


Robust Sleep Stage Prediction from Electroencephalogram with Label Noise Using Multimodal Large Language Models

Xihe Qiu^{1*}, Chen Zhan^{1*}, Gengchen Ma¹, Jingjing Huang², and Xiaoyu Tan³

¹ Shanghai University of Engineering Science, Shanghai 201600, China

² Eye & ENT Hospital, Fudan University, Shanghai 200031, China

³ INFLY TECH(Shanghai)Co.,Ltd, Shanghai, 200030, China

txywilliam1993@outlook.com


Abstract. Sleep stage prediction is a critical task in medical diagnostics, such as for sleep disorders like Obstructive Sleep Apnea-Hypopnea Syndrome (OSAHS). Traditionally, this task involves analyzing Electroencephalogram (EEG) signals and classifying the stages based on general features, often relying on medical expertise. However, this process is prone to bias and variance, as clinicians incorporate subjective experience into their predictions. In recent years, multimodal large language models (MLLMs) have demonstrated significant advancements, particularly in medical applications, outperforming traditional methods in many domains. Despite their promising potential, MLLMs are sensitive to high memorization effects and require high-quality, well-labeled data for fine-tuning. Label noise, commonly present in real-world datasets, can severely hinder their performance and robustness. Consequently, directly applying MLLMs to sleep stage prediction using noisy EEG labels presents a challenge. In this paper, we introduce a novel framework for sleep stage prediction using EEG data under label noise, leveraging the power of MLLMs. Our approach integrates multi-perspective agreement techniques to identify high-quality samples based on the prior knowledge embedded in MLLMs. We then employ a self-training method to enhance prediction accuracy despite the presence of label noise. We validate our framework using real patient EEG data in sleep stage prediction tasks, and the results demonstrate that our approach is both robust and accurate under label noise, outperforming other state-of-the-art robust learning methods. Our code will be made publicly available at <https://github.com/Leonard-zc/MICCAI2025-RSSP>.

Keywords: Sleep stage · MLLM · Label noise · Robust learning.

1 Introduction

Sleep staging constitutes the gold-standard methodology for assessing sleep architecture in clinical sleep medicine[14, 18]. Conventional practice requires sleep

* Equal contribution

 Corresponding author

specialists to manually classify 30-second polysomnography (PSG) epochs following the American Academy of Sleep Medicine (AASM) criteria[1, 22]. As the principal biosignal for sleep staging[17], single-channel electroencephalogram (EEG) recordings offer practical advantages by minimizing sleep-state interference caused by multi-sensor PSG setups[9, 20]. While deep learning approaches employing single-channel EEG data have undergone extensive investigation[5], their clinical translation remains constrained by pervasive label noise in medical datasets.

Label noise originates from two primary sources: (1) inter-rater variability inherent in multi-expert annotation protocols[11], and (2) intrinsic feature heterogeneity leading to misclassification of phenotypically similar categories, generating instance-dependent noise patterns[16, 4, 30]. Recent progress in multimodal large language models (MLLMs) demonstrates significant potential for medical applications[8, 24, 21], particularly through enhanced multimodal reasoning capabilities evident in medical diagnosis and imaging analysis tasks[19, 23]. Nevertheless, MLLMs applications in EEG-based sleep staging remain underexplored despite their success in related medical imaging domains[3, 15].

The operational effectiveness of MLLMs fundamentally depends on precise cross-modal alignment, yet noise-contaminated datasets can disrupt alignment mechanisms and degrade inference performance[26, 28]. These models exhibit notable memorization effects, with representation learning capabilities demonstrating strong label precision dependence[2, 12]. When trained on noisy labels, MLLMs frequently manifest hallucinations characterized by semantic mismatches between generated interpretations and input signals, severely compromising output reliability[10, 31].

To address this challenge, this paper proposes a strategy combining multi-perspective agreement with co-optimized self-training. In complex data environments, particularly in multimodal medical data, label noise and annotation inconsistency are often inevitable, which may cause models to learn erroneous or inconsistent features, leading to hallucinations. We argue that selecting high-quality samples as reliable learning signals can effectively reduce the negative impact of label noise on model performance. Specifically, first, high-quality samples are filtered through multi-perspective agreement. Multi-perspective agreement helps eliminate potential label noise, thereby ensuring training data quality. Second, a self-training strategy is employed to fine-tune the model based on high-quality samples and expand the high-quality sample set through model self-prediction. Subsequently, the multi-perspective agreement technique is reapplied to filter new high-quality samples, which are added to the training set for secondary fine-tuning, further enhancing model performance. Experimental results show that our method effectively reduces hallucinations, particularly in noisy environments, significantly improving model robustness and generalization capabilities. This approach provides a new technical pathway for complex data analysis in the medical field and demonstrates the great potential of MLLMs in applications such as sleep staging.

