CLARA: Can AI Assess Developmental Appropriateness in Children's Stories?

Sijing Yin[1], Zirui Wang[1], Qian Liu[1], Jiamou Liu[1]*

[1]University of Auckland

syin565@aucklanduni.ac.nz, jiamou.liu@auckland.ac.nz

Abstract

Assessing the developmental suitability of children's narratives is important for educational recommendation and developmental literacy research, yet such assessment typically relies on subjective and difficult-to-scale human judgment. This raises an important question: *Can AI systems approximate human developmental judgments of children's stories?* To study this problem, we introduce CLARA, a cognitively grounded framework for developmental narrative understanding through structured annotation across cognitive (COG), language (LAN), and social-emotional (SEL) dimensions, together with a bilingual benchmark resource containing 1107 Chinese–English children's stories with normalized silver developmental references and structured developmental annotations. We evaluate CLARA through benchmark comparison, component analysis, translated bilingual consistency analysis, and blinded human evaluation with educators. Experimental results show that structured developmental annotation achieves substantially stronger alignment with developmental references and human judgments than readability-based methods and direct prompting baselines. Overall, our findings suggest that AI systems can approximate certain aspects of human developmental judgment when guided by structured developmental annotation, while also highlighting the importance of interpretability and human oversight in educational NLP.

1 Introduction

Assessing the developmental suitability of children's narratives is important for educational recommendation, curriculum design, and developmental literacy research. In practice, however, such assessment typically relies on human judgment from educators, parents, or publishers, which is often subjective, culturally dependent, and difficult to scale (Gilardi et al., 2023). Existing computational approaches mainly estimate reading difficulty through readability formulas or direct age prediction, focusing on lexical and syntactic complexity rather than the developmental abilities required for narrative understanding (Flesch, 1948; Kincaid et al., 1975; Dale and Chall, 1948). Yet understanding children's stories often requires substantially more than surface language processing, including causal reasoning, temporal understanding, emotional interpretation, social reasoning, and perspective taking (Piaget, 1952; Vygotsky, 1978; Astington et al., 1988; Wellman and Liu, 2004). Consequently, two stories with similar readability may still require very different levels of developmental understanding.

This raises an important question: *Can AI systems approximate human developmental judgments of children's stories?* More specifically, can language models identify the cognitive, linguistic, and social-emotional abilities required for understanding narratives in a way that aligns with human educational reasoning? Addressing this problem requires not only predictive models, but also structured developmental representations, interpretable annotation spaces, and cognitively grounded narrative analysis frameworks capable of modeling developmental narrative understanding beyond readability alone (Rashkin et al., 2018; Mostafazadeh et al., 2020; van Dijk et al., 2023).

To study this problem, we introduce CLARA (**C**ognitive **L**abeling and **A**ge **R**easoning for N**A**rratives), a cognitively grounded framework for developmental narrative understanding through structured annotation. Rather than directly predicting developmental stages from stories, CLARA decomposes developmental understanding into interpretable developmental representations across three dimensions: cognitive skills (COG), language abilities (LAN), and social-emotional abilities (SEL) (Wei et al., 2022; Wang et al., 2023;

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 5390–5403
October 24-29, 2026 ©2026 Association for Computational Linguistics

Zhou et al., 2023). The framework consists of four components: (1) a developmental annotation taxonomy, (2) a prompt-based annotation protocol, (3) a lightweight developmental aggregation mechanism, and (4) a bilingual benchmark resource for evaluation. Importantly, the current implementation uses GPT-4o as one possible annotation engine rather than a required component of the framework itself.

To support systematic evaluation, we construct a bilingual benchmark containing 1107 Chinese–English children's stories with normalized silver developmental references and structured developmental annotations. Here, "bilingual" refers to aligned Chinese–English versions of the same stories, where the English narratives are translated from the Chinese sources. Accordingly, the bilingual analysis evaluates consistency under translation rather than cross-lingual or cross-cultural generalization. Existing narrative understanding resources mainly focus on story comprehension or commonsense reasoning rather than developmental cognitive representations (Xu et al., 2022; Tikhonov et al., 2021; Mostafazadeh et al., 2016). Experimental results show that CLARA substantially outperforms readability-based and direct prompting baselines, improving overlap accuracy from 0.603 (FKGL) and 0.730 (GPT-4o Direct) to 0.904 while reducing mean developmental distance to 0.096. Human evaluation further shows 0.81 agreement with educator developmental rankings and an interpretability score of 4.42/5. These findings suggest that developmental narrative understanding cannot be captured through surface linguistic complexity alone, but instead requires explicit modeling of cognitive, linguistic, and social-emotional developmental signals (Rashkin et al., 2018; Mostafazadeh et al., 2020).

Overall, our findings suggest that AI systems can approximate certain aspects of human developmental judgment when guided by structured developmental annotation, while also highlighting the importance of interpretability and human oversight in educational NLP (Gilardi et al., 2023).

Contributions.

- We formulate developmental narrative understanding as a structured and interpretable NLP problem centered on approximating human developmental judgments of children's stories.
- We introduce CLARA, a cognitively grounded framework that models developmental understanding through structured annotations across cognitive, language, and social-emotional dimensions.
- We construct a bilingual benchmark resource containing 1107 Chinese–English children's stories with normalized silver developmental references and structured developmental annotations.
- We conduct benchmark comparison, component analysis, translated bilingual consistency analysis, and blinded human evaluation, showing that structured developmental annotation aligns more closely with developmental references and human judgments than readability-based and direct prompting approaches.

2 Related Work

2.1 Readability and Developmental Assessment

Traditional approaches to reading difficulty and age estimation mainly rely on readability formulas (Flesch, 1948; Kincaid et al., 1975; Dale and Chall, 1948). Recent work further applies machine learning and neural language models to text readability and educational assessment (Martinc et al., 2021; Gilardi et al., 2023). However, these approaches mainly model surface linguistic complexity and provide limited developmental interpretability.

