SleepGPT: A Sleep Stage Language Model for Efficient Sleep Assessment

Tianyou Yu[1,3†], Zhenghui Gu[1,3†], Zhenfu Wen[1,3†], Rui Huang[1], Fei Wang[9,3], Man Li[1,3], Jingang Yu[1,3], Zhuliang Yu[1,3], Jun Zhang[10,3], Yan Xu[11], Haiteng Jiang[4,5], Wenjuan Liu[4,6], Guifeng Deng[4,6], Zhengrun Gao[7], Yiwen Wu[8], Jun Liu[8], Yu Zhang[12], Matt W Jones[13], Yuanqing Li[1,3*], Jun Xiao[2,3*], Wei Wu[7*]

[1]School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510641, China.
[2]School of Electric Power Engineering, South China University of Technology, Guangzhou, 510641, China.
[3]Pazhou Lab, Guangzhou, 510330, China.
[4]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital and Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, 310058, China.
[5]School of Brain Science and Brain Medicine, MOE Frontier Science Center for Brain Science and Brain-machine Integration, State Key Laboratory of Brain-machine Intelligence, Zhejiang University, Hangzhou, 311121, China.
[6]College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, 310058, China.
[7]Songjiang Hospital & Songjiang Research Institute, Shanghai Key Laboratory of Emotions and Affective Disorders, Shanghai Jiao Tong University School of Medicine, Shanghai, 201600, China.
[8]Department of Neurology & Institute of Neurology, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China.
[9]School of Artificial Intelligence, South China Normal University, Guangzhou, 510631, China.
[10]School of Information Engineering, Guangdong University of Technology, Guangzhou, 510320, China.

[11]Sleep Medicine Center, Department of Psychiatry, Nanfang Hospital Affiliated to Southern Medical University, Guangzhou, 510515, China.
[12]Department of Neurology, Songjiang Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, 201699, China.
[13]School of Physiology, Pharmacology & Neuroscience, University of Bristol, Bristol, BS8 1QU, UK.

*Corresponding author(s). E-mail(s): auyqli@scut.edu.cn; junxiao@scut.edu.cn; weiwuneuro@sjtu.edu.cn;
†These authors contributed equally to this work.

Abstract

Accurate and scalable sleep assessment is crucial for diagnosing sleep disorders and advancing personalized medicine. However, current approaches heavily rely on manual scoring of polysomnography (PSG) recordings, which is labor-intensive, costly, and difficult to scale. Here, we introduce SleepGPT, the first sleep-specific language model trained on over 5.8 million sleep stage annotations from 5,793 whole-night recordings to model the sequential dynamics of sleep macrostructure. We demonstrate that SleepGPT consistently enhances the performance of multiple state-of-the-art sleep staging models across diverse independent datasets (total N=1,320,654), and enables high staging accuracy from low-density wearable EEG recordings (N=120,095). We further show that SleepGPT supports robust diagnosis of sleep disorders when embedded into a sequence-level Transformer framework. The model accurately identifies abnormal sleep stage dynamics and Type-1 narcolepsy across multiple clinical datasets (total N=685). Attention-based visualization of model outputs revealed interpretable macrostructural patterns associated with diagnostic labels. Together, these results demonstrate that SleepGPT serves as a general-purpose foundation model for sleep sequence modeling, offering a scalable and clinically translatable tool for automated sleep staging and sleep disorder diagnosis.

Keywords: Sleep language model, automated sleep staging, sleep disorder diagnosis, hypnogram, generative pre-trained transformers

Introduction

Sleep is essential for health and well-being, with disrupted sleep linked to serious cardiometabolic and neuropsychiatric disorders that collectively affect hundreds of millions worldwide [1, 2]. Comprehensive sleep assessment is foundational to understanding sleep architecture and its contributions to these disorders. Typically, sleep assessment heavily relies on manual scoring of nocturnal polysomnography (PSG) recordings, which is a time-consuming process that requires visually inspecting each

30-second epoch of an entire night's recording. This results in the generation of a sleep stage sequence, or hypnogram, that characterizes sleep macrostructure, including cycles between rapid eye movement (REM) and non-REM stages, stage durations, transitions, latency, and efficiency. This scoring process is subjective, with inter-scorer reliability often reaching only 82.6% [3], leading to variability and inefficiencies. Consequently, there is a strong demand for more scalable and objective sleep staging solutions.

Recent advances in machine learning, particularly deep learning, have shown promise for automating sleep staging. Deep neural networks, such as convolutional neural networks (CNNs)[4, 5], recurrent neural networks (RNNs)[6, 7], transformers[8], and hybrid networks[9–12], have been successfully applied to process raw PSG signals or its time-frequency representations to classify sleep stages. However, staging performance often remains suboptimal when contextual information from adjacent epochs was not fully utilized [6, 7]. Sleep staging involves the classification of highly correlated time-series data, where adjacent epochs contain critical contextual information. While recurrent models like long short-term memory (LSTM) networks and gated recurrent units (GRUs) have been employed to address this by capturing long-term dependencies[6, 7, 13], they still face challenges related to dataset-specific biases and limited generalizability across diverse populations.

