# RG-RCT: Real-Governed Representation-Calibrated Training for Low-Resource Medical ASR

Sri Charan Devarakonda, Ravi Sastry Kolluru, Srihari Bandarupalli, Bhavana Akkiraju, Anil Kumar Vuppala
LTRC, International Institute of Information Technology Hyderabad
{sricharan.d, kolluru.s, srihari.bandarupalli, bhavana.akkiraju}@research.iiit.ac.in, anil.vuppala@iiit.ac.in

***Abstract*—Developing Automatic Speech Recognition (ASR) systems for specialized medical domains is challenging in low-resource settings due to the limited availability of annotated speech data. Synthetic speech generated using text to speech (TTS) systems is often used to augment training data, but directly mixing synthetic and real speech can introduce distribution mismatch that degrades encoder representations. This work introduces Real-Governed Representation-Calibrated Training (RG-RCT), a training strategy that regulates the influence of synthetic speech at the representation level. The proposed framework combines reliability-based confidence estimation, representation alignment, and layer-wise representation governance to stabilize the encoder feature space while preserving acoustic diversity from multi-speaker synthetic data. Experiments on medical-domain speech datasets using Wav2Vec2 and Whisper demonstrate that RG-RCT consistently outperforms conventional strategies such as direct mixing and confidence-aware training. The proposed method achieves the lowest Word Error Rate (WER), reaching 17.1% and 16.6% for Telugu and 23.3% and 25.6% for Kannada. These results indicate that regulating encoder representation geometry enables more effective utilization of synthetic speech for low-resource medical ASR.**

***Index Terms*—ASR, TTS, RG-RCT, Confidence-Aware**

## I. Introduction

Recent progress in automatic speech recognition (ASR) has been largely driven by deep neural architectures and self-supervised learning methods. Representation learning models such as wav2vec 2.0 [1], HuBERT [2], and WavLM [3] learn meaningful acoustic features from large-scale unlabeled speech corpora, enabling strong performance across a variety of languages and acoustic conditions. Similarly, large-scale weakly supervised systems such as Whisper demonstrate that training on massive multilingual speech datasets can significantly improve robustness and generalization [4]. In addition, data augmentation techniques such as SpecAugment have been widely adopted to improve the robustness of ASR models under limited data conditions [5]. Despite these advances, building accurate ASR systems remains challenging for low-resource languages where annotated speech data is scarce. This problem becomes even more pronounced for morphologically rich languages such as Telugu and Kannada, particularly in specialized domains such as healthcare where domain-specific speech data is limited.

One widely explored solution for addressing limited training data is synthetic speech augmentation. Early studies demonstrated that augmenting training datasets with speech generated through text-to-speech systems can improve recognition performance in end-to-end ASR models [6]. Subsequent work expanded this idea by generating large-scale synthetic datasets from text corpora, enabling models to learn from substantially larger training collections [7]. Multilingual representation learning efforts for Indian languages have also shown that cross-lingual pretraining can transfer useful acoustic knowledge across related languages and improve performance in low-resource scenarios [8]. Complementary research has explored semi-supervised training approaches where pseudo-labels generated from large pretrained ASR models are used to improve model robustness in low-resource settings [9], [10].

Recent research has further explored synthetic speech as a mechanism for domain adaptation. Some approaches investigate training ASR systems primarily using synthetic speech when real recordings are limited [11]. Other studies show that combining TTS with voice conversion (VC) can significantly increase speaker diversity, allowing models to learn more robust representations that generalize better across speakers [12]. While these approaches increase the quantity and diversity of available training data, they also introduce a critical challenge: synthetic speech often follows a different acoustic distribution than real speech.

To mitigate the mismatch between synthetic and real speech, several techniques have been proposed. Some approaches attempt to reduce the synthetic–real gap by carefully selecting or filtering synthetic data used for training. For example, Liu *et al.* [13] propose a method for selecting synthetic speech generated from TTS systems based on its similarity to real speech, allowing only acoustically reliable synthetic samples to be used for ASR training. Other methods attempt to reduce this mismatch through model adaptation strategies. Task arithmetic methods adjust model parameters using knowledge transferred from real speech data [14], while representation alignment approaches such as Real-to-Synthetic (R2S) learning aim to bring embeddings derived from synthetic speech closer to those obtained from real recordings [15]. Studies examining purely synthetic training datasets further highlight that recognition performance strongly depends on the acoustic realism and diversity of generated speech [16].

