# LEVERAGING GPT CONTINUAL FINE-TUNING FOR IMPROVED RNA EDITING SITE PREDICTION

**Zohar Rosenwasser**
Bar-Ilan University
Israel
rosenwz@biu.ac.il

**Erez Y. Levanon**
Bar-Ilan University
Israel
erez.levanon@biu.ac.il

**Michael Levitt**
Stanford University
United States
levittm@stanford.edu

**Gal Oren**
Stanford University, Technion
United States
galoren@stanford.edu

## ABSTRACT

RNA editing is a critical regulatory process that diversifies the transcriptome by altering nucleotide sequences in messenger RNA molecules. We propose a novel framework for predicting adenosine-to-inosine (A-to-I) RNA editing sites by leveraging a specialized fine-tuned GPT-4o-mini model and a tissue-specific liver dataset. Grounding our approach in the high expression levels of ADAR1 in liver tissue, we avoid confounding factors from other ADAR isoforms and complex multi-tissue data. We categorize editing levels into progressively narrower thresholds (1%, 5%, 10%, and 15%) and introduce continual fine-tuning (CFT) to guide the model step-by-step from low-editing (1%) to high-editing (15%) scenarios. Compared to static fine-tuning (SFT) on a single threshold, our multi-stage method incrementally refines the model's ability to distinguish editing features and demonstrates superior performance over base GPT-3.5/4o-mini models across various configurations. We further show that employing strict, non-overlapping threshold bins facilitates clearer distinctions between edited and non-edited sites, consequently improves performance. In contrast, reducing the distinction between edited and non-edited classes significantly degrades classification accuracy. These findings underscore the importance of biologically appropriate data partitioning and continual, threshold-based fine-tuning in enhancing the predictive power of generative language models for RNA editing. Our study paves the way for future work on building more nuanced models that incorporate tissue-specific constraints, ultimately broadening the applicability of generative AI in post-transcriptional regulation analysis. [1]

## 1 INTRODUCTION

RNA editing is one of the post-transcriptional mechanisms that modify pre-RNA sequences encoded in the genome. Editing events can lead to changes in amino acids, alternative splicing, gene silencing, and alterations in RNA stability and localization Brennicke et al. (1999); Gott & Emeson (2000). The most common type of RNA editing in animals is adenosine-to-inosine (A-to-I) deamination, catalyzed by the enzyme adenosine deaminase acting on RNA (ADAR) family of proteins Bass (2002); Levanon et al. (2004). ADARs bind to double-stranded RNA structures and have the ability to change specific adenosines (A) to inosines (I), which are recognized during translation as guanosines (G). ADAR has two major structural motifs: the double-stranded RNA-binding domains (dsRBD), which binds to double-stranded RNA, and the deaminase domain, which catalyzes the hydrolytic deamination Nishikura (2010). ADAR enzymes can efficiently modify RNA

[1]The sources of this work are available at our repository: https://github.com/Scientific-Computing-Lab/CFT-RNA-Editing-Detection-using-GPT.

sequences at specific positions. The vast majority of mRNA editing in humans occurs within *Alu* repetitive elements. These ∼300bp long elements are abundant with over a million copies and make up approximately 10% of the human genome. Binding of neighboring, reversely aligned, *Alu* elements results in stable intramolecular secondary structures, which are targeted by ADARs Bazak et al. (2014); Kim et al. (2004); Athanasiadis et al. (2004). The specificity and efficiency of editing are influenced by the structural and sequence characteristics of double-stranded RNA. However, the underlying mechanisms by which ADAR efficiently targets specific adenosine (A) are not well characterized.