In addition to accurate sleep staging, effective sleep assessment should also support the detection and diagnosis of sleep disorders. Traditional diagnostic procedures, such as the multiple sleep latency test (MSLT) for narcolepsy, require significant expertise and are resource-intensive, often leading to diagnostic delays of 7–10 years [14]. Moreover, manual analysis by sleep specialists introduces inter- and intra-operator variability [15], further complicating the diagnostic process. Although automated approaches to sleep disorder diagnosis have been proposed to mitigate these issues, traditional machine learning methods often require extensive feature engineering and suffer from poor generalization to new data. To date, no studies have employed raw sleep stage annotation sequences to diagnose sleep disorders.

Sleep macrostructure is driven by fundamental neurophysiological mechanisms that guide predictable sequential transitions between sleep stages. Leveraging this sequential structure has the potential to enhance both sleep staging and disorder diagnosis [16]. Inspired by the success of language models in analyzing sequential biological data [17, 18], we introduce SleepGPT, a novel sleep language model designed to model sleep macrostructure and improve sleep staging and disorder diagnosis (see Figure 1). By capturing the sequential dynamics and transition patterns within sleep stage sequences, SleepGPT represents a departure from traditional methods. To our knowledge, this is the first study to use a language model for sleep stage sequence modeling, with the potential to significantly enhance the accuracy and generalizability of sleep disorder detection.

Results

Development of SleepGPT

We developed SleepGPT based on the GPT-2 architecture [28], leveraging its capacity to model long-range dependencies in sequential data. The model was trained using the SHHS-1 dataset [19] (N≈5.8M), comprising 5,793 subjects' overnight sleep stage annotations based on PSG recordings. Each recording was organized as a sequence of 30-second annotations, ranging from 360 to 1,199 epochs per subject (mean = 1,012), resulting in a total of 5,863,207 sleep annotations. To capture the temporal dynamics of sleep macrostructure, SleepGPT was trained using a next-stage prediction objective: given a sequence of preceding sleep stages, the model learns to predict the subsequent stage (Figure 2A). This task enables the model to learn the latent patterns, transitions, and dependencies characteristic of human sleep architecture. Through this training approach, SleepGPT effectively internalizes the sequential organization of sleep stage sequences, forming the basis for downstream applications in sleep staging and disorder diagnosis.

We first evaluated the pretrained SleepGPT model on the sleep staging task (Figure 2B) using three external datasets: SleepEDF [20] (N=199,352), MASS [21] (N=229,040), and Physio2018 [22] (N=892,262). In these evaluations, state-of-the-art sleep staging models were used to predict sleep stages from PSG signals, and SleepGPT was used as an add-on module to further correct the predicted stage sequences. We compared the predictions with and without SleepGPT post-hoc corrections to assess the enhancement enabled by SleepGPT. To assess generalizability, we conducted inter-cohort validation across the SleepEDF, MASS, and Physio2018 datasets. In addition, we employed the YASA toolbox [29]—which provides pretrained, publicly available models—for blinded out-of-sample evaluation, where all model parameters were fixed prior to testing. To examine the clinical translational potential of SleepGPT, we further evaluated its performance on the BOAS dataset (N=120,095) [23], which contains simultaneous PSG and low-density wearable EEG recordings from 128 subjects. This enabled a direct comparison of sleep staging accuracy between wearable and standard PSG modalities, with and without SleepGPT integration.

To evaluate SleepGPT's utility for sleep disorder diagnosis, we incorporated the model as a local feature extractor within a hierarchical transformer network (HTN) (Figure **S1**). This architecture captures both local sleep stage patterns and global sequence-level features. The HTN model was first cross-validated on the CAP dataset (N=108) [24] to distinguish normal from abnormal sleep stage sequences. For external validation, we then tested generalization performance by retraining the model on the whole CAP dataset and testing the trained model on three independent datasets: ISRUC (N=86) [26], MNC (N=407) [25], and HANG7 (N=84) [27]. To assess the model's diagnostic performance in a more clinically challenging task, we evaluated its ability to distinguish Type-1 narcolepsy from other hypersomnia disorders and healthy controls. This analysis was conducted via cross-validation on MNC and external validation on HANG7. Finally, we visualized the learned global attention weights of the HTN model to identify salient sequence-level features that may serve as candidate biomarkers.

Enhancing sleep staging with SleepGPT

We evaluated SleepGPT's ability to enhance automated sleep staging by post-hoc correction of predictions made by established models, including TinySleepNet [5] and XSleepNet [12]. Performance was assessed in terms of accuracy, macro-averaging F1-score (MF1), and Cohen's kappa coefficient (κ) across multiple external datasets. Note that SleepGPT was pretrained on the SHHS-1 dataset and was not fine-tuned on any of the downstream datasets used in this evaluation. To assess generalization, we applied several state-of-the-art sleep staging models trained on one dataset to unseen data from another, and tested whether SleepGPT could improve their out-of-sample predictions. Specifically, sleep staging models (TinySleepNet and XSleepNet) trained on the MASS dataset (N=229,040) were tested on the Physio2018 dataset (N=892,262), and vice versa, enabling cross-dataset validation with matching EEG configurations (C3-A2/C4-A1). We also applied SleepGPT to enhance predictions made by YASA (Yet Another Spindle Algorithm) [29], a widely used open-source sleep staging toolbox with fixed pretrained models, on the SleepEDF and MASS datasets.