2.2 Developmental Narrative Understanding

Developmental psychology studies how children's cognitive, linguistic, and social-emotional abilities evolve over time (Piaget, 1952; Vygotsky, 1978; Collaborative for Academic, Social, and Emotional Learning, 2012). Prior work on children's narratives further highlights the importance of causal reasoning, perspective taking, emotional interpretation, narrative coherence, and discourse organization in developmental language understanding (Astington et al., 1988; Wellman and Liu, 2004; Reese et al., 2011; Gagarina et al., 2019; Gagarina and Bohnacker, 2022). Research on bilingual and developmental narratives additionally studies narrative macrostructure, referential coherence, and discourse-level developmental progression across children with different linguistic backgrounds (Yang et al., 2022; Kupersmitt and Armon-Lotem, 2019). Recent NLP-based studies further explore theory-of-mind reasoning and developmental signals in children's narratives (van

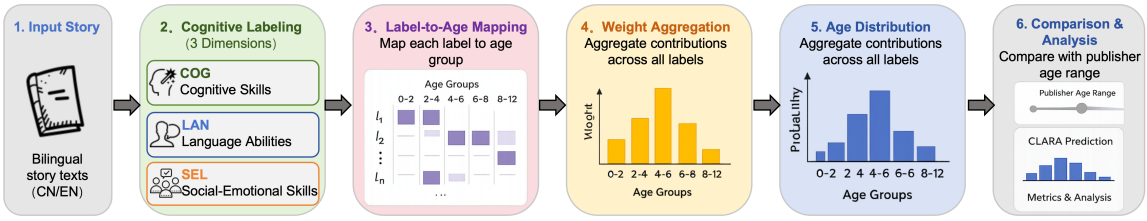

Figure 1: Overview of the CLARA framework. Given bilingual story narratives, CLARA generates structured developmental annotations across COG, LAN, and SEL dimensions, aggregates developmental evidence into developmental-stage estimates, and compares the resulting representations with human and publisher developmental judgments.

Dijk et al., 2023). However, these developmental frameworks are rarely operationalized into scalable computational annotation systems for NLP.

2.3 Structured Story Understanding

Narrative understanding has been widely studied in NLP, including story comprehension, commonsense reasoning, character-centered inference, and structured narrative explanation (Mostafazadeh et al., 2016; Rashkin et al., 2018; Sap et al., 2019; Mostafazadeh et al., 2020). Several works further introduce structured cognitive annotations for stories, such as Story Commonsense (Rashkin et al., 2018) and GLUCOSE (Mostafazadeh et al., 2020). Educational narrative resources such as FairytaleQA (Xu et al., 2022) and StoryDB (Tikhonov et al., 2021) additionally connect narrative understanding with educational settings. Unlike prior work, CLARA studies whether AI systems can approximate human developmental judgments of children's narratives through structured cognitive, linguistic, and social-emotional annotation.

3 Developmental Annotation Framework

CLARA studies whether AI systems can approximate human developmental judgments of children's narratives through structured developmental annotation. Rather than directly predicting developmental stages, CLARA represents narratives through COG, LAN, and SEL developmental signals.

CLARA consists of four components: (1) a developmental taxonomy, (2) a prompt-based annotation protocol, (3) a lightweight aggregation mechanism, and (4) a bilingual benchmark resource. GPT-4o is used as one possible annotation engine rather than a required component of the framework. The framework is model-agnostic and can be instantiated with different annotation models under the same developmental taxonomy and prompting protocol. Importantly, the taxonomy defines the allowable developmental label space and associated age groups, but does not determine which labels are activated for a given story. The open-ended inference step is performed by the annotation model, which identifies the cognitive, linguistic, and social-emotional abilities required for narrative comprehension.

3.1 Problem Formulation

Given a narrative text S, CLARA derives a developmental label set $\mathcal{L}(S)$, where each label belongs to one of three dimensions:

$$\text{COG}, \ \text{LAN}, \ \text{SEL}.$$

Based on these annotations, CLARA estimates a distribution over developmental age groups:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_K\}.$$

3.2 Structured Developmental Annotation

CLARA uses a predefined developmental taxonomy inspired by developmental psychology and educational theory. The taxonomy is hierarchically organized from foundational abilities, such as vocabulary acquisition and emotional expression, to higher-level reasoning abilities including causal reasoning, narrative inference, and perspective taking.

During annotation, the model is constrained to select labels only from the predefined taxonomy. For each story, the annotation model determines which developmental abilities are required for comprehension and activates the corresponding COG, LAN, and SEL labels. Each label is associated with one or more developmental age groups and corresponding activation criteria. The complete taxonomy is provided in Appendix K.

Age Groups	COG cognitive Skills	LAN Language Abilities	SEL Social-Emotional Skills
8-12 years	• multi-step planning • hypothetical thinking • perspective taking • ...	• figurative language • argument understanding • sophisticated expression • ...	• moral reasoning • identity understanding • conflict resolution • ...
6-8 years	• rule application • cause-effect • problem solving • ...	• story comprehension • sequence understanding • detail description • ...	• cooperation • intention understanding • simple negotiation • ...
4-6 years	• seriation • conservation • working memory • ...	• vocabulary growth • simple narratives • question answering • ...	• understanding feelings • simple rules • basic empathy • ...
2-4 years	• sorting • early memory • imitation • ...	• frequent words • naming objects • simple commands • ...	• parallel play • routines • turn taking • ...
0-2 years	• object tracking • sensory exploration • cause-effect (basic) • ...	• single words • babbling • understanding words • ...	• attachment • smiling / responding • basic emotions • ...

Figure 2: Hierarchical developmental annotation framework across COG, LAN, and SEL dimensions.