We summarize our contributions as follows:

- (1) **Proposed a novel framework for EEG sleep stage prediction under label noise:** First introduced an innovative framework combining MLLMs with label noise handling techniques, aiming to improve sleep stage prediction accuracy using EEG data with label noise in real clinical environments.
- (2) **Innovative noise-robust learning strategy:** Innovatively combined multiple-perspective agreement and self-training strategies, effectively reducing the impact of hallucinations in multimodal large models under noisy data.
- (3) **Comprehensive clinical validation:** Demonstrated superior performance over state-of-the-art methods across benchmark datasets and real-world clinical cohorts.

2 Method

We propose a novel framework combining Multiple-perspective agreement technology and Self-training strategy, aiming to address the application problem of single-channel EEG data with label noise in sleep stage prediction. The core concept of this method is to dynamically construct high-quality training sets based on a consensus sample selection mechanism in noisy data, while further enhancing model robustness and generalization capabilities through iterative self-training strategies. The workflow of the entire framework is shown in Fig 1.

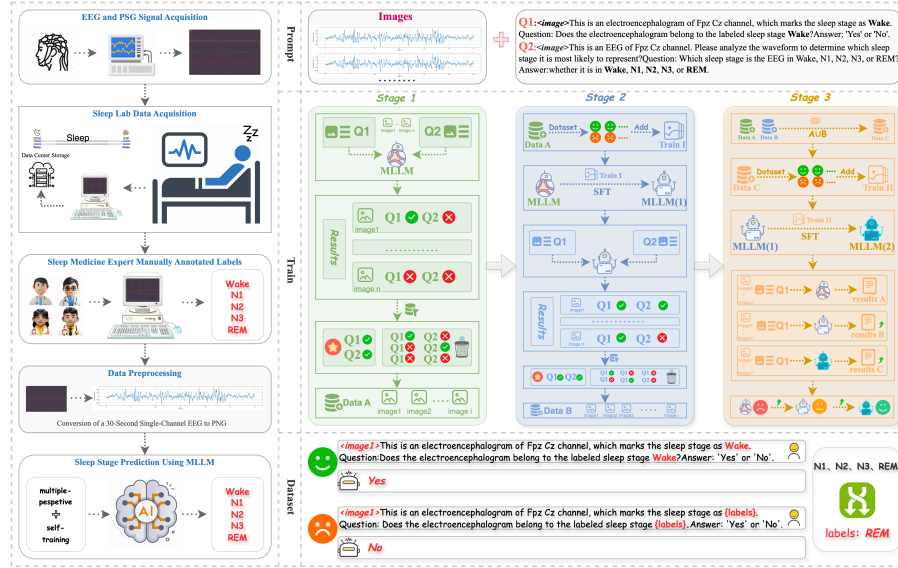


Fig. 1. Sleep stage classification framework using MLLM based on EEG Image representation.

2.1 Multi-perspective Agreement for High-quality Sample Selection

Problem Formulation and Input Data. Consider the original dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i denotes the EEG image and $y_i \in \{\text{Wake, N1, N2, N3, REM}\}$ is the associated label. We employ a multi-perspective agreement technique to assess each sample from various perspectives, determining its qualification as a high-quality sample. For each EEG image x_i , we formulate two prompts for a question-answering task:

Question 1 (Q1): Directly assess whether x_i corresponds to the current label y_i : The model outputs $\hat{y}_{i,1} \in \{\text{Yes, No}\}$.

Question 2 (Q2): Identify the most probable label for the current EEG image from all sleep stages: The model outputs $\hat{y}_{i,2} \in \{\text{Wake, N1, N2, N3, REM}\}$.

Consistency Judgment Rules. Leveraging the MLLM for each sample (x_i, y_i) , we evaluate the following consistency condition:

$$c_i = \begin{cases} 1, & \text{if } \hat{y}_{i,1} = \text{Yes and } \hat{y}_{i,2} = y_i, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The consistency evaluation logic is: if the model's responses to both prompts are correct, the sample is regarded as consistent ($c_i = 1$), indicating high reliability of its label y_i ; otherwise ($c_i = 0$), the sample is considered potentially noisy.

