CoSy: Conversational Synthesis for Grounded Question Answering

Patrick Huber, Arash Einolghozati, Rylan Conway, Kanika Narang,
Matt Smith, Waqar Nayyar, Adithya Sagar, Ahmed Aly and **Akshat Shrivastava**
Meta Reality Labs

Abstract

High-quality, large-scale conversational datasets are scarce, making it difficult to train on-device language models (OD-LLMs, ∼1B parameters) as effective assistants. We introduce **CoSy** (**Co**nversational **Sy**nthesis), a novel framework for generating diverse, steerable, multi-turn conversations at scale. CoSy combines three key mechanisms: (1) conversational graphs that ensure natural dialogue flow, (2) turn-based prompt augmentations for diversity, and (3) explicit linguistic phenomena for coherence. We evaluate CoSy on conversational grounded reasoning tasks (i.e., answering questions based on contextual information), a core on-device use case. Our on-device sized models trained on CoSy-synthesized data achieve competitive performance with human-annotated baselines and outperform instruction-tuned models of up to 70B parameters in zero-shot settings.

1 Introduction

Existing high-quality conversational datasets are rare, small, and often cover only narrow domains (Duan et al., 2023). Furthermore, obtaining human annotations for multi-turn data is difficult, since every sample requires a valid and meaningful conversational history. As a result, annotating multi-turn datasets is either significantly more resource-intensive (at the same scale) or yield much smaller datasets at the same resource contraint, e.g., CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), or OpenAssistant (Köpf et al., 2023). This leads to a shortage of suitable training resources for conversational models in terms of both volume and diversity. Here, we address this gap with our novel **Co**nversational **Sy**nthesis (**CoSy**) approach, which synthesizes diverse and steerable multi-turn conversations. Using our high-quality, high-quantity conversational data, we train on-device sized language models (around 1B parameters) as assistants. In contrast to large language models, these smaller alternatives generally do not exhibit the same level of generalization and cannot be easily prompted or few-shot trained (Fu et al., 2023). Instead, they require large amounts of high-quality training data to explicitly learn conversational abilities.

Training on-device sized language models on our synthetic conversation data, we find that the resulting models achieve competitive results with models fine-tuned on human-annotated datasets and consistently outperform similar-sized instruction-tuned baselines in zero-shot settings.

To summarize, our work makes the following contributions:

- We propose CoSy, a novel framework for synthesizing diverse, multi-turn conversations using conversational graphs, turn-based augmentations, and explicit linguistic phenomena.
- We demonstrate that 1.4B parameter models trained on CoSy data close 75–92% of the gap between single-turn and human multi-turn baselines on CoQA and QuAC.
- We show that CoSy-trained models outperform instruction-tuned baselines up to 50× larger in zero-shot conversational grounded reasoning.

2 Method

We propose CoSy, a data synthesis methodology that enables OD-LLMs to function as conversational assistants. Unlike prior data synthesis work focused on single-turn instruction tuning (e.g., WizardLM (Xu et al., 2023), Alpagasus (Chen et al.), Alpaca (Taori et al., 2023)), CoSy generates *multi-turn conversations* using a turn-by-turn paradigm. This design directly addresses two limitations of LLM-based synthesis: limited diversity and lack

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1–10

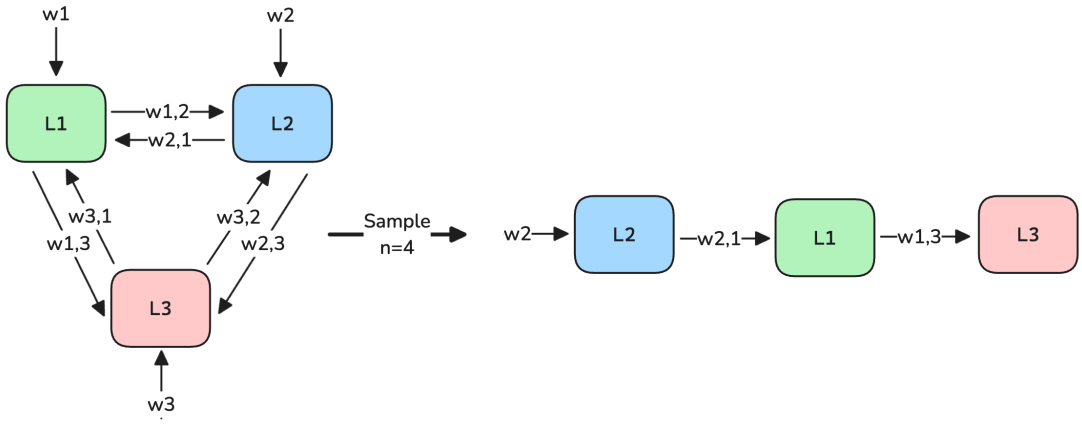

Figure 1: Conversational Graph Generation Example. Left: General conversational graph. Right: Rolled-out version of a sampled graph of length $n = 3$.