3.3 Developmental Representation Aggregation

Given the annotated label set $\mathcal{L}(S)$, CLARA first derives an unnormalized developmental score:

$$\tilde{P}(a_i \mid S) = \sum_{l \in \mathcal{L}(S)} w_l \cdot \mathbf{1}(a_i \in g(l)),$$

where $g(l)$ denotes the developmental groups associated with label l.

The scores are normalized into a developmental distribution:

$$P(a_i \mid S) = \frac{\tilde{P}(a_i \mid S)}{\sum_{a_j \in \mathcal{A}} \tilde{P}(a_j \mid S)}.$$

In the current implementation, all labels are uniformly weighted ($w_l = 1$) for simplicity and interpretability. The aggregation step is intentionally kept lightweight and transparent so that the final developmental estimate remains directly traceable to the activated developmental labels, rather than introducing an additional opaque prediction stage. The strong performance of "LogReg over CLARA Labels" suggests that the developmental label space itself contains meaningful predictive signals, while CLARA further improves alignment without supervised training. Alternative aggregation strategies are discussed in Appendix C.

The resulting developmental representation is:

$$\mathbf{P}(S) = [P(a_1 \mid S), \ldots, P(a_K \mid S)],$$

and the dominant developmental stage is estimated through:

$$\hat{a} = \arg\max_{a_i \in \mathcal{A}} P(a_i \mid S).$$

The full distribution $\mathbf{P}(S)$ is the primary developmental representation produced by CLARA; the argmax is used only when a single developmental stage is required for comparison with categorical baselines and reference labels.

4 Benchmark Resource Construction

To study whether AI systems can approximate human developmental judgments of children's narratives, we construct a bilingual benchmark containing 1107 Chinese–English children's stories collected from storybooks and educational materials in PDF and PPT formats.

Each story includes aligned bilingual narrative text, normalized silver developmental references, and structured annotations across COG, LAN, and SEL dimensions. Among them, 1043 stories contain publisher-provided age annotations and are primarily used for quantitative evaluation.

We will release annotation prompts, developmental labels, derived annotations, evaluation splits, and metadata under appropriate licenses. Original story text will only be redistributed where legally permitted. The released resources will be available at: `https://anonymous.4open.science/r/anonymous_story_dataset`.

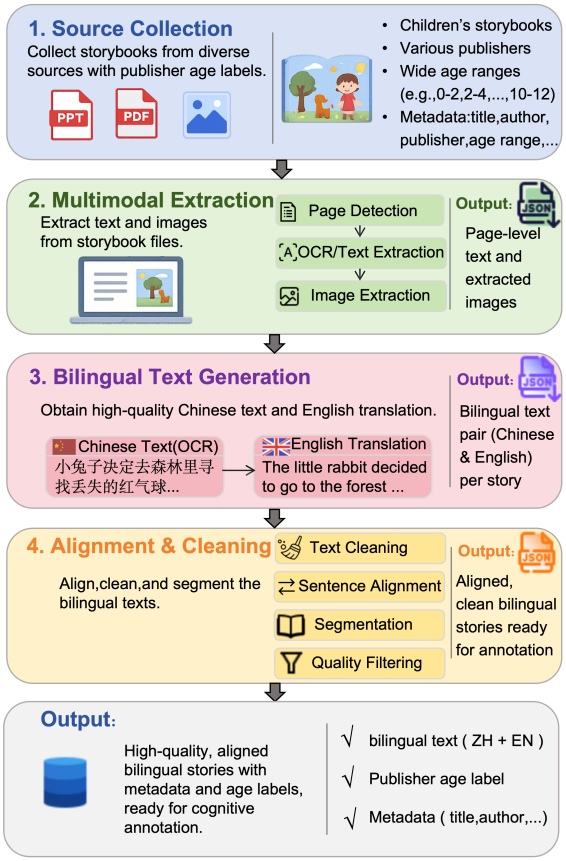

Figure 3: Benchmark construction pipeline.

Figure 3 illustrates the benchmark construction pipeline, including multimodal story extraction, bilingual narrative generation, developmental normalization, and structured annotation.

4.1 Story Extraction and Data Construction

The original materials contain both visual and textual content stored in PDF and PPT formats. We use Qwen3-VL-32B-Instruct to extract narrative text from storybook pages and convert it into structured Chinese narratives, which are subsequently translated into English to construct aligned bilingual story pairs.

Since individual stories may span multiple pages or slides, related pages are grouped into unified narrative representations during preprocessing. Most stories target children between ages 2 and 8 and fall within the medium-length narrative range appropriate for developmental analysis.

4.2 Data Quality Verification

Because the benchmark is constructed through multimodal extraction and bilingual translation, we conduct manual quality verification on a randomly sampled subset of 100 stories.

Age Group	#Mapped Stories	Share
0–2	30	2.9%
2–4	544	52.2%
4–6	687	65.9%
6–8	510	48.9%
8–12	345	33.1%
Annotated stories	1043	94.2%
Unknown age	64	5.8%
Total stories	1107	100.0%

Table 1: Developmental age distribution after normalization.

We evaluate extraction correctness, story grouping correctness, and translation adequacy. Manual inspection shows that 95% of sampled stories contain acceptable extraction quality, 93% preserve correct story grouping, and 91% maintain adequate bilingual semantic consistency.

4.3 Silver Developmental References

Publisher-provided age labels are treated as noisy silver developmental references rather than definitive developmental ground truth. We normalize publisher age ranges into five developmental groups:

$$\{0\text{–}2,\ 2\text{–}4,\ 4\text{–}6,\ 6\text{–}8,\ 8\text{–}12\}.$$

Broad ranges are assigned to all overlapping developmental groups. For example, a story labeled as 5–10 is mapped to 4–6, 6–8, and 8–12.

