Persona-Based Prompting: Enhancing Readability and Understanding in AI Responses for children

Jordano de Castro[1]
Supervisor(s): Sole Pera[1] **Hrishita Chakrabarti**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Jordano de Castro
Final project course: CSE3000 Research Project
Thesis committee: Sole Pera, Hrishita Chakrabarti, Catholijn Jonker

Abstract

Large language models (LLMs) are increasingly used by children, yet their responses are often not tailored to young users' reading levels or cognitive development. Previous attempts to improve content readability through prompt modifications such as adding "for kids" have shown limited success. This project explores an alternative strategy: persona-based prompting. Rather than directly specifying the target audience, we instruct the LLM to role-play a teacher as a familiar figure to children. Using real child-authored queries, we evaluate whether this role-based approach leads to more readable and comprehensible responses across different LLMs. Readability and comprehension were measured using established metrics, including Flesch-Kincaid formulas and Age of Acquisition data. Our results show that for 4 out of the four evaluated models, persona-based prompting consistently produces responses that are more readable and accessible across all readability metrics and some comprehensibility metrics compared to standard or intended-user prompting. This finding suggests that persona-based prompting is a promising strategy for improving the suitability of LLM outputs for young audiences.

1 Introduction

Children are increasingly interacting with large language models (LLMs) to explore topics, complete schoolwork, or engage in creative inquiry [1]. However, these systems often generate responses that exceed children's reading and comprehension levels. While some models can detect the type of audience they are addressing, they do not consistently tailor their outputs to suit that audience's needs [2].

Prior attempts to improve readability have included modifying prompts like, adding "for kids" [3] but such strategies have shown limited and inconsistent success.

This project explores an alternative approach: persona-based prompting. LLMs are known to adopt roles effectively and align their responses with specified personas [4]. We hypothesize that adopting a familiar role may lead to outputs that better match children's comprehension abilities.

In Human-Computer Interaction (HCI) and Information Retrieval (IR), a persona is typically a fictional character derived from research, used to guide user-centered design [5]. In our work, we define a persona as a communicative role (e.g., teacher or caregiver) embedded in a context that sets tone, language, or situation, for example, "a friendly teacher explaining something to a 10-year-old in Dutch."

Persona-based prompting, then, refers to instructing an LLM to respond as such a persona within a specific context to influence tone, style, and accessibility.

Using real child-authored queries, we test several open-source LLMs with and without persona instructions. We evaluate the resulting outputs using standard readability and age-of-acquisition metrics to measure alignment with children's reading abilities.

Our findings show that across all models, persona-based prompting consistently improves readability and comprehensibility. This suggests that role-based prompting is a robust, effective method for tailoring LLM responses to child audiences.

2 Methodology

This section describes the key components of our study design, including the data sources, the strategies used for data collection, the metrics applied for evaluation, and the statistical methods used for analysis.

2.1 Data

To center our study around the real information needs of children, we use a dataset of real search queries collected by Madrazo-Azpiazu et al. [6]. This dataset consists of 301 genuine queries submitted by children aged 6 to 13, making it well-suited for studying how language models can better respond to child-authored prompts in realistic settings.

2.2 Strategy

In our experimental setup, we use three types of prompting strategies:

- Persona-Based Prompting: We will instruct the LLM to adopt a role (such as a parent or teacher) tailored to the specific age category. This method involves guiding the model to generate content as if speaking from the perspective of someone who is best suited to explain the concept to children.

  prompt:

  > "You are a friendly teacher explaining this to a child between the ages of 6 and 13. Please answer the following question in a clear, simple, and engaging way: `«INSERT QUERY»`"

- Intended User Prompting: This strategy involves specifying the target demographic (children within the selected age groups), without instructing the LLM to assume any particular role. This method will allow us to compare how well the LLM performs when informed only of the intended user, in contrast to the persona-based prompting approach. Essentially, We are replicating the approach used in Rooein et al. [3], but with newer models to ensure a fair assessment of the improvements in LLM capabilities.

  prompt:

  > "`«INSERT QUERY»` For children aged 6-13"

- Standard Prompting: In this approach, we will prompt the LLM with the standard query without additions, allowing us to compare how the LLMs perform when no audience specific information is provided.

In this study, we aim to evaluate the performance of several state-of-the-art open-source language models. While Rooein et al. [3] laid the foundation for investigating how LLMs respond to demographic-specific prompts, we build on their work by utilizing newer, more advanced models for a fair assessment of the improvements in LLM performance. The selected models represent a diverse range of architectures and sizes, reflecting recent developments in open-source LLM research and providing a meaningful comparison across different design approaches. Additionally, all three models are publicly accessible and widely used, making our findings relevant for both researchers and practitioners. The models we'll use in this study are:

- DeepSeek-R1
- Qwen2.5
- Mistral 7B

2.3 Metrics

To evaluate the readability and age-appropriateness of model outputs, we selected metrics that reflect both surface-level readability and deeper linguistic accessibility. These measures help determine how well responses align with children's language development and comprehension abilities.

