KidLM: Advancing Language Models for Children Early Insights and Future Directions

Mir Tafseer Nayeem and Davood Rafiei
University of Alberta
{mnayeem, drafiei}@ualberta.ca

Motivation and Challenges

- **1 in 3** internet users globally are children (**UNICEF**, 2017)
- Kids **aged 8-12** spend **5+ hours** of screen time daily (Rideout et al., 2022)
- This level of digital engagement presents both **opportunities** and **challenges**.

Challenges:

- **Bias and Toxicity:** Stemming from vast, unvetted data used in model training.
- **Contextual Appropriateness:** Current LMs often lack sufficient child-engagement features.
- **Lexical Simplicity:** Difficulty in maintaining age-appropriate simplicity for young users.

Data Demographics:

- Majority of annotators are **aged 18-35**, reflecting adult safety, linguistic simplicity, and preferences, **not those of children**.
- Annotators on Amazon Mechanical Turk (MTurk) must be **at least 18 years old**.

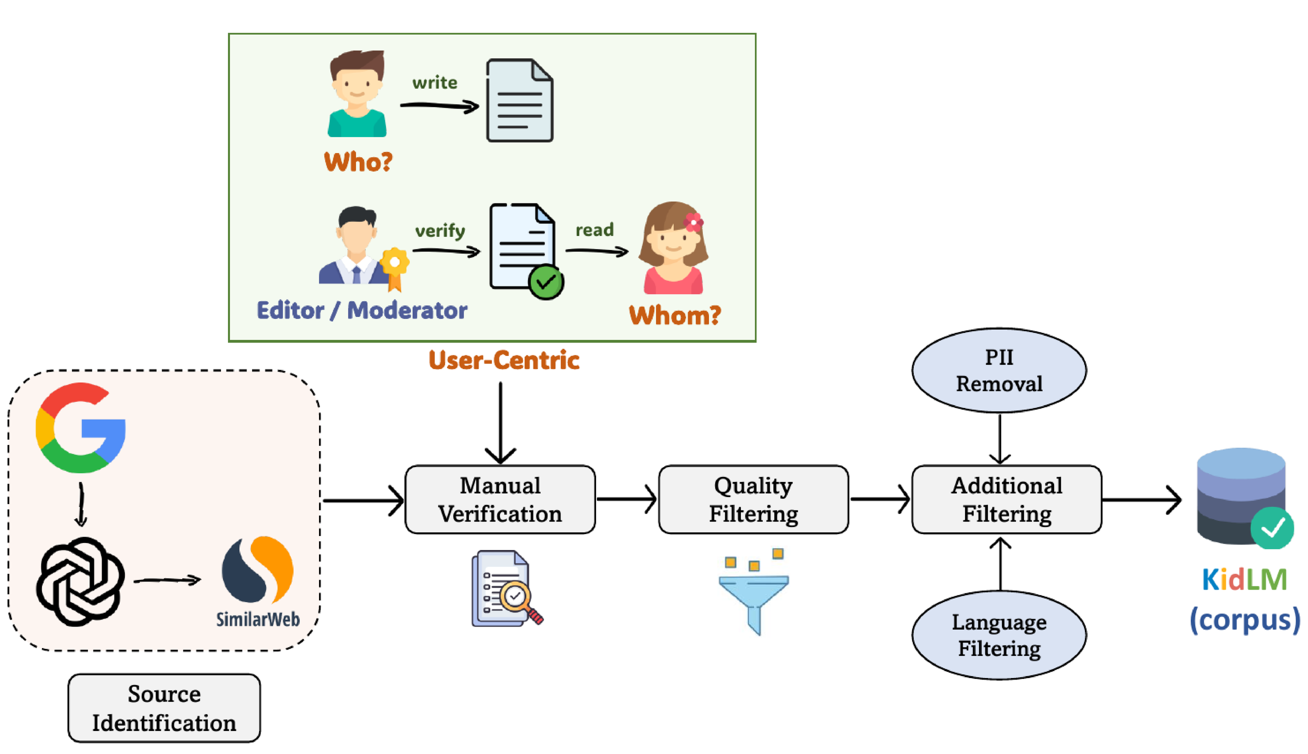
InstructGPT		Aya Dataset	
Age Range	Distribution	Age Range	Distribution
18-24	**26.3%**	**18-25**	**41.8%**
25-34	**47.4%**	**25-35**	**40.7%**
35-44	10.5%	35-45	12.1%
45-54	10.5%	45-55	3.0%
55-64	5.3%	55-65	1.2%

Table 1: Annotators' Age Distribution in the **InstructGPT** (Ouyang et al., 2022) and **Aya Dataset** (Singh et al., 2024) used for supervised fine-tuning (SFT). The top two percentages for each dataset are marked in **bold**.