After normalization and filtering ambiguous labels, 1030 stories are retained for quantitative evaluation. Benchmark evaluation should therefore be interpreted as measuring alignment with conventional developmental recommendations rather than absolute developmental correctness.

4.4 Translated Bilingual Consistency

The bilingual design supports consistency evaluation under translated bilingual settings by comparing developmental predictions and annotations across Chinese and English narrative pairs. Since the English narratives are translations of the corresponding Chinese stories, this analysis evaluates consistency under translation rather than cross-lingual or cross-cultural generalization.

4.5 Developmental Annotation Analysis

The taxonomy contains 122 developmental labels distributed across COG, LAN, and SEL dimensions. Stories targeting younger children emphasize perceptual abilities and emotional expression,

whereas stories for older children increasingly involve causal reasoning, narrative inference, emotional regulation, and perspective taking.

5 Evaluation and Analysis

We evaluate CLARA from the perspective of developmental narrative understanding through developmental label validation, annotation model comparison, translated bilingual consistency analysis, benchmark comparison, ablation studies, and human evaluation. Our evaluation focuses on whether structured developmental annotation can capture meaningful developmental progression patterns, align with human developmental judgments, remain stable across translated bilingual narratives, and provide advantages over readability-based and direct prompting approaches.

5.1 Evaluation Setup

The benchmark contains 1107 stories, among which 1030 are retained for quantitative evaluation after normalization and filtering. We additionally construct a balanced 100-story validation subset for annotation model comparison and consistency analysis.

All LLM-based experiments use temperature 0, and annotation model selection is repeated three times. Table 2 reports annotation-model comparison over repeated runs on the 100-story validation subset, whereas Table 4 reports the final full-benchmark result using the selected GPT-4o annotation engine. Because the full-benchmark evaluation uses temperature 0 with deterministic downstream aggregation and evaluation, Table 4 reports a single final result. Because publisher age labels are noisy silver references rather than definitive developmental ground truth, the evaluation should be interpreted as measuring alignment with conventional developmental recommendations. Results involving the 0–2 developmental group should be interpreted cautiously because this category contains only 30 stories and developmental suitability at this age may depend more strongly on illustrations, repetition, sensory engagement, and caregiver interaction, which are not modeled in the current text-only formulation.

We evaluate developmental alignment using overlap accuracy and mean developmental distance. For annotation model comparison, we additionally report mapped accuracy, label stability (L-Stab.), and age stability (A-Stab.). Detailed metric definitions are provided in Appendix B.

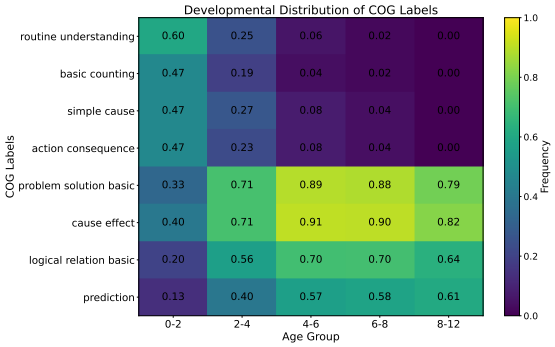

Figure 4: Developmental distribution of representative COG labels across age groups.

5.2 Empirical Validation of Developmental Labels

We analyze representative label distributions across developmental stages using annotations generated by GPT-4o, Gemini-2.5-Pro, Qwen-Max, and DeepSeek-Chat.

Figures 4, 5, and 6 show developmental progression trends across COG, LAN, and SEL dimensions. Stories targeting younger children emphasize perceptual abilities and emotional expression, whereas stories for older children increasingly involve causal reasoning, narrative inference, emotional regulation, and perspective taking.

These trends remain relatively consistent across annotation models, suggesting that the observed structures capture meaningful developmental signals rather than artifacts of a specific annotation engine. Some foundational labels also remain active in older age groups because developmental abilities are cumulative rather than mutually exclusive. Interpretation of the 0–2 developmental group should nevertheless be treated cautiously due to the substantially smaller number of stories in this category.

5.3 Annotation Model Comparison

We compare GPT-4o, Gemini-2.5-Pro, Qwen-Max, and DeepSeek-Chat on the 100-story validation subset using developmental alignment, label coverage, and annotation stability metrics.

As shown in Table 2, Gemini-2.5-Pro achieves the highest mapped accuracy but tends to over-label developmental signals, whereas DeepSeek-Chat produces excessively dense annotations. GPT-4o achieves the most balanced tradeoff between developmental alignment and annotation stability, and

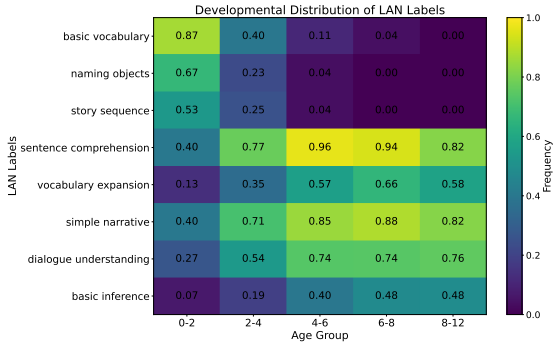

Figure 5: Developmental distribution of representative LAN labels across age groups.

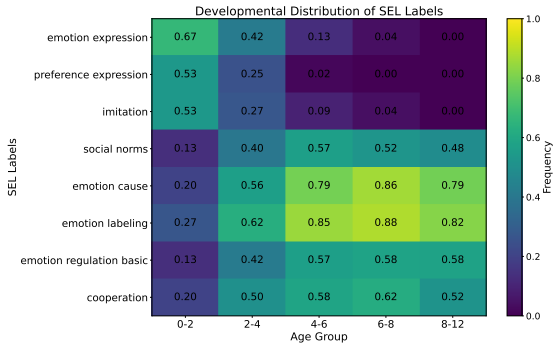

Figure 6: Developmental distribution of representative SEL labels across age groups.

is therefore selected as the primary annotation engine. Additional annotation validation results on a manually annotated educator subset are provided in Appendix E.