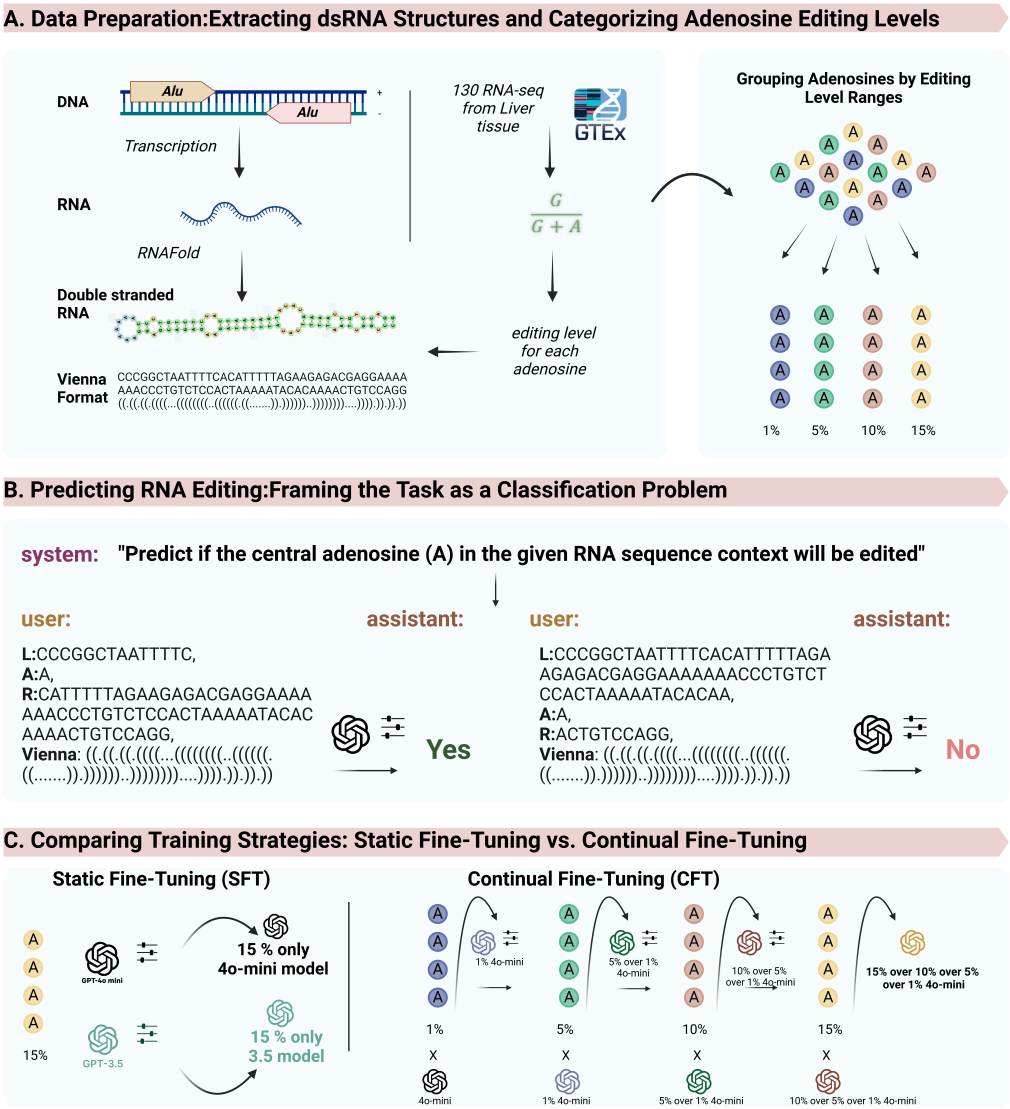


Figure 1: Data Collection, Preprocessing, and Model Training Approaches for RNA Editing Prediction. (A) Extraction of double-stranded RNA structures from *Alu* pairs, annotation of editing levels from GTEx liver samples, and grouping of adenosines by editing level. (B) Framing RNA editing prediction as a classification task, determining whether a given adenosine is edited. (C) Comparison of static fine-tuning (training on a single threshold) vs. continual fine-tuning (progressively refining thresholds), highlighting improved classification performance.

Beyond its natural regulatory role, A-to-I RNA editing is increasingly recognized as a promising tool for therapeutic applications Reautschnig et al. (2024); Pfeiffer & Stafforst (2023); Katrekar et al. (2022); Montiel-Gonzalez et al. (2019). Unlike permanent genomic DNA modifications, RNA editing enables reversible and programmable sequence alterations, making it a safer alternative for

gene therapy. One emerging strategy for harnessing endogenous ADAR activity involves the design of guide RNAs (gRNAs) that hybridize with target transcripts, forming dsRNA structures that recruit ADAR enzymes to catalyze site-specific A-to-I conversions Merkle et al. (2019). This approach allows precise RNA modifications while minimizing off-target effects, thereby offering a potential strategy for correcting pathogenic mutations and treating genetic disorders. Currently, much of the effort in this field is focused on designing guide RNAs that will generate the most optimal dsRNA structure to maximize ADAR enzyme activity at the desired editing site.

There has been extensive research into modeling RNA sequences and their various modifications using artificial intelligence, leveraging different computational techniques to understand and predict RNA behavior (see section 2). Predicting RNA editing sites, particularly A-to-I modifications, is a challenging task due to the complexity of RNA secondary structures and the nuanced activity of the ADAR family of proteins. Over the years, various AI techniques have been employed to tackle this problem, including support vector machines, random forests, convolutional neural networks, and recurrent neural networks. These methods have shown potential but often suffer from limitations such as low success rates, limited data specificity, and inadequate generalization across different tissue types and RNA contexts. Recently, transformer-based models and generative AI approaches using large language models (LLMs) like GPT-3.5 have been explored, with methods treating RNA sequences as natural language and capturing long-range dependencies without extensive feature engineering. However, previous efforts primarily employed static fine-tuning, which restricts the models' ability to dynamically adapt to new patterns in RNA editing data. For these reasons, improving the predictive accuracy and versatility of models in this domain remains a significant research challenge.

**Contribution:**

In this paper, we introduce a novel framework for predicting A-to-I RNA editing sites within *Alu* elements by leveraging advanced LLMs, specifically using the GPT-4o-mini model (Figure 1). We enhance existing approaches that frame RNA editing site prediction as a classification task, where a model determines whether a highlighted potential editing site has undergone editing based on RNA secondary structure information in the Vienna format, a standard notation for representing RNA secondary structures using base-pairing symbols, which GPT models are familiar with Rosenwasser et al. (2024). In this work, we enhance this approach by employing continual fine-tuning (CFT) of GPT-4o-mini and introducing hierarchical editing thresholds, allowing the model to progressively learn from different editing levels. Additionally, we adapt the model specifically for liver tissue, where ADAR1 is the dominant enzyme, enabling a more accurate representation of tissue-dependent RNA editing patterns. These advancements lead to a significant improvement in prediction accuracy and provide deeper insights into the mechanistic factors influencing RNA editing specificity.

Specifically, we ask and answer the following research questions:

- RQ1: Is there an alternative dataset that can better explain the enzymatic activity in the RNA editing site prediction problem than previously devised?
- RQ2: Does using a more advanced GPT model (GPT-4o-mini) further improve performance?
- RQ3: Does distinguishing between different levels of RNA editing (1%, 5%, 10%, 15%) and using continual fine-tuning (CFT) improve insights?
- RQ4: How does reducing the difference between edited and non-edited sites affect model performance?

The remainder of this paper is organized as follows: In section 2, we provide a detailed literature review, highlighting key advancements in RNA editing prediction and the application of generative AI models in this domain. section 3 describes our methodology, including data collection, preprocessing, and model training approaches. Experimental results and a discussion of our findings are presented in section 4, structured in the form of answers to the RQs. We conclude in section 5 by summarizing our contributions and discussing potential avenues for future research.

## 2 Previous Work

RNA can be conceptualized as a language, with its nucleotide sequences forming a code that carries biological information (see Appendix A). Generative AI models have demonstrated significant

potential in analyzing, predicting, and even generating biologically relevant RNA sequences. Since RNA consists of both a sequence and a secondary structure, it is well-suited for models capable of learning and generating linguistic patterns.