Another line of research focuses on estimating the reliability of training samples through confidence estimation. Entropy-based methods quantify prediction uncertainty in end-to-end ASR models to identify unreliable outputs [17]. More advanced approaches incorporate linguistic and alignment information when estimating confidence, such as the Temporal

Lexeme Similarity (TeLeS) score [18] and the TruCLeS confidence framework [19]. Although these techniques help identify unreliable predictions, they typically operate at the output level and do not explicitly regulate how synthetic speech affects internal encoder representations during training.

Beyond confidence estimation, several works investigate representation learning strategies for improving cross-domain robustness. Domain-adversarial training encourages the learning of domain-invariant features by aligning representations from different data distributions [20]. Similarly, contrastive representation learning methods such as Contrastive Predictive Coding learn invariant speech representations by maximizing mutual information between latent features across temporal contexts [21]. However, these methods typically focus on global representation alignment and do not explicitly control how synthetic data influences intermediate encoder representations during training.

In this work, we introduce a Real-Governed Representation-Calibrated Training (RG-RCT) framework for domain-specific ASR. The central idea of the proposed approach is that synthetic speech should contribute acoustic variability while the encoder representation space remains governed by real speech data. The framework combines reliability estimation, representation alignment, and layer-wise representation governance to regulate how synthetic samples influence encoder representations during training.

## II. Proposed Methodology

Consider a low-resource ASR setting where training data consists of real speech together with synthetic speech generated using TTS systems. Although synthetic speech increases the amount of available training data, it often follows a different acoustic distribution than natural speech:

$$P_{\text{synthetic}}(x|y) \neq P_{\text{real}}(x|y) \tag{1}$$

where $x$ denotes the speech signal and $y$ represents the corresponding transcript. When synthetic speech dominates training, the encoder may learn artifacts specific to synthetic generation rather than robust linguistic representations. This effect becomes more pronounced in multi-speaker TTS systems, where increased speaker variability can distort the representation space.

The RG-RCT framework regulates how synthetic samples influence encoder representations during training while allowing synthetic speech to provide acoustic diversity.

### A. Problem Setup

Let the real speech dataset be

$$R = \{(x_i^r, y_i)\} \tag{2}$$

where $x_i^r$ denotes real audio and $y_i$ its corresponding transcript.

Synthetic speech generated using TTS is divided into aligned and unaligned sets:

$$S_a = \{(x_i^a, y_i)\}, \qquad S_u = \{(x_j^u, y_j)\} \tag{3}$$

Aligned samples share transcripts with real recordings, while unaligned samples contain additional sentences.

Let the ASR encoder contain $L$ layers. The hidden representation at layer $l$ is defined as

$$h_i^{(l)} = f_\theta^{(l)}(x_i) \tag{4}$$

The objective is to utilize synthetic speech while maintaining encoder representations that remain governed by real speech.

### B. Confidence Scoring Mechanism

Synthetic samples may vary in acoustic quality and reliability. A confidence scoring mechanism is therefore used to estimate the reliability of each training sample and regulate its influence during training.

*1) Static Acoustic Reliability:* Following prior work [22], a static reliability score is assigned to each synthetic sample based on acoustic realism:

$$C_{static,i} = \alpha S_{perceptual,i} + \beta S_{sim,i} + \gamma S_{wer,i} \tag{5}$$

where $S_{perceptual}$ measures acoustic feature consistency, $S_{sim}$ represents MFCC similarity between real and synthetic speech, and $S_{wer}$ evaluates recognition reliability using an external ASR model. The weights satisfy $\alpha + \beta + \gamma = 1$.