5.4 Consistency under Translated Bilingual Narratives

To evaluate whether structured developmental representations remain stable under bilingual settings, we compare annotations generated from Chinese and English versions of the same 100 stories using GPT-4o.

As shown in Table 3, developmental representations remain relatively stable across translated bilingual narratives, with overall annotation similarity of 0.772 and developmental-stage agreement of 86.0%. The relatively low strict exact full-set match reflects the difficulty of obtaining identical fine-grained developmental annotations across translated bilingual narratives. However, the substantially higher Jaccard overlap and developmental-stage agreement suggest that higher-level developmental reasoning signals remain comparatively stable despite variation in individual labels.

Model	Acc.	Cov.	L-Stab.	A-Stab.
GPT-4o	0.71±0.01	17.73±0.15	0.87	0.89
Gemini-2.5-Pro	**0.87**±0.01	29.56±0.65	0.66	0.79
Qwen-Max	0.73±0.02	26.86±0.71	0.77	0.79
DeepSeek-Chat	0.69±0.02	62.41±0.48	**0.91**	**0.91**

Table 2: Annotation model comparison on the 100-story validation subset.

Metric	Score
Overall Label Jaccard	0.772
COG Jaccard	0.793
LAN Jaccard	0.774
SEL Jaccard	0.775
Age Agreement	0.860
Strict Exact Full-Set Match	0.180

Table 3: Consistency under translated bilingual narratives.

5.5 Benchmark Comparison with Readability and Prompting Baselines

We compare CLARA against readability-based baselines, direct prompting baselines, rubric-guided prompting, and a logistic regression classifier trained over CLARA developmental labels.

As shown in Table 4, readability metrics substantially underperform LLM-based approaches, suggesting that surface linguistic complexity alone is insufficient for developmental narrative understanding. The rubric-guided baseline improves over direct prompting, indicating that intermediate reasoning contributes to developmental estimation, while the strong performance of "LogReg over CLARA Labels" further shows that the structured developmental annotations themselves contain meaningful developmental signals. CLARA nevertheless achieves the strongest alignment with normalized developmental references, suggesting that explicit developmental annotation and structured aggregation provide additional benefits beyond implicit reasoning alone.

5.6 Contribution Decomposition

To better understand which components contribute most to CLARA's performance, we conduct a contribution decomposition study.

Table 5 shows progressive improvement as additional structured components are introduced. The taxonomy itself contributes substantially to developmental alignment, while explicit developmental age mapping and structured aggregation further improve performance. These results suggest that CLARA's gains do not originate solely from prompt engineering, but from the interaction

Method	Overlap ↑	Mean Distance ↓
Dale–Chall	0.469	0.882
FRE	0.517	0.681
FKGL	0.603	0.502
GPT-4o	0.730	0.672
Gemini-2.5-Pro	0.768	0.497
Qwen-Max	0.775	0.467
DeepSeek-Chat	0.787	0.451
GPT-4o + Rubric	0.835	0.284
LogReg + CLARA Labels	0.874	0.173
CLARA (ours)	**0.904**	**0.096**

Table 4: Benchmark comparison using normalized developmental reference labels.

between structured developmental representation, constrained annotation, and developmental aggregation.

5.7 Taxonomy Quality Ablation

To evaluate the contribution of the developmental taxonomy, we construct several degraded taxonomy variants.

As shown in Table 6, degraded taxonomies substantially reduce developmental alignment quality, suggesting that CLARA benefits from meaningful developmental organization rather than prompt structure alone. Additional dimension-level ablation results are provided in Appendix D.

5.8 Human Evaluation

We conduct a blinded human evaluation study to assess whether the proposed developmental annotations align with educational judgments from early-childhood educators. We recruit eight educators with early-childhood education experience and randomly sample 60 stories covering multiple developmental stages from the benchmark, where each story is independently evaluated by three educators, resulting in 180 evaluation instances.

As shown in Table 7, educational interpretability receives the highest average score, suggesting that the proposed annotation space provides meaningful developmental explanations beyond direct age prediction alone. Educators further noted that structured developmental labels better expose emotional reasoning, social interaction, and implicit causal understanding, while the average pairwise annotation agreement across the three evaluators per story reaches 0.81, supporting the reliability of the proposed developmental annotation framework.

Variant	Overlap ↑	Mean Distance ↓
GPT-4o Direct	0.730	0.672
Free-form + Agg.	0.791	0.411
Taxonomy Labels	0.852	0.231
Taxonomy + Mapping	0.889	0.148
Full CLARA	**0.904**	**0.096**

Table 5: Contribution decomposition analysis for CLARA.

Variant	Overlap ↑	Distance ↓
Random Mapping	0.611	0.593
Dimension-swapped	0.742	0.366
Coarse Taxonomy	0.861	0.181
Full Taxonomy	**0.904**	**0.096**

Table 6: Taxonomy quality ablation study.

Criterion	Score
COG Label Quality	4.24
LAN Label Quality	4.18
SEL Label Quality	4.31
Educational Interpretability	**4.42**
Predicted Age Appropriateness	4.17
Average	4.26

Table 7: Blinded human evaluation results.

5.9 Human Developmental Ranking Validation

To provide a more independent evaluation beyond publisher-provided developmental references, we conduct a blinded human developmental ranking study using 60 story pairs covering both adjacent and distant developmental stages, and compare human pairwise rankings against readability metrics, direct prompting baselines, rubric-guided prompting, and CLARA developmental representations.

As shown in Table 8, CLARA achieves the highest agreement with human developmental judgments, suggesting that structured developmental annotation better captures relative developmental complexity than readability-based or direct prompting approaches.