A key measure is Age of Acquisition (AoA) ratings which indicate the typical age at which a word is learned using the ratings from Brysbaert et al. [7]. This allows us to go beyond general readability and directly assess vocabulary suitability for a younger audience.

The following metrics were used:

- **Flesch-Kincaid Reading Ease (FKRE)**: Widely used to assess how easy a text is to understand. Higher scores indicate more readable language.
- **Gunning Fog Index**: Estimates the years of formal education needed to understand the text. Complements FKRE by emphasizing word complexity.
- **AoA - Average**: Computes the mean AoA of all non-stopwords in a response. Lower values indicate language typically learned at younger ages.
- **AoA - Threshold**: Calculates the proportion of words with AoA ratings above 13, highlighting the presence of potentially advanced vocabulary.

These metrics were chosen to provide both a broad and targeted view of how accessible LLM outputs are for children aged 6-13.

2.4 Research Question

Can persona-based prompting improve the readability and comprehension of large language model responses for children?

2.5 Statistical Analysis

To evaluate differences in readability metrics across the three prompting strategies (Persona-Based, Intended-User, and Standard), we employed a one-way Analysis of Variance (ANOVA) test. ANOVA was chosen because it allows for the comparison of means across multiple independent groups to determine whether at least one group differs significantly from the others. This test is appropriate for our study since we are analyzing continuous readability scores (e.g., Flesch-Kincaid Reading Ease, Gunning Fog Index) across three categorical prompting methods.

A significance level of $p < 0.05$ was used to determine statistical significance. After getting the ANOVA result, we conducted pairwise comparisons using Tukeyâs Honestly Significant Difference (HSD) test. Tukey HSD controls the family-wise error rate and allows us to identify which specific prompting strategies differ significantly from each other. This two-step approach ensures robust and interpretable statistical conclusions regarding the impact of prompting strategy on readability outcomes.

3 Experimental Setup

The 301 search queries from Madrazo-Azpiazu et al. [6] were transformed using the three prompting strategies. Each transformed prompt was input into the language models to generate responses.

For each prompting strategy, we calculated average scores over all responses to compare performance across strategies. Additionally, we evaluated each prompting strategy separately for each language model to determine which strategies performed best with each model.

This setup allowed us to draw conclusions about the relative effectiveness of prompting strategies for different LLMs.

4 Results

This section presents the outcomes of our experiments, comparing the performance of different prompting strategies across the evaluated language models. We analyze the results to identify patterns and differences in how each model responds to the various prompt types.

4.1 Mistral 7B

Figure 1 presents FKRE scores for each prompting strategy. Higher scores indicate greater ease of reading. Persona-Based Prompting led to the most readable outputs, followed by Intended-User Prompting, with Standard Prompting yielding the least readable content.

The statistical analysis confirms these observations. A one-way ANOVA found significant differences in FKRE scores across the three prompting strategies ($F = 166.72$, $p < 0.05$). Post-hoc pairwise comparisons using Tukey's HSD test (Table 2) indicate that Persona-Based prompting is significantly more readable than both Intended-User and Standard prompting, while no significant difference was found between Intended-User and Standard prompting.

Table 1: Pairwise comparisons of FKRE scores between prompting strategies using Tukey's HSD test for the Mistral model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	18.60	< 0.05	15.89	21.31
Intended-User vs Standard	0.73	0.80	-1.98	3.44
Persona-Based vs Standard	-17.87	< 0.05	-20.58	-15.16

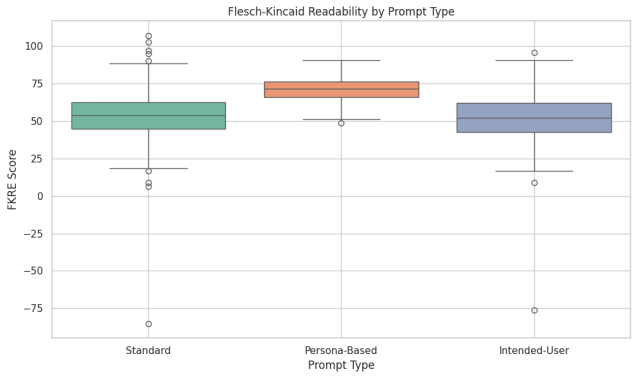

Figure 1: Flesch-Kincaid Reading Ease scores across prompting strategies for the Mistral model

Figure 2 presents Gunning Fog Index scores across the three prompting strategies. Lower Fog scores indicate simpler, more accessible language. Persona-Based Prompting consistently produced the lowest Fog scores, suggesting it most effectively made the model's response more readable. Intended-User Prompting yielded intermediate scores, while Standard Prompting resulted in the highest scores, reflecting the most complex output.

