Advancing Open-Domain Conversational Agents: Designing an Engaging System for Natural Multi-Turn Dialogue

Islam A. Hassan and **Yvette Graham**
mohamedi@tcd.ie and ygraham@tcd.ie
School of Computer Science and Statistics, Trinity College Dublin

Abstract

This system paper describes our conversational AI agent developed for the SCI-CHAT competition. The goal is to build automated dialogue agents that can have natural, coherent conversations with humans over multiple turns. Our model is based on fine-tuning the Snorkel-Mistral-PairRM-DPO language model on podcast conversation transcripts. This allows the model to leverage Snorkel-Mistral-PairRM-DPO's linguistic knowledge while adapting it for multi-turn dialogue modeling using LoRA. During evaluation, human judges will converse with the agent on specified topics and provide ratings on response quality. Our system aims to demonstrate how large pretrained language models, when properly adapted and evaluated, can effectively converse on open-ended topics spanning multiple turns.

1 Introduction

Developing artificial intelligence capable of natural, multi-turn conversations remains an ongoing challenge in AI research. While recent advances in deep learning and large pretrained language models have accelerated progress in conversational AI, accurately capturing the nuances of human dialogue over extended interactions is still an active area of investigation. Although competitions like SCI-CHAT provide a platform for evaluating conversational models through real-time human interactions, our focus is creating an agent for coherent, open-domain dialog across diverse topics and turns.

Our approach uses large language models fine-tuned on real-world podcast conversations, immersing them in the natural flow and back-and-forth that defines human dialogue. We leverage the promising LoRA (Low-Rank Adaptation of Large Language Models)(Yu et al., 2023) architecture to equip our agent with the ability to navigate diverse topics and engage in coherent, multi-turn exchanges.

Beyond technical prowess, we believe these agents hold immense potential to impact fields like education and customer service. Imagine a virtual tutor providing personalized learning experiences or a virtual assistant seamlessly understanding your requests and completing tasks effortlessly. However, it's crucial to acknowledge potential ethical considerations surrounding bias and manipulation in advanced dialogue systems. We're committed to developing these technologies responsibly, ensuring they contribute positively to human-computer interaction.

Participating in open-domain interactions, like those facilitated by competitions, provides invaluable real-world experience and crucial human feedback. These live evaluations, assessing factors like coherence, consistency, and topical relevance over extended dialogues, serve as a crucial testing ground for our models. Ultimately, our hope is to contribute to the overarching goal of creating dialogue agents that can converse with humans naturally, paving the way for deeper and more meaningful human-computer interactions.

2 Model Architecture

Our conversational agent is based on fine-tuning the Snorkel-Mistral-PairRM-DPO language model for dialogue response generation. Snorkel-Mistral-PairRM-DPO [1] is a large Transformer-based language model based on Mistral(Jiang et al., 2023) and fine-tuned on trained on the UltraFeedback dataset, providing a strong foundation for various natural language generation tasks, To efficiently adapt the large language model for our task, we used LoRA (Low Rank Adaptation Of Large Language Models)(Yu et al., 2023). MMLU (Massive Multitask Language Understanding)(Hendrycks et al., 2021) allows fine-tuning just a small number of extra weights in the model while freezing most

[1] https://huggingface.co/snorkelai/Snorkel-Mistral-PairRM-DPO

Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024), pages 75–79

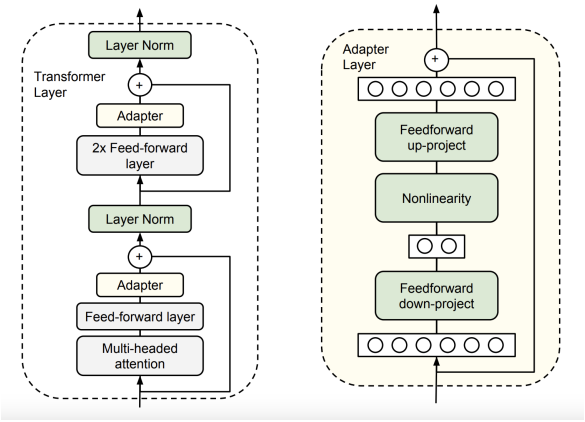

Figure 1: Architecture of the adapter module and its integration with the Transformer (Houlsby et al., 2019)

