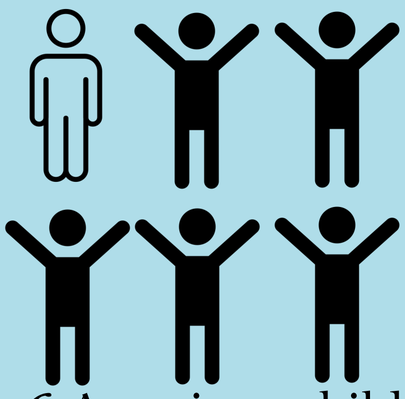
How to Teach a Large Language Model to Tell Us About Childhood Development

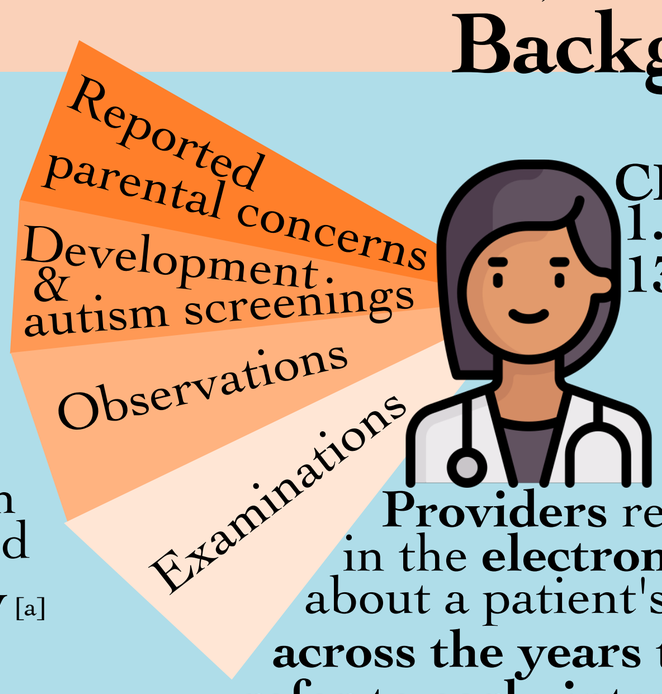
Raven (Ruiwen) Tang [School of Engineering and Applied Science '27]
Ian Campbell, MD, PhD [Perelman School of Medicine, Pediatrics]

Thank you to: Matt Donati, PhD [Children's Hospital of Philadelphia, Biomedical & Health Informatics]
Xinwei Zhao, PhD [Children's Hospital of Philadelphia, Biomedical & Health Informatics]

Background

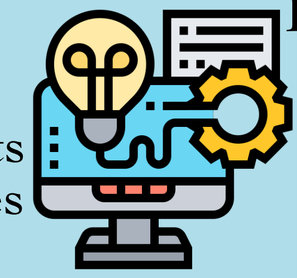


1 in 6 American children aged 3-17 has a diagnosed **developmental disability** [a]



CHOP EHR:
1.5 million patients
133.2 million notes

Providers record information in the **electronic health record** about a patient's development **across the years** to **flag delays** & **refer to early intervention services**



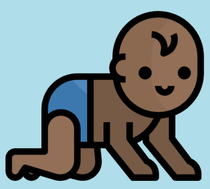
Recent meteoric rise of **LLMs**
Large = Trained on a massive amount of data
Language = Capable of processing & generating natural language
Model = Built on neural networks

Good at & can be trained to be **even better at:**
- **Comprehending** free text
- **Extracting** details
- **Summarizing** in narrative style

Objective

Starting Point: Base Model (Meta Llama 3 8B Instruct+ One epoch of CHOP EHR)

Has: Read **CHOP's EHR notes**, so it knows *something* about patients
Has Not: Been trained for **medical assistance purposes**
So: Neither completely **helpful** nor completely **trustworthy** for pediatrics



Goal Endpoint: Fine Tuned Model

Asked about a specific patient, will be capable of

Extracting 5 milestone achievement ages:
- Sitting independently
- Speaking first words
- Walking independently
- Handing objects to others
- Speaking in sentences (3+ words)

Providing a developmental summary:
- Diagnosed disorders
- Screening results
- Areas of concern
- Loss, regression, plateauing
- Early intervention services
- Potential resolutions

Eventually with citations!
Generated responses **preferred by pediatricians** over base model

Methods

LLM Fundamentals
- **Python** language & packages
- Secure work in **Arcus** (connected clinical & research data)
- Deployment of jobs on **Respublica** high performance computing cluster with GPUs
- Practice with **embeddings**, **tokenizers**, **inference**, **chunking strategies**, **vector search**, & **retrieval augmented generation**

Clinical Understanding
- **Human research protection** training
- **Research** into childhood development
- **Consultation** of clinical subject matter experts & LLM collaborative
- Acquaintaince with **Epic EHR notes**
- Definition of **model task**

Training Data Generation
- Annotation **tool** & **workflow** development
- **Annotation guide** creation & maintenance
- **Broad coverage search** for edge patients
- Manual **extraction** & **summarization**
- **De & reidentification** of patients & notes
- **Prompt engineering** & **chat templating**

Data Pipelining + Fine Tuning
- **SQL** querying
- **JSON**, **dataframe**, and **text file** dumping
- **Dataset format compatibility** with machine learning packages
- **Training** epoch runs
- **Low-rank adaptation** technique

Benchmarking + Error Analysis
- LoRA **parameter experimentation**
- **Sample answer generation** with base & fine tuned models for evaluation
- Human, non-expert, & clinician **review** & **judgement**

Results

6-page electronic health record **annotation guide** for childhood development, **reviewed** by pediatricians & health informatics specialists

84 **annotated patients**, each with **5 extracted milestones** and a **50-600 word developmental summary**, accompanied by **note citations**

Scripts for:
Annotation assistance
Data pipelines
Fine tuning training
Inference

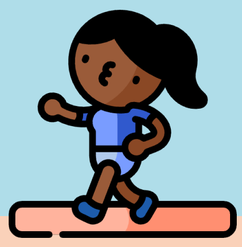
1,008 dataset rows for training, containing patient metadata & corresponding example chats

Generated post-training examples:
+Matched style +Relevance -Inaccurate

Discussion + Conclusions

Takeaway
It's hard to train a model to be **relevant *and* correct** (and to know when the model is doing a "good job")

Next Steps (getting from **relevant** to **relevant and correct**)
- Increased **quantity** & **inter-annotator reliability** of training data
- Addition of **citation capabilities**
- Potential integration with **retrieval augmented generation**
- **Reinforcement** learning
- Development of more precise **benchmarking standards**
- Reduction of **hallucinations**



PURM Penn Undergraduate Research Mentoring Program
PennCURF CENTER *for* UNDERGRADUATE RESEARCH & FELLOWSHIPS

Icons designed by Freepik
Resources from Flaticon.com

[a] Benjamin Zablotsky, Lindsey I. Black, Matthew J. Maenner, Laura A. Schieve, Melissa L. Danielson, Rebecca H. Bitsko, Stephen J. Blumberg, Michael D. Kogan, Coleen A. Boyle; Prevalence and Trends of Developmental Disabilities among Children in the United States: 2009–2017. Pediatrics October 2019; 144 (4): e20190811. 10.1542/peds.2019-0811