Statistical analysis supports these findings. A one-way ANOVA analysis confirmed a significant effect of prompting strategy on Fog scores ($F = 186.69$, $p < 0.05$). Post-hoc Tukey's HSD tests show that Persona-Based prompting significantly reduces Fog scores compared to both Intended-User (mean difference = -4.356, $p < 0.05$) and Standard prompting (mean difference = 4.150, $p < 0.05$), whereas the difference between Intended-User and Standard prompting is not statistically significant (mean difference = -0.206, $p = 0.696$).

Table 2: Pairwise comparisons of Gunning Fog Index scores between prompting strategies using Tukey's HSD test for the Mistral model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-4.3560	< 0.05	-4.9532	-3.7589
Intended-User vs Standard	-0.2064	0.6961	-0.8035	0.3908
Persona-Based vs Standard	4.1496	< 0.05	3.5525	4.7468

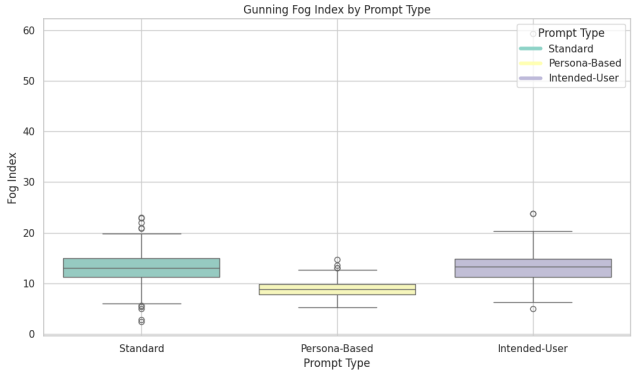

Figure 2: Gunning Fog across prompting strategies for the Mistral model

Figure 3 presents the average Age of Acquisition (AoA) rating for words used in responses across each prompting strategy. Lower AoA ratings indicate that the words are typically learned at a younger age, reflecting simpler and more accessible vocabulary. Persona-Based Prompting resulted in responses with the lowest average AoA rating, followed by Intended-User Prompting, while Standard Prompting produced outputs with the highest AoA values. This pattern closely matches the results observed for both the FKRE and Fog Index, further supporting the effectiveness of persona-based instructions in generating language that is more appropriate for children.

Importantly, the average AoA ratings for all strategies were below 13, indicating that the vocabulary used in all cases is, in principle, accessible to children. However, the Persona-Based Prompting strategy produced responses with even lower AoA values, making the content especially suitable for even younger audiences.

Statistical analysis supports these findings: a one-way ANOVA revealed significant differences in average AoA across the three prompting strategies ($F = 417.4060$, $p < 0.05$). Post-hoc pairwise comparisons using Tukey's HSD test (Table 3) show that Persona-Based Prompting yields significantly lower AoA values than both Intended-User and Standard Prompting, and that Intended-User Prompting also results in significantly lower AoA values compared to Standard Prompting. This indicates that both Persona-Based and Intended-User Prompting improve the accessibility of model outputs over the approach of just sending the query without adjustments.

Table 3: Pairwise comparisons of average Age of Acquisition ratings between prompting strategies using Tukey's HSD test for the Mistral model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-0.9632	< 0.05	-1.0774	-0.849
Intended-User vs Standard	0.4053	< 0.05	0.2911	0.5196
Persona-Based vs Standard	1.3685	< 0.05	1.2543	1.4828

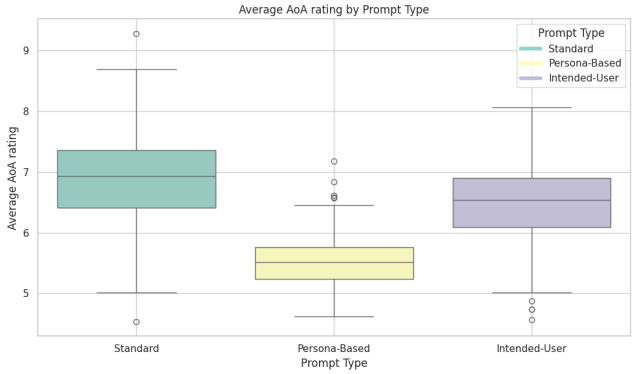

Figure 3: average Age of Acquisition ratings across prompting strategies for the Mistral model

Figure 4 shows the proportion of words in model responses with an Age of Acquisition (AoA) rating above 13 for each prompting strategy. A lower proportion indicates that fewer words are likely to be unfamiliar or advanced for children. Persona-Based Prompting resulted in the lowest proportion of words above this threshold, followed by Intended-User Prompting, with Standard Prompting yielding the highest proportion. again supporting the effectiveness of persona-based instructions in generating language that is more appropriate for children.