5.10 Analysis of Developmental Label Discrepancies

We analyze disagreement cases between publisher developmental references and CLARA developmental estimates using 30 randomly sampled stories evaluated by blinded educators.

As shown in Table 9, educators more frequently prefer CLARA developmental estimates over publisher recommendations. Many disagreement cases involve abstract reasoning, emotional nuance, or social understanding that publisher labels may underrepresent. Disagreements may arise from several sources, including broad publisher recommendations, stories containing mixed developmental

Method	Human Agreement
FKGL	0.58
GPT-4o Direct	0.67
GPT-4o Rubric Direct	0.73
CLARA (ours)	**0.81**

Table 8: Agreement with blinded human developmental ranking judgments.

Preference	Human Evaluators
Prefer CLARA	61%
Prefer Publisher	19%
Both Reasonable	20%

Table 9: Human evaluation on disagreement cases between publisher developmental references and CLARA developmental estimates.

signals across adjacent age bands, and occasional over- or under-activation of developmental labels by the annotation model.

6 Conclusion

In this paper, we study whether AI systems can approximate human developmental judgments of children's narratives through structured developmental annotation. We introduce CLARA, a cognitively grounded approach for developmental narrative understanding across COG, LAN, and SEL dimensions, together with a translated bilingual benchmark containing 1107 Chinese–English children's stories and normalized silver developmental references. Experimental results show that structured developmental annotation achieves substantially stronger alignment with developmental references and blinded human educational judgments than readability-based methods and direct prompting baselines, suggesting that developmental narrative understanding cannot be fully captured through surface linguistic complexity alone. Additional analyses further demonstrate relatively stable developmental representations across translated bilingual narratives and annotation models. Overall, our findings indicate that AI systems can approximate certain aspects of human developmental reasoning when guided by interpretable developmental representations, while also highlighting the importance of transparency, human oversight, and cognitively grounded evaluation in educational NLP.

Limitations

Despite promising results, this work has several limitations. Developmental narrative understanding is inherently subjective and may vary across educational, cultural, and linguistic contexts. The proposed framework relies on structured developmental annotations and silver developmental references, which may not fully capture the complexity and diversity of human developmental reasoning. In addition, the current study mainly focuses on text-based narrative analysis and does not explicitly model multimodal storytelling elements such as illustrations, visual narrative structure, repetition, or caregiver interaction. This limitation is particularly important for the 0–2 age group, which contains only 30 stories and may rely more heavily on non-textual developmental cues. Moreover, because the English narratives are translations of the corresponding Chinese stories, the bilingual analysis should be interpreted as consistency under translation rather than evidence of broad cross-lingual or cross-cultural generalization. Future work may further explore more diverse developmental settings, multimodal evaluation, and stronger reasoning-oriented approaches for developmental narrative understanding.

Ethics Statement

This work studies developmental narrative understanding for educational and research purposes. The proposed framework is intended to support interpretable analysis of children's narratives rather than automated educational judgment, psychological assessment, or clinical decision-making. Because developmental suitability is inherently influenced by cultural, linguistic, educational, and individual factors, the developmental annotations and predictions produced by CLARA should be interpreted as supportive computational signals rather than authoritative developmental conclusions. In addition, the framework relies on large language model annotations and silver developmental references, which may introduce biases and inconsistencies inherited from pretrained models, prompting strategies, and publisher conventions. The story corpus consists of published children's narratives and educational materials rather than personal-user data, and no personally identifying information is included in the released benchmark. Human evaluation was conducted anonymously with voluntary participation from educators, and no personal or sensitive information was collected. Participants were recruited through direct invitation based on their early-childhood education experience and participated voluntarily without monetary compensa-

tion. All participants provided informed consent prior to participation and were informed that their anonymized responses would be used for research purposes. Overall, CLARA is intended as a research framework for studying developmental narrative understanding and interpretable educational NLP.

Acknowledgments

We thank the anonymous reviewers, educators, and evaluators for their valuable feedback and developmental perspectives on children's narrative understanding, as well as all contributors who provided discussions, annotation support, and technical assistance throughout this project. Code, prompts, derived annotations, metadata, and permitted benchmark components will be released upon acceptance to support future research in developmental narrative understanding and educational NLP.

References

Janet W. Astington, Paul L. Harris, and David R. Olson, editors. 1988. *Developing Theories of Mind*. Cambridge University Press, Cambridge.

Collaborative for Academic, Social, and Emotional Learning. 2012. 2013 CASEL guide: Effective social and emotional learning programs—preschool and elementary school edition. Technical report, Collaborative for Academic, Social, and Emotional Learning, Chicago, IL.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–20, 28.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Natalia Gagarina and Ute Bohnacker. 2022. A new perspective on referentiality in elicited narratives: Introduction to the special issue. *First Language*, 42(2):171–190.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: Multilingual assessment instrument for narratives – revised. *ZAS Papers in Linguistics*, 63:20.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command, Millington, TN, USA.

Judy R. Kupersmitt and Sharon Armon-Lotem. 2019. The linguistic expression of causal relations in picture-based narratives: A comparative study of bilingual and monolingual children with tld and dld. *First Language*, 39(3):319–343.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: Generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Jean Piaget. 1952. *The Origins of Intelligence in Children*. International Universities Press, New York, NY.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Elaine Reese, Catherine A. Haden, Lynne Baker-Ward, Patricia J. Bauer, Robyn Fivush, and Peter A. Ornstein. 2011. Coherence of personal narratives across the lifespan: A multidimensional model and coding method. *Journal of Cognition and Development*, 12(4):424–462.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035. AAAI Press.

