On-Device LLMs: Unleashing the Power of Conversational AI for Mobile and Wearable Devices

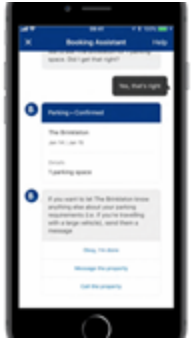
Work done with Patrick Huber, Akshat Srivastava, Arash Einolghozati, Rylan Conway et al.
Paper link: https://arxiv.org/abs/2408.11219

Kanika Narang, Meta AI

Agenda

- Importance of On-device LMs
- Related Work in Data Distillation
- Conversational Distillation Methodology
- Distilling Conversational Abilities
- Experiments and Models
- Results and Analysis
- Conclusion and Future Work

Innovative Applications of Virtual Assistants



airbnb

AI in Travel

AI enhances travel with personalized recommendations and automated booking systems.

Copilot

Amazon Code Whisperer

AI in Coding

AI tools assist in code generation, debugging, and optimization for developers.

wysa

medwhat

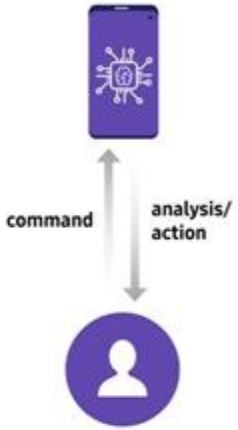
Woebot

AI in Healthcare

AI improves diagnostics, patient care, and operational efficiency in healthcare.

Server LLM

On-Device LM



Benefits of On-Device LM

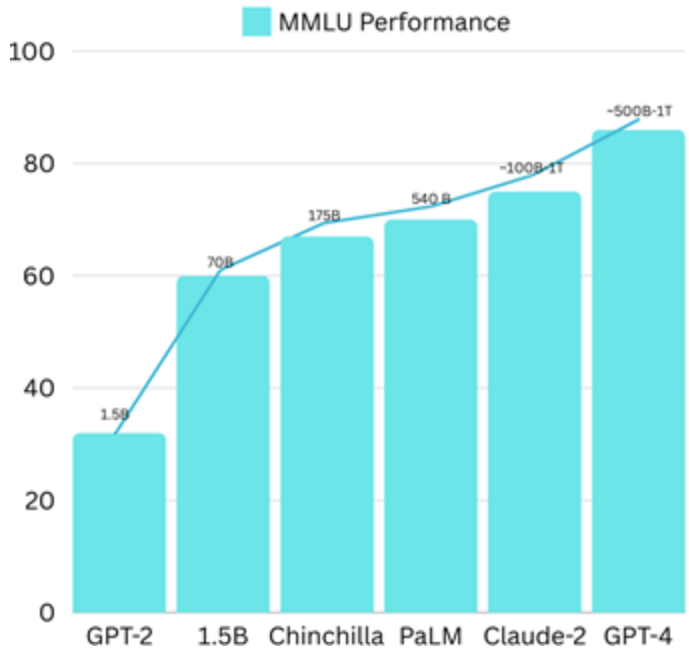
Privacy Protection

Low Latency

Low Power & Low Cost

Task-specific LMs

- On-Device Language Models (ODLMs) (<1.5B) face challenges in generalization due to limited parameters, impacting their conversational capabilities.
- They are better suited as task specialists rather than generalists, focusing on specific tasks like quick reply, summarization, instruct messaging etc.
- Task-specific training allows for improved performance in targeted applications.



Training Data Quality

- Multi-turn conversation reasoning is not at par with humans for large language models.
- The main reason is lack of good quality, large-scale multi-turn conversational data. High-quality conversational datasets are scarce, often domain-specific, and resource-intensive to create.
- Data distillation from larger models is a cost-effective way to generate large-scale and high-quality training data.

Data Distillation Requirements

Quality in Data Distillation

Gunasekar et al. emphasize refining datasets for ODLMs' effective learning.

Diversity of Datasets

Wei et al. advocate for diverse samples to improve model performance.

Synthesis Techniques

Innovative techniques such as topic switches, text rewriting and discourse units should be used create large-scale datasets, enhancing ODLM training.

CoDi: our novel data distillation framework synthesizes large-scale, diverse and conversation-style datasets.

By training on these high-quality datasets, CoDi improves the performance of On-Device Language Models in specific conversational tasks. This enables ODLMs to better understand human conversation, enhancing applications like customer service, virtual assistants, and chatbots.