of steerability at the conversation level. Our framework introduces three key mechanisms:
Conversational Graphs (§2.1): Ensure valid, natural conversation structure at the marco-level
Conversational Links (§2.2): Enable per-turn diversity and steerability
Linguistic Phenomena (§2.3): Tie turns together in a natural linguistic style

2.1 Conversational Graph Generation

Synthesizing diverse yet naturally flowing conversations is essential for imitating human interactions. To provide guarantees on conversation validity, we propose a conversational graph generation approach inspired by Markov Chains. Specifically, we model conversation structure as a Markov process over *conversation links*—reusable templates for generating individual turns. Like Markov Chains, our approach defines transition probabilities between states (links), enabling us to sample valid conversation trajectories while maintaining diversity.

Formally, a conversational graph $G = (V, E)$ consists of vertices $V = \{\emptyset, L_1, L_2, \ldots, L_k\}$ representing conversation links, and edges E representing transition probabilities between links. The special vertex $\emptyset$ denotes the conversation start. To generate a conversation of length n:

1. Initialize at $\emptyset$
2. Sample an outgoing edge according to transition probabilities
3. Instantiate the corresponding link to generate a turn
4. Repeat until n turns are generated

Figure 1 illustrates an example. The graph defines a set of conversation links (vertices), each serving as a blueprint for prompting a conversational turn. These are connected by transition probabilities (edges) used to sample a *conversational chain*—a sequence of links representing a valid and natural multi-turn conversation.

For instance, given a graph $G = (V, E)$ with vertices $V = (\emptyset, L_1, L_2, L_3)$, generation begins by sampling from the valid edges leaving $\emptyset$ (e.g., $\{w_{\emptyset,1}, w_{\emptyset,2}, w_{\emptyset,3}\}$). If L_2 is selected, the next step samples from $\{w_{2,1}, w_{2,3}\}$. Once the target conversation length n is reached (e.g., $n = 3$ in Figure 1), the traversal ends and the resulting conversational blueprint is returned for synthesis.

2.2 Conversational Links

Once a conversational chain is sampled, links are executed in order (e.g., in Figure 1: $L_2 \to L_1 \to L_3$). While the conversational graph defines the "macro-level" conversation structure, individual links define the conversational turns themselves. Each link contains: (1) a prompt to steer the conversation, and (2) additional seed data to support diversity in the conversational chain or to directly incorporate external information (e.g., context; see Figure 3)[1]. The prompt and seed are configurable per link, providing the primary mechanism for encoding explicit conversational phenomena. Figure 2 shows the data flow in a single generation step.

2.3 Linguistic Phenomena

Without additional constraints, the generated turns could still be a sequence of independent, single-

[1]The prompt can also contain "Chain-of-Thought" reasoning steps (Wei et al., 2023).

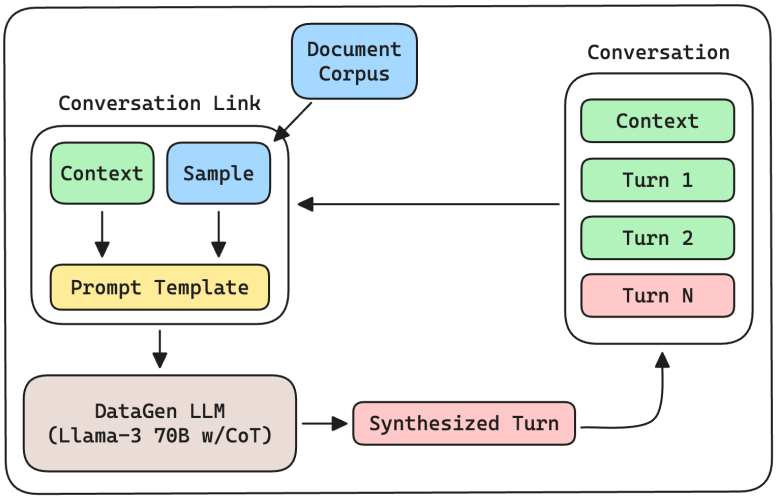

Figure 2: A single generation step to synthesize a new turn in the conversation.