Specifically, A-to-I RNA editing, mediated by ADAR enzymes, has been extensively studied through both experimental and computational approaches. In experimental studies, large-scale synthesis of RNA sequences has been used to probe the specificity of the ADAR substrate systematically Uzonyi et al. (2021); Zambrano-Mila et al. (2023). In parallel, computational frameworks have been developed to predict A-to-I editing sites, utilizing distinct methodologies such as traditional machine learning, deep learning architectures, and more recently, transformer-based models.

Early machine learning approaches utilized algorithms such as Random Forests, Support Vector Machines, and probabilistic models to predict editing sites based on sequence and structural characteristics Tac et al. (2021); Ouyang et al. (2018). Recent studies employing XGBoost, an advanced tree-based ensemble method, have shown promise in predicting RNA editing efficiency but struggled with generalization across different substrates Liu et al. (2021); Jiang et al. (2024). Generally, while these methods improved upon earlier sequence motif-based approaches, they still faced accuracy limitations due to the complexity of editing site determination.

The advent of deep learning brought significant advancements to the field. Convolutional Neural Networks trained on large datasets of confident editing sites demonstrated good performance in predicting A-to-I editing events. EditPredict, for instance, employed CNN models to predict RNA editing in both *Alu* and non-*Alu* regions Wang et al. (2021). Another notable contribution was DeepAdar, a convolutional and recurrent neural network-based model, which was designed with a more structured feature extraction approach Jung et al. (2024). It incorporated predefined sequence motifs, secondary structure predictions, and base-pairing probabilities as input features. While DeepAdar demonstrated strong performance on curated RNA-seq datasets, its reliance on explicit feature engineering constrained its adaptability to novel sequence contexts. Additionally, its static feature representation limited its ability to dynamically learn from new data without retraining the entire model.

More recently, transformer-based models have emerged as promising tools for this task, treating RNA sequences akin to natural language, thereby improving the capture of complex sequence interactions. One of the first applications of large language models (LLMs) to this problem was a GPT-3.5-based approach, which treated RNA sequences as natural language and leveraged transformer-based sequence modeling Rosenwasser et al. (2024). Unlike previous deep learning methods, this approach captured long-range dependencies in RNA sequences without relying on handcrafted features. By framing the prediction task within a generative modeling context, it learned implicit sequence patterns and improved upon traditional classification-based models. Additionally, dynamic thresholding mechanisms and data augmentation techniques were incorporated to enhance model robustness. However, the model was trained on a generalized dataset that aggregated editing information across multiple tissues, which introduced inter-tissue variability and limited specificity. Moreover, it lacked continual learning capabilities, making it challenging to systematically adapt to new data distributions.

Building upon the strengths and addressing the limitations of these prior approaches, our current study introduces a new framework based on GPT-4o-mini. This model enhances RNA editing prediction through two key strategies (see section 3): (1) CFT across progressively refined datasets and (2) tissue-specific training focused on liver RNA editing sites.

Focusing on liver tissue provides a controlled biological context for analyzing ADAR1-mediated editing with minimal interference from other ADAR isoforms. The choice of liver tissue is motivated by its relatively high expression of ADAR1 compared to other tissues, while ADAR2 and ADAR3 levels remain particularly low. This distinction allows us to isolate ADAR1 activity and reduce confounding effects from other isoforms. Since RNA-binding protein expression varies across different tissues, selecting a single tissue for analysis helps neutralize variability in their expression levels across multiple tissues. By concentrating on a well-defined biological system, we improve both the specificity and interpretability of RNA editing site predictions.

Compared to DeepAdar, which requires manual feature extraction and has limited adaptability, GPT-4o-mini dynamically learns from raw RNA sequences and secondary structures. This flexibility makes it a more scalable solution. Relative to the previous GPT-3.5-based model, this study achieves

enhanced predictive accuracy and efficiency by leveraging dataset specificity and continual learning. As a result, our approach not only surpasses prior models in predictive performance but also provides a more interpretable and biologically grounded understanding of A-to-I RNA editing patterns.

## 3 DATA CURATION AND TRAINING METHODOLOGY

### 3.1 DATA COLLECTION AND PREPROCESSING – LIVER GTEX DATASET

To ensure high-confidence structure determination, we applied strict selection criteria: for each UTR containing *Alu* elements, we selected the closest oppositely oriented *Alu* pair, maximizing the likelihood of dsRNA formation. This resulted in a dataset of 905 *Alu* pairs, whose secondary structures were predicted using RNAfold from the ViennaRNA package Lorenz et al. (2011).

Unlike previous work, which analyzed editing across multiple tissues Rosenwasser et al. (2024), we specifically focused on 131 liver samples from the GTEx database Lonsdale et al. (2013), extracting editing levels for each adenosine within the selected dsRNA structures. This resulted in a dataset with annotated editing levels for all adenosines within these regions.

To handle data complexity and improve model performance, we employed two data manipulation strategies (Figure 1-A):

- **Overlapping Sites**: This subset included 16,752 training samples and 4,188 validation samples, totaling 20,940 samples. Editing sites were allowed to fall into multiple predefined editing level categories (e.g., sites with 1-5%, 5-10%, 10-15%, and above 15% editing).
- **Non-overlapping Sites**: This subset included 5,201 training samples and 1,301 validation samples, making a total of 6,502 samples. Each editing site was uniquely categorized into a single editing level range to prevent overlap between editing levels (e.g., the 1% group contained only sites with 1-5% editing).

To systematically analyze RNA editing patterns, we framed the problem as a classification task, determining whether a given site had been edited (Figure 1-B), similarly to Rosenwasser et al. (2024).

### 3.2 MODEL TRAINING APPROACHES — CONTINUAL FINE-TUNING (CFT)

CFT represents an advanced training methodology, progressively refining models through sequential tasks that mirror an evolving data distribution. In contrast to static fine-tuning (SFT) – which optimizes for a specific task at a single instance – CFT promotes continuous adaptation to new data and tasks while mitigating the forgetting of previously learned information Ke & Liu (2022). This approach is particularly beneficial in the era of LLMs, where constant updates and evolving datasets are the norm. CFT simplifies optimization objectives by focusing on better adaptation and less forgetting McCloskey & Cohen (1989), enhancing the model's resilience to the introduction of new data Shi et al. (2024).