Conversational Distillation Methodology

Conversational Graph Generation

Inspired by Markov Chains, defines 'conversation links' or turns as vertices connected to each other.

Each edge has transition probabilities to ensure diverse multi-turn conversations.

Final sampled graph is used as a blueprint for prompting the teacher model to generate diverse and natural conversations.

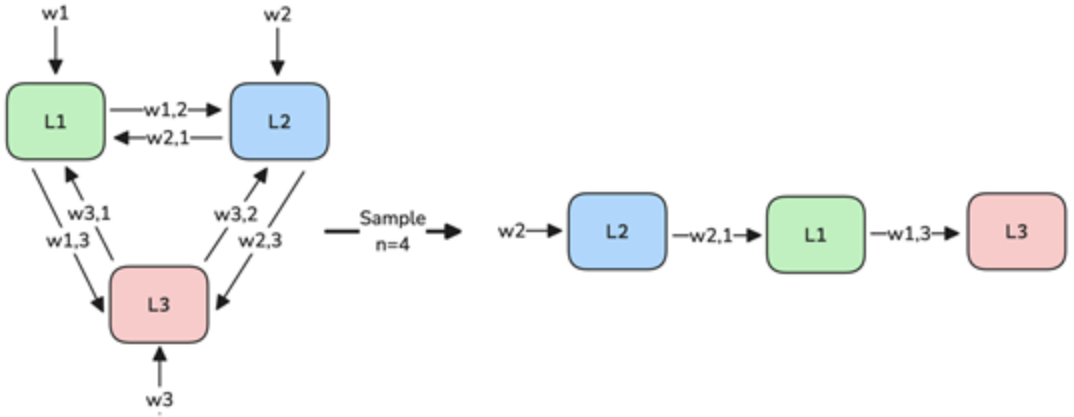


Figure 1 Conversational Graph Generation Example. Left: General conversational graph, Right: Rolled out version of a sampled graph at length n=4

Conversational Distillation Methodology

Link Execution

Executes each link sequentially, using prompts to guide the model in generating the next conversational turn, ensuring dynamic flow.

Uses additional seed data to support diversity in the conversational chain or,

Directly add external information as auxiliary data(e.g., context).

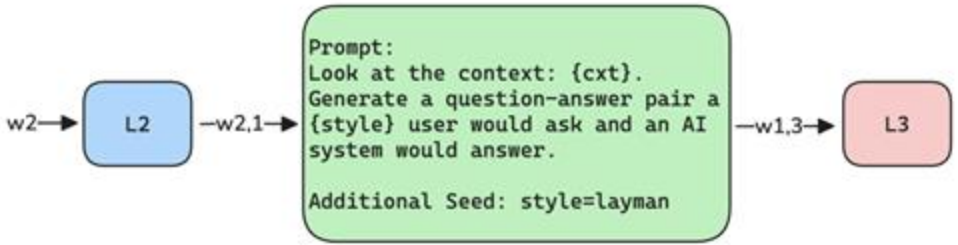


Figure 3 Per-turn conversational link augmentation with prompt and (optional) seed data

Conversational Distillation Methodology

Linguistic Phenomena

Incorporate linguistic features, such as coreference and discourse markers, to create natural, human-like interactions.

This ties together conversation turns, enhancing coherence and engagement.

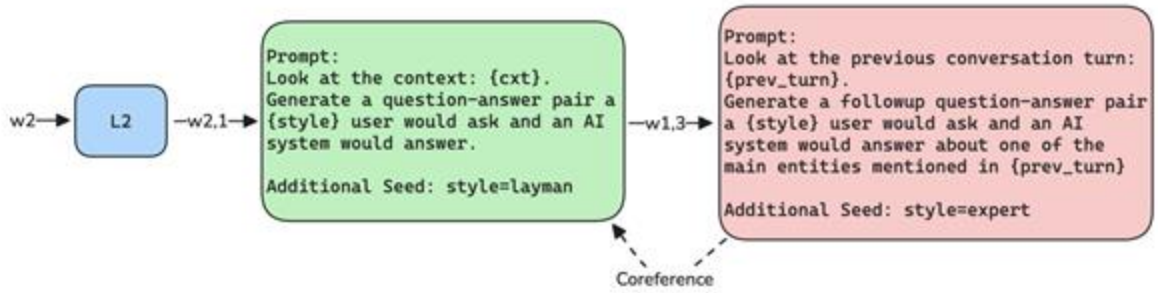


Figure 4 Example of linguistic phenomena used in the final turn prompt.

