RNA-GPT: Multimodal Generative System for RNA Sequence Understanding

Yijia Xiao[1], **Edward Sun**[1], **Yiqiao Jin**[2], **Wei Wang**[1]

[1]University of California, Los Angeles,
[2]Georgia Institute of Technology,
{yijia.xiao,wei.wang}@cs.ucla.edu, edwardsun12895@g.ucla.edu,
yjin328@gatech.edu

Abstract

RNAs are vital molecules that carry genetic information essential for life, with significant implications for drug development and biotechnology. However, RNA research is often slowed by the vast amount of literature. To address this, we introduce RNA-GPT, a multi-modal RNA chat model that simplifies RNA discovery by leveraging extensive RNA literature. RNA-GPT combines RNA sequence encoders with linear projection layers and state-of-the-art large language models (LLMs) for precise representation alignment. This enables it to process user-uploaded RNA sequences and provide concise, accurate responses. Our scalable training pipeline, powered by RNA-QA, automatically gathers RNA annotations from RNACentral using a divide-and-conquer approach with GPT-4o and latent Dirichlet allocation (LDA) to handle large datasets and generate instruction tuning samples. Experiments show RNA-GPT effectively handles complex RNA queries, streamlining RNA research. We also introduce RNA-QA, a 407,616 RNA dataset for modality alignment and instruction tuning.

1 Introduction

Large language models (LLMs) trained on internet-scale corpora have been shown to perform extraordinarily well on a large array of tasks from Olympiad-level mathematical and scientific reasoning to planning long-term tasks for robotic systems [1, 2, 3]. Recent advances in the biological and medical fields have enabled the adaptation of powerful models to accelerate research, significantly reducing reliance on traditional experiments. Since proteins, RNAs, and DNAs can be represented as character strings and a vast amount of sequenced data is readily available, this has created an ideal environment for training language models to predict and generate protein, DNA, and RNA structures and sequences. Protein language models like ESM have successfully encoded protein sequence and structure information, inspiring works such as ProteinGPT and ProtSt, which adapt protein representations into a language-based format, enabling natural language querying of protein data [4, 5, 6, 7, 8, 9, 10]. Similar to ESM-2, works like RiNALMo and RNA-FM have utilized the flexible capabilities of LLMs to learn and predict RNA structure and functions [11, 12].

Much like the motivation behind protein research, where proteins are represented as strings of characters, RNAs—with their sequences of five unique nucleotides—have also sparked interest in computational RNA and DNA research using large language models (LLMs).

While models like ProteinGPT, ProtST, ProteinChat, and ProteinCLIP, have made significant progress in aligning protein sequences and structures with textual descriptions, advancements in the DNA and RNA domains are far less advanced [9, 13, 14, 10, 15, 16]. Previous efforts, such as RiNALMo

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

and RNA-FM have mainly focused on specific tasks like promoter or enhancer prediction, and structure and function analysis [12, 11, 17]. ChatNT is among the few models striving to bridge the gap between RNA comprehension and natural language [18]. However, its emphasis is more on performing biological tasks as a conversational agent rather than providing deep RNA understanding and comprehensive dialogue. As a result, there is a notable gap in RNA chat models that offer in-depth knowledge. However, applying multimodal LLMs to RNA modeling presents unique challenges, especially in integrating diverse modalities such as textual descriptions, RNA sequences, and structural data.

To overcome these challenges, we propose a two-step approach to RNA-GPT. First, we utilize the RNA-FM sequence encoder to embed RNA sequences, followed by aligning these sequence representations with natural language through a large, automatically curated QA dataset from RNA Central [12, 19]. Secondly, to ensure our model generates concise and accurate responses, we break down RNA-QA's abstract summaries into individual QA pairs for instruction tuning, enhancing the model's ability to deliver clear and relevant answers. We utilize Meta AI's flagship Llama-3 8B Instruction as our backbone LLM to provide solid general language understanding [20]. More specifically, our contributions are as follows:

- **Novel Framework.** RNA-GPT is one of the first multi-modal RNA sequence chat models that enables deep, interactive RNA-focused conversations, significantly enhancing the understanding of RNAs for biological research.
- **Large-scale Dataset and Collection Pipeline.** We introduce RNA-QA, a QA dataset derived from the RNA Central Dataset for modality alignment instruction tuning of RNA chat models [19]. We also present our highly scalable collection pipeline that automates the scraping and summarizing of relevant literature on RNA. Using a divide-and-conquer summarization strategy, we ensure that research details are preserved effectively. For over 407,616 RNA samples, we create QA pairs, each accompanied by a comprehensive research summary based on available literature, and between 5 and 14 QA pairs. The depth and diversity of these annotations make RNA-QA an excellent resource for instruction tuning.