Statistical analysis confirms these differences: a one-way ANOVA revealed significant variation in the proportion of words above AoA 13 across prompting strategies ($F = 30.3678$, $p < 0.05$). Post-hoc pairwise comparisons using Tukey's HSD test (Table 4) show that Persona-Based Prompting yields a significantly lower proportion than both Intended-User and Standard Prompting, and Intended-User Prompting also results in a significantly lower proportion compared to Standard Prompting. This demonstrates that both Persona-Based and Intended-User Prompting improve the accessibility of model outputs compared to unmodified queries.

Table 4: Pairwise comparisons of proportion AoA ratings above 13 between prompting strategies using Tukey's HSD test for the Mistral model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-0.0043	< 0.05	-0.0072	-0.0014
Intended-User vs Standard	0.0053	< 0.05	0.0024	0.0082
Persona-Based vs Standard	0.0096	< 0.05	0.0067	0.0125

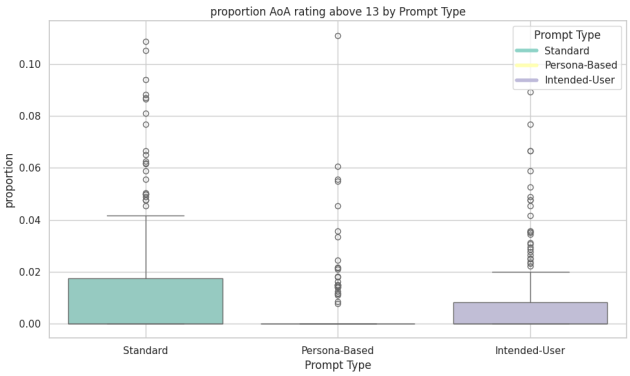

Figure 4: proportion AoA ratings above 13 across prompting strategies for the Mistral model

4.2 DeepSeek-R1

Figure 5 shows the Flesch-Kincaid Reading Ease (FKRE) scores for each prompting strategy with the DeepSeek model. For DeepSeek, Persona-Based Prompting produced the most readable outputs, with significantly higher FKRE scores than both Intended-User and Standard Prompting. Intended-User Prompting did not result in a significant improvement over Standard Prompting, as their FKRE scores were not significantly different. This pattern suggests that, for DeepSeek, only the persona-based approach reliably enhances the readability of responses for children.

The statistical analysis supports these observations. A one-way ANOVA revealed significant differences in FKRE scores across the three prompting strategies ($F = 125.8466$, $p < 0.05$). Post-hoc pairwise comparisons using Tukey's HSD test (Table 5) show that Persona-Based Prompting yields significantly higher FKRE scores than both Intended-User and Standard Prompting, while the difference between Intended-User and Standard Prompting is not statistically significant.

Table 5: Pairwise comparisons of FKRE scores between prompting strategies using Tukey's HSD test for the DeepSeek model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	12.2515	< 0.05	10.0898	14.4132
Intended-User vs Standard	-0.7649	0.684	-2.9266	1.3968
Persona-Based vs Standard	-13.0165	< 0.05	-15.1781	-10.8548

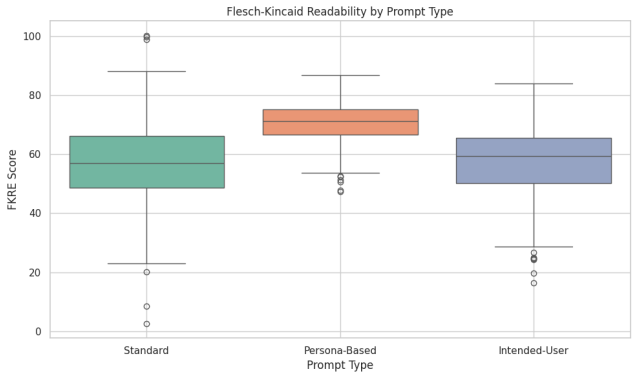

Figure 5: FKRE scores across prompting strategies for the DeepSeek model

Figure 6 displays the Gunning Fog Index scores for each prompting strategy using the DeepSeek model. For DeepSeek, Persona-Based Prompting produced responses with the lowest Fog Index values, indicating the highest readability among the strategies. Intended-User Prompting also resulted in lower Fog Index scores than Standard Prompting. Persona-based instructions led to the simplest language, followed by intended-user, with standard prompts producing the most complex outputs.