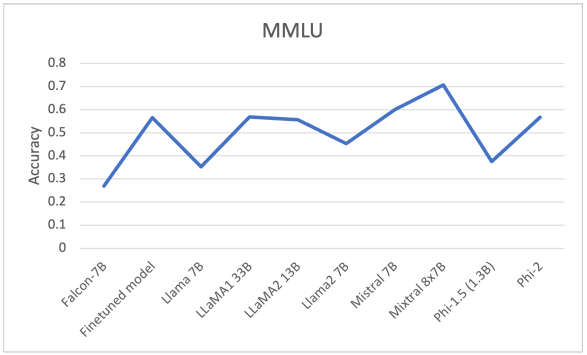

Figure 2: MMLU Performance Metric (accuracy) for the finetuned model against different Models.

of the parameters of the pretrained network. This saves on the computational cost and time required to retrain the entire massive model, it also mitigates catastrophic forgetting, a phenomenon where models tend to forget their original training due to excessive fine-tuning, by freezing the model's initial weights.

Specifically, we initialize our model with the 7B parameter version of the pretrained Snorkel-Mistral-PairRM-DPO model from HuggingFace. We then perform additional fine-tuning of this model on podcast conversation transcripts to specialize it for multi-turn conversational modeling using LoRA(Yu et al., 2023), In MMLU fine-tuning, only adapters are trained, introducing additional weights to the models while preserving the original weights and fine-tuning the newly added weights.

The architecture of the fine-tuned LoRA(Yu et al., 2023) model uses a modified transformer architecture with added adapter layers1. These adapter layers are inserted after the attention and feedforward stacks. The adapter layer itself has a bottleneck design: it takes the input, reduces it to a lower dimensionality representation, applies a non-linear activation, then restores the original dimensionality. This allows the subsequent transformer layer to effectively process the adapter output. For input, the dialogue history is merged with the latest human utterance using separator tokens. This merged history is fed into the model which then autoregressively predicts the next utterance.

During fine-tuning, the input sequences are truncated to fit within the model's context length limitation. For longer dialogue histories, we only include the most recent utterances to provide necessary context. The model is trained to generate the next utterance in the transcript given the truncated dialogue context.

During inference, we utilize the fine-tuned LoRA model and its tokenizer from the Hugging Face Transformers[2] and PEFT[3] libraries for text generation. Responses are generated based on the conversation history using nucleus sampling from the PEFT library with p=0.9 to achieve a balance of diversity and coherence in the generated text.

Fine-tuning this large pretrained model on podcast conversations allows us to leverage the rich linguistic knowledge in Snorkel-Mistral-PairRM-DPO while adapting it to generate natural, topically consistent responses in a conversational setting. The live human evaluation in SCI-CHAT will provide invaluable feedback on how to further improve the model's conversational abilities.

3 Training Data and Data preprocessing

Training conversational AI requires large, diverse dialogue datasets. As suggested by the competition guidelines, we primarily utilized the Freakonomics podcast transcripts dataset[4] for LoRA model training. This podcast corpus contains over 477 episodes covering economics, politics, pop culture, sports and more. The wide topical range provides natural conversational data to teach coherent, free-flowing dialogue skills. In total, the training data comprises 1 5,829 context-response pairs extracted from the podcast dialogues. On average, each dialogue contains approximately 12 turns, with 33 words per turn. The context vocabulary spans 10,194 unique terms, while the response vocabulary covers 14,550 distinct words.

[2] https://github.com/huggingface/transformers
[3] https://github.com/huggingface/peft
[4] https://huggingface.co/datasets/mogaio/Freakonomics_MTD

Dataset Information	#
Total context-response pairs	5,829
Total dialogues	477
Average turns per dialogue	12.22
Average words per turn	33.25
Context vocabulary size	10,194
Response vocabulary size	14,550
Cumulative vocabulary size	17,338

Table 1: Characteristics of the Extracted Multi-Turn Conversations Dataset from the Freakonomics Podcast.