2 Methodology

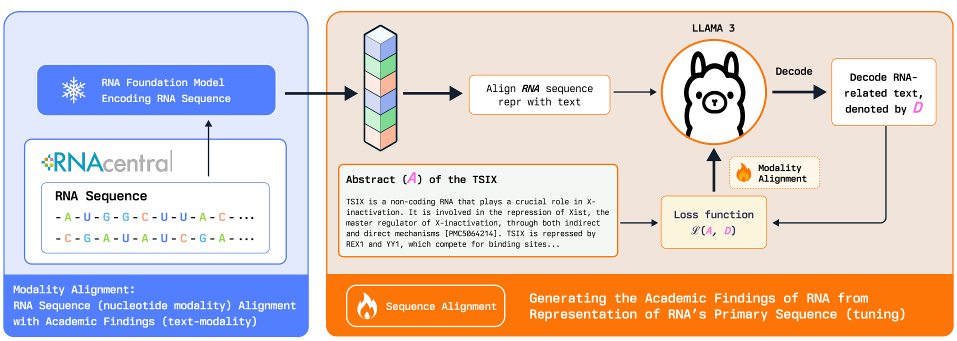

Figure 1: RNA-GPT Modality Fusion & Alignment Stage: we freeze the sequence encoder block and train the linear projection layer to learn how to align RNA sequence representations with text. In the alignment stage, the input to the training is only the projected RNA representation. No text prompts are incorporated in this stage.

2.1 Model Architecture

RNA-GPT uses the pre-trained RNA-FM sequence encoder (Figure 1 and Figure 2) to embed RNA sequences, which are then passed through a linear projection layer. This layer learns to map the RNA embeddings to a shared representation space with natural language, enabling alignment with a backbone LLM, for which we chose Meta's Llama-3 8B model. The training process is divided into two stages: 1) *Sequence and Modality Alignment*, where RNA and natural language representations are aligned, and 2) *Instruction Tuning*, where the model is fine-tuned for task-specific QA generation.

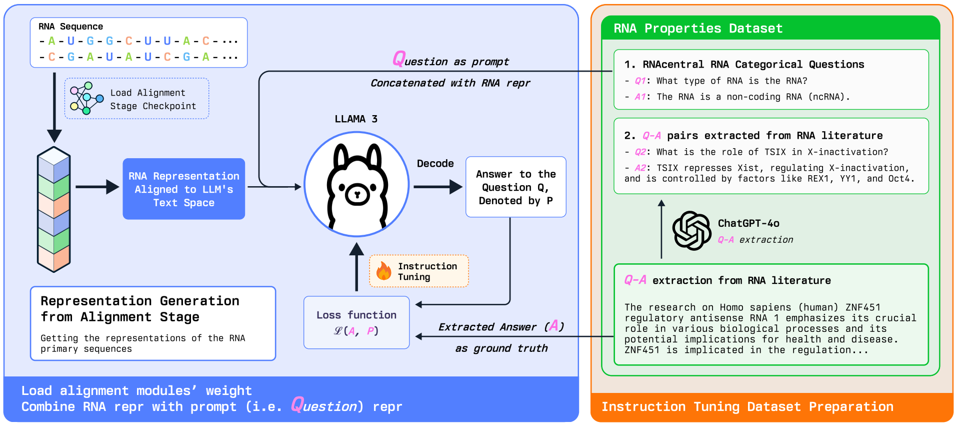

Figure 2: RNA-GPT Modality Fusion & Alignment Stage: we freeze the sequence encoder block and train the linear projection layer to learn how to align RNA sequence representations with text. In the alignment stage, the input to the training is only the projected RNA representation. No text prompts are incorporated in this stage.

Modality Alignment Stage (Stage 1)

RNA sequences in the form of strings are first fed into the pre-trained sequence encoder, featuring 12 transformer layers trained with 23 million RNAs from the RNA Central database via self-supervised learning [19, 12]. We utilize a specialized token <RNAHere> for RNA-text modality alignment:

$$\textbf{Q}: < RNA >< RNAHere >< /RNA >< QuestionPrompts >$$
$$\textbf{A}: < Description >$$

The embedded sequence information is encoded into the soft prompts and prepended to the question prompt. In stage 1 training, the question **Q** is left empty to prioritize learning the abstract description from the RNA representation. The description tag **A** is replaced with the full annotation from RNA Central [19] to train the linear projection layer to align an RNA with its full abstract annotation.