*2) Representation Reliability for Aligned Speech:* For aligned synthetic samples, representation consistency with real speech is evaluated using encoder embeddings:

$$R_i = \|h_{i,a}^{(L)} - h_{i,r}^{(L)}\| \tag{6}$$

The representation reliability is defined as

$$C_{repr,i} = \exp(-R_i) \tag{7}$$

The final confidence score becomes

$$C_i^{aligned} = C_{static,i} \cdot C_{repr,i} \tag{8}$$

*3) Distribution Reliability for Unaligned Speech:* For unaligned synthetic samples, reliability is estimated based on their distance from the distribution of real speech representations. Following distribution-based out-of-distribution detection approaches that use Mahalanobis distance in feature space [23], we model the real speech embedding distribution using its mean and covariance.

$$\mu_r = E[h_r^{(L)}], \quad \Sigma_r = Cov(h_r^{(L)}) \tag{9}$$

The Mahalanobis distance between the representation of an unaligned synthetic sample and the real speech distribution is computed as

$$D_i = (h_u^{(L)} - \mu_r)^T \Sigma_r^{-1} (h_u^{(L)} - \mu_r) \tag{10}$$

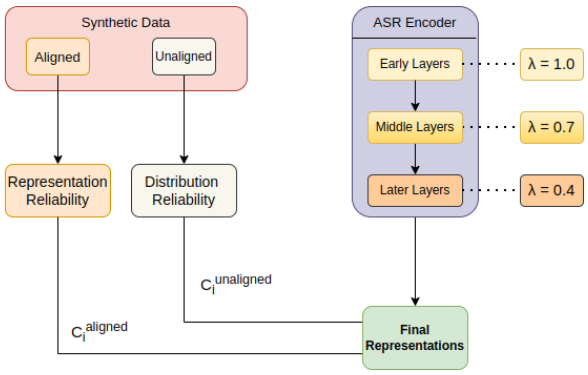


Fig. 1. Overview of the proposed RG-RCT methodology.

This distance measures how far the synthetic representation deviates from the real speech manifold in the encoder embedding space. Larger distances indicate a higher likelihood that the synthetic sample lies outside the distribution of natural speech representations.

The distribution reliability score is therefore defined as

$$C_{dist,i} = \exp(-D_i) \tag{11}$$

which assigns higher reliability to synthetic samples that are closer to the real speech representation distribution.

The final confidence score for unaligned synthetic samples is computed by combining the static acoustic reliability score with the distribution reliability:

$$C_i^{unaligned} = C_{static,i} \cdot C_{dist,i} \tag{12}$$

Real speech samples are considered fully reliable and are therefore assigned a confidence value of

$$C_{real} = 1 \tag{13}$$

### C. Layer-wise Representation Governance

To maintain real-governed representations, the influence of synthetic speech is controlled across encoder layers using layer-dependent coefficients:

$$\lambda_l = \begin{cases} 1.0 & l \leq \frac{L}{3} \\ 0.7 & \frac{L}{3} < l \leq \frac{2L}{3} \\ 0.4 & l > \frac{2L}{3} \end{cases} \tag{14}$$

Synthetic representations are interpolated with real representations as

$$\tilde{h}_s^{(l)} = \lambda_l h_s^{(l)} + (1 - \lambda_l) h_r^{(l)} \tag{15}$$

The resulting representations are further weighted using the confidence scores:

$$\hat{h}_s^{(l)} = C_i \cdot \tilde{h}_s^{(l)} \tag{16}$$

### D. Representation Stabilization Strategy

To prevent synthetic speech from distorting the encoder representation space, stabilization constraints are applied. These constraints align synthetic representations with real speech and reduce the distribution drift caused by synthetic data.

*1) Real-Anchored Representation Alignment:* For aligned synthetic samples, the encoder representations are encouraged to remain close to the corresponding real speech representations.

$$L_{anchor}^{(l)} = \mathbb{E}_i \left[ \| h_{i,a}^{(l)} - h_{i,r}^{(l)} \|^2 \right] \tag{17}$$

The total anchor loss across all encoder layers is computed as

$$L_{anchor} = \sum_{l=1}^{L} w_l L_{anchor}^{(l)} \tag{18}$$

*2) Distribution Drift Stabilization:* To reduce representation drift caused by unaligned synthetic speech, a distribution alignment loss is introduced:

$$L_{drift} = \| \mu_u - \mu_r \|^2 \tag{19}$$

*3) Training Objective:* The final objective combines the standard ASR loss with the proposed stabilization terms:

$$L_{total} = L_{ASR} + \lambda_{anchor} L_{anchor} + \lambda_{drift} L_{drift} \tag{20}$$

where $L_{ASR}$ denotes the standard ASR training loss (CTC [24] for Wav2Vec2 or cross-entropy for Whisper). This formulation enables synthetic speech to contribute acoustic diversity while maintaining encoder representations anchored to real speech.