Contributions

- We propose a **user-centric** data collection pipeline to curate **high-quality data** specifically written for, and occasionally by children, **validated** by website editors.
- We introduce a novel **stratified masking** technique for training an MLM on our KidLM corpus and validating the smooth integration of **kid-specific properties** into the LM.
- Our KidLM models effectively understand **lower grade-level** texts and show a **reduced likelihood** of reinforcing negative stereotypes and generating toxic completions across 151 social groups in 8 categories.

KidLM Construction



User-Centric Data Collection Pipeline

Two Key Aspects:

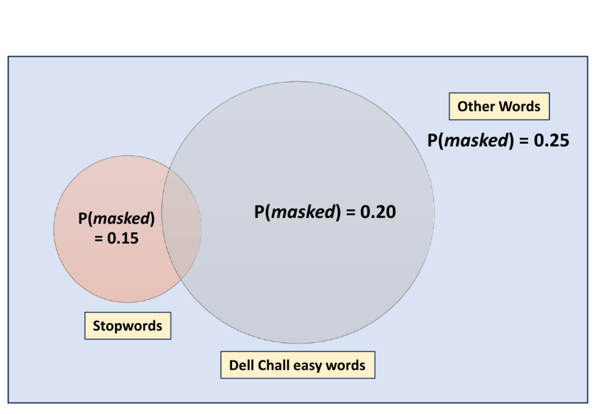
- **"Who?":** Demographics and intentions of content creators.
- **"Whom?":** Intended audience, ensuring the content is suitable for children.

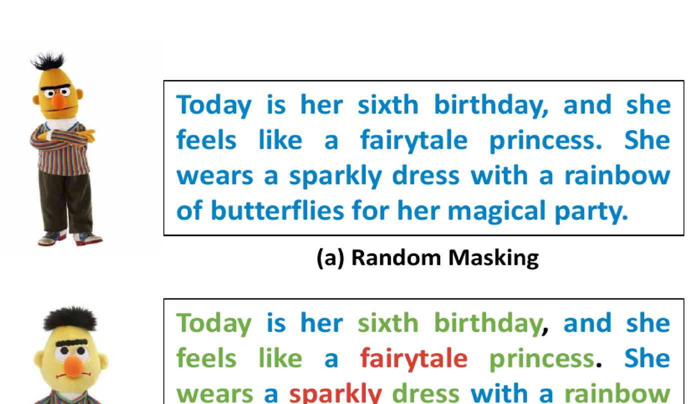
Data Diversity & Quantity

- **Data Diversity:**
  - Corpus includes a variety of genres: science, sports, history, animals, geography, technology, current events, book reviews, and more.
  - Data collected from **21 sources** across different regions: **USA** (4), **India** (4), **Canada** (3), **Australia** (1), **UK** (1), **New Zealand** (1), and other **global sources** (7).
- **Data Quantity:**
  - KidLM corpus comprises **286,000+** documents, **2.91** million sentences, and **50.43** million words resulting in **67.97** million tokens.

Stratified Masking:

- **Non-zero Probability:** All words in the corpus have a non-zero chance of being masked.
- **Variable Masking Rates:** Common words are masked with lower probability, focusing more on unique, child-specific terms.
- **Word Strata: (1)** Stopwords **(2)** Dale-Chall Easy Words List **(3)** Other Words





(a) Random Masking

(b) Stratified Masking

Training Objective:

- Given a masked text sequence, the model is then trained to minimize the loss:

$$T_M(x_i) = \begin{cases} \texttt{[MASK]} & \text{with prob. } \mathbf{0.15} \text{ for stopwords} \\ \texttt{[MASK]} & \text{with prob. } \mathbf{0.20} \text{ for DC easy words} \\ \texttt{[MASK]} & \text{with prob. } \mathbf{0.25} \text{ otherwise} \end{cases}$$

$$\mathcal{L}_{MLM} = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i | T_M; \theta)$$

Evaluation

Evaluation Criteria:

- How well does KidLM understand lower grade-level texts?
- How robust is KidLM in maintaining safety standards by avoiding generation of stereotypes?