Statistical analysis reinforces these findings. A one-way ANOVA showed significant differences in Fog Index scores across all three strategies ($F = 114.1349$, $p < 0.05$). Tukey's HSD post-hoc comparisons (Table 6) confirm that each strategy is significantly different from the others: Persona-Based Prompting yields significantly lower Fog Index values than Intended-User Prompting, which in turn is significantly lower than Standard Prompting. This demonstrates that both persona-based and intended-user instructions improve the readability of DeepSeek's responses, with the greatest benefit seen from persona-based prompting

Table 6: Pairwise comparisons of Gunning Fog index between prompting strategies using Tukey's HSD test for the DeepSeek model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-2.2615	< 0.05	-2.7141	-1.8089
Intended-User vs Standard	0.4591	< 0.05	0.0065	0.9117
Persona-Based vs Standard	2.7206	< 0.05	2.268	3.1732

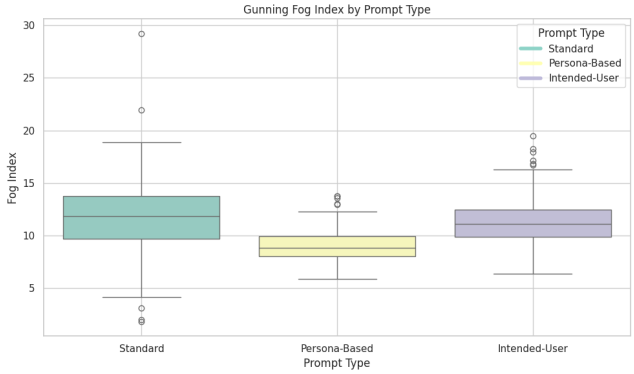

Figure 6: Gunning fog index across prompting strategies for the DeepSeek model

Figure 7 illustrates the average Age of Acquisition (AoA) for words used in DeepSeek's responses under each prompting strategy. Among the three strategies, Persona-Based Prompting consistently resulted in the lowest average AoA, indicating the simplest and most age-appropriate language. Intended-User Prompting produced the second lowest AoA values, while Standard Prompting led to the highest AoA values.

It is important to note, again, that all strategies produced responses with average AoA values below 13, meaning that the outputs are technically age-appropriate for children. However, Persona-Based Prompting stands out by generating language that is especially suitable for even younger audiences.

These differences were statistically significant, as confirmed by a one-way ANOVA ($F = 277.5818$, $p < 0.05$). Tukey's HSD post-hoc analysis (Table 7) revealed that Persona-Based Prompting produced significantly lower AoA scores than Intended-User Prompting, which in turn was significantly lower than Standard Prompting.

Table 7: Pairwise comparisons of Average AoA ratings between prompting strategies using Tukey's HSD test for the DeepSeek model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-0.5712	< 0.05	-0.6759	-0.4664
Intended-User vs Standard	0.4789	< 0.05	0.3741	0.5836
Persona-Based vs Standard	1.05	< 0.05	0.9453	1.1548

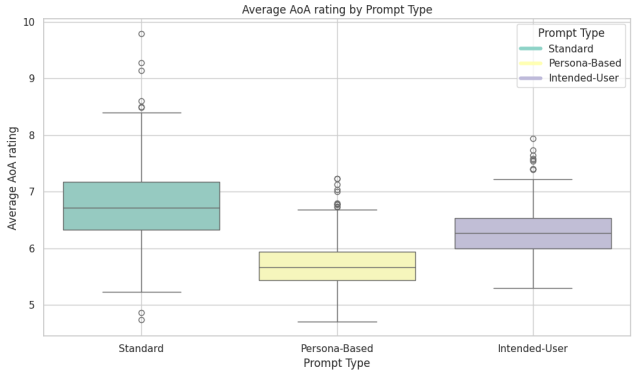

Figure 7: Average AoA ratings across prompting strategies for the DeepSeek model

Figure 8 presents the proportion of words in DeepSeek's responses with an Age of Acquisition (AoA) rating higher than 13 for each prompting strategy. Both Persona-Based and Intended-User Prompting resulted in a significantly lower proportion of difficult words compared to Standard Prompting. However, there was no significant difference between Persona-Based and Intended-User strategies.

Statistical analysis confirms these observations. A one-way ANOVA revealed significant differences in the proportion of words above the AoA 13 threshold across strategies ($F = 26.4475$, $p < 0.05$). Tukey's HSD post-hoc tests (Table 8) showed that both Persona-Based and Intended-User Prompting produced significantly lower proportions than Standard Prompting, but the difference between Persona-Based and Intended-User was not statistically significant. Both specifying the intended user and using a persona help make the response easier by lowering the number of complex words, bot one is not better then the other.