Alexey Tikhonov, Igor Samenko, and Ivan P. Yamshchikov. 2021. StoryDB: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*,

pages 32–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bram van Dijk, Marco Spruit, and Max van Duijn. 2023. Theory of mind in freely-told children's narratives: A classification approach. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12979–12993, Toronto, Canada. Association for Computational Linguistics.

Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Henry M. Wellman and David Liu. 2004. Scaling of theory-of-mind tasks. *Cognition*, 92(3):167–190.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Jingdan Yang, Jae-Hyun Kim, Outi Tuomainen, and Nan Xu Rattanasone. 2022. Bilingual mandarin-english preschoolers' spoken narrative skills and contributing factors: A remote online story-retell study. *Frontiers in Psychology*, 13:797602.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

A AI Usage Statement

AI tools were used only for literature search assistance, data processing support, and annotation generation during the research process. The writing, organization, interpretation, and final wording of the manuscript were completed and reviewed by the authors. The authors take full responsibility for the content of this paper.

B Detailed Evaluation Metrics

We evaluate developmental alignment using overlap accuracy and mean developmental distance.

Given predicted developmental groups $\hat{\mathcal{A}}$ and silver reference groups $\mathcal{A}^*$, overlap accuracy is defined as:

$$\text{Overlap} = \frac{|\hat{\mathcal{A}} \cap \mathcal{A}^*|}{|\hat{\mathcal{A}} \cup \mathcal{A}^*|}.$$

Mean developmental distance measures the average ordinal distance between predicted and reference developmental groups after mapping developmental stages into ordered indices.

For annotation model comparison, mapped accuracy measures whether the dominant predicted developmental group overlaps with any of the normalized silver reference groups.

Label stability (L-Stab.) is computed using Jaccard similarity between repeated annotation runs:

$$\text{L-Stab} = \frac{|\mathcal{L}_1 \cap \mathcal{L}_2|}{|\mathcal{L}_1 \cup \mathcal{L}_2|},$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ denote developmental label sets generated from repeated runs.

Age stability (A-Stab.) measures agreement between dominant developmental-stage predictions across repeated annotation runs.

C Aggregation Sensitivity Analysis

CLARA currently adopts a uniform weighting strategy in which all developmental labels contribute equally during developmental aggregation. This design is intentionally lightweight and interpretable, allowing the framework to focus on structured developmental representations rather than manually engineered weighting heuristics.

To examine whether the framework is overly sensitive to the specific aggregation strategy, we additionally explored several alternative weighting schemes, including frequency-normalized weighting and dimension-balanced weighting across COG, LAN, and SEL dimensions.

Preliminary experiments showed that alternative weighting strategies produced relatively similar developmental alignment trends, while also introducing additional hyperparameters and reduced interpretability. We therefore adopt uniform weighting in the main experiments for simplicity, transparency, and reproducibility.

These observations further suggest that the primary contribution of CLARA lies in the structured

Variant	Overlap ↑	Mean Distance ↓
Full CLARA	**0.904**	**0.096**
w/o COG	0.861	0.164
w/o LAN	0.873	0.151
w/o SEL	0.842	0.183

Table 10: Dimension-level ablation study on different developmental dimensions in CLARA.

Metric	Score
COG F1	0.84
LAN F1	0.81
SEL F1	0.79
Representative Macro F1	0.73
Representative Micro F1	0.78
Human–Human Cohen's κ	0.76

Table 11: Gold annotation validation on a manually annotated subset.

developmental annotation space itself rather than a particular aggregation heuristic.

D Dimension-Level Ablation Study

To further analyze the contribution of different developmental dimensions in CLARA, we conduct a dimension-level ablation study by removing one developmental dimension at a time, including COG, LAN, and SEL.

Table 10 shows that removing any developmental dimension reduces developmental alignment and increases developmental distance. Among the three variants, removing SEL leads to the largest performance degradation, suggesting that social-emotional development provides important signals beyond linguistic and cognitive complexity alone. Removing COG and LAN also negatively affects performance, indicating that developmental reasoning benefits from integrating complementary developmental dimensions.

E Gold Annotation Validation

To directly evaluate the developmental annotations, we construct a manually annotated gold subset for label-level comparison.

We randomly sample 60 stories from the benchmark. Three educators from the evaluation pool independently annotate developmental dimensions and representative developmental labels for each story.

The results demonstrate relatively strong agreement between CLARA annotations and human developmental judgments. The human–human agreement score further suggests that developmental interpretation remains partially subjective even among expert evaluators.

Metric	Score
Exact Agreement	0.68
Adjacent Agreement	0.89
Cohen's κ	0.72

Table 12: Agreement between publisher developmental references and educator judgments.

Story Excerpt
A cloth bear receives a letter inviting him to visit the big bears when the moon is full. On the journey, the bear takes a bus, climbs a mountain path, and finally joins the celebration with the other bears.
Developmental Labels
COG: simple cause; action consequence; routine understanding **LAN:** basic vocabulary; story sequence **SEL:** emotion expression; preference expression; imitation
Publisher Reference
2–5 (mapped to 2–4 and 4–6)
CLARA Prediction
4–6

Table 13: Representative example from the CLARA dataset.

F Publisher–Educator Agreement Validation

Because publisher-provided developmental references constitute the primary benchmark target, we evaluate whether these references align with independent educator judgments.

We randomly sample 120 stories from the benchmark and recruit five educators with early-childhood education experience to assign developmental age groups without observing the original publisher recommendations.

The results show that publisher recommendations exhibit reasonably strong alignment with independent educator judgments despite ambiguity in stories involving broad developmental ranges or abstract reasoning. These findings support the use of publisher age recommendations as noisy but meaningful silver developmental references.

G Representative Dataset Example

Table 13 presents a representative example from the CLARA dataset.

H Representative Disagreement Example

Table 14 presents a representative disagreement case between publisher-provided age recommendations and CLARA predictions.