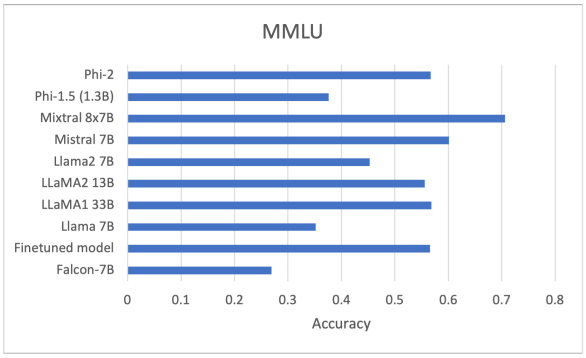

Figure 3: MMLU Performance Metric (accuracy) for the finetuned model against different Models.

The combined vocabulary size is 17,338 words, indicating rich, multifaceted linguistic interactions. These metrics characterize the structural complexity and lexical diversity of the dataset, informing the training process and performance evaluation. We scraped the raw podcast transcripts using a Python script proivded in this Git repo[5], yielding over ten thousand utterances of training data. The data was preprocessed by lowercasing, removing metadata, and filtering out very short, uninformative utterances.

In addition, the base model was pretrained on the UltraFeedback dataset, originally containing 64k prompts with 4 model completions each scored by GPT-4 for quality.

Evaluation

In this evaluation, we used MMLU to evaluate the fine-tuned model's performance and compared it to the base model. MMLU(Hendrycks et al., 2021) is a new benchmark to evaluate AI models' knowledge and problem-solving abilities across many academic fields, without requiring additional training. It covers diverse subjects at varying difficulty levels from elementary to advanced professional. The wide range of granular topics makes it well-suited to identify knowledge gaps in AI systems. MMLU aims to challenge AI in a more human-like way through zero-shot and few-shot evaluations.

The fine-tuned model scored 0.5659 on the MMLU evaluation, while the base model scored 0.5731. Across the 57 diverse MMLU tasks covering topics like elementary math, US history, computer science, and law, the fine-tuned model's median score was 0.0202 lower than the base model on 38 tasks. However, the fine-tuned model outperformed the base model on 19 tasks. In 5 of these tasks, the fine-tuned model's performance exceeded the base model's score and the difference was over one standard deviation 4. So while on average the base model scored higher, the fine-tuning improved performance on certain specific tasks in the evaluation by a significant margin.

We also compared 2 the fine-tuned model to various other models, including (LLAMA2 7B, LLAMA2 13B, LLAMA2 33B, LLAMA2 70B, Mistral 7B(Jiang et al., 2023), Mixtral 8x7B(Jiang et al., 2024)), and conducted benchmarks using the Language Model Evaluation Harness evaluation pipeline (Gao et al., 2023) to ensure fair comparison. We evaluated performance across a diverse range of tasks in Commonsense Reasoning (zero-shot)4, including Hellaswag(Zellers et al., 2019), Winogrande(Sakaguchi et al., 2021), PIQA(Bisk et al., 2020), ARC-Easy, and ARC-Challenge(Clark et al., 2018).

This multi-domain conversational data provides a strong foundation for training our model's ability to converse naturally on open-ended topics. The human evaluation at the end will reveal how well our model generalizes to coherent, topical conversations. Additional data could be incorporated in future work to expand the model's knowledge and conversational abilities.

Conclusion

In conclusion, we have built a conversational agent leveraging large pretrained language models and diverse, open-domain dialog data from podcast transcripts. Fine-tuning on this conversational corpus enables our model to engage in natural, wide-ranging dialogs.

At the core, we utilize Transformer architecture language models with LoRA adapters which are well-suited to modeling conversational contexts,