Table 8: proportion AoA ratings above 13 across prompting strategies for the Deepseek model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-0.0013	0.4618	-0.004	0.0013
Intended-User vs Standard	0.0064	< 0.05	0.0037	0.009
Persona-Based vs Standard	0.0077	< 0.05	0.0051	0.0104

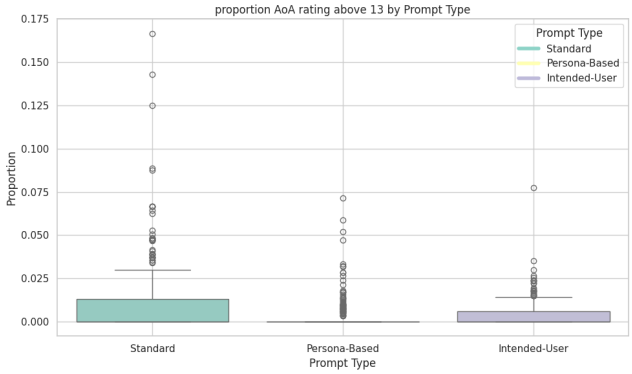

Figure 8: proportion AoA ratings above 13 across prompting strategies for the Deepseek model

4.3 Qwen2.5

Figure 9 illustrates the Flesch-Kincaid Reading Ease (FKRE) scores for Qwen under each prompting strategy. Persona-Based Prompting led to the highest FKRE scores, suggesting that this strategy most effectively improved readability. Intended-User Prompting showed slightly higher scores than the Standard condition, but the difference was not statistically meaningful.

A one-way ANOVA confirmed that prompting strategy had a significant effect on FKRE scores ($F = 377.6226$, $p < 0.05$). Tukey's HSD test (Table 9) revealed that Persona-Based Prompting significantly increased FKRE scores compared to both Intended-User and Standard Prompting. However, the difference between Intended-User and Standard Prompting was not statistically significant. This suggests that, for Qwen, only the persona-based strategy consistently improved readability outcomes.

Table 9: Pairwise comparisons of FKRE scores between prompting strategies using Tukey's HSD test for the Qwen

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	22.5133	< 0.05	20.28781	24.7388
Intended-User vs Standard	-0.0971	0.9942	-2.3226	-2.3226
Persona-Based vs Standard	-22.6104	< 0.05	-24.8359	-20.3849

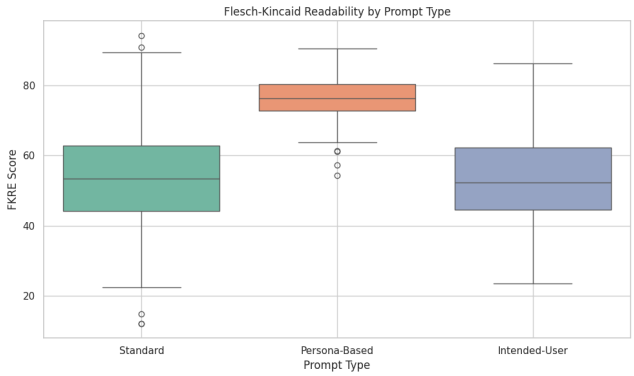

Figure 9: FKRE scores across prompting strategies for the Qwen model

Figure 10 displays the Gunning Fog Index results for the Qwen model under each of the three prompting strategies. Among the three strategies, Persona-Based Prompting led to the lowest Fog Index values, indicating that this approach produced the most accessible language. Intended-User Prompting fell between the two extremes, while Standard Prompting resulted in the highest scores, suggesting more complex and potentially less suitable output for a child audience.

The statistical analysis reinforces these observations. A one-way ANOVA revealed a significant main effect of prompting strategy on Fog scores ($F = 333.4136$, $p < 0.05$). Tuke's HSD test identified significant differences between Persona-Based and both other strategies. Specifically, Persona-Based prompting yielded significantly lower scores than Intended-User (mean difference = -4.6658, $p < 0.05$) and Standard prompting (mean difference = 4.7677, $p < 0.05$). The difference between Intended-User and Standard prompting was not statistically significant (mean difference = 0.1019, $p = 0.8793$).