CFT technique aligns perfectly with the needs of RNA editing site prediction, where the evolving biological data can benefit from continual model updates to capture new patterns and improve predictive accuracy over time. Initially, our CFT model (Figure 1-C) was trained on GPT-4o-mini[2] using a dataset representing lower editing thresholds (e.g., $1\% \overset{\text{FT}}{\rightarrow} GPT$ editing). Subsequently, the model underwent progressive fine-tuning across increasingly strict thresholds (i.e., $5\% \overset{\text{FT}}{\rightarrow} 1\% \overset{\text{FT}}{\rightarrow} GPT$, $10\% \overset{\text{FT}}{\rightarrow} 5\% \overset{\text{FT}}{\rightarrow} 1\% \overset{\text{FT}}{\rightarrow} GPT$, $15\% \overset{\text{FT}}{\rightarrow} 10\% \overset{\text{FT}}{\rightarrow} 5\% \overset{\text{FT}}{\rightarrow} 1\% \overset{\text{FT}}{\rightarrow} GPT$), allowing it to retain prior knowledge while refining its classification decisions. This is in contrast to the SFT model, which was trained separately on a single threshold subset (e.g., only $\geq 15\%$ editing, or $15\% \overset{\text{FT}}{\rightarrow} GPT$) for GPT-4o-mini and GPT-3.5 (mainly for comparison with previous results). Training and evaluation specifics, including loss and mean token accuracy, are in Appendix B.

[2]Azure OpenAI GPT-4o-mini fine-tuning: https://github.com/MicrosoftDocs/azure-ai-docs/blob/main/articles/ai-services/openai/tutorials/fine-tune.md

## 4 RESULTS AND ANALYSIS

**RQ1: Is there an alternative dataset that can better explain the enzymatic activity in the RNA editing site prediction problem than previously devised?**

To evaluate whether the liver-specific dataset provides a more informative representation of enzymatic activity in RNA editing, we adopted an empirical testing approach. Specifically, we used GPT-3.5 as a benchmark model, following the methodology applied in a previous study ( Rosenwasser et al. (2024)), which utilized a dataset derived from multiple tissues. By fine-tuning GPT-3.5 on our liver-specific dataset and comparing the performance metrics to those obtained in the previous research (which also used GPT-3.5 SFT), we aimed to determine whether this dataset indeed offers superior predictive features.

The results, presented in Table 1, indicate that the liver dataset led to a substantial improvement in predictive performance across multiple evaluation metrics. Notably, accuracy increased from 64.8% to 71.5%, and the F1 score improved from 64.8% to 69.4%. These improvements suggest that the liver-specific dataset, where ADAR1 is the predominant enzyme, captures more relevant biological features that influence RNA editing. The fact that GPT-3.5 performed better on this dataset – without any change in model architecture or training methodology – supports the conclusion that the alternative dataset provides a clearer and more biologically relevant signal for predicting RNA editing sites.

Table 1: Comparison of model performance metrics for RNA editing site prediction using datasets from combined tissues versus liver tissue. ACC: Accuracy, PRE: Precision, REC: Recall, F1: F1-score.

| Data | Fine-tuning | Base model | ACC | PRE | REC | F1 |
|---|---|---|---|---|---|---|
| **Combined Tissues** | SFT | GPT-3.5 | 64.8% | 64.7% | 65.0% | 64.8% |
| **Liver Tissue** | SFT | GPT-3.5 | 71.5% | 74.2% | 65.1% | 69.4% |

**RQ2: Does using a more advanced GPT model (GPT-4o-mini) further improve performance?**

To assess whether a more advanced model improves RNA editing site prediction, we designed two dataset configurations, while each group contained 6,502 sites, equally distributed between the "Yes" and "No" classes:

- Overlapping Editing Sites — where sites were classified based on whether their editing level was above or below a fixed threshold (e.g., 15%), ensuring an equal number of "Yes" and "No" labels. We tested multiple thresholds (1%, 5%, 10%, and 15%) to evaluate model robustness.
- Non-Overlapping Editing Sites — where sites were strictly separated into four distinct groups, preventing overlap between editing levels (e.g., the 1% group contained only sites with 1–5% editing, and so on). This partitioning allowed us to test whether a model could better generalize when given more clearly defined editing categories.

Using these datasets, we compared GPT-4o-mini against GPT-3.5. For the overlapping dataset, GPT-4o-mini achieved 71% accuracy, 73% precision, 65% recall, and a 77% F1-score (Figure 2-15% S[FT]), while GPT-3.5 had nearly identical results (72% accuracy, 74% precision, 65% recall, and a 69% F1-score) (Figure 2-15% S[FT] GPT-3.5). This indicates that the model's complexity had little effect in this setting, likely because the overlapping nature of the data limited its ability to learn distinct patterns.

However, for the non-overlapping dataset, GPT-4o-mini significantly outperformed GPT-3.5. It achieved 90% accuracy, 93% precision, 85% recall, and an 89% F1-score (Figure 3-15% S[FT]), compared to GPT-3.5's 85% accuracy, 95% precision, 73% recall, and an 83% F1-score (Figure 3-15% S[FT] GPT-3.5). The primary gain was in recall (+12%), demonstrating that the advanced model was better at identifying true editing sites when provided with well-separated categories. This suggests that while model improvements alone may not help in complex, overlapping cases, a well-structured dataset enables a more advanced model to extract clearer biological signals, leading to significant performance gains.

**RQ3: Does distinguishing between different levels of RNA editing (1%, 5%, 10%, 15%) and using a continual fine-tuning (CFT) improve insights?**

To answer this question, we examined whether a gradual, structured learning approach using CFT improves the model's ability to identify RNA editing patterns. We designed two training strategies: **(i)** CFT, where the model is iteratively trained on datasets with increasing editing thresholds (1%, 5%, 10%, and 15%), progressively refining its understanding, and **(ii)** static fine-tuning, where the model is trained directly on the 15% dataset without prior exposure to lower-threshold datasets.