Final Distillation Architecture

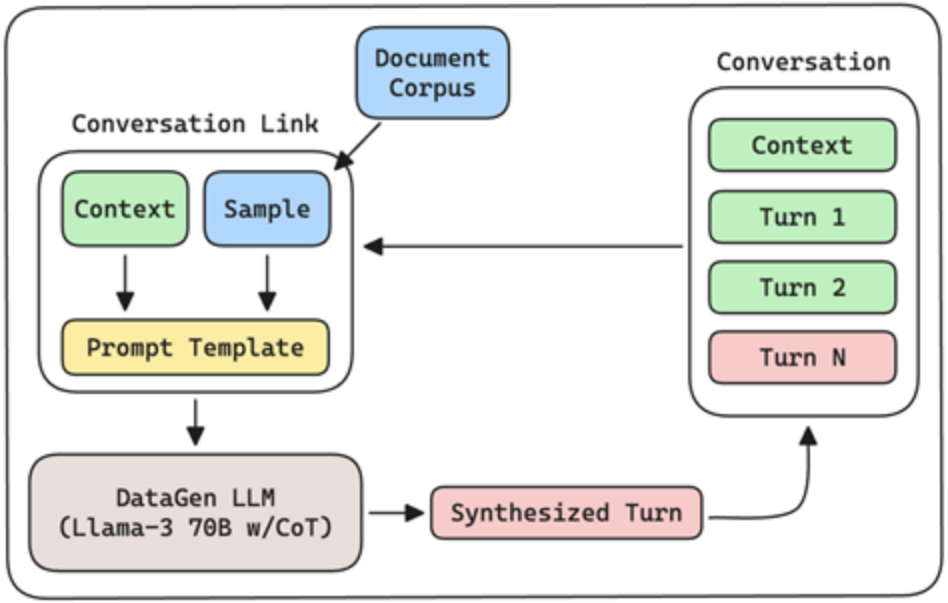


Figure 2 A single generation step to synthesize a new turn in the conversation.

Vancouver is [...] the most populous city in the province, the 2016 census recorded 631,486 people in the city, up from 603,502 in 2011. The Greater Vancouver area had a population of 2,463,431 in 2016, making it the third-largest metropolitan area in Canada. Vancouver has the highest population density in Canada with over 5,400 people per square kilometer. With over 250,000 residents, Vancouver is the fourth-most densely populated city in North America behind New York City, San Francisco, and Mexico City according to the 2011 census. Vancouver is one of the most ethnically and linguistically diverse cities in Canada according to that census; 52% of its residents have a first language other than English...

Q: Which country is Vancouver in?
A: Canada
Q: What is Vancouver's ranking in terms of population density in North America?
A: fourth
Q: What cities are ahead of it in terms of population density?
A: New York City, San Francisco, and Mexico City
Q: What is the population of the Greater Vancouver area?
A: 2,463,431
Q: What is its population density?
A: over 5,400 people per square kilometer

Figure 5 CoDi synthesized example. Context document taken from the CoQA training corpus.

Data Format Update

Now supports flexible roles for dynamic interactions beyond USER and AGENT.

Includes CONTEXT role for essential background information to improve accuracy.

Introduces Role Weighting to emphasize important roles in training process.

```
[CONTEXT]<User Message
          Inbox>[/CONTEXT]

 [USER] Any new messages for
me? [/USER]

[AGENT] You have 2 messages from
Alex. He asks about your weekend
    plans. Reply? [/AGENT]
```

Experiments and Models

Experimental Setup

We focused on the task of grounded reasoning, where ODLMs interact with on-device context to provide accurate responses. This scenario is ideal for evaluating conversational abilities.

Teacher and Student Models

The teacher model utilized was the 70B Llama3 instruction-tuned checkpoint, while the student model was a 1.4B Llama2-style model, allowing for effective knowledge distillation.

Synthesis Datasets

We synthesized diverse conversational data using CoDi, leveraging intra-domain sources from CoQA and QuAC to create a robust training dataset for ODLMs. Also generated zero-shot large-scale dataset generation using web documents.

Evaluation Datasets

To assess performance, we employed the original test sets from CoQA and QuAC, comparing models against human-annotated baselines in both single-turn and multi-turn settings.