Table 10: Pairwise comparisons of Gunning Fog index between prompting strategies using Tukey's HSD test for the Qwen model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-4.6658	< 0.05	-5.161	-4.1705
Intended-User vs Standard	0.1019	0.8793	-0.3933	0.5971
Persona-Based vs Standard	4.7677	< 0.05	4.2725	5.2629

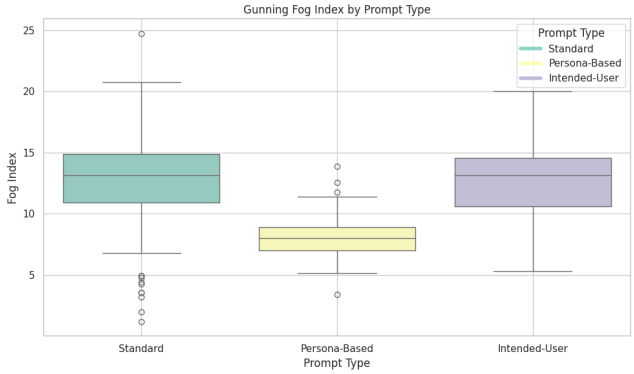

Figure 10: Gunning fog index across prompting strategies for the Qwen model

Figure 11 shows the average Age of Acquisition (AoA) scores for the words used in Qwen's responses across the three prompting strategies. As with the other models, Persona-Based Prompting resulted in the lowest average AoA, reflecting the use of simpler, more age-appropriate language. Intended-User Prompting produced slightly lower AoA values than Standard Prompting, which had the highest scores overall.

Similar to the trends observed in the Mistral and DeepSeek models, all three strategies yielded outputs with average AoA values below 13, indicating that the language used is broadly appropriate for children aged 6-13. However, the Persona-Based strategy once again distinguished itself by producing the most developmentally accessible responses, likely to be understood by even younger children within that range.

A one-way ANOVA confirmed a statistically significant effect of prompting strategy on AoA scores ($F = 854.8247$, $p < 0.05$). Post-hoc analysis using Tukey's HSD test (Table 11) showed that all pairwise differences were statistically significant: Persona-Based Prompting produced significantly lower AoA scores than Intended-User, which in turn outperformed Standard Prompting.

Table 11: Pairwise comparisons of Average AoA ratings between prompting strategies using Tukey's HSD test for the Qwen model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-1.3645	< 0.05	-1.4731	-1.256
Intended-User vs Standard	0.4855	< 0.05	0.3764	0.5946
Persona-Based vs Standard	1.85	< 0.05	1.7409	1.9591

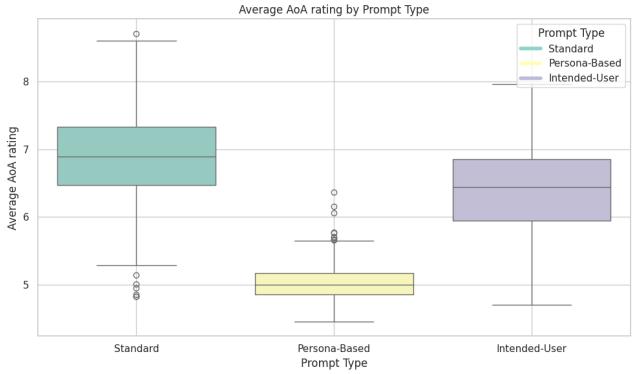

Figure 11: Average AoA ratings across prompting strategies for the Qwen model

Figure 12 presents the proportion of words in Qwen's responses that have an Age of Acquisition (AoA) rating above 13 across the three prompting strategies. Among the strategies tested, Persona-Based Prompting consistently resulted in the lowest proportion of such words, indicating the most accessible vocabulary. Intended-User Prompting also reduced the proportion compared to Standard Prompting, though to a lesser extent.

Statistical analysis supports this trend. A one-way ANOVA showed a significant effect of prompting strategy on the proportion of high-AoA words ($F = 42.8664$, $p < 0.05$). Tukey's HSD test (Table 12) revealed that Persona-Based Prompting led to a significantly lower proportion of words above the AoA threshold compared to both Intended-User and Standard Prompting. Additionally, Intended-User Prompting also produced significantly fewer high-AoA words than Standard Prompting. These findings suggest that both targeted prompting strategies contribute to improving the linguistic accessibility of Qwen's outputs, with Persona-Based Prompting showing the strongest effect.

Table 12: Pairwise comparisons of proportion AoA ratings above 13 between prompting strategies using Tukey's HSD test for the Qwen model

Comparison	Mean Difference	p-value	95% CI Lower	95% CI Upper
Intended-User vs Persona-Based	-0.0038	< 0.05	-0.0061	-0.0015
Intended-User vs Standard	0.0053	< 0.05	0.003	0.0076
Persona-Based vs Standard	0.0091	< 0.05	0.0068	0.0114