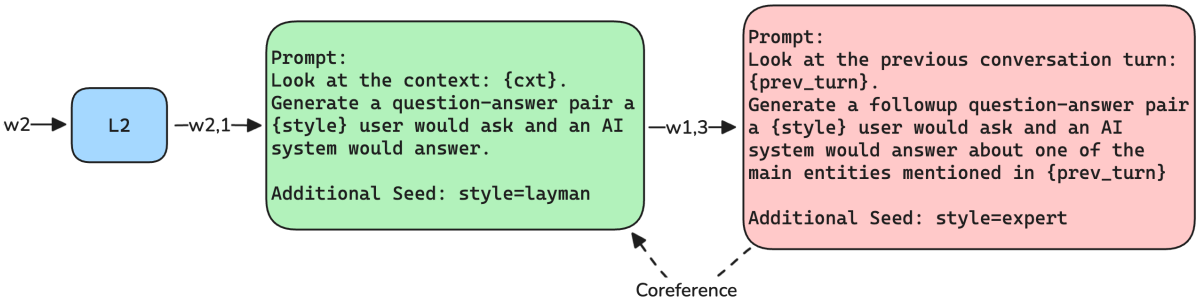

Figure 3: Example of linguistic phenomena used in the final turn prompt.

turn utterances. Inspired by human conversations, we use explicit linguistic phenomena to naturally tie turns together. Figure 3 shows the details of links supporting coreference to refer back to prior conversational turns. This way, we explicitly connect entity mentions in the context and synthesize semantic follow-ups, completing our steerable and diverse conversation synthesis framework.

3 Experiments

To demonstrate the effectiveness of CoSy, we evaluate it on assistant-style grounded reasoning. This is a natural testbed for our synthesis approach using small, specialist language models at on-device scale, where the interaction with on-device context (e.g., summarizing notes) is a primary use case.

3.1 Models

In this paper, we use two main models:
LLM: The 70B Llama3 instruction-tuned checkpoint as a large-scale baseline.
OD-LLM: A 1.4B Llama 2-style model trained on our synthesized data.

We use additional model sizes and architectures for ablation experiments. Specifically, we also employ the 70B Llama 2 instruction-tuned checkpoint as an alternative large language model (Touvron et al., 2023), and a pre-trained 500M Llama 2-style model as an alternative OD-LLM. As instruction-tuned baselines, we explore 7B (Touvron et al., 2023), 1.4B, and 500M Llama 2-style models, as well as Phi-3 (Abdin et al., 2024). Additional baselines are taken from the literature and cited in the relevant sections.

3.2 CoSy Seed Data

To evaluate our CoSy data synthesis approach, we explore two synthesis scenarios: in-domain and zero-shot.

In-Domain Synthesis: We compare the Llama 2 1.4B model trained on our synthetic conversations against gold datasets, using the same set of available documents during training (see in-domain contexts (blue) in Figure 4). Specifically, we use two common multi-turn grounded question answering tasks: Conversational Question Answering (CoQA)

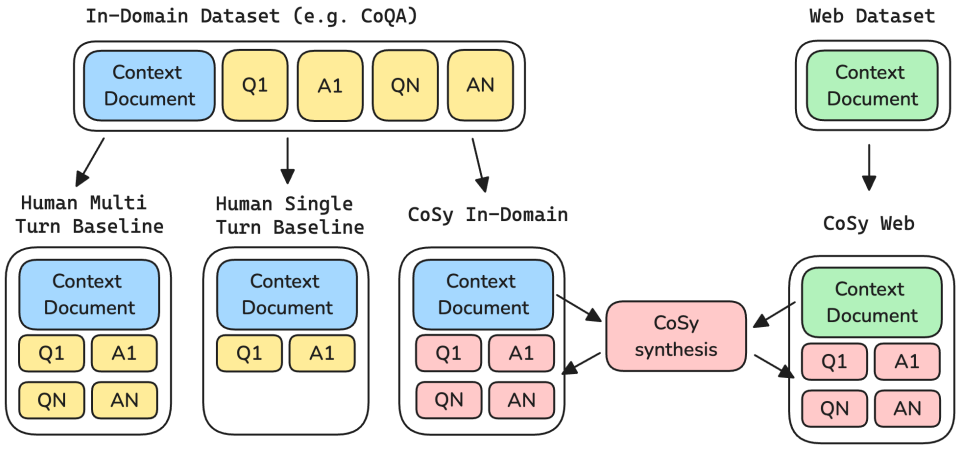

Figure 4: Training setup of human-annotated baselines and in-domain / web-based CoSy versions.