## III. Experimental Setup

### A. Medical Domain Speech Database

*1) Real Medical Speech Data:* Real medical-domain speech datasets were used for both Telugu and Kannada. For Telugu, an in-house corpus consisting of recordings from 87 speakers (40 male, 47 female) was collected, containing healthcare-related utterances such as disease names, symptoms, and treatment instructions. The recordings were captured in controlled conditions and resampled to 16 kHz. From this corpus, 16 hours were used for training and 5 hours for testing.

For Kannada, real speech data was obtained from a publicly available Kaggle dataset. From this dataset, 8 hours were used for training and 2.5 hours were reserved for evaluation. All audio samples were normalized and resampled to 16 kHz to maintain consistency with the Telugu dataset.

TABLE I
WER (%) COMPARISON OF DIFFERENT TRAINING STRATEGIES FOR TELUGU AND KANNADA MEDICAL-DOMAIN ASR.

| **Training Configuration** | **Telugu** | | **Kannada** | |
|---|---|---|---|---|
| | **Wav2Vec2** | **Whisper** | **Wav2Vec2** | **Whisper** |
| *Real Speech Only* | | | | |
| Real Only (Fine-tuning) | 25.8 | 27.1 | 30.1 | 34.6 |
| *Synthetic Speech Only* | | | | |
| Synthetic Only – Multi Speaker | 30.7 | 32.0 | 36.9 | 38.5 |
| Synthetic Only – Single Speaker | 34.2 | 35.6 | 40.8 | 42.3 |
| *Real + Synthetic (Direct Mixing)* | | | | |
| Real + Synthetic – Multi Speaker | 19.9 | 19.8 | 27.5 | 27.6 |
| Real + Synthetic – Single Speaker | 21.6 | 21.8 | 29.8 | 29.0 |
| *Confidence-Aware Training* | | | | |
| Confidence Aware – Multi Speaker | 19.3 | 24.1 | 27.1 | 33.4 |
| Confidence Aware – Single Speaker | 18.4 | 22.5 | 26.2 | 31.8 |
| *Proposed Method* | | | | |
| RG-RCT (Real + Synthetic Multi Speaker) | 17.1 | **16.6** | **23.3** | 25.6 |

*2) Synthetic Speech Generation:* To augment the limited real data, synthetic speech was generated using two TTS systems: IndicTTS [25] and GlowTTS [26]. For Telugu, 35 hours of synthetic speech were generated, including 17 hours using IndicTTS and 18 hours using GlowTTS. For Kannada, 18 hours of synthetic speech were generated, consisting of 9 hours each from IndicTTS and GlowTTS. The synthetic data includes both sentences aligned with the real corpus and additional domain-specific sentences to increase acoustic variability and lexical coverage.

*3) Voice Conversion Augmentation:* To further increase speaker diversity, voice conversion was applied to the synthetic speech. For Telugu, 15 reference speakers (8 male, 7 female) were selected from the recorded speaker pool, while for Kannada, 7 reference speakers (5 male, 2 female) were used. The synthetic utterances were converted to these speaker styles using the FreeVC voice conversion model [27], implemented through the Coqui TTS framework.

### B. Implementation Details

The ASR experiments were conducted using two pretrained models, Wav2Vec2-Large (317M parameters) and Whisper-Medium (769M parameters), which were fine-tuned on the medical-domain datasets. Training was performed with a learning rate of $10^{-3}$ using cosine annealing, a batch size of 8 per GPU, and the AdamW optimizer on six NVIDIA GeForce RTX 2080 Ti GPUs. Models were trained for up to 60 epochs with early stopping based on validation performance. The GlowTTS model used for synthetic speech generation was trained separately prior to data synthesis. The confidence scoring weights were set to $\alpha = 0.4$, $\beta = 0.3$, and $\gamma = 0.3$ based on validation experiments.