Story Excerpt
A man is rescued by the Nine-Colored Deer after falling into a river. Although he promises to keep the deer secret, he later betrays the deer in exchange for a reward. The story follows the consequences of betrayal, trust, and moral responsibility.
Developmental Labels
COG: cause effect; prediction; sequence order **LAN:** narrative inference; discourse understanding **SEL:** emotion cause; perspective taking; rule awareness
Publisher Reference
3–8 (mapped to 2–4, 4–6, 6–8)
CLARA Prediction
6–8

Table 14: Representative disagreement case between publisher-provided age recommendations and CLARA predictions.

I Human Evaluation Questionnaire

Human evaluators were asked to assess stories and corresponding CLARA annotations using a 1–5 Likert scale.

- **COG Label Quality:** Do the cognitive labels accurately reflect the reasoning abilities required by the story?
- **LAN Label Quality:** Do the language-related labels appropriately describe the language complexity of the story?
- **SEL Label Quality:** Do the social-emotional labels reflect the emotional and interpersonal understanding required by the story?
- **Educational Interpretability:** Are the developmental labels educationally meaningful and interpretable?
- **Predicted Age Appropriateness:** Does the predicted developmental age group reasonably match the story content?

Scores range from 1 (strongly disagree) to 5 (strongly agree).

J Prompt Templates

J.1 CLARA Annotation Prompt

You are an expert in developmental psychology and children's narrative understanding.

Given a children's story, analyze the developmental abilities required for comprehension.

Assign developmental labels from the predefined taxonomy across the following dimensions:

- Cognitive Skills (COG)
- Language Abilities (LAN)
- Social-Emotional Skills (SEL)

Select only labels that are strongly supported by the story content. Avoid unsupported, redundant, or weakly related labels.

Return the selected developmental labels in structured JSON format.

J.2 Direct Prompting Baseline

Given a children's story, predict the most appropriate developmental age group from the following categories:

0–2, 2–4, 4–6, 6–8, 8–12.

Base your prediction on the developmental abilities required to understand the story, including reasoning complexity, language understanding, emotional interpretation, and social understanding.

Return only the predicted age group.

J.3 Rubric-Guided Prompting Baseline

Given a children's story, first analyze the following aspects:

- Cognitive complexity
- Language complexity
- Social-emotional understanding

Based on these factors, predict the most appropriate developmental age group from:

0–2, 2–4, 4–6, 6–8, 8–12.

Do not use predefined developmental labels.

Return only the predicted age group and a brief explanation.

J.4 LLM-as-Judge Prompt

You are asked to compare two developmental age assessments for the same children's story.

One assessment is provided by the publisher. The other assessment is generated

by a structured developmental annotation system.

Evaluate which assessment better matches the developmental abilities required for understanding the story.

Consider:

- reasoning complexity
- language understanding
- emotional understanding
- social and moral interpretation

Choose one of the following options:

- Prefer CLARA
- Prefer Publisher
- Both Reasonable

Provide a brief explanation for the decision.

J.5 Human Ranking Evaluation Prompt

Given two children's stories, determine which story requires more advanced developmental abilities for comprehension.

Consider:

- reasoning complexity
- language understanding
- emotional interpretation
- social understanding

Choose one:

- Story A requires more advanced developmental abilities
- Story B requires more advanced developmental abilities
- Both are similar

K Full developmental label taxonomy

Some labels intentionally preserve compact annotation-oriented naming conventions to maintain consistency across visualization figures, prompts, and annotation outputs. Table **??** presents the complete developmental label inventory used in CLARA. The taxonomy contains over one hundred labels across five developmental age groups and three dimensions: COG, LAN, and SEL.

0–2 years (Sensorimotor stage)
COG: visual tracking, sound response, face recognition, object recognition, object permanence, color recognition, shape recognition, size discrimination, pattern recognition, sensory integration **LAN:** sound association, babbling pattern, word recognition, word-object mapping, phonetic awareness **SEL:** emotion reaction, emotion recognition, attachment security, social smiling, caregiver response
2–4 years (Early preoperational stage)
COG: simple cause, action consequence, routine understanding, object function, basic counting, category learning, similarity detection, difference detection **LAN:** basic vocabulary, two-word sentence, sentence imitation, naming objects, story sequence, picture description, symbolic play language **SEL:** emotion expression, emotion labeling basic, imitation, parallel play, role play, self-awareness basic, preference expression, turn-taking basic
4–6 years (Late preoperational stage)
COG: cause effect, prediction, sequence order, temporal understanding, spatial relation, problem solution basic, categorization, attribute sorting, logical relation basic, multi attribute comparison **LAN:** sentence comprehension, grammar structure, vocabulary expansion, simple narrative, character tracking, event linking, story retelling, dialogue understanding, basic inference, question answering **SEL:** emotion labeling, emotion cause, emotion regulation basic, friendship, sharing behavior, cooperation, rule awareness, rule following, social norms, group participation
6–8 years (Concrete operational stage)
COG: logical reasoning, multi-step reasoning, causal chain understanding, problem solving, problem decomposition, conservation understanding, reversibility understanding, hierarchical categorization **LAN:** reading comprehension, early reading, independent reading, text inference, summarization, story structure, narrative coherence, multi event sequence **SEL:** emotion regulation, empathy basic, perspective taking basic, conflict resolution, responsibility, cooperation advanced, self control, decision making basic
8–12 years (Formal operational stage)
COG: abstract thinking, analogy reasoning, hypothesis generation, counterfactual thinking, concept generalization, symbolic reasoning, rule abstraction, meta cognition **LAN:** critical reading, theme understanding, figurative language, metaphor understanding, argumentation, perspective language, cross text analysis **SEL:** empathy, perspective taking, moral reasoning, ethical reasoning, fairness, honesty, self reflection, goal setting, identity awareness, prosocial behavior

Table 15: Full developmental label taxonomy used in CLARA.