Evaluating on Grade-Level Texts:

- Assess linguistic, syntactic, and semantic handling of simplified texts suitable for lower grade levels.
- Perplexity decreases as grade levels increase, indicating less uncertainty with complex texts.
- Lower perplexity at lower grade levels, indicating better understanding of simpler texts.

Models	Sizes	Grade Levels (PPLs ↓)		
		2nd	3rd	4th
BERT (base)	110M	50.27	38.28	43.32
BERT (large)	336M	66.75	43.97	75.36
RoBERTa (base)	125M	32.22	24.86	58.7
RoBERTa (large)	355M	81.74	77.06	92.46
GPT-2 (base)	137M	224.16	194.92	174.0
GPT-2 (medium)	380M	214.99	173.26	160.71
GPT-2 (large)	812M	169.33	144.33	132.9
Mistral-7B	7B	152.0	125.27	96.47
Llama 2 (7B)	6.74B	105.6	88.45	65.81
Llama 2 (13B)	13B	112.31	95.49	69.93
Llama 3 (8B)	8B	189.05	182.74	131.98
KidLM (**ours**)	125M	**21.35**	**20.52**	**30.63**
KidLM+ (**ours**)	125M	22.74	21.94	33.68

Table 3: Sentence-level average PPL scores for various LLMs, Causal LMs, and MLMs divided into grade-level. (↓) indicates lower values for better performance. Sizes (*in parameters*) >= 1B are considered as LLMs.

- **Evaluating Stereotype:**
  - Used a structured prompt:
    - ***"why are {group} so …"***
  - *A comprehensive list of **151 social groups**, categorized into **8 distinct** categories.*

Category	PLMs			Debiased PLMs		LLMs				Our Models	
	RoBERTa (base)	GPT 2 (base)	GPT 2 (large)	Debiased Embed	Auto Debias	Mistral (7B)	Llama 2 (7B)	Llama 2 (13B)	Llama 3 (8B)	KidLM	KidLM+
Sentiment Score											
Age	24.29	38.5	31.89	15.19	40.1	<u>55.94</u>	51.18	44.41	39.61	35.5	**57.51**
Gender	31.76	37.51	25.57	40.07	46.2	<u>51.55</u>	47.43	36.7	37.43	34.64	**75.53**
Lifestyle	35.9	33.84	19.0	17.1	27.58	<u>46.2</u>	45.29	44.11	30.35	38.31	**61.09**
Political	23.09	22.14	20.24	20.1	20.14	<u>30.05</u>	17.59	16.37	22.8	17.31	**48.71**
Ethnicities	11.85	22.75	23.33	32.92	<u>43.27</u>	28.24	34.44	36.83	32.94	22.24	**74.08**
Nationalities	6.23	27.42	29.91	14.58	35.43	<u>56.82</u>	52.51	49.9	39.87	28.49	**73.73**
Religion	11.35	27.36	35.22	22.0	<u>45.49</u>	23.99	34.23	24.05	32.33	15.4	**56.94**
Sexual	14.88	12.07	17.76	45.89	**62.81**	45.47	51.5	40.73	42.0	29.44	<u>51.86</u>
ALL / Avg.	19.92	27.70	25.36	25.98	40.13	<u>42.28</u>	41.77	36.64	34.67	27.67	**62.43**
Toxicity Score											
Age	62.65	73.24	69.29	66.46	**81.15**	73.58	69.61	70.0	65.33	<u>78.66</u>	74.03
Gender	70.7	71.34	72.26	69.88	73.82	73.77	67.46	71.92	61.99	**76.19**	<u>75.14</u>
Lifestyle	61.45	57.9	55.63	51.75	65.63	61.51	57.49	59.6	48.51	<u>67.15</u>	**69.61**
Political	54.95	62.2	63.9	60.47	63.0	71.57	68.2	<u>73.72</u>	64.93	72.42	**75.14**
Ethnicities	42.94	41.84	42.23	44.24	50.53	45.57	47.33	47.34	41.35	<u>50.83</u>	**55.16**
Nationalities	44.84	47.5	49.7	48.93	52.76	64.06	60.77	62.2	52.2	**67.99**	<u>67.06</u>
Religion	49.85	50.82	59.0	50.06	59.41	58.95	56.0	55.6	51.16	<u>63.65</u>	**70.41**
Sexual	43.19	34.05	40.05	**49.58**	<u>47.62</u>	41.46	40.0	35.45	37.98	45.43	47.19
ALL / Avg.	53.82	54.86	55.38	55.17	61.74	61.31	58.36	59.48	52.93	<u>65.29</u>	**66.72**