Dataset	CoQA	QuAC
Dialogs	8,199	13,594
Questions	121,300	98,407
Passage length	270	401
Avg. Turns	15	7

Table 1: Dataset dimensions of the in-domain grounded reasoning training portion.

(Reddy et al., 2019) and Question Answering in Context (QuAC) (Choi et al., 2018). Both tasks test a system's ability to generate responses to a query based on a given context and conversational history. Dataset statistics are shown in Table 1. While both datasets target conversational question answering, CoQA answers are generally factoid-style (i.e., short and precise), whereas QuAC responses are more elaborate and longer. We generate synthetic conversations using the approach described above and train the Llama 2 1.4B model with the in-domain synthesized data, hereafter called **"CoSy In-Domain"**.

Zero-Shot Synthesis: As shown in the bottom right of Figure 4, in this setting we use web documents as the context for question answering. Given the automatic nature of our approach, we are limited only by the amount of available seed data. As a result, we can leverage large-scale data collections to scale CoSy across several orders of magnitude. To demonstrate this scaling capability, we use readily available web data, generate conversations, and train a Llama3 1.4B pre-trained checkpoint. We name this setting **"CoSy Web"**.

3.3 Evaluation Tasks

We evaluate all models on two human test sets provided as part of the CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) corpora[2]. We use the original context along with the complete grounded question-answering conversation, analogous to the human multi-turn training set shown in the top left of Figure 4. As a held-out, zero-shot evaluation scenario, we also report results on abstractive summarization using two popular datasets: CNN/DM (Nallapati et al., 2016) and XSum (Narayan et al., 2018).

Due to the nature of LLM-synthesized conversations, our generations are more aligned with the long-form responses in QuAC than the factoid-style CoQA answers. For this reason, we opt for the less strict recall metric on CoQA to avoid penalizing results based on output length. To ensure that model candidates do not exploit the recall metric by generating excessively long responses, we supplement our results with response length statistics (Table 3). Given the similar human-like style of QuAC gold answers and our generations, we report the standard F1 scores for QuAC.

3.4 Evaluation Protocol

We evaluate models in two settings that capture different aspects of conversational ability:

- **Gold History:** Each turn is predicted given the ground-truth conversation history. This isolates single-turn response quality.

[2]As is common practice, we use the validation portion of the datasets as our evaluation set, while sampling a validation set from the original training split.

- **Predicted History:** Each turn uses the model's own prior predictions as history. This tests end-to-end conversation coherence, where early errors propagate.

3.5 Baselines

Human Single Turn: This baseline uses the in-domain, single-turn, human-annotated question answering dataset. We take the original multi-turn human-annotated dataset and remove all turns past the first interaction (see top right in Figure 4).

Human Multi Turn: This baseline consists of the original, human multi-turn conversations provided as part of the CoQA and QuAC datasets (see top left in Figure 4). It is a superset of the single-turn setup and is expected to perform significantly better on tasks that require conversational reasoning.

Instruction-Tuned: A common alternative to task-specialist models are instruction-tuned checkpoints. Given that these models are solely prompted to solve a specific task (e.g., as done in Liu et al. (2024b)), instruction-tuned models serve as an important baseline in domains where training data is sparse or non-existent. We use a range of instruction-tuned baselines at different parameter sizes to verify the benefit of CoSy.

4 Results

Our experiments address two questions: (1) Can synthesized conversations match human-annotated data? (2) Can synthesis scale improve zero-shot performance? Table 2 summarizes our main findings.

The top portion shows the in-domain setting, where we synthesize grounded conversations from documents in the CoQA and QuAC training sets and evaluate on the respective test sets (blue "in-domain" examples in Figure 4). This experiment tests whether synthesized data can close the gap between human-annotated single-turn and multi-turn baselines. We find that CoSy In-Domain closes the single-turn–multi-turn gap by 92% (91% with gold context) on CoQA and 75% (40% with gold context) on QuAC, despite using exclusively synthesized conversations.