Training Set Generation - Construction of Positive and Negative Cases.

For the consistent sample set $\mathcal{D}_A = \{(x_i, y_i) \mid c_i = 1\}$, we construct training data based on the following rules:

1) Positive Case: For each EEG image x_i , paired with its correct label y_i a positive training sample is constructed:

$$\mathcal{T}_{\text{pos}} = \{(x_i, \text{'Yes'}) \mid (x_i, y_i) \in \mathcal{D}_A\}, \quad (2)$$

2) Negative Case: Randomly choose an incorrect label $y_j \neq y_i$ from the remaining labels to construct negative training samples:

$$\mathcal{T}_{\text{neg}} = \{(x_i, \text{'No'}) \mid (x_i, y_i) \in \mathcal{D}_A, y_j \neq y_i\}, \quad (3)$$

In the generation of negative cases, the following constraints are introduced to ensure data balance and diversity:

3) Balance between Positive and Negative Cases: To maintain the balance of the training set, the number of positive and negative cases for each sample x_i is equal:

$$n_{\text{neg}}(x_i) = n_{\text{pos}}(x_i), \quad (4)$$

where $n_{\text{neg}}(x_i)$ and $n_{\text{pos}}(x_i)$ denote the number of negative and positive cases for sample x_i respectively.

4) Uniform Distribution of Wrong Labels: All wrong labels $y_j \in \mathcal{Y}_{\text{neg}}$ are selected uniformly, i.e., for each y_j :

$$n_{\text{neg}}(x_i, y_j) = \frac{n_{\text{neg}}(x_i)}{|\mathcal{Y}_{\text{neg}}|}, \quad (5)$$

where $|\mathcal{Y}_{\text{neg}}|$ represents the size of the wrong label set \mathcal{Y}_{neg} (in this paper, the size of the wrong label set should be 4). $n_{\text{neg}}(x_i, y_j)$ represents the number of negative cases with the wrong label y_i for sample x_i .

The final training dataset comprises positive and negative cases:

$$\mathcal{T}_A = \mathcal{T}_{\text{pos}} \cup \mathcal{T}_{\text{neg}}. \quad (6)$$

This approach ensures that the training dataset includes high-quality positive cases while effectively utilizing negative cases to enhance the model’s discriminative capability.

2.2 Self-training and Preliminary Model Fine-tuning

Preliminary Model Fine-tuning. The MLLM f_{MLLM} is initially fine-tuned using the filtered training dataset \mathcal{T}_A , resulting in the first-round updated model $f_{\text{MLLM}}^{(1)}$. The objective of this phase is to allow the model to learn robust classification boundaries from the initial high-quality samples and mitigate the impact of label noise. The loss function for the fine-tuning process is defined as:

$$\mathcal{L}_A = -\frac{1}{|\mathcal{T}_A|} \sum_{(x,y) \in \mathcal{T}_A} \log P(y | x; f_{\text{MLLM}}), \quad (7)$$

where $P(y | x; f_{\text{MLLM}})$ is the predicted probability of the model for the sample (x, y) .

Data Update and Consistency Screening. The original dataset \mathcal{D} is re-screened for consistency using the preliminarily fine-tuned model $f_{\text{MLLM}}^{(1)}$, producing the second-round high-quality sample set \mathcal{D}_B .

The first-stage filtered sample set \mathcal{D}_A and the new sample set \mathcal{D}_B are combined and deduplicated to create the updated high-quality sample set:

$$\mathcal{D}_C = \mathcal{D}_A \cup \mathcal{D}_B. \quad (8)$$

Training Set Update. For the updated sample set \mathcal{D}_C , the updated training set \mathcal{T}_C is built following the previously established training set generation rules.

