']	
		11	ear→['IH','R','IY','R'], air→['EH','R'], ure→['Y','UH','R'], er→['ER']	
5-6	5	13	s|z→['ZH'], wh→['W'], ph→['F']	oh, their, people, Mr, Mrs, looked, called, asked
		14	ay→['EY'], a.e→['EY'], eigh→['EY'], ey→['EY'], ei→['EY']	
		15	ea→['IY'], e.e→['IY'], ie→['IY'], ey→['IY'], y→['IY']	
		16	ie→['AY'], i.e→['AY'], y→['AY'], i→['AY']	
		17	ow→['OW'], o.e→['OW'], o→['OW'], oe→['OW']	
		18	ew→['UW'], ue→['UW'], u.e→['UW'], u→['UH'], oul→['UH']	
		19	aw→['AO'], au→['AO'], al→['AO','L'], our→['AO','R']	
		20	ir→['ER'], er→['ER'], ear→['ER']	
		21	ou→['AW'], oy→['OY']	
		22	ere→['IH','R'], eer→['IH','R'], are→['EH','R'], ear→['EH','R']	
		23	c→['K'], k→['K'], ck→['K'], ch→['K']	
		24	ce→['S'], ci→['S'], cy→['S'], sc→['S'], stl→['S'], se→['S']	
		25	ge→['JH'], gi→['JH'], gy→['JH'], dge→['JH']	
		26	le→['L','AH','L'], mb→['M'], kn→['N'], gn→['N'], wr→['R']	
		27	tch→['CH'], chlc→['SH'], ea→['EH'], a→['AA'], o→['AH']	

Table 3: Phoneme sets and progression mapping used in this study, adapted from Pearson's Phonics Progression Chart (Pearson Education, n.d.) Each set groups grapheme-phoneme correspondences and associated target age ranges, and is assigned a phonics phase corresponding to instructional progression. The phoneme progression consists of 26 sets comprising a total of 112 grapheme-phoneme (GP) correspondences. Phase 4 (Set 12) is omitted, as it serves as a consolidation phase focusing on blending and fluency using previously introduced correspondences, rather than introducing new GP mappings. Tricky words indicate words to be learned at that phase but may contain certain graphemes that conflict with target GP mappings.

B.4 Metric 4: Maximum Phonics Level Deviation

Measures the deviation of the highest phonics phase deviation per word, averaged across the story. Note that tricky words listed in 3 were counted towards that phonics phase. Values closer to 0 indicate better alignment.

$$\mathrm{M}_4 = \left| \frac{1}{|W|} \sum_{w \in W} \max_{\ell \in L(w)} \ell - PP_t \right| \qquad (4)$$

where

B.5 Metric 5: Mean Age of Acquisition Deviation

Measures the deviation of the average word-level Age of Acquisition (AoA) values from the target age range. Words within the target range incur no penalty, while words outside the range contribute proportionally to their distance.

$$\mathrm{M}_5 = \delta \left(\frac{1}{|\mathcal{A}|} \sum_{w \in \mathcal{A}} A(w) \right) \qquad (5)$$

$$\delta(a) = \begin{cases} a_{\min} - a, & a < a_{\min} \\ 0, & a_{\min} \le a \le a_{\max} \\ a - a_{\max}, & a > a_{\max} \end{cases} \quad (6)$$

B.6 Metric 6: Maximum Age of Acquisition Deviation

Measures the deviation of the AoA of the most difficult word in the text from the target age range. This captures the worst-case lexical difficulty.

$$\mathrm{M}_6 = \delta\left(\max_{w \in \mathcal{A}} A(w)\right) \quad (7)$$

B.7 Metrics 7-10: Readability

These metrics assess surface-level linguistic complexity based on sentence length, word difficulty, and character-level features. Lower complexity generally corresponds to better suitability for early readers, but readability threasholds are also presented in C.

Let:

- $\mathrm{ASL} = \frac{\text{words}}{\text{sentences}}$ (average sentence length)
- $\mathrm{ASW} = \frac{\text{syllables}}{\text{words}}$ (average syllables per word)
- $\mathrm{PHW} = \frac{\text{hard words}}{\text{words}}$ (proportion of complex words)

Syllable count was obtained from 'cmudict' in 'nltk.corpus' but when the word was not found in the dataset, the 'syllables' library in Python was used to estimate it. The number of words whose syllable count were estimated was noted as 'Est Syl' in the dataset.