The bottom portion of Table 2 shows zero-shot comparisons of our large-scale CoSy Web model, trained on one million generated conversations, against instruction-tuned Llama 2 baselines (Touvron et al., 2023) and the 3.8B Phi-3 model (Abdin et al., 2024). CoSy Web consistently outperforms instruction-tuned baselines at similar and larger scales: it surpasses the 1.4B, 7B, and 70B instruction-tuned Llama 2 models. It also outperforms the Phi-3 baseline (2× larger) in three of four settings.

Comparing across sub-tables, we find that zero-shot synthesis data outperforms human-curated multi-turn data on CoQA, while underperforming on QuAC. Examining the performance gap between history settings ("Gold" vs. "Pred"), we observe a consistent degradation for predicted history, as expected. However, this gap is generally smaller for CoSy models, suggesting more coherent conversational trajectories across multi-turn conversations. To validate that the recall results in Table 2 are not inflated, we present a supplementary word count analysis in Table 3, confirming that our approach does not exploit the recall metric through excessively long generations.

5 Analysis

To better understand the properties of our synthetic dataset, we conduct a range of ablation experiments. Unless stated otherwise, we sample 10,000 synthesized conversations from our zero-shot web seed.

5.1 Synthesis Scale

While small-scale, human-annotated conversational datasets exist, scaling them is resource-intensive. Using CoSy, we can synthesize large amounts of diverse conversational data across several orders of magnitude. Table 4 shows the influence of the number of synthesized conversations on model performance. The trend is clear: larger synthesis scales improve performance near-linearly up to one million samples.

5.2 Student Model Size

This ablation compares the 1.4B student checkpoint used in the main results with a 500M Llama 2-style checkpoint, both trained on the full dataset. Table 5 shows a clear quality regression when moving from 1.4B to 500M parameters. However, even at 500M, our zero-shot CoSy Web model only slightly underperforms the model trained on human multi-turn data.

5.3 Per-Turn Performance

To gain deeper insight into our synthesized conversations, we perform a per-turn analysis on the

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
CoSy 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
Instruction-Tuned	7B	85.0	82.8	25.99	17.58
Instruction-Tuned†	70B	–	–	32.47	–
CoSy Web	1.4B	86.3	**84.2**	**38.66**	**35.51**
Phi-3	3.8B	**89.0**	78.8	34.56	16.24

Table 2: Conversational question answering performance. **Top:** In-domain comparison where all models train on documents from CoQA/QuAC. **Bottom:** Zero-shot comparison where CoSy Web trains on web documents only. "Gold" uses ground-truth conversation history; "Pred" uses model-predicted history. All models are Llama 2-based unless noted. †Results from Liu et al. (2024b).

Model	# Params	Average		Median		90P	
Context		Gold	Pred	Gold	Pred	Gold	Pred
In Domain							
Human Single Turn	1.4B	2.1	2.2	1	1	4	4
Human Multi Turn	1.4B	2.4	2.5	2	2	5	5
CoSy In-Domain	1.4B	4.8	4.9	3	3	10	11
Zero Shot							
CoSy Web	1.4B	4.7	4.7	3	3	11	10
Instruction-Tuned	1.4B	65.9	29.5	28	7	154	117
Phi-3	3.8B	89.1	16.3	49	5	209	27
Instruction-Tuned	7B	12.2	7.6	7	3	25	19
Gold Answer	–	2.52	–	2	–	5	–

Table 3: Response length in words on CoQA. All models are Llama 2-based unless otherwise noted.

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

Table 4: Zero-shot results on CoQA and QuAC across three synthesis scales. All models are Llama 2-based.

Eval Dataset		CoQA		QuAC	
Metric	# Params	Recall		F1-score	
Context		Gold	Pred	Gold	Pred
Human Single Turn	500M	55.5	52.2	29.28	25.50
CoSy Web	500M	72.3	70.1	30.76	29.06
Human Multi Turn	500M	**72.9**	**70.9**	**39.12**	**32.58**
CoSy Web	1.4B	80.8	78.4	36.69	31.50

Table 5: Zero-shot results on CoQA/QuAC using the 500M Llama 2-style model on the full (1M) dataset.

CoQA evaluation set using the full CoSy Web dataset. Figure 5 shows that our model matches the gold multi-turn baseline for short conversations and near-consistently outperforms it for longer ones. The single-turn baseline performs consistently worse, with the gap generally widening in later turns.

