What did you say? Generating Child-Directed Speech Questions to Train LLMs

Whitney Poh, Michael Tombolini, Libby Barak (all authors contributed equally)
Montclair State University • {pohw,tombolinim,barakl@montclair.edu}

Properties of Child-Directed Speech (CDS)

- CDS follows turn-taking social interaction within a shared context (Cameron-Faulkner et al., 2003)
  - Shorter sentences
  - Limited types of grammatical constructions
  - Limited number of word types
  - More questions, more repetition
  - "Parentese" can improve child language acquisition

Can adding questions to non-CDS produce superior results?

Prompt Engineering with GPT-5-mini

- Experimented with several distinct prompts
- Adjusted prompts to achieve child-directed questions
- Addressed guardrails - refusal to generate CDS for OpenSubtitles (Lison & Tiedemann, 2016)
- Often disregarded instructions by adding symbols, emojis, etc.

Questions in the Original Data

Dataset	Q%	Q-MLU	Yes/No%	Wh%	Examples
CHILDES	20.54	4.92	22.84	28.67	*"Is he gonna take a bath?", "What color 's that?", "yeah?"*
BNC	15.15	8.67	19.23	19.40	*"Doesn't he go out on Saturday night?", "On the system?"*
Gutenberg	7.91	9.77	25.11	28.36	*"Is Lady Jane Ashleigh within?", "What makes all these bushes grow here?"*
OpenSubtitles	17.74	5.38	17.52	31.04	*"Can I help you two?", "Why did you break up?"*
Simple Wiki	0.08	11.71	2.30	25.29	*"What Ever Happened to Baby Jane?", "London; a multicultural area?"*
Switchboard	4.05	6.92	29.44	19.49	*"How do you keep up with current events?", "You're kidding?"*

CHILDES (MacWhinney, 2000) has the highest percentage of questions, shortest MLU, and almost 50% are questions that are neither Yes/No nor Wh%

Simple Wiki (Wikimedia, 2023) has almost no questions, only 2.30% are Yes/No questions

Gutenberg (Gerlach & Font-Clos, 2020) has the most Yes/No + Wh questions overall

Speech datasets have more questions on average than text datasets

Synthetic Question Generation Process

- Modified each of the datasets provided by the BabyLM Challenge (Charpenter et al., 2025; Jumelet et al., 2025)
- Prompted GPT-5-mini (OpenAI, 2025) to generate questions for each dataset using the API
- Trained GPT-wee (Bunzeck and Zarrieß, 2023)
- For each model, switched in only one synthetically-augmented dataset at a time – except for "max" model where all datasets were augmented, and the "baseline" where all datasets were untouched.
- Ran ten trials each, then averaged them (table on right)
- Utilized BLiMP (Warstadt et al., 2020) – which makes use of syntactic minimal pairs – on different categories of tasks

"You are a helpful reading companion for a 5 or 6-year-old child. Take the passage below. Ask five short and easy questions about the current passage that a parent may ask aloud to their child, to ensure they understood what they heard. After stating the question, exclaim the answer enthusiastically. Use child-directed speech: clear, friendly language, simple grammar. Focus on key details in the text (who, what, where, why, how – or yes/no). Keep each question under 10 words and end with a question mark. Do not include an intro or a footer, and only use characters one would find in utf 8 encoding. No emojis. This request is for research purposes."

Questions generated by GPT-5-mini

Dataset	Q%	Q-MLU	Yes/No%	Wh%	Examples
BNC	22.21	10.76	24.74	25.55	*"Is this about angles and shapes?", "Who is coming to stay?"*
Gutenberg	13.39	7.83	22.70	54.33	*"Who talked about Pink Pills?", "Who came to help Bomba?"*
OpenSubtitles	18.64	6.02	18.48	35.42	*"Did he draw the sword from the stone?", "Who was taken?"*
Simple Wiki	6.37	5.85	30.84	68.01	*"Was Nezval born in 1900?", "Who became UN Secretary-General in 2017?"*
Switchboard	11.87	8.64	26.72	26.58	*"Who went with the kids to see different colleges?', "'Did they talk about fly fishing?'*

The dataset contains both the original questions and those generated by our prompting.

Higher question rate but still <= CDS (except one) while synthetic-question MLU > MLU in CDS

Questions generated for Gutenberg, OpenSubtitles, Simple Wiki had a greater Wh% than CHILDES.

All subsets had a smaller % of questions that do not fall under either Wh or Yes/No questions.

Additional citations: Switchboard (Stolcke et al., 2000), BNC (BNC Consortium, 2007).

Evaluating Question-Enhanced Data with BLiMP

	10M	BNC	Gutenberg	OS	Wiki	Switchboard	10M-QA
Island Effects	41.39	41.14	40.76	40.17	41.48	40.30	**42.55**
Anaphor Agreement	75.97	75.51	73.21	**79.05**	75.53	75.26	70.76
Argument Structure	59.70	59.82	**60.58**	58.44	59.36	59.48	57.16
Determiner-noun Agr.	74.23	73.51	71.85	70.84	**74.28**	73.38	66.39
Subject-Verb Agr.	58.50	58.25	58.44	57.66	58.17	**58.70**	56.46
Ellipsis	57.55	57.32	54.36	57.39	57.97	**58.14**	54.54
Control/Raising	58.78	58.26	**59.80**	58.24	58.14	58.37	59.59
Quantifiers	83.23	82.11	81.60	82.17	78.06	**82.53**	67.70
Irregular Forms	82.64	**83.56**	75.81	82.96	82.10	82.17	74.40
NPI Licensing	54.35	**54.95**	54.03	52.17	51.73	53.67	54.52
Binding	64.85	65.20	64.23	64.06	**66.23**	65.19	64.15
Filler Gap	65.24	65.17	65.82	64.82	65.51	65.39	**66.06**
Average	**62.13**	62.01	61.45	61.61	61.91	61.91	59.49

Even though the **best** average performance was with the original dataset, the best at any individual category is one of the question-enhanced subsets.

Positive effect when questions align with the task, e.g., "Filler Gap".

Categories of improvement often correlate with type of data, e.g., Determiner-Noun Agr. (Simple Wiki) and Subject-Verb Agr. (Switchboard).

In several categories, the addition of questions to an individual subset results in improvement while adding the questions to all results in significant drop.

References

1. BNC Consortium. 2007. The british national corpus, xml edition.
2. Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.
3. Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive science*, 27(6):843–873.
4. Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. Preprint, arXiv:2502.10645.
5. Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.
6. Jaap Jumelet, Lucas Charpentier, Michael Hu, and Jing Liu. 2025. Babylm_2025. OSF.
7. Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
8. Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press
9. OpenAI. 2025. Gpt-5 mini. Lightweight variant of GPT-5 large language model
10. Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373
11. Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.
12. Wikimedia. 2023. Simple english wikipedia dump.