Metric 7: Flesch Reading Ease Measures how easy the text is to read. Higher values indicate simpler and more accessible text.

$$\mathrm{M}_7 = 206.835 - 1.015 \cdot \mathrm{ASL} - 84.6 \cdot \mathrm{ASW} \quad (8)$$

Metric 8: Flesch-Kincaid Grade Level Estimates the U.S. grade level required to understand the text. Lower values indicate simpler text.

$$\mathrm{M}_8 = 0.39 \cdot \mathrm{ASL} + 11.8 \cdot \mathrm{ASW} - 15.59 \quad (9)$$

Metric 9: Gunning Fog Index Measures readability based on sentence length and proportion of complex words. Lower values indicate easier readability.

$$\mathrm{M}_9 = 0.4 \cdot (\mathrm{ASL} + 100 \cdot \mathrm{PHW}) \quad (10)$$

Metric 10: Automated Readability Index Estimates readability using character-level and sentence-level features. Lower values indicate easier text.

$$\mathrm{M}_{10} = 4.71 \cdot \frac{\text{characters}}{\text{words}} + 0.5 \cdot \frac{\text{words}}{\text{sentences}} - 21.43 \quad (11)$$

B.8 Metrics 11-12: Narrative Quality

Narrative quality is evaluated using an LLM-as-a-judge framework based on predefined rubrics (A1-4, B1-4) as stated in the prompt in Figure 11. Each criterion is scored on a 3-point scale: 0 = Not present; 1 = Partially present; 2 = Clearly present. The scores for A1-4 and B1-4 were also noted in the dataset.

Metric 11: Coherence Measures meaning-level organization, including clarity of setup, logical progression, causal relationships, and resolution.

$$\mathrm{M}_{11} = \frac{1}{4} \sum_{k=1}^{4} \frac{A_k}{2} \quad (12)$$

Metric 12: Cohesion Measures surface-level connectedness across sentences, including referential clarity and linguistic consistency.

$$\mathrm{M}_{12} = \frac{1}{4} \sum_{k=1}^{4} \frac{B_k}{2} \quad (13)$$

C Readability Thresholds

Metric	Ideal	Acceptable	Too Hard
FRE	80-100	70-80	< 60
FKGL	0-2	3-4	≥ 5
Fog	≤ 6	7-8	≥ 9
ARI	0-2	3-4	≥ 5

Table 4: Readability thresholds for early readers.

D Prompt-Level Figures

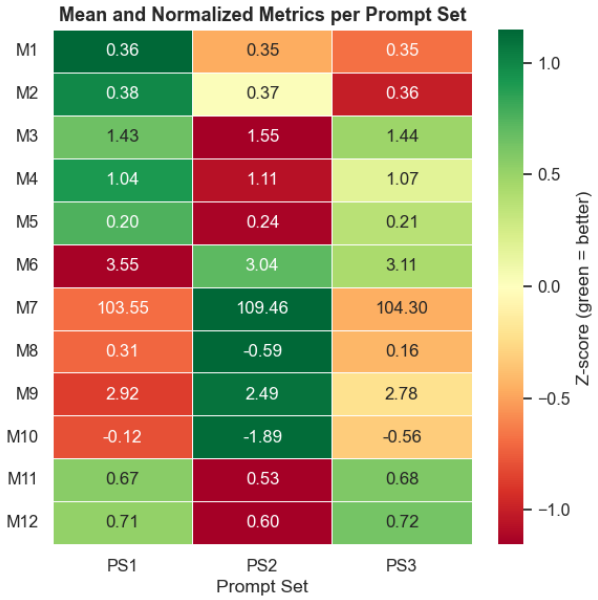

Figure 5: Performances of prompt sets across all metrics, averaged (annotation=mean) and normalized (color=Z-score) over all prompt configurations.