2.3 Multi-round Optimization to Enhance Model Performance

Second-round Fine-tuning. The second round of fine-tuning is conducted on the updated training set \mathcal{T}_C , where the model $f_{\text{MLLM}}^{(1)}$ is fine-tuned to yield the final model $f_{\text{MLLM}}^{(2)}$. Through multiple rounds of optimization, the model gradually learns more robust classification boundaries, enhancing its generalization ability across different datasets. The loss function for the second-round fine-tuning is defined as:

$$\mathcal{L}_C = -\frac{1}{|\mathcal{T}_C|} \sum_{(x,y) \in \mathcal{T}_C} \log P(y | x; f_{\text{MLLM}}^{(1)}), \quad (9)$$

Multi-round Optimization and Model Convergence. The process of data screening and model fine-tuning is repeated, progressively expanding the training set and optimizing model parameters until one of the following convergence criteria is satisfied:

1) The number of high-quality samples in the dataset ceases to increase significantly:

$$|D_{k+1}| - |D_k| < \epsilon, \quad (10)$$

2) The model’s performance on the validation set stabilizes:

$$\text{Accuracy}_{k+1} - \text{Accuracy}_k < \delta, \quad (11)$$

The final model $f_{\text{MLLM}}^{(k)}$ serves as the ultimate prediction model for the sleep staging task on the actual test set.

Table 1. Performance Comparison of Different Methods on Sleep-EDF and P-EDF Datasets.

Method	Dataset	Manual	EEG Channels	Test Epochs	Overall ACC(%)
GCE[29]	Sleep-EDF	R&K	Fpz-Cz	1000	82.5
L1-loss[25]	Sleep-EDF	R&K	Fpz-Cz	1000	82.4
co-teaching+[27]	Sleep-EDF	R&K	Fpz-Cz	1000	82.8
ours	Sleep-EDF	R&K	Fpz-Cz	1000	85.2
GCE[29]	P-EDF	AASM	C4-M1	500	80.8
L1-loss[25]	P-EDF	AASM	C4-M1	500	80.0
co-teaching+[27]	P-EDF	AASM	C4-M1	500	77.6
ours	P-EDF	AASM	C4-M1	500	82.6

3 Experiments and Results

3.1 Datasets and Baselines

Datasets. This study employs the public **Sleep-EDF** [13] dataset and a proprietary clinical dataset, **P-EDF**, collected from the Eye & ENT Hospital of Fudan University, for validation. The Sleep-EDF dataset comprises PSG recordings annotated following Rechtschaffen and Kales (R&K) standards, segmented into 30-second epochs. Our analysis specifically employs Fpz-Cz lead signals sampled at 100 Hz. The P-EDF dataset consists of de-identified PSG recordings from patients with mild OSAHS. These recordings, staged according to AASM guidelines, were analyzed using C4-M1 lead signals acquired at a 200 Hz sampling frequency for comprehensive validation.

To ensure data quality, segments containing motion artifacts or undefined stages were removed, particularly pruning the initial and final 30-minute periods susceptible to electrode placement effects; Standardized staging criteria by

Table 2. Per-class Accuracy (ACC) Comparison of Different Methods on Sleep-EDF and P-EDF Datasets.

Methods	Datasets	Overall	Per-class ACC (%)									
			Wake		N1		N2		N3		REM	
			Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
GCE[29]	Sleep-EDF	82.5%	74	26	85	25	76	24	94	6	83.5	16.5
L1-loss[25]	Sleep-EDF	82.4%	78.5	21.5	85	25	76.5	23.5	92	8	80	20
co-teaching+[27]	Sleep-EDF	82.8%	70.5	29.5	81.5	18.5	80.5	19.5	92.5	7.5	89	11
ours	Sleep-EDF	85.2%	73.5	26.5	84	16	86	14	94	6	88.5	11.5
GCE[29]	P-EDF	80.8%	78	22	83	17	73	27	90	10	80	20
L1-loss[25]	P-EDF	80.0%	75	25	80	20	72	28	88	12	85	15
co-teaching+[27]	P-EDF	77.6%	77	23	81	19	77	23	68	32	85	15
ours	P-EDF	82.6%	75	25	84	16	75	25	93	7	86	14

merging N3-N4 stages from R&K standards into AASM-compliant N3 stage[6]; Converted 30-second EEG epochs into waveform plots (2000×300 pixels, PNG format), with detailed conversion workflow shown in the data preprocessing module of Figure1.

Baselines. This investigation benchmarks three state-of-the-art noise-robust learning approaches: **1) Generalized Cross Entropy (GCE) [29]:** Implements confidence-aware loss weighting with exponential decay to progressively filter noisy samples by attenuating gradient contributions from low-confidence predictions. **2) L_1 -loss [25]:** Employs Manhattan distance to quantify absolute differences between predicted and annotated probability distributions, demonstrating inherent robustness to label outliers. **3) Co-teaching+ [27]:** Utilizes dual-network architecture with cross-parameter updates and dynamic sample exchange to preserve model diversity while counteracting memorization bias.