For non-overlapping editing sites, CFT on GPT-4o-mini achieved superior performance on the 15% dataset with an accuracy of 91%, precision of 97%, recall of 84%, specificity of 98%, and an F1-score of 90%. Notably, the peak performance occurred at the 10% threshold with an F1-score of 96%, accuracy of 96%, precision of 99%, recall of 93%, and specificity of 99% (Figure 3-$15\% \xrightarrow{FT} 10\% \xrightarrow{FT} 5\% \xrightarrow{FT} 1\%$). In contrast, static fine-tuning resulted in lower performance (accuracy: 90%, precision: 93%, recall: 85%, specificity: 94%, F1-score: 89%) (Figure 3-15% S[FT]). For overlapping editing sites, CFT also outperformed static fine-tuning, achieving an accuracy of 79%, precision of 82%, recall of 74%, specificity of 84%, and an F1-score of 77% (Figure 2-$15\% \xrightarrow{FT} 10\% \xrightarrow{FT} 5\% \xrightarrow{FT} 1\%$). Static fine-tuning, by comparison, yielded an accuracy of 71%, precision of 73%, recall of 65%, specificity of 77%, and an F1-score of 69% (Figure 2-15% S[FT]).

These results indicate that CFT allows the model to better capture hierarchical editing patterns, particularly in the non-overlapping dataset, where structured learning progressively enhances predictive accuracy. Training dynamics across all experiments (Figure 5) confirm stable convergence, with mean token accuracy ranging between 0.85-0.95. However, non-overlapping data introduces higher noise during early iterations, likely due to the sharper distinction between groups, which initially complicates training but ultimately results in superior performance.

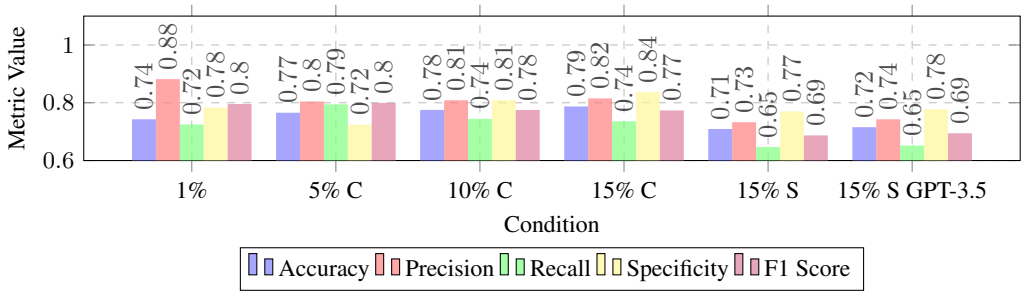


Figure 2: **Overlapping Editing Sites dataset training results using GPT-4o-mini vs. baselines:** Metrics evaluated for models trained using continual fine-tuning across increasing RNA editing thresholds (1%, 5% C[FT], 10% C[FT], 15% C[FT]), as well as a model trained statically on the 15% dataset (15% S[FT]) and a GPT-3.5 S[FT] baseline for the overlapping dataset. The figure illustrates the effect of continual fine-tuning versus static fine-tuning on classification performance.

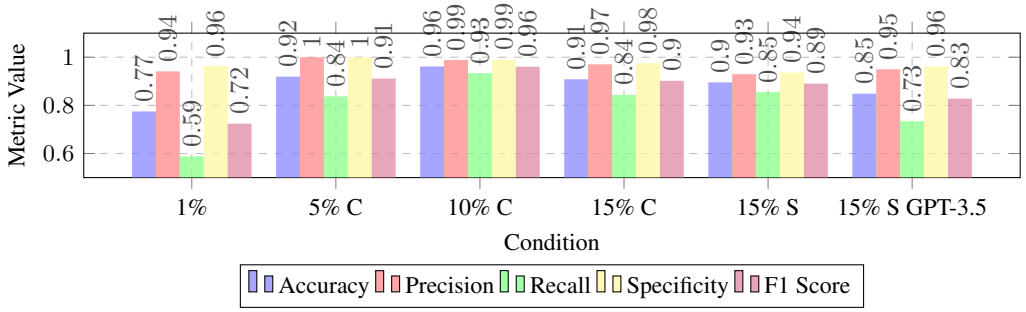


Figure 3: **Non-Overlapping Editing Sites dataset training results using GPT-4o-mini vs. baselines:** Metrics evaluated for models trained using continual fine-tuning across non-overlapping RNA editing level groups (1%, 5% C[FT], 10% C[FT], 15% C[FT]), as well as a model trained statically on the 15% dataset (15% S[FT]) and a GPT-3.5 S[FT] baseline. The figure illustrates the impact of training strategy on model performance when editing thresholds are distinct across groups.

**RQ4: How does reducing the difference between edited and non-edited sites affect model performance?**

One possible explanation for the model's strong performance in the non-overlapping experiments is the substantial difference between sites labeled as "Yes" and "No" in the dataset. In the 15% group, all "Yes" sites had editing levels $> 15\%$, whereas most "No" sites had editing levels well below $1\%$. This distribution arose because sites with editing levels between $1\%$ and $15\%$ were assigned to other groups, creating a sharp contrast between edited and non-edited sites that likely facilitated classification.

To investigate whether reducing this difference affects model performance, we restructured the dataset into three groups, gradually minimizing the gap between "Yes" and "No" classes:

- **1–5% group:** "Yes" for sites with $1\% \leq$ editing level $< 5\%$, and "No" for sites with $0\% \leq$ editing level $< 1\%$.
- **5–10% group:** "Yes" for sites with $5\% \leq$ editing level $< 10\%$, and "No" for sites with $1\% \leq$ editing level $< 5\%$.
- **10–15% group:** "Yes" for sites with $> 15\%$, and "No" for sites with $5\% \leq$ editing level $< 10\%$.

Using static fine-tuning on the GPT-4o-mini model for these newly defined groups, we observed a marked decline in performance compared to previous experiments with clearer distinctions between edited and non-edited sites. In the prior context, the stark contrast between highly edited sites ($> 15\%$) and completely unedited sites ($< 1\%$) provided a clear separation, leading to better classification accuracy. However, in this experiment, the model struggled to distinguish between sites with closer editing levels, resulting in substantially lower performance (Figure 4).