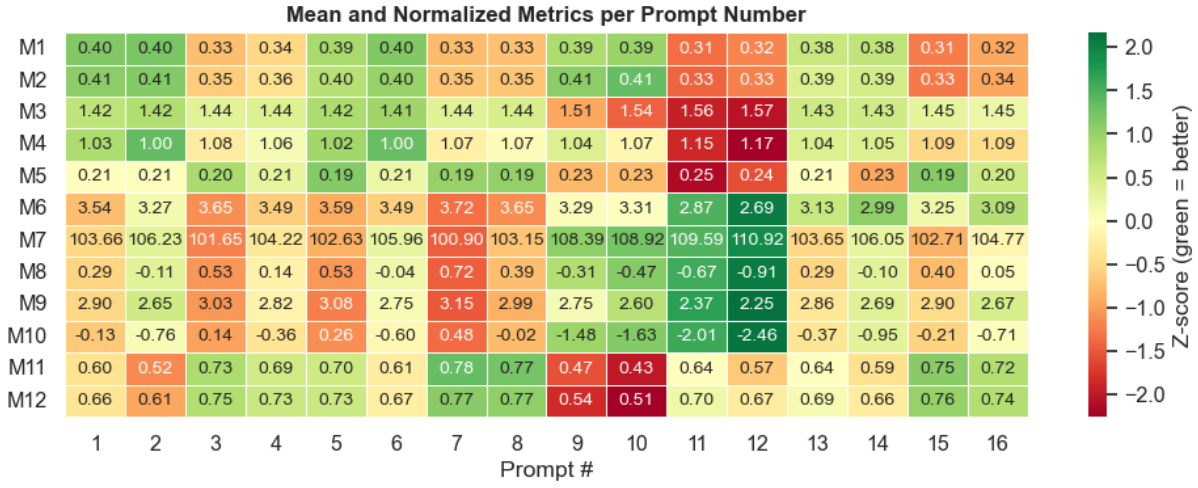

Figure 6: Performances of each specific prompt across all metrics, averaged (annotation=mean) and normalized (color=Z-score) over all prompt configurations.

E Model-Level Figures

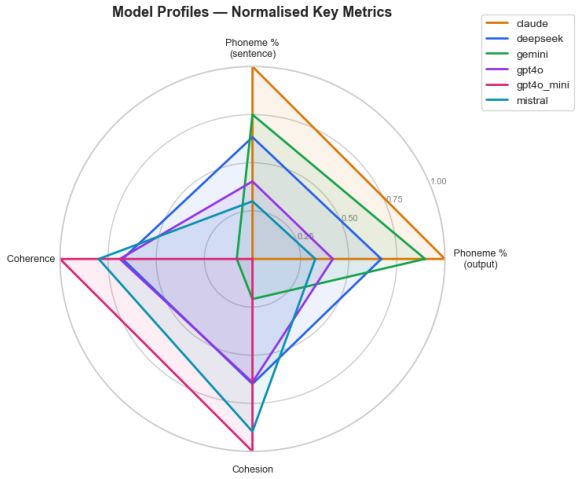

Figure 7: Model profiles for phoneme density and narrative quality show trade-offs between model performance.

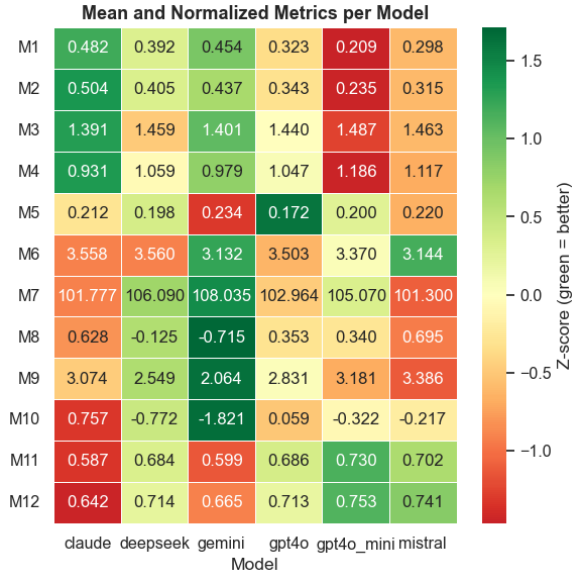

Figure 8: Performances of models across all metrics, averaged (annotation=mean) and normalized (color=Z-score) over all prompt configurations.