Instruction Tuning Stage (Stage 2)

In stage 2, we instruction-tune the model using our curated RNA-QA dataset. Previous protein-related chat models rely on fully annotated abstracts, frequently resulting in excessively lengthy and irrelevant responses. We take a different approach to address this by breaking down the full annotations into targeted QA samples with concise answers to specific questions as prediction targets. This allows the chat model to provide more relevant and accurate responses.

We augment the full abstract annotation dataset from stage 1 using GPT-4o-mini to generate explicit QA pairs for this stage. The prompts from stage 1 are adapted to the Llama-3 style ("$\#\#\#Human: \ldots$" and "$\#\#\#Assistant: \ldots$"), with **Q** replaced by explicit questions from RNA-QA, such as "What regulatory role does the RNA have along with other RNAs?" The model then generates descriptive yet concise answers such as "The RNA is involved in transcript splicing regulation along with RNVU1-18 and CLK1" as **A**.

2.2 RNA-QA Dataset

To achieve modality alignment, we constructed a large-scale dataset from the RNA Central database [19], comprising 407,616 RNA sequences paired with abstract descriptions.

Divide and Conquer RNA Literature Summarization

We begin by filtering RNA sequences from RNA Central [19], focusing on those indexed with "Lit Scan," yielding around 420,000 RNAs with associated research papers. We refine this set to

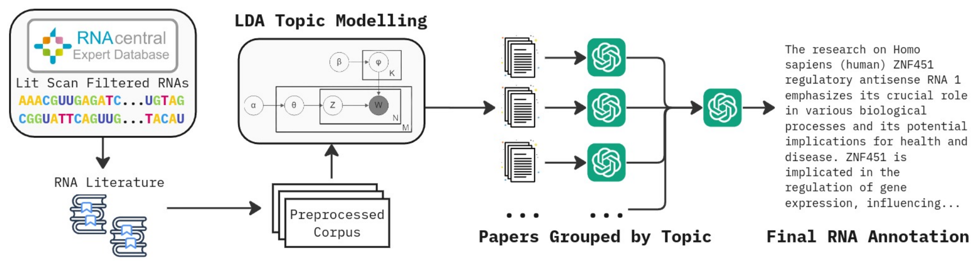

Figure 3: RNA-QA uses an automated pipeline to scrape and summarize existing RNA literature. We apply latent Dirichlet allocation (LDA) to group the vast literature on each RNA, and then we summarize each group individually using GPT-4o-mini. These summaries are then combined and refined to produce the final RNA annotation.

include only sequences up to 1024 nucleotides, the maximum input length for our sequence encoder. For the remaining 407,616 RNAs, we scrape and extract abstracts from all relevant literature. As shown in Figure 3, we apply LDA topic modeling to group papers by topic, summarizing each group individually. This ensures each summarization focuses on a narrower, cohesive subject area, minimizing information loss. We have found that summarizing broad topics often causes key details to be missed, as the model struggles to condense diverse information. Grouping similar topics allows for more precise, detailed summaries that retain essential context. The final annotation is created by combining these summaries in a final round of summarization. This divide-and-conquer approach improves accuracy and efficiently handles large datasets. Moreover, it overcomes the token limits of GPT models, allowing for detailed, information-dense annotations of large RNA research profiles.

2.3 Data Augmentation

Similar works in protein chat alignment often use the entire protein annotation for instruction tuning [10, 13, 21, 9], which often result in verbose and irrelevant responses. To address this, RNA-GPT decomposes the rich RNA annotations of RNA-QA into more specific QA-pairs for instruction tuning using GPT-4o-mini so that user instructions can be concisely answered. Concretely, we prompt GPT-4o-mini to generate both open-ended and close-ended QA pairs with the context of the RNA-QA annotation to decompose the abstract into atom-level QA pairs.

3 Experiments

We trained RNA-GPT using the flagship Llama-3 8B model architecture [20] using a smaller 5K RNA, 121K QA samples subset for our initial model. We are in the process of training the large RNA-GPT that uses all 407,616 RNAs of RNA-QA with millions of QA samples.