As the editing thresholds increased, performance further deteriorated, with the most substantial decline occurring in the 10–15% group, where the model failed to generalize effectively (Figure 4). This suggests that when the editing levels between the "Yes" and "No" classes are closer, the distinguishing features become less prominent, making pattern recognition more difficult.

From a biological perspective, these findings are consistent with expectations. Sites with high editing levels ($> 15\%$) likely have distinct sequence and structural properties, making them strong ADAR substrates and easily separable from unedited sites. Conversely, sites with intermediate editing levels (e.g., 5–10%) are less optimal ADAR targets, leading to weaker differentiation between "Yes" and "No" classes and, consequently, decreased model performance. This insight highlights the limitations of classification models when the distinction between classes is subtle, suggesting that predictive performance is strongly influenced by the inherent separability of the biological features associated with RNA editing levels.

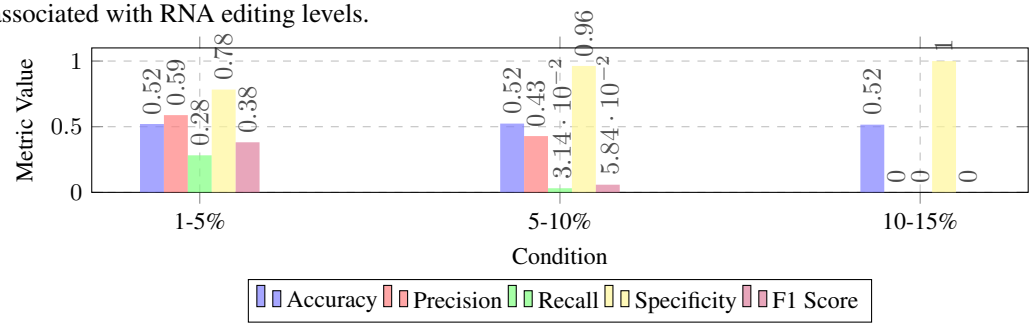


Figure 4: Performance metrics for classification with minimal editing level differences. Metrics evaluated for models trained on datasets where the difference between edited and non-edited sites is minimized (1–5%, 5–10%, 10–15%). Significant decline in model performance when distinguishing between sites with similar editing levels, highlighting the challenge of classification at intermediate editing thresholds.

## 5 DISCUSSION & CONCLUSION

Our study demonstrates the effectiveness of a liver-specific dataset in improving RNA editing site prediction, attributed to its high ADAR1 and low ADAR2/ADAR3 expression levels, which provide

clearer editing patterns. Stratifying editing levels into distinct thresholds (1%, 5%, 10%, 15%) facilitated better model training. The CFT approach, starting with lower editing levels and progressing to higher ones, proved highly effective. Future work will extend this analysis to additional tissues and across species to assess the model's generalizability and integrate more advanced reasoning models, such as GPT-o1-mini.

## ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health award R35GM122543 (M.L.), National Natural Science Foundation of China grant No. 31770776 (F.Z.), Foundation Fighting Blindness (grants TA-GT-0620-0790-HUJ and PPA-0923-0865-HUJ), grants from the Israeli Ministry of Science (grant 3-17916) and by Israel Science Foundation (2637/2). Michael Levitt is the Robert W. and Vivian K. Cahill Professor of Cancer Research.

## REFERENCES

Alekos Athanasiadis, Alexander Rich, and Stefan Maas. Widespread a-to-i rna editing of alu-containing mrnas in the human transcriptome. *PLoS biology*, 2(12):e391, 2004.

Brenda L Bass. Rna editing by adenosine deaminases that act on rna. *Annual review of biochemistry*, 71(1):817–846, 2002.

Lily Bazak, Ami Haviv, Michal Barak, Jasmine Jacob-Hirsch, Patricia Deng, Rui Zhang, Farren J Isaacs, Gideon Rechavi, Jin Billy Li, Eli Eisenberg, et al. A-to-i rna editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome research*, 24(3): 365–376, 2014.

Axel Brennicke, Anita Marchfelder, and Stefan Binder. Rna editing. *FEMS microbiology reviews*, 23(3):297–316, 1999.

Francesco Calvanese, Camille N Lambert, Philippe Nghe, Francesco Zamponi, and Martin Weigt. Towards parsimonious generative modeling of rna families. *Nucleic Acids Research*, 52(10): 5465–5477, 2024.

Jonatha M Gott and Ronald B Emeson. Functions and mechanisms of rna editing. *Annual review of genetics*, 34(1):499–531, 2000.

Yue Jiang, Lina R Bagepalli, Bora S Banjanin, Yiannis A Savva, Yingxin Cao, Lan Guo, Adrian W Briggs, Brian Booth, and Ronald J Hause. Generative machine learning of adar substrates for precise and efficient rna editing. *bioRxiv*, pp. 2024–09, 2024.

Andrew J Jung, ALICE J. GAO, Leo J Lee, and Brendan Frey. DeepADAR: A deep learning approach to model regulatory elements of ADAR-based RNA editing and its application to gRNA design. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. URL `https://openreview.net/forum?id=D8wVEAKFRA`.

Mehran Karimzadeh, Amir Momen-Roknabadi, Taylor B Cavazos, Yuqi Fang, Nae-Chyun Chen, Michael Multhaup, Jennifer Yen, Jeremy Ku, Jieyang Wang, Xuan Zhao, et al. Deep generative ai models analyzing circulating orphan non-coding rnas enable detection of early-stage lung cancer. *Nature Communications*, 15(1):10090, 2024.

Dhruva Katrekar, James Yen, Yichen Xiang, Anushka Saha, Dario Meluzzi, Yiannis Savva, and Prashant Mali. Efficient in vitro and in vivo rna editing via recruitment of endogenous adars using circular guide rnas. *Nature biotechnology*, 40(6):938–945, 2022.

Zixuan Ke and Bing Liu. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*, 2022.