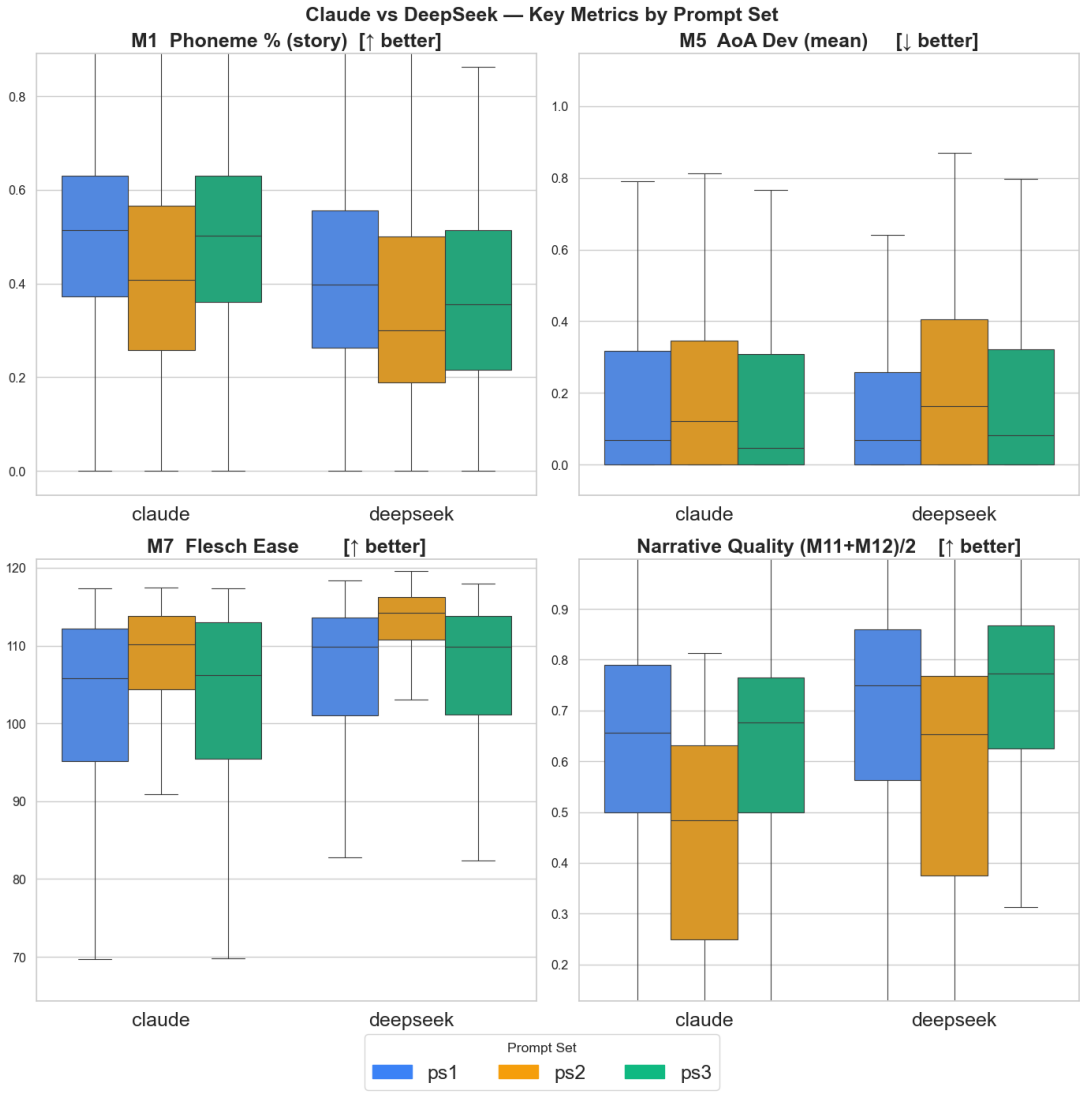

Figure 9: Performance of `claude` vs `deepseek` over metric categories using each prompt set.