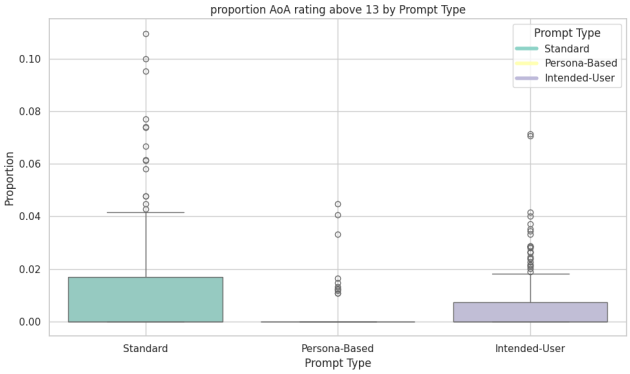

Figure 12: proportion AoA ratings above 13 across prompting strategies for the Mistral model

5 Responsible Research

To support transparency, reproducibility, and responsible research practices, all CSV files used for the calculation of readability and Age of Acquisition metrics have been included in the corresponding project repository. These files contain the model outputs, associated prompt variants, and computed scores, allowing others to inspect, verify, and build upon our analysis. By making these materials publicly available, we aim to contribute to an open research culture and facilitate further work in evaluating language models for child-directed communication. This paper was not written with the use of generative AI. However, grammar assistance tools were used to improve the clarity and correctness of the text.

6 Discussion & Future Work

As AI systems increasingly participate in collaborative tasks, it is essential to design them for effective interaction. The Observability, Predictability, and Directability (OPD) framework by Johnson et al. [8] highlights these principles. Our study focused on enhancing directability and guiding models to generate responses aligned with readability goals.

Although the tested models already produced reasonably readable outputs, persona-based prompting consistently improved results across all metrics and models. This shows that prompt engineering can effectively steer LLMs toward more age-appropriate language, making them more suitable for child-facing applications.

6.1 Limitations

Despite improvements in consistency, we observed one case where the model responded in Greek without being prompted, suggesting that persona and context instructions may need to be more explicit for robust control.

We defined persona as a communicative role combined with a fixed context (e.g., tone or language). For the purpose of this experiment, we pre-defined a single, reasonable context to isolate the effect of persona-based prompting. Future work could explore how varying context influences model behavior.

Additionally, the 301 child-authored queries used in this study were collected for search engine use [6], which may not fully reflect how children interact with LLMs. Future research should gather LLM-specific child inputs to better align with natural use cases.

7 Conclusion

This study investigated the effectiveness of different prompting strategies, Standard, Intended-User, and Persona-Based, on improving the readability and accessibility of responses generated by LLMs for child audiences. We evaluated these strategies across three models: Mistral 7B, Qwen 2.5, and DeepSeek-R1, using four complementary metrics: Flesch-Kincaid Reading Ease (FKRE), Gunning Fog Index, average Age of Acquisition, and the proportion of words with an AoA rating above 13.

Across all models, Persona-Based Prompting consistently outperformed both Intended-User and Standard Prompting for FKRE, Gunning Fog Index, and average AoA scores. For the proportion of words with an AoA rating above 13, both Mistral and Qwen showed significant improvements with Persona-Based Prompting compared to the other strategies. However, DeepSeek was an exception: while Persona-Based Prompting reduced the proportion of developmentally advanced words relative to Standard Prompting, it did not yield a statistically significant improvement over Intended-User Prompting. This suggests that the impact of prompting strategies may vary slightly across model architectures.

Overall, these results support the conclusion that persona-based prompting is an effective and scalable technique for enhancing the accessibility of LLM outputs for children, though further refinement may be needed to maximize its effectiveness across all models and metrics.

References

[1] S. Quan, Y. Du, and Y. Ding, "Young children and chatgpt: Parents' use of chatgpt in parenting," *ResearchGate*, 2024.

[2] Y. Choi, E. J. Kang, S. Choi, M. K. Lee, and J. Kim, "Proxona: Supporting creators' sensemaking and ideation with llm-powered audience personas," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, (New York, NY, USA), Association for Computing Machinery, 2025.

[3] D. Rooein, A. Cercas Curry, and D. Hovy, "Know your audience: Do llms adapt to different age and education levels?," *arXiv preprint arXiv:2312.02065*, 2023.

[4] N. Chen, Y. Deng, and J. Li, "The oscars of ai theater: A survey on role-playing with language models," *arXiv preprint arXiv:2407.11484v4*, 2024.

[5] R. F. Dam and Y. S. Teo, "Personas â a simple introduction," 2025. Accessed: 2025-06-21.

[6] I. Madrazo Azpiazu, N. Dragovic, O. Anuyah, and M. S. Pera, "Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, (New York, NY, USA), p. 92â101, Association for Computing Machinery, 2018.

[7] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior Research Methods*, vol. 46, no. 3, pp. 904–911, 2014.

[8] M. Johnson, J. Bradshaw, P. J. Feltovich, C. Jonker, M. Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Human Robot-Interaction*, vol. 3(1), 03 2014.