3.2 Experimental Setup and Results

Experimental Details. We conducted experiments using LLaMA-3.2-11B Vision-Instruct[7] as the base model. Parameter-efficient fine-tuning was achieved through Low-Rank Adaptation (LoRA) methodology[32].The optimization protocol initialized with learning rate 1e-6, implementing DeepSpeed ZeRO-3 optimization strategy across dual NVIDIA RTX 4090 GPUs to accomplish three training epochs. Model efficacy was quantified using accuracy as primary evaluation criterion, where task-specific performance on Question1 (Q1 accuracy) served as definitive performance metric.

Experimental Results. As shown in Tables 1 and 2, our method achieves overall accuracies of 85.2% and 82.6% on Sleep-EDF and P-EDF datasets respectively, significantly outperforming baseline models.Granular analysis reveals optimal performance in N3 and REM stage classification, with notable improvements in the easily confused N1 stage compared to contrastive methods, validating the effectiveness of its noise-resistant design.Experimental results demonstrate that this method combines robustness and high accuracy across datasets,

multiple annotation standards, and complex sleep stage classification tasks, providing a reliable solution for sleep staging in noisy environments.

3.3 Ablation Analysis

We perform ablation studies on three key components: multi-perspective agreement filtering for noise reduction, positive-negative sample balancing for feature optimization, and self-training iteration for performance improvement. The results highlight each component’s unique contribution to system robustness.

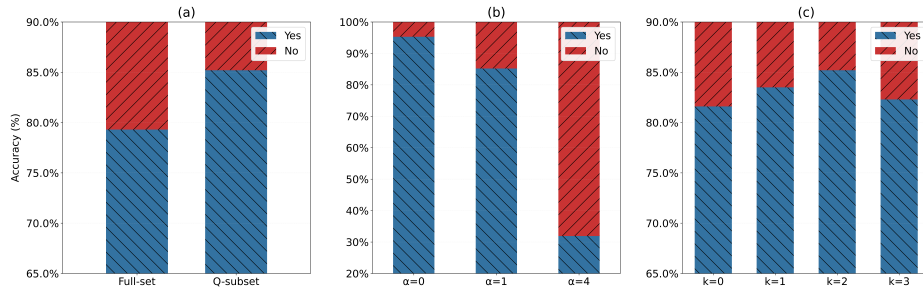


Fig. 2. Ablation Analysis of Multi-perspective Agreement Filtering, Sample Balancing, and Self-Training in Sleep Stage Classification.

1) Multi-perspective Agreement Filtering Framework: As demonstrated in Figure 2(a), the agreement-filtered high-quality subset (Q-subset) achieves a 5.9% accuracy improvement over the original dataset (0.793 vs. 0.852), effectively mitigating label noise propagation through selective sample purification. **2) Positive-Negative Sample Balancing:** Figure 2(b) illustrates the impact of the class-balancing coefficient α on model performance during training. The model achieves the highest accuracy of 85.2% when $\alpha = 1$, indicating a balanced 1:1 class ratio. However, when α is set to 0 or 4, an imbalance in the positive and negative samples occurs, leading to feature confusion and a negative impact on model performance. This emphasizes the significance of maintaining balanced sample representation in model training. **3) Self-training Optimization:** The convergence analysis in Figure 2(c) reveals peak accuracy (0.852) at two training iterations ($k = 2$); the third round results in a decline, representing 1.7% and 3.6% improvements over single iteration ($k = 1$) and baseline ($k = 0$), respectively.

4 Conclusion

We present a novel noise-robust sleep staging framework leveraging MLLMs, which systematically resolves label noise challenges in EEG analysis through

synergistic integration of cross-modal feature alignment and adaptive curriculum learning paradigms. Capitalizing on MLLMs’ embedding space priors, the framework employs multi-perspective agreement learning for sample purification, coupled with iterative curriculum optimization to progressively refine model parameters through dynamic learning schedules. Comprehensive experimental evaluations reveal substantial superiority of our framework over conventional noise-robust methodologies across multiple evaluation metrics. The findings not only validate MLLMs’ capability in decoding complex biosignals, but also establish a transferable noise-resilient framework that advances analytical methodologies for clinical time-series data processing.

Acknowledgments. This work is supported by Shanghai Municipal Natural Science Foundation (23ZR1425400) and Shanghai Soft Science Project (25692114700).