Dennis DY Kim, Thomas TY Kim, Thomas Walsh, Yoshifumi Kobayashi, Tara C Matise, Steven Buyske, and Abram Gabriel. Widespread rna editing of embedded alu elements in the human transcriptome. *Genome research*, 14(9):1719–1725, 2004.

Erez Y Levanon, Eli Eisenberg, Rodrigo Yelin, Sergey Nemzer, Martina Hallegger, Ronen Shemesh, Zipora Y Fligelman, Avi Shoshan, Sarah R Pollock, Dan Sztybel, et al. Systematic identification of abundant a-to-i editing sites in the human transcriptome. *Nature biotechnology*, 22(8):1001–1005, 2004.

Huaqing Liu, Peiyi Chen, Xiaochen Zhai, Ku-Geng Huo, Shuxian Zhou, Lanqing Han, and Guoxin Fan. Ppb-affinity: Protein-protein binding affinity dataset for ai-based protein drug discovery. *Scientific Data*, 11(1):1–11, 2024.

Xin Liu, Tao Sun, Anna Shcherbina, Qin Li, Inga Jarmoskaite, Kalli Kappel, Gokul Ramaswami, Rhiju Das, Anshul Kundaje, and Jin Billy Li. Learning cis-regulatory principles of adar-based rna editing from crispr-mediated mutagenesis. *Nature communications*, 12(1):2165, 2021.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6:1–14, 2011.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Tobias Merkle, Sarah Merz, Philipp Reautschnig, Andreas Blaha, Qin Li, Paul Vogel, Jacqueline Wettengel, Jin Billy Li, and Thorsten Stafforst. Precise rna editing by recruiting endogenous adars with antisense oligonucleotides. *Nature biotechnology*, 37(2):133–138, 2019.

Maria Fernanda Montiel-Gonzalez, Juan Felipe Diaz Quiroz, and Joshua JC Rosenthal. Current strategies for site-directed rna editing using adars. *Methods*, 156:16–24, 2019.

Kazuko Nishikura. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry*, 79(1):321–349, 2010.

Zhangyi Ouyang, Feng Liu, Chenghui Zhao, Chao Ren, Gaole An, Chuan Mei, Xiaochen Bo, and Wenjie Shu. Accurate identification of rna editing sites from primitive sequence with deep neural networks. *Scientific Reports*, 8(1):6005, 2018.

Laura S Pfeiffer and Thorsten Stafforst. Precision rna base editing with engineered and endogenous effectors. *Nature Biotechnology*, 41(11):1526–1542, 2023.

Philipp Reautschnig, Carolin Fruhner, Nicolai Wahn, Charlotte P Wiegand, Sabrina Kragness, John F Yung, Daniel T Hofacker, Jenna Fisk, Michelle Eidelman, Nils Waffenschmidt, et al. Precise in vivo rna base editing with a wobble-enhanced circular cluster guide rna. *Nature biotechnology*, pp. 1–13, 2024.

Aidan T Riley, James M Robson, and Alexander A Green. Generative and predictive neural networks for the design of functional rna molecules. *bioRxiv*, 2023.

Zohar Rosenwasser, Erez Levanon, Michael Levitt, and Gal Oren. Detection of rna editing sites by gpt fine-tuning. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

Huseyin Avni Tac, Mustafa Koroglu, and Ugur Sezerman. Rddsvm: accurate prediction of a-to-i rna editing sites from sequence using support vector machines. *Functional & Integrative Genomics*, 21(5):633–643, 2021.

Anna Uzonyi, Ronit Nir, Ofir Shliefer, Noam Stern-Ginossar, Yaron Antebi, Yonatan Stelzer, Erez Y Levanon, and Schraga Schwartz. Deciphering the principles of the rna editing code via large-scale systematic probing. *Molecular cell*, 81(11):2374–2387, 2021.

Jiandong Wang, Scott Ness, Roger Brown, Hui Yu, Olufunmilola Oyebamiji, Limin Jiang, Quanhu Sheng, David C Samuels, Ying-Yong Zhao, Jijun Tang, et al. Editpredict: prediction of rna editable sites with convolutional neural network. *Genomics*, 113(6):3864–3871, 2021.

Yijia Xiao, Edward Sun, Yiqiao Jin, and Wei Wang. Rna-gpt: Multimodal generative system for rna sequence understanding. *arXiv preprint arXiv:2411.08900*, 2024.

Marlon S Zambrano-Mila, Monika Witzenberger, Zohar Rosenwasser, Anna Uzonyi, Ronit Nir, Shay Ben-Aroya, Erez Y Levanon, and Schraga Schwartz. Dissecting the basis for differential substrate specificity of adar1 and adar2. *Nature Communications*, 14(1):8212, 2023.

He Zhang, Hailong Liu, Yushan Xu, Yiming Liu, Jia Wang, Yan Qin, Haiyan Wang, Lili Ma, Zhiyuan Xun, Timothy K Lu, et al. Deep generative models generate mrna sequences with enhanced translation capacity and stability. *bioRxiv*, pp. 2024–06, 2024.

Yichong Zhao, Kenta Oono, Hiroki Takizawa, and Masaaki Kotera. Generrna: A generative pretrained language model for de novo rna design. *PLoS One*, 19(10):e0310814, 2024.

## A APPENDIX: PREVIOUS WORK — GENERATIVE AI FOR RNA TASKS

RNA can be conceptualized as a language, with its nucleotide sequences forming a code that carries biological information. Just as natural language models have been able to process and generate meaningful text, generative AI models have the potential to analyze, predict, and even generate biologically relevant RNA sequences. The structured nature of RNA, including its sequence, motifs, and secondary structure, suggests that models capable of learning and generating linguistic patterns could be leveraged for RNA-related tasks.

Recent advancements in the field have demonstrated the efficacy of generative models in handling RNA-specific challenges. For instance, GenerRNA Zhao et al. (2024), a Transformer-based model inspired by large language models, has shown remarkable success in de novo RNA design. This model, pre-trained on approximately 16-30 million RNA sequences, can generate novel RNA sequences with stable secondary structures while ensuring distinctiveness from existing sequences.