F Statistical Analysis

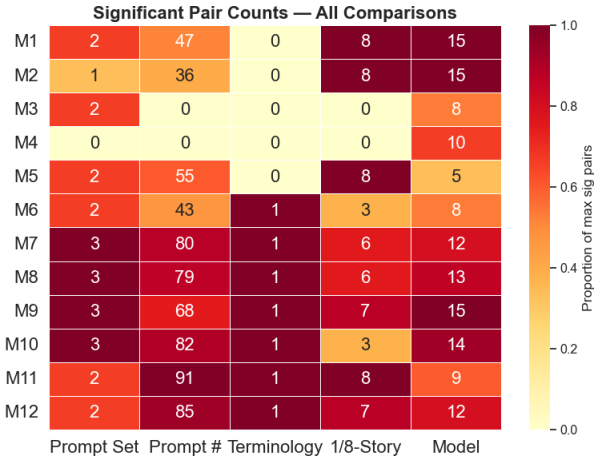

Figure 10: Count of pairs (by prompt set, prompt number, terminology, story count, and model) with significant difference ($p \leq 0.05$) across all metrics

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
1. Story-Level Phoneme Density (↑)	8688	0.360	0.222	0.000	0.184	0.346	0.522	1.000
2. Sentence-Level Phoneme Density (↑)	8688	0.373	0.219	0.000	0.202	0.367	0.538	1.000
3. Mean Phonics Level Deviation (↓)	8688	1.440	0.997	0.000	0.433	2.049	2.372	3.000
4. Max Phonics Level Deviation (↓)	8688	1.053	0.728	0.000	0.310	1.094	1.666	3.000
5. Mean AoA Deviation (↓)	8688	0.206	0.275	0.000	0.000	0.085	0.328	2.394
6. Max AoA Deviation (↓)	8688	3.378	1.637	0.000	2.254	3.245	4.380	10.920
7. Flesch Reading Ease (↑)	8688	104.206	11.721	18.377	98.841	107.487	112.760	119.867
8. Flesch-Kincaid Grade Level (↓)	8688	0.196	1.681	-2.880	-0.954	-0.143	0.985	12.129
9. Gunning Fog Index (↓)	8688	2.848	1.217	0.933	2.160	2.640	3.183	21.010
10. Automated Readability Index (↓)	8688	-0.386	2.320	-8.466	-1.855	-0.448	0.966	10.723
11. Coherence (↑)	8688	0.665	0.231	0.000	0.500	0.719	0.844	1.000
12. Cohesion (↑)	8688	0.704	0.167	0.000	0.625	0.750	0.828	1.000

Table 5: Full descriptive statistics across all generated stories. Arrows indicate whether higher (↑) or lower (↓) values are generally preferable.

G Narrative Evaluation Framework

LLM Evaluation Prompt for Narrative Quality

You are an expert in children's phonics stories for speech development in early childhood education (ages 4-6). You understand that these stories are short (exactly 4 sentences), simple, and designed for early readers.
You must evaluate the narrative quality of each story based ONLY on structural coherence and cohesion.
Do NOT evaluate creativity, vocabulary richness, or literary sophistication. Do NOT penalize simplicity. Focus only on whether each story is logically clear and well-connected.

INSTRUCTIONS:

1. Input Requirements: - The user will provide multiple numbered phonics stories, each with a title and 4 sentences.
2. Evaluation Criteria

Evaluate each story using the following criteria.

A. COHERENCE (Meaning-Level Organization) Score each criterion: 0 = Not present 1 = Partially present 2 = Clearly present

A1. Situation Setup Clarity - Does the first sentence clearly introduce who or what the story is about?
A2. Logical Event Progression - Do the sentences follow a logical order? - Does each sentence build naturally from the previous one?
A3. Causal or Motivational Connection - Are actions connected by cause, goal, or intention? - Is it clear why events happen?
A4. Resolution or Completion - Does the final sentence complete or resolve the situation?

B. COHESION (Surface-Level Connections) Score each criterion: 0 = Not present 1 = Partially present 2 = Clearly present

B1. Referential Clarity - Are pronouns clearly linked to nouns? - Is there ambiguity in who is performing actions?
B2. Lexical Consistency - Are key words repeated or semantically related? - Is the topic maintained across all sentences?
B3. Use of Connective Devices - Are simple connectors used appropriately? (e.g., and, then, so, but)
B4. Absence of Contradiction - Are there inconsistencies or sudden unexplained changes?