Table 1: RNA-QA (**AIS**): RNA Sequence (left), Modality Fusion (middle), and RNA-GPT (right). Embedding base models are BERT, PubMedBERT, and OpenAI's GPT text-embedding-3-large.

	RNA Sequence			Modality Fusion			RNA-GPT		
Metric	S_{BERT}	S_{Pub}	S_{GPT}	S_{BERT}	S_{Pub}	S_{GPT}	S_{BERT}	S_{Pub}	S_{GPT}
Precision	0.7372	0.5528	0.5219	0.6929	0.6507	0.6655	0.8602	0.7384	0.7848
Recall	0.7496	0.5270	0.5474	0.8028	0.6082	0.6603	0.8404	0.7208	0.7561
F1 Score	0.7424	0.5387	0.5339	0.7403	0.6283	0.6627	0.8494	0.7293	0.7700

We conducted a series of experiments to assess RNA-GPT's effectiveness both quantitatively and qualitatively along with ablation studies to ascertain the importance of various modules at different stages. These included the original model (LLM with RNA sequence as text input), the modality-aligned model, and the final instruction-tuned model.

Table 2: RNA-QA (**AIS**): ROUGE Score with RNA Sequence, Modality Fusion, and RNA-GPT.

Metric	RNA Sequence			Modality Fusion			RNA-GPT		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
ROUGE	0.2364	0.0935	0.2037	0.2239	0.1364	0.2091	0.5031	0.3667	0.4747

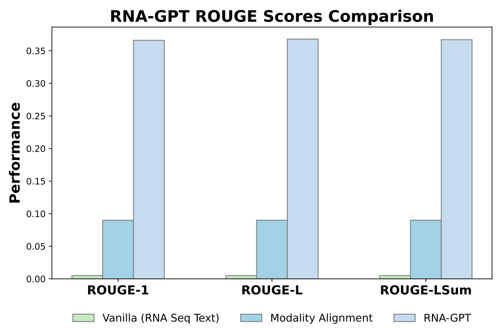

Figure 4: ROUGE Score Comparison

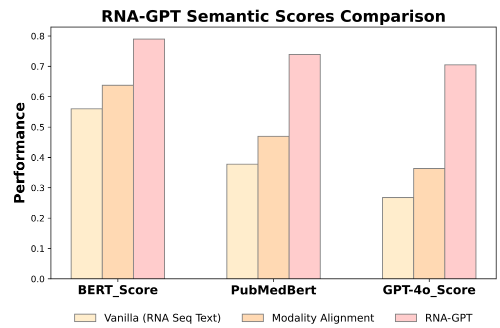

Figure 5: Semantic Score Comparison

Table 3: RNA-QA (**D&C**): RNA Sequence (left), Modality Fusion (middle), and RNA-GPT (right). Embedding base models are BERT, PubMedBERT, and OpenAI's GPT text-embedding-3-large.

Metric	RNA Sequence			Modality Fusion			RNA-GPT		
	S_{BERT}	S_{Pub}	S_{GPT}	S_{BERT}	S_{Pub}	S_{GPT}	S_{BERT}	S_{Pub}	S_{GPT}
Precision	0.7612	0.5498	0.5479	0.6884	0.6201	0.6676	0.8620	0.7173	0.7568
Recall	0.7654	0.5512	0.5649	0.8187	0.5830	0.6602	0.8623	0.7161	0.7554
F1 Score	0.7625	0.5501	0.5561	0.7466	0.6005	0.6637	0.8609	0.7165	0.7560

Table 4: RNA-QA (**D&C**): ROUGE Score with RNA Sequence, Modality Fusion, and RNA-GPT.

Metric	RNA Sequence			Modality Fusion			RNA-GPT		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
ROUGE	0.2472	0.0964	0.2182	0.0922	0.0393	0.0799	0.4791	0.2690	0.4405

4 Conclusions

We present RNA-GPT, a multimodal chat model for RNA sequences that enhances LLM-based question-answering and accelerates RNA discovery by providing concise, accurate responses to complex queries. RNA-GPT aligns RNA embeddings from the RNA-FM encoder with natural language in LLMs like Llama-3 using a learnable projection layer. We optimize instruction tuning with GPT-4o-mini to ensure high-quality, precise answers. We also introduce RNA-QA, a 407,616 RNA question-answering dataset derived from the extensive RNA research literature. Our scalable framework, which uses topic modeling and divide-and-conquer summarization, enables efficient RNA-to-language dataset curation.