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Berry, R.B., Budhiraja, R., Gottlieb, D.J., Gozal, D., Iber, C., Kapur, V.K., Marcus, C.L., Mehra, R., Parthasarathy, S., Quan, S.F., et al.: Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine. *Journal of clinical sleep medicine* **8**(5), 597–619 (2012)
2. Biderman, S., Prashanth, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., Raff, E.: Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems* **36**, 28072–28090 (2023)
3. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280* (2024)
4. Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., Liu, Y.: Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347* (2020)
5. Eldele, E., Chen, Z., Liu, C., Wu, M., Kwok, C.K., Li, X., Guan, C.: An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **29**, 809–818 (2021)
6. Fiorillo, L., Favaro, P., Faraci, F.D.: Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE transactions on neural systems and rehabilitation engineering* **29**, 2076–2085 (2021)
7. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024)
8. Han, T., Adams, L.C., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressen, K.K.: Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247* (2023)

9. Hassan, A.R., Bashar, S.K., Bhuiyan, M.I.H.: On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram. In: 2015 International conference on advances in computing, communications and informatics (ICACCI). pp. 2238–2243. IEEE (2015)
10. Huang, W., Liu, H., Guo, M., Gong, N.Z.: Visual hallucinations of multi-modal large language models. arXiv preprint arXiv:2402.14683 (2024)
11. Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z.: Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging* **41**(6), 1533–1546 (2022)
12. Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R.: Challenges and applications of large language models. arXiv preprint arXiv:2307.10169 (2023)
13. Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A., Obery, J.J.: Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering* **47**(9), 1185–1194 (2000)
14. Lan, K.C., Chang, D.W., Kuo, C.E., Wei, M.Z., Li, Y.H., Shaw, F.Z., Liang, S.F.: Using off-the-shelf lossy compression for wireless home sleep staging. *Journal of neuroscience methods* **246**, 142–152 (2015)
15. Liu, F., Zhu, T., Wu, X., Yang, B., You, C., Wang, C., Lu, L., Liu, Z., Zheng, Y., Sun, X., et al.: A medical multimodal large language model for future pandemics. *NPJ Digital Medicine* **6**(1), 226 (2023)
16. Liu, Y.: Understanding instance-level label noise: Disparate impacts and treatments. In: International Conference on Machine Learning. pp. 6725–6735. PMLR (2021)
17. Malekzadeh, M., Hajibabae, P., Heidari, M., Berlin, B.: Review of deep learning methods for automated sleep staging. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). pp. 0080–0086. IEEE (2022)
18. Patel, A.K., Reddy, V., Shumway, K.R., Araujo, J.F.: Physiology, sleep stages. In: StatPearls [Internet]. StatPearls Publishing (2024)
19. Patel, S.B., Lam, K.: Chatgpt: the future of discharge summaries? *The Lancet Digital Health* **5**(3), e107–e108 (2023)
20. Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J.M., Taheri, S., Guan, Y., Fernandez-Luque, L.: The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ digital medicine* **3**(1), 42 (2020)
21. Qiu, X., Wei, Y., Tan, X., Xu, W., Wang, H., Ma, J., Huang, J., Fang, Z.: Mimarosa: Enhancing obstructive sleep apnea diagnosis through multimodal data integration and missing modality reconstruction. *Pattern Recognition* p. 111917 (2025)
22. Rundo, J.V., Downey III, R.: Polysomnography. *Handbook of clinical neurology* **160**, 381–392 (2019)
23. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. *Nature medicine* **29**(8), 1930–1940 (2023)
24. Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T.: Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)
25. Yang, A.Y., Sastry, S.S., Ganesh, A., Ma, Y.: Fast ℓ_1 -minimization algorithms and an application in robust face recognition: A review. In: 2010 IEEE international conference on image processing. pp. 1849–1852. IEEE (2010)

26. Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., Zhuang, Y.: Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12944–12953 (2024)
27. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International conference on machine learning. pp. 7164–7173. PMLR (2019)
28. Yue, Z., Zhang, L., Jin, Q.: Less is more: Mitigating multimodal hallucination from an eos decision perspective. arXiv preprint arXiv:2402.14545 (2024)
29. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* **31** (2018)
30. Zhao, P., Yang, J.J., Buu, A.: Applied statistical methods for identifying features of heart rate that are associated with nicotine vaping. *The American Journal of Drug and Alcohol Abuse* pp. 1–8 (2025)
31. Zheng, K., Chen, J., Yan, Y., Zou, X., Hu, X.: Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. arXiv preprint arXiv:2408.09429 (2024)
32. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)