*a) Confidence-Aware Training Baseline.:* In addition to the proposed RG-RCT framework, we implement a confidence-aware training baseline using the reliability score defined in Eq. (5). In this setting, each training sample is weighted by its confidence value during optimization. Specifically, higher-confidence samples contribute more strongly to the ASR loss, while lower-confidence synthetic samples are down-weighted. This baseline does not include representation alignment, layer-wise governance, or distribution-based stabilization, and serves to evaluate the effectiveness of confidence-based sample weighting alone.

## IV. RESULTS AND ANALYSIS

### A. Comparison of Training Strategies

Table I summarizes the Word Error Rate (WER) obtained using different training strategies across the Telugu and Kannada medical-domain datasets. The experiments compare real-only training, synthetic-only training, direct mixing of real and synthetic speech, confidence-aware training, and the proposed Real-Governed Representation-Calibrated Training (RG-RCT) framework.

Training with only real speech results in relatively high WER due to the limited availability of domain-specific medical speech data. The real-only baseline achieves WER values of 25.8% and 27.1% for Telugu and 30.1% and 34.6% for Kannada using Wav2Vec2 and Whisper respectively. When models are trained using only synthetic speech, performance deteriorates further, with single-speaker synthetic data producing the highest WER values across both languages. Although multi-speaker synthetic speech slightly improves performance compared to single-speaker synthesis due to increased acoustic variability, synthetic-only training remains significantly worse than real-only training, confirming that synthetic speech alone cannot capture the acoustic characteristics of natural speech.

Combining real and synthetic speech through direct mixing significantly improves recognition performance. The multi-speaker direct mixing configuration achieves WER values of 19.9% and 19.8% for Telugu and 27.5% and 27.6% for Kannada. Multi-speaker synthetic augmentation performs better than single-speaker augmentation because voice conversion introduces a wider range of speaker characteristics, allowing the encoder to observe a broader acoustic space and learn more robust representations.

Confidence-aware training attempts to reduce the synthetic–real mismatch by weighting training samples according to reliability scores derived from perceptual features, acoustic similarity, and recognition accuracy. While this approach improves performance in the single-speaker setting, it becomes

less effective when multi-speaker synthetic data is introduced. Multi-speaker synthetic speech increases the geometric spread of encoder representations in the feature space, making static confidence metrics less reliable. As a result, useful synthetic samples may receive lower confidence scores, leading to sub-optimal weighting during training and higher WER compared to direct mixing.

The proposed RG-RCT framework achieves the best performance across both languages and architectures, obtaining WER values of 17.1% and 16.6% for Telugu and 23.3% and 25.6% for Kannada. The improvement arises from representation-level regulation rather than simple data weighting. RG-RCT evaluates the reliability of synthetic samples using both representation similarity and distribution distance measures. Aligned synthetic samples are constrained to remain close to their real counterparts through representation alignment, while unaligned samples are regulated using Mahalanobis distance to prevent distribution drift.

Furthermore, the layer-wise representation governance mechanism stabilizes the encoder representation space by interpolating synthetic and real representations across encoder layers. Early layers retain acoustic diversity from synthetic speech, while deeper layers progressively enforce stronger alignment with real speech representations. This encoder-level fusion allows the model to benefit from the diversity of multi-speaker synthetic data while preventing synthetic artifacts from dominating higher-level representations. Consequently, RG-RCT maintains a stable representation geometry anchored to real speech while still leveraging the variability introduced by synthetic augmentation.

Overall, the results demonstrate that although synthetic data augmentation improves ASR performance in low-resource settings, naive mixing or static confidence weighting alone cannot fully address the mismatch between synthetic and real speech distributions. By explicitly regulating representation geometry and anchoring encoder embeddings to real speech, the RG-RCT framework achieves consistent improvements for both Telugu and Kannada medical-domain ASR tasks.