5.4 Language Modeling Evaluations

We further evaluate our final model (trained on the full 1M dataset) on a commonly used subset of language model evaluations for small language models (e.g., as used in Liu et al. (2024a); Allal et al. (2024)). Specifically, we evaluate on ARC-easy and ARC-challenge (Clark et al., 2018), BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), OBQA (Mihaylov et al., 2018), and Winogrande (Sakaguchi et al., 2021). The goal is to assess whether our grounded-reasoning-based training approach degrades performance on this diverse set of language understanding tasks compared to pre-trained and instruction-tuned alternatives. As shown in Table 6, the CoSy Web model on average

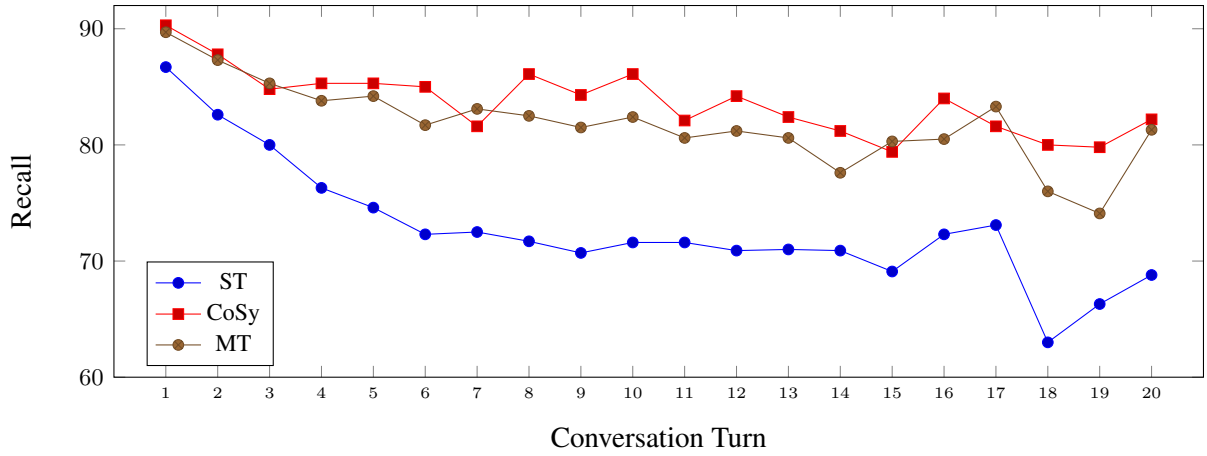

Figure 5: Average per-turn model recall on CoQA.

still underperforms instruction-tuned versions of the same base model, but improves over the pre-trained checkpoint without any prompt adjustments during evaluation. The fact that grounded reasoning training does not regress—and even slightly improves—language modeling performance suggests promise for such models to serve as generalists.

5.5 Zero-Shot Summarization

Lastly, we evaluate on zero-shot abstractive summarization using CNN/DM (Nallapati et al., 2016) and XSum (Narayan et al., 2018). Table 7 compares instruction-tuned 1.4B and 7B Llama 2-style baselines, the 3.8B Phi-3 checkpoint, and our CoSy Web model. We also include two supervised fine-tuned models for reference (fine-tuned 1.4B Llama 2, BART-Large (Lewis et al., 2019)). CoSy Web outperforms both Llama 2-style instruction-tuned baselines. While Phi-3 achieves higher scores on CNN/DM, the margin is small given the significant size difference. On XSum, CoSy Web outperforms all other zero-shot models, despite their size advantage.

6 Related Work

High-Quality Data Distillation Recently, there has been a strong push toward curating high-quality datasets for training small language models. With the intuition that small models are more sensitive to low-quality data, recent research has (1) filtered datasets based on quality, (2) rewritten data samples to improve quality, and (3) synthesized new, diverse samples to teach the model desired behaviors. For example, Gunasekar et al. (2023) use code data from the web and refine it to "textbook-style" samples for pre-training small-scale decoder-only models. Similarly, Zhou et al. (2023) argue for the importance of high-quality datasets, even for alignment purposes. Along similar lines, Wei et al. (2022); Longpre et al. (2023) show that dataset diversity along the task axis plays a crucial role in model training. In the creative writing domain, Ravi et al. (2024) show that small language models can learn difficult concepts, such as humor, when distilled in an interactive manner.