Experiments with our initial model trained on a 5K subset of RNA-QA show that RNA-GPT generates very promising responses already, achieving high semantic and lexical scores on unseen RNA data. The full RNA-QA provides even more value in the development of multimodal RNA LLMs. RNA-GPT with RNA-QA will inspire further innovations in LLM-based RNA research, driving rapid advancements in the field.

References

[1] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024.

[2] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

[3] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models, 2022.

[4] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pages 2020–12, 2020.

[5] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[6] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[7] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[8] Yijia Xiao, Jiezhong Qiu, Ziang Li, Chang-Yu Hsieh, and Jie Tang. Modeling protein using large-scale pretrain language model. *arXiv preprint arXiv:2108.07435*, 2021.

[9] Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*, 2024.

[10] Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts, 2023.

[11] Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks, 2024.

[12] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions, 2022.

[13] Han Guo, Mingjia Huo, and Pengtao Xie. Proteinchat: Towards enabling chatgpt-like capabilities on protein 3d structures. 2023.

[14] Kevin E Wu, Howard Chang, and James Zou. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, pages 2024–05, 2024.

[15] Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. Instructprotein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*, 2023.

[16] Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. Protllm: An interleaved protein-language llm with protein-as-word pre-training, 2024.

[17] Yekaterina Shulgina, Marena I. Trinidad, Conner J. Langeberg, Hunter Nisonoff, Seyone Chithrananda, Petr Skopintsev, Amos J. Nissley, Jaymin Patel, Ron S. Boger, Honglue Shi, Peter H. Yoon, Erin E. Doherty, Tara Pande, Aditya M. Iyer, Jennifer A. Doudna, and Jamie H. D. Cate. Rna language models predict mutations that improve rna function. April 2024.

[18] Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie P Lopez, Alexander Laterre, Maren Lang, Ugur Sahin, Karim Beguir, and Thomas Pierrot. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. May 2024.

[19] Blake A Sweeney, Anton I Petrov, Carlos E Ribas, Robert D Finn, Alex Bateman, Maciej Szymanski, Wojciech M Karlowski, Stefan E Seemann, Jan Gorodkin, Jamie J Cannone, Robin R Gutell, Simon Kay, Steven Marygold, Gil dos Santos, Adam Frankish, Jonathan M Mudge, Ruth Barshir, Simon Fishilevich, Patricia P Chan, Todd M Lowe, Ruth Seal, Elspeth Bruford, Simona Panni, Pablo Porras, Dimitra Karagkouni, Artemis G Hatzigeorgiou, Lina Ma, Zhang Zhang, Pieter-Jan Volders, Pieter Mestdagh, Sam Griffiths-Jones, Bastian Fromm, Kevin J Peterson, Ioanna Kalvari, Eric P Nawrocki, Anton S Petrov, Shuai Weng, Philia Bouchard-Bourelle, Michelle Scott, Lauren M Lui, David Hoksza, Ruth C Lovering, Barbara Kramarz, Prita Mani, Sridhar Ramachandran, and Zasha Weinberg. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220, October 2020.

[20] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris

Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[21] Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. Protchatgpt: Towards understanding proteins with large language models, 2024.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[23] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training, 2018.

A Training Details

We conducted initial training of RNA-GPT on a 5K subset of the RNA-QA dataset to generate the initial model checkpoints.

During the modality alignment (***MA***) stage (Stage 1), we optimized the projection layer over 10 epochs using a batch size of 1, weight decay of 0.05, and 2,048 warm-up steps. The AdamW optimizer [22] was employed with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. A learning rate scheduler with a linear warm-up followed by cosine annealing was applied. Specifically, the initial learning rate was set to 1×10^{-4}, the minimum learning rate to 8×10^{-5}, and the warm-up learning rate to 1×10^{-6}. Automatic Mixed Precision (AMP) [23] was employed to enhance training efficiency.

For the instruction tuning (***IT***) stage (Stage 2), the RNA-GPT architecture was further trained for an additional 10 epochs, maintaining the same batch size of 1, weight decay of 0.05, and reducing the warm-up steps to 200. The configuration of the AdamW optimizer and the use of AMP remained consistent with those used in Stage 1.

The training process was executed on two Nvidia A100 GPUs (40GB) and required a total of 1 week for both stages combined.

We will extend the training process by scaling RNA-GPT to a larger configuration using the full 407,616 dataset from RNA-QA.