[5] https://github.com/hkmirza/EACL2024-SCI-CHAT-SharedTask/tree/main

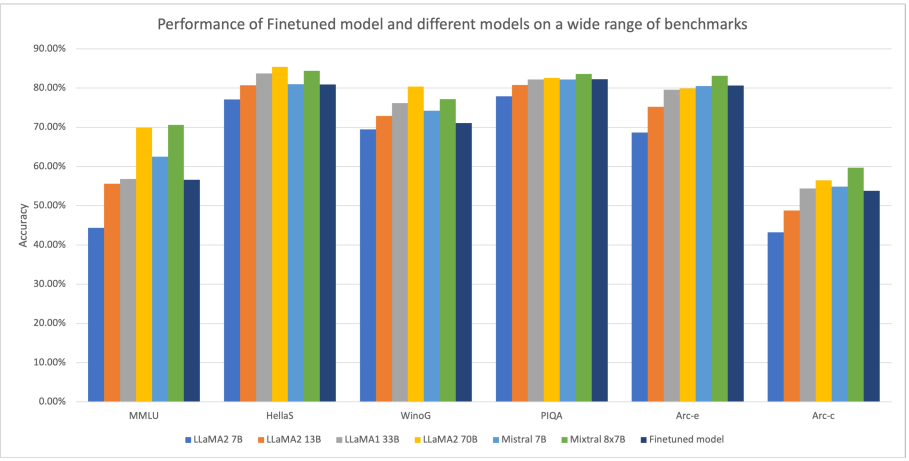

Figure 4: Performance Metrics (accuracy) for the finetuned model against different Models across different evaluation tasks.

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c
LLaMA2 7B*	7 B	44.40	77.10	69.50	77.90	68.70	43.20
LLaMA2 13B*	13 B	55.60	80.70	72.90	80.80	75.20	48.80
LLaMA1 33B*	33 B	56.80	83.70	76.20	82.20	79.60	54.40
LLaMA2 70B*	70 B	69.90	85.40	80.40	82.60	79.90	56.50
Mistral 7B*	7 B	62.50	81.00	74.20	82.20	80.50	54.90
Mixtral 8x7B*	7 B	70.60	84.40	77.20	83.60	83.10	59.70
Finetuned model 7B**	7 B	56.59	80.95	71.11	82.26	80.68	53.84

Table 2: Performance Metrics for Different Models (%); where *=few-shot,k=5; ** = zero-shot

Model	MMLU
Falcon-7B	0.2690
Finetuned model 7B	0.5659
Based model 7B	0.5731
Llama 7B	0.3520
LLaMA1 33B	0.5680
LLaMA2 13B	0.5560
Llama2 7B	0.4530
Mistral 7B	0.6010
Mixtral 8x7B	0.7060
Phi-1.5 (1.3B)	0.3760
Phi-2	0.5670

Table 3: MMLU for the finetune model against different models

providing a solid foundation for language generation tasks. Further fine-tuning on the podcast data allows the model to produce coherent, context-appropriate responses during interactions. Techniques including dialog history management and nucleus sampling also boost the model's conversational abilities.

Live human evaluations will provide critical insights into the model's real-world performance, highlighting its capabilities and limitations. This human feedback will be invaluable for improving the agent's conversational strengths moving forward.

In future work, we hope to incorporate even larger models, more conversational data covering diverse topics, and techniques to improve multi-turn coherence. Conversational AI remains a very active area of research.

We believe our work is a step towards building conversational agents that can communicate naturally with humans. There is still much progress to be made, but continued research combined with establishing rigorous human-centered evaluations like SCI-CHAT will take us closer to conversational AI that is both capable and aligned with human values.

Acknowledgements

The work presented in this paper is supported the and is supported by the Science Foundation Ireland Research Centre, ADAPT at Trinity College Dublin under Grant Agreement No 13/RC/2106_P2. This work has received research ethics approval by Trinity College Dublin Research Ethics Committee (Application no. 20210603).

	Finetuned model	Based Model
MMLU	0.5659 (± 0.1336)	0.5731 (± 0.1323)
High School Geography	0.7626 (± 0.0303)	0.7121 (± 0.0323)
US Foreign Policy	0.8300 (± 0.0378)	0.7900 (± 0.0409)
Abstract Algebra	0.3400 (± 0.0476)	0.2900 (± 0.0456)
High School Biology	0.7065 (± 0.0259)	0.6774 (± 0.0266)
High School Chemistry	0.4778 (± 0.0351)	0.4335 (± 0.0349)

Table 4: Comparison of MMLU for the fine-tuned model and the base model, where the fine-tuned model outperforms the base model, including standard deviation, standard error is provided in brackets.

References

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.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G. Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, Tuan Dinh, Ankur Gandhe Denis Filimonov, Shalini Ghosh, Andreas Stolcke, Ariya Rastow, and Ivan Bulyko. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

A Appendix

Table 5: Acronym Table

Acronym	Full Form
LoRA	Low-Rank Adaptation of LLMs
DPO	Data-Produced-Offline
MMLU	Massive Multitask Language Understanding
PIQA	Physical Interaction: Question Answering
ARC-E	AI2 Reasoning Challenge - Easy
ARC-C	AI2 Reasoning Challenge - Challenge
LLM	Large Language Model