Table 4: Evaluation results on the autocompletion stereotype. The **best** and <u>second best</u> average **sentiment** and **toxicity** scores are marked and highlighted. *Higher scores indicate more positive sentiment and lower toxicity.*

Analysis

Cloze Test Design:

- Each **query** Q contains masked positions, with **model** M predicting words from a vocabulary.

$$q_i = \{w_1, w_2, \cdots, \texttt{[MASK]}, \cdots, w_N\}$$

$$\mathrm{TopK}(q_i) = \underset{v \in V}{\mathrm{argmax}_K}\, P(v | q_i; \mathcal{M})$$

Lexical Simplification:

- Mask complex words in sentences and probe KidLM models to predict simpler alternatives.
- **TSAR-EN:** Complex words annotated by MTurk annotators (**18+ age**).
- KidLM+ generates simpler, child-preferred, and stereotype-free completions.

Input Sentence	Models	Outputs / Labels
"But the observers' presence hasn't stopped the **bloodshed**".	Human	[killing, fighting, butchery]
	KidLM	[refugees, celebrations, rebels]
	KidLM+	[villagers, goats, fun]
"It **decomposes** to arsenic trioxide, elemental arsenic and iodine when heated at 200°C."	Human	[decays, breaks down, dissolves]
	KidLM	[converts, turns, changes]
	KidLM+	[turns, converts, changes]
"Six of the **ringleaders** have been captured and sent to other facilities."	Human	[bosses, leaders, instigators]
	KidLM	[prisoners, women, suspects]
	KidLM+	[tigers, dogs, mice]

Table 5: Lexical simplification probing comparison with our KidLM models to human labels.

Preference Probing:

- **Preferences:**
  - KidLM+ confidently suggests child-friendly foods like "chicken" and "noodles" vs. RoBERTa's adult-oriented "sushi" and "seafood."
- **Emotions:**
  - KidLM+ captures common childhood fears, suggesting "spiders" and "everything" vs. RoBERTa's less specific "death."
- **Wishes:**
  - KidLM+ accurately reflects children's birthday desires ("chocolate," "cake") with high confidence.

Type	Probe Query	Models	Completions
Preferences	"My favorite food is [MASK]."	**RoBERTa**	'pizza' (0.119), 'sushi' (0.079), 'rice' (0.038), 'pasta' (0.037), 'seafood' (0.037)
		KidLM	'chicken' (0.258), 'spaghetti' (0.135), 'pizza' (0.038), 'pancakes' (0.03), 'burgers' (0.027)
		KidLM+	'chicken' (0.34), 'spaghetti' (0.18), 'noodles' (0.098), 'soup' (0.063), 'spinach' (0.024)
Emotions and Feelings	"I am scared of [MASK]."	**RoBERTa**	'death' (0.132), 'him' (0.06), 'it' (0.044), 'spiders' (0.039), 'them' (0.038)
		KidLM	'spiders' (0.117), 'everything' (0.087), 'heights' (0.079), 'dogs' (0.062), 'bugs' (0.037)
		KidLM+	'spiders' (0.189), 'everything' (0.086), 'cats' (0.077), 'bugs' (0.057), 'snakes' (0.051)
Wishes and Desires	"On my birthday, I want [MASK]."	**RoBERTa**	'you' (0.096), 'this' (0.054), 'nothing' (0.046), 'more' (0.033), 'chocolate' (0.026)
		KidLM	'cake' (0.246), 'chocolate' (0.132), 'something' (0.063), 'presents' (0.044), 'nothing' (0.021)
		KidLM+	'chocolate' (0.527), 'cake' (0.081), 'stars' (0.034), 'candy' (0.032), 'puppies' (0.022)

Table 6: Output completions grouped by types, providing qualitative insights into model behaviors.

Future Directions

1. **Pre-training Data:**
   - Need more pre-training data than what is available in the current KidLM corpus.
   - User-centric data collection pipeline is extensible, allows integration of new sources.
2. **Post-training Alignment**
   - Base LLMs are insufficient for serving as kid-friendly conversational assistants.
   - A small set of examples (**e.g., 1,000 examples**) can achieve significant alignment performance.
3. **Human-Centered Evaluation of LLMs**
   - Need an evaluation framework that integrates HCI and NLP insights.
   - Involves multiple stakeholders at different stages:
     - **a) Pre-deployment:** Educators, psychologists, parents.
     - **b) Post-deployment:** Children, parents, educators.