B Case Studies

RNA Central ID: URS000000F6A7_645133
RNA Sequence: GGCUAGCUCAAUCGGUAGAGCGUGAGACUCUUAAUCUCAAGGU-UGUGGGUUCGACCCGUAAGCCU

User: What type of RNA is the RNA?

Vanilla LLM: *RNA Sequence Type:* The RNA sequence provided is a DNA sequence. *DNA*

RNA-GPT (MA): The RNA in question is a non-coding RNA (ncRNA) that is a component of the ribonucleic acid (RNA) family. Specifically, it is a small nuclear RNA (snRNA) that plays a crucial role in the regulation of gene expression and is involved in the processing of other RNAs.

RNA-GPT (IT): The RNA is a non-coding RNA (ncRNA).

User: What species does this RNA belong to and what are this RNA's various forms?

Vanilla LLM: (multiple answer)\nIt belongs to Homo sapiens (Humans) and its various forms include...

RNA-GPT (MA): The RNA belongs to the species Homo sapiens (Human). The various forms of this RNA are: primary transcripts, mature transcripts, and non-coding RNAs.

RNA-GPT (IT): The RNA belongs to *Plasmodium falciparum* and its various forms include tRNA-Phe.

User: What biological mechanisms does the RNA play a role in?

Vanilla LLM: The RNA sequence provided is a piece of ribosomal RNA (rRNA).

RNA-GPT (MA): The RNA is involved in various biological processes, particularly in gene regulation and epigenetic modification.

RNA-GPT (IT): The RNA plays a role in the regulation of gene expression and the maintenance of cellular homeostasis.

Figure 6: Interaction between user and RNA-GPT on URS000000F6A7_645133

RNA Central ID: URS0000012D5F_15368
RNA Sequence: GUUGCUGUGGUGUAGUGGUUAUCACGUCAGUCUUACACACU-GAAGGUCUCCAGUUCGAUCCUGGGCAGCAACA

User: What significance does mitochondrial ribosomal RNA (rRNA) have in relation to the RNA?

RNA-GPT: Mitochondrial rRNA is significant in relation to the RNA, as it is involved in protein synthesis and has implications for genetic engineering and biotechnology applications.

User: What methods are used to analyze the sequence and structural variations of the RNA?

RNA-GPT: The methods used to analyze the sequence and structural variations of the RNA include ambivalent covariance models and non-negative matrix factorization (NMF).

Figure 7: Interaction between user and RNA-GPT on URS0000012D5F_15368

RNA Central ID: URS000006DA05_9925
RNA Sequence: UGAAGUGCUGUGGAUUUCUUUGUGAAUCACCAUAU-CUAAGCUAAUGUGGUGGUGGUUUACAAAGUAAUUCAUAGUGCUUCA

User: What diseases are associated with the RNA?

RNA-GPT: The RNA is associated with osteosarcoma, breast cancer, and epilepsy.

User: How is the RNA implicated in cholangiocarcinoma (CCA)?

RNA-GPT: The RNA is significantly upregulated in CCA and may serve as a biomarker for early detection and diagnosis.

User: What is the role of the RNA in CCA cell growth?

RNA-GPT: In CCA cell growth, the RNA acts as a tumor suppressor by regulating BCL-XL expression.

User: What is the role of circular RNAs (circRNAs) in relation to this RNA?

RNA-GPT: CircRNAs play a significant role in the regulation of gene expression and have been linked to various diseases, including cancer.

Figure 8: Interaction between user and RNA-GPT on URS000006DA05_9925 (Part 1)

User: How does the knockdown of CDR1as affect gastric cancer (GC) cells?

RNA-GPT: The knockdown of CDR1as in GC cells inhibits cell proliferation, migration, and invasion, indicating its potential as a therapeutic target.

User: How does the RNA function as a tumor suppressor in CCA treatment?

RNA-GPT: The RNA functions as a tumor suppressor by regulating BCL-XL and BCL-2 expression, which are key regulators of apoptosis.

User: What impact does the RNA have on myotube area and fusion index in DM1?

RNA-GPT: The RNA has a significant impact on myotube area and fusion index in DM1, indicating its potential role in muscle regeneration.

User: What species does this RNA belong to and what are this RNA's various forms?

RNA-GPT: The RNA belongs to the chimpanzee (*Pan troglodytes*) and its various forms include miRNAs, long non-coding RNAs, and circular RNAs.

Figure 9: Interaction between user and RNA-GPT on URS000006DA05_9925 (Part 2)