The application of fully generative models, such as GPT-like architectures Xiao et al. (2024), to RNA tasks, has also yielded promising results. RNA-GPT, a multimodal generative system, combines RNA sequence encoders with linear projection layers and state-of-the-art LLMs for precise representation alignment. This approach enables the processing of user-uploaded RNA sequences and provides concise, accurate responses to RNA-related queries.

These models have successfully captured the representation of RNA sequences, allowing for their utilization in various tasks. Some examples of these additional applications include protein binding prediction, where BERT-RBP Liu et al. (2024) adapts the BERT architecture pre-trained on a human reference genome to predict RNA-protein interactions. In enhanced mRNA design, GEMORNA Zhang et al. (2024), a deep generative model, has been developed to optimize mRNA coding sequences and untranslated regions, improving translation efficiency for therapeutic applications. In RNA family modeling, Edge Activation Direct Coupling Analysis (eaDCA) Calvanese et al. (2024) provides a generative framework for understanding RNA sequence variation and structural constraints. Additionally, Generative Adversarial RNA Design Networks (GARDN) Riley et al. (2023) have been used to generate realistic and functional RNA molecules, advancing synthetic RNA design. In cancer research, the deep generative model Orion Karimzadeh et al. (2024) has been applied to analyze circulating orphan non-coding RNAs (oncRNAs) for early cancer detection and tumor classification. These diverse applications highlight the versatility and potential of generative AI models in advancing RNA research across multiple domains.

# B APPENDIX: TRAINING PERFORMANCES OF FINE-TUNED GPT MODELS

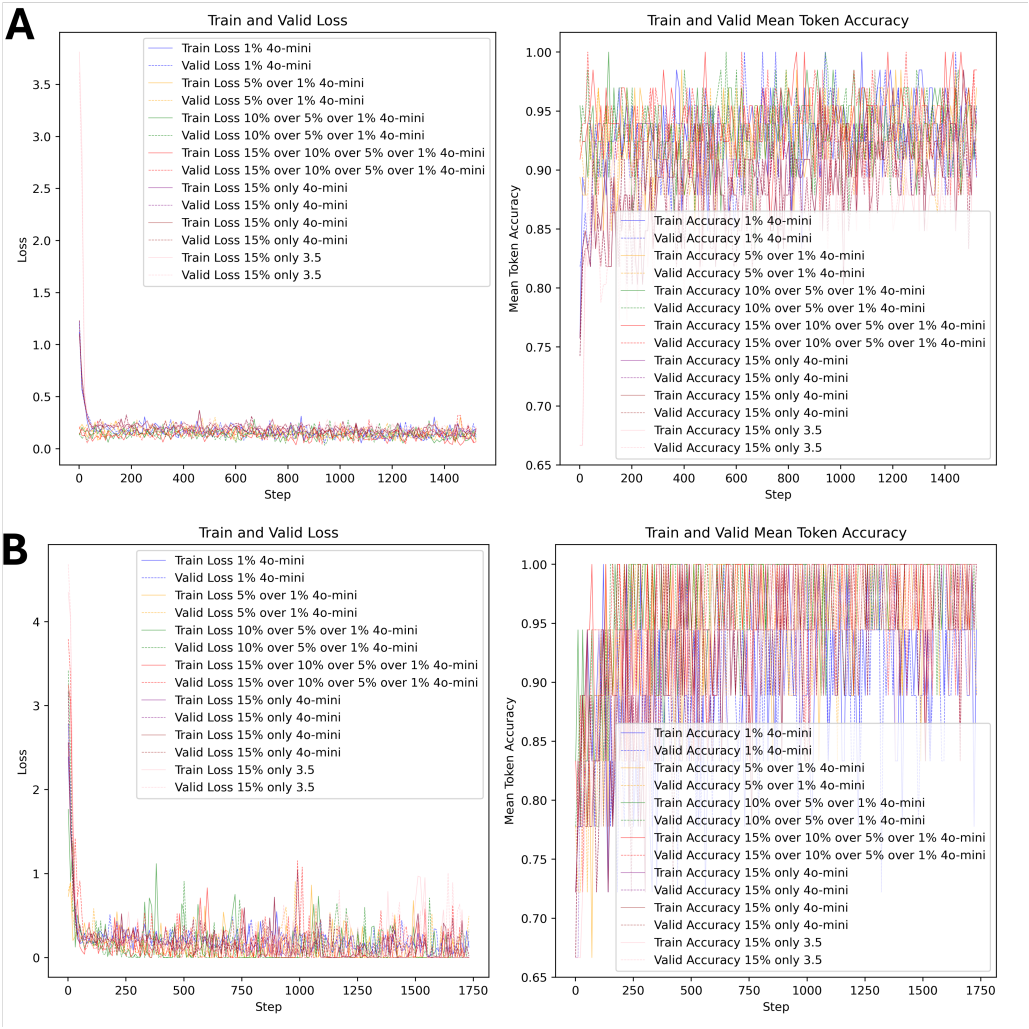


Figure 5: Training and validation performance of fine-tuned GPT-4o-mini and GPT-3.5 models for RNA editing site prediction. (A) overlapping sites experiments. (B) non-overlapping sites experiments. In each panel, the left plot shows the training and validation loss across steps for different models, including GPT-4o-mini and GPT-3.5, trained under various fine-tuning strategies. The right plot presents the train and validation mean token accuracy for the same models. The trends illustrate the convergence behavior and performance stability of fine-tuned models across different experimental conditions.

In this study, we intentionally used a small batch size of 6 due to the large instance size, ensuring that training captured variability in the data while preventing overfitting. The added noise from smaller batches promotes more generalized learning by preventing the model from converging too quickly to suboptimal solutions. Additionally, we limited training to only two epochs, as further iterations did not yield meaningful improvements in loss reduction or token accuracy. The mean token accuracy remained relatively high throughout training, indicating that the model effectively learned the underlying patterns early on. This approach balances computational efficiency with model generalization, avoiding excessive training that could lead to diminishing returns.