Conversational Question Answering Conversational question answering has been explored extensively given the importance of the task. Liu et al. (2024b) propose a family of conversational question answering models at large scale by adding a dense retrieval module. (Feng et al., 2020) propose a method to create conversational datasets using discourse units, while Anantha et al. (2021) propose a question rewriting method in the conversational context. Adlakha et al. (2022) publish a dataset for conversational question answering focusing on topic switches. Compared to these approaches, our conversational synthesis framework is more scalable while maintaining data diversity and steerability.

Prompting Paradigms In black-box LLM distillation, the human-curated prompt plays a major role in downstream performance. Among many proposed approaches, "Chain-of-Thought" (CoT) prompting is one of the most popular paradigms for achieving high-quality results (Wei et al., 2023). We follow the approach in Wei et al. (2023) and prompt our per-turn conversational links using a CoT flavor, asking the model to produce a reasoning trace.

Model	Arc-E	Arc-C	BoolQ	PIQA	SIQA	Hellaswag	OBQA	Winogrande	Avg
Pre-Trained	64.27	39.30	61.95	**73.75**	45.87	63.08	47.42	59.98	56.95
CoSy Web	**65.38**	38.83	67.97	73.54	46.19	61.60	47.66	60.47	57.71
Instruction-Tuned	63.65	**40.00**	**68.90**	73.54	**47.31**	**63.13**	**49.61**	**61.41**	**58.44**

Table 6: Language model evaluations of our CoSy-trained student model compared to the 1.4B Llama 2-style pre-trained and instruction-tuned baselines.

Eval Dataset	#Params	CNN/DM			XSUM		
Metric		R-1	R-2	R-L	R-1	R-2	R-L
Instruction-Tuned	1.4B	11.26	3.71	7.47	4.56	0.99	3.42
CoSy Web	1.4B	24.71	7.73	17.36	**17.66**	**2.60**	**13.49**
Phi-3	3.8B	**28.48**	**9.04**	**18.05**	12.59	2.59	9.04
Instruction-Tuned	7B	17.31	5.72	11.17	7.88	1.90	5.65
Fine-Tuned	1.4B	41.03	17.42	29.59	28.14	10.95	23.04
BART large†	406M	44.16	21.28	40.90	45.14	22.27	37.25

Table 7: Abstractive summarization zero-shot results on CNN/DM and XSum. All models are Llama 2-based unless otherwise noted. †Results from (Lewis et al., 2019).

Small Language Models Given the strong generalist performance of LLMs such as the GPT (OpenAI et al., 2024) and Llama (Touvron et al., 2023) series, the question of how much these abilities can be distilled into OD-LLMs has become an important research question. For example, Mukherjee et al. (2023) show promising performance at the 13B scale when distilling data from GPT-4 using explanation traces. Similarly, the Phi series (Gunasekar et al., 2023) demonstrates strong performance of even smaller language models when trained on code data. Lastly, OpenELM (Mehta et al., 2024) shows similar results.

7 Conclusion

In this paper, we present a novel conversational synthesis method applied to the challenging task of conversational grounded reasoning for question answering. We show that using CoSy to generate diverse, steerable, and conversational question-answer traces can significantly close the in-domain performance gap compared to human-curated multi-turn conversations. Furthermore, our synthesis approach improves zero-shot question answering and summarization performance compared to similarly sized instruction-tuned models, and even outperforms models of significantly larger size. These results make a compelling case for using CoSy to synthesize data from diverse seeds instead of relying on resource-intensive human annotation or scaling up model size.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, and Hany Awadalla et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context. *Preprint*, arXiv:1808.07036.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *Preprint*, arXiv:2310.13650.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *Preprint*, arXiv:2301.12726.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment. *Preprint*, arXiv:2304.07327.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024a. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *Preprint*, arXiv:2402.14905.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. Chatqa: Surpassing gpt-4 on conversational qa and rag. *Preprint*, arXiv:2401.10225.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. Openelm: An efficient language model family with open training and inference framework. *Preprint*, arXiv:2404.14619.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Preprint*, arXiv:1602.06023.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, and Diogo Almeida et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sahithya Ravi, Patrick Huber, Akshat Shrivastava, Aditya Sagar, Ahmed Aly, Vered Shwartz, and Arash Einolghozati. 2024. Small but funny: A feedback-driven approach to humor distillation. *Preprint*, arXiv:2402.18113.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Preprint*, arXiv:1808.07042.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.