3. Output Format

Return ONLY a JSON array with one object per story, in the same order as the input:

```
{
"story": <story number>,
"title": "<story title>",
"coherence": {
"A1_situation_setup": <0-2>,
"A2_event_progression": <0-2>,
"A3_causal_connection": <0-2>,
"A4_resolution": <0-2>
},
"cohesion": {
"B1_referential_clarity": <0-2>,
"B2_lexical_consistency": <0-2>,
"B3_connectives": <0-2>,
"B4_no_contradiction": <0-2>
}
}
```

Figure 11: Prompt used for narrative quality scoring.

H Sample of Generated Stories

#	Target GP Mapping	Prompt-Model	Story	$M_1(\uparrow)$	$M_2(\uparrow)$	$M_3(\downarrow)$	$M_4(\downarrow)$	$M_{11}(\uparrow)$	$M_{12}(\uparrow)$
1	le→['L', 'AH L']	6-claude	**Little Turtle's Rumble** Little Turtle rumbles in a puddle. The gentle turtle nibbles on a purple apple. Noble turtle tumbles while juggling a paddle. The little turtle cuddles with his uncle.	0.677	0.715	2.142	1.000	0.375	0.625
2	le→['L', 'AH L']	6-gemini	**Kyle's Little Kettle** Kyle likes to juggle apples. Purple bubbles flew from his kettle. He tickled a beetle with his little toe. Kyle giggled, "Simple, the kettle is silly!"	0.586	0.536	2.262	1.000	0.375	0.250
3	le→['L', 'AH L']	6-deepseek	**The Little Turtle's Puzzle** The little turtle waddles. He finds a purple pebble. It sparkles like a bubble. He giggles with a chuckle.	0.522	0.540	2.264	1.304	0.375	0.625
4	le→['L', 'AH L']	6-gpt4o	**Little Mice Fable** A gentle turtle found an apple. He saw little mice by the table. They played and giggled in a circle. "Share the apple," said wise turtle.	0.379	0.414	2.069	1.034	0.750	0.875
5	le→['L', 'AH L']	6-mistral	**The Little Puddle** A little puddle lay near the apple tree. The beetle and the ant took a stroll by the little puddle. The beetle said, "Let's settle near the little puddle and have a picnic with our crumble." They had a wonderful time near the little puddle, enjoying their picnic under the apple tree.	0.309	0.371	2.315	1.618	1.000	1.000
6	le→['L', 'AH L']	6-gpt4o_mini	**The Little Blue Turtle** The little blue turtle took a stroll. She saw a lovely bubble roll and roll. "Look at that bubble!" she said with glee. Then she joined it in the deep blue sea.	0.194	0.232	2.321	1.694	1.000	1.000
7	a→['AA', 'AE', 'AH']	1-claude	**Sam and Dan's Grand Plan** Sam and Dan had a plan. A cat ran past their fast car. Dan had a snack at the stand. Pat and Matt clapped at their last act.	0.909	0.918	0.197	0.212	0.250	0.250
8	a→['AA', 'AE', 'AH']	1-deepseek	**Sam and the Cat** Sam has a black cat. The cat sat on a mat. Sam pats the fat cat. The cat naps in his lap.	0.731	0.743	0.000	0.000	0.875	0.875
9	o→['AA']	14-claude	**On the Spot** Tom got a hot dog from Mom. Ron and I stop at the pond. Rob drops the mop on the cot. The frog hops on top of a rock.	0.594	0.601	0.000	0.000	0.125	0.125
10	o→['AA']	13-deepseek	"*The Frog on the Log** The frog hops on the log. The dog spots the frog. The frog jumps off the log. The dog stops and sits.	0.333	0.333	0.007	0.037	1.000	0.875

Table 6: Sample stories from some models showcase tradeoffs between phonological alignment (metrics 1-4) and narrative quality (metrics 11-12). Arrows indicate whether higher (↑) or lower (↓) values are preferable. The full dataset of stories and scores is available at https://huggingface.co/datasets/monicamanlises/zero-shot-phonics