Conversation Grounded Reasoning Performance

In the intra-domain setting, model trained on CoDi synthesized data closes the gap between single-turn and multi-turn human baselines.

Even in zero-shot settings, CoDi Web models consistently outperforms instruction-tuned baselines (incl. Phi-3) at similar scale and above.

Small gap between the conversation history settings (i.e. “Gold” and “Pred”) suggests a more coherent conversational trajectory across multi-turn conversations.

Eval Dataset		CoQA		QuAC	
Metric	#Params	Recall		F1-score	
Context		Gold	Pred	Gold	Pred
In-Domain					
Human Single Turn	1.4B	77.3	73.7	36.73	30.80
CoDi In-Domain	1.4B	84.5	81.8	40.97	38.42
Human Multi-Turn	1.4B	**85.2**	**82.5**	**47.20**	**41.02**
Zero-Shot					
Instruction-Tuned	1.4B	79.7	68.9	21.37	18.04
CoDi Web	1.4B	**86.3**	**84.2**	**38.66**	**35.51**
Phi-3	3.8B	89.0	78.8	34.56	16.24
Instruction-Tuned	7B	85.0	82.8	25.99	17.58
CoDi Web	7B	**91.0**	**89.3**	**39.63**	**37.41**
Instruction-Tuned†	70B	–	–	32.47	–

Eval Dataset		CoQA		QuAC	
Metric	# Params	Recall		F1-score	
Context		Gold	Pred	Gold	Pred
CoDi Web 10k	1.4B	80.8	78.4	36.69	31.50
CoDi Web 100k	1.4B	83.1	80.9	37.51	33.92
CoDi Web 1M	1.4B	**86.3**	**84.2**	**38.66**	**35.51**

Larger synthesized datasets lead to better model performance.

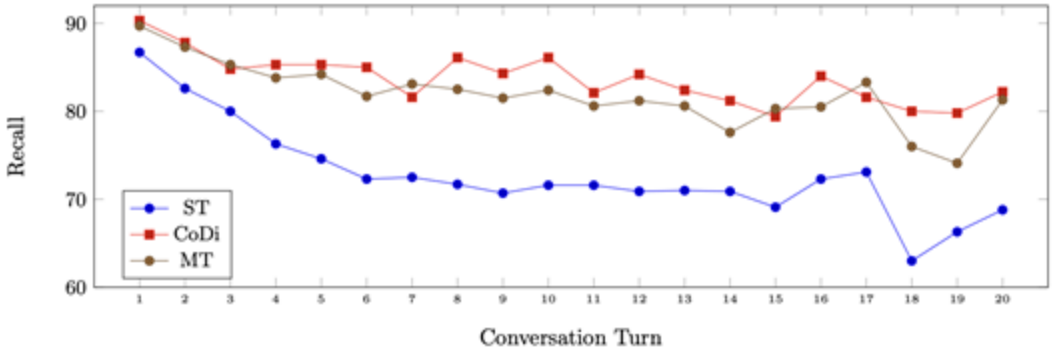


Figure 7 Average Per-Turn Model Recall on CoQA.

CoDi outperforms in longer conversations.

Multimodal Grounded Reasoning

Eval Dataset		VQAv2	VizWiz
Metric	#Params	Acc	Acc
Flamingo	3B	49.2	28.9
Flamingo	9B	51.8	28.8
Blip2	3.4B	**62.3**	29.4
Instruction-Tuned	1.4B	49.5	28.7
CoDi Web	1.4B	51.8	**32.4**
SG MM-LLM	70B	67.8	41.3

Table 5: Multimodal GQA Results on visual QA tasks.

We replace the language backbone of the multimodal model with CoDi Web fine tuned model and then train the vision adaptor on the captioning data.

The model is evaluated on vision grounded conversational datasets in zero shot settings.

CoDi Web models consistently outperforms instruction-tuned baselines that are double its size.

Conclusion and Future Work

CoDi demonstrates significant effectiveness in synthesizing diverse and steerable conversational data, enhancing the capabilities of On-Device Language Models (ODLMs).

The framework achieves competitive results compared to models trained on human-annotated datasets, addressing the limitations of ODLMs in conversational settings.

Future research could focus on expanding CoDi's applications to additional domains, improving synthesis quality, and exploring cross-lingual capabilities.

Investigating the integration of more sophisticated linguistic features and adaptability for specific user interactions remains a key area for enhancement.

THANK YOU

EMAIL :

kanika13@meta.com

LINKEDIN:

@kanika_narang