## V. Conclusion

This work presented Real-Governed Representation-Calibrated Training (RG-RCT), a representation-level training strategy for improving low-resource medical-domain ASR using both real and synthetic speech data. Unlike conventional approaches based on direct data mixing or simple confidence weighting, the proposed method explicitly regulates how synthetic speech influences the encoder representation space. By combining reliability-based confidence estimation with representation alignment and distribution-aware constraints, RG-RCT stabilizes encoder geometry while preserving the acoustic diversity introduced by multi-speaker synthetic augmentation. Experimental results using Wav2Vec2 and Whisper demonstrate that the proposed framework consistently outperforms conventional training strategies on Telugu and Kannada medical speech datasets, achieving the lowest WER across both languages. The results highlight that controlling representation geometry enables more effective utilization of synthetic speech for low-resource ASR training. The proposed framework can be extended to other languages and specialized domains where annotated speech data is limited.

## Acknowledgment

This work was supported by the Department of Science and Technology (DST), Government of India, under Project Sanction No. DST/CSRI/2024/1139.

## References

[1] A. Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," arXiv:2006.11477, 2020.
[2] W. Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," IEEE/ACM TASLP, 2021.
[3] S. Chen et al., "WavLM: Large-Scale Self-Supervised Pretraining for Full Stack Speech Processing," IEEE Journal of Selected Topics in Signal Processing, 2022.
[4] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," ICML, 2023.
[5] D. Park et al., "SpecAugment: A Simple Data Augmentation Method for ASR," Interspeech, 2019.
[6] A. Rosenberg et al., "Data Augmentation Using Text-to-Speech for End-to-End ASR," Interspeech, 2019.
[7] A. Fazel et al., "SynthASR: Unlocking Synthetic Data for Speech Recognition," arXiv:2106.07803, 2021.
[8] T. Javed et al., "Towards Building ASR Systems for the Next Billion Users," arXiv:2111.03945, 2021.
[9] A. Baevski et al., "Self-Training and Pre-Training are Complementary for Speech Recognition," ICASSP, 2021.
[10] Q. Xu et al., "Self-Training with Noisy Student Improves Automatic Speech Recognition," in Proc. ICASSP, 2020.
[11] M. Tran et al., "A Domain Adaptation Framework for Speech Recognition Systems with Only Synthetic Data," arXiv:2501.12501, 2025.
[12] S. Ogun et al., "Evaluation of TTS and Voice Conversion Based Data Augmentation for ASR," arXiv:2503.08954, 2025.
[13] S. Liu et al., "Towards Selection of Text-to-Speech Data to Augment ASR Training," Interspeech, 2022.
[14] H. Su et al., "Task Arithmetic Can Mitigate Synthetic-to-Real Gap in ASR," EMNLP, 2024.
[15] M. Tran et al., "R2S: Real-to-Synthetic Representation Learning," Interspeech, 2025.
[16] B. Hilmes et al., "On the Effect of Purely Synthetic Training Data for ASR Architectures," 2024.
[17] A. Laptev and B. Ginsburg, "Entropy-Based Methods of Word-Level Confidence Estimation," IEEE SLT, 2023.
[18] N. Ravi et al., "TeLeS: Temporal Lexeme Similarity Score," arXiv:2401.03251, 2024.
[19] N. Ravi et al., "ASR Confidence Estimation using True Class Lexical Similarity Score," Interspeech, 2025.
[20] Y. Ganin and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," JMLR, 2016.
[21] A. van den Oord et al., "Representation Learning with Contrastive Predictive Coding," arXiv:1807.03748, 2018.
[22] S. C. Devarakonda et al., "Enhancing ASR Performance in the Medical Domain for Dravidian Languages," *NCC*, pp. 596–601, 2026.
[23] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. International Conference on Machine Learning (ICML)*, 2006.
[25] G. K. Kumar, P. S. V, P. Kumar, M. M. Khapra, and K. Nandakumar, "Towards Building Text-To-Speech Systems for the Next Billion Users," arXiv preprint arXiv:2211.09536, 2023.
[26] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," arXiv:2005.11129, 2020.
[27] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," arXiv preprint arXiv:2210.15418, 2022.