As shown in Figure 2, integrating SleepGPT consistently improved the performance of all baseline models across all datasets and model architectures. YASA's accuracy improved by 4.2% on SleepEDF and 1.6% on MASS. TinySleepNet's accuracy increased by 2.9% and 1.6%, and XSleepNet's by 2.3% and 1.9%, during cross-dataset validation between the Physio2018 and MASS datasets. Consistent improvements in MF1, κ, and stage-specific accuracies are also observed. Notably, models trained on the larger Physio2018 dataset (N=892,262) and tested on the smaller MASS dataset (N=229,040) demonstrated superior staging performance than the reverse, suggesting that the size and composition of the training dataset, particularly the distribution of sleep stages, have a considerable impact on model generalizability. Examples of the corrections made by SleepGPT during cross-dataset validation are shown in Figure 2F (more detailed results are provided in Table **S1** of the Supplementary Appendix), illustrating the ability of SleepGPT in correcting the erroneous sleep stage predictions made by baseline models. Overall, these results demonstrate SleepGPT's robustness and utility as a general-purpose add-on module that improves sleep staging performance across diverse datasets and model architectures.

Assessing translatability of SleepGPT-powered models on wearable EEG data

Recent advances in wearable EEG devices have created new opportunities for low-burden, at-home sleep monitoring. To assess the translational potential of SleepGPT, we evaluated its ability to enhance sleep staging performance using data from the BOAS dataset, which contains simultaneous PSG and headband EEG recordings from 128 subjects. The TinySleepNet [5] and XSleepNet [12] models, with and without SleepGPT integration, were cross-validated on the PSG and the headband EEG data, respectively. As shown in Figure 3, both models exhibit comparable performance across the PSG and headband EEG data, and the addition of SleepGPT consistently improved accuracy, MF1, and κ scores across both modalities. Notably, the sleepGPT-powered XSleepNet model achieved the highest accuracy of 84.3% and 83.9% for the

PSG and headband EEG data, respectively. More detailed results are provided in Table **S2** of the Supplementary Appendix. These results underscore the robustness of the SleepGPT models in enhancing sleep staging performance across EEG modalities and support their potential for integration into wearable sleep monitoring applications, enabling adoption of low-burden, low-cost, longitudinal sleep monitoring without significantly compromising accuracy.

Sleep disorder diagnosis enhanced by SleepGPT

We next evaluated SleepGPT as a feature extractor for sleep disorder diagnosis, comparing its performance to state-of-the-art diagnostic methods. Specifically, we integrated SleepGPT into a hierarchical transformer network (HTN) that takes a subject's full-night sleep stage sequence as input, enabling it to leverage the temporal dynamics of sleep stage transitions for disorder identification. Receiver operating characteristic (ROC) curves and confusion matrices were used to assess diagnostic performance across the modified CAP [24], ISRUC [26], MNC [25], and HANG7 datasets. Figure 4 presents results for two empirical feature-based XGBoost classifiers (based on Hypnogram [29] and Hypnodensity [25]), a baseline end-to-end neural network (BaseNet), the proposed HTN trained from scratch, and the HTN model initialized with pretrained SleepGPT parameters. ROC curves and confusion matrices for each method across the four datasets are shown in Figure 4A and 4B, with detailed results provided in Table **S3** and **S4**.

On the CAP dataset, in which the task was to distinguish between normal sleep and abnormal sleep (including sleep disorders such as insomnia, disordered breathing, narcolepsy, etc.), the XGBoost classifier trained with hypnogram features achieved a balanced accuracy of 90.91%, with a sensitivity of 95.65% and a specificity of 86.17%. While end-to-end models like BaseNet can learn useful features from hypnograms, they may struggle to capture the sequential attributes of sleep stages if the model architecture is not carefully designed, as evidenced by a lower balanced accuracy of 84.97%, sensitivity of 86.96%, and specificity of 82.98%. The SleepGPT-based HTN model, however, significantly improved performance in an end-to-end fashion. This highlights the HTN's capacity to capture both localized sequential sleep transitions and broader contextual patterns in sleep hypnograms, enhancing diagnostic performance. Notably, initializing the HTN with pretrained SleepGPT parameters yielded the highest performance, achieving a balanced accuracy of 96.27%, with a sensitivity of 98.91% and a specificity of 93.62%. This pretrained HTN demonstrated a marked improvement in sensitivity over its non-pretrained counterpart, with an increase of over 10% on the CAP dataset.

Similar trends were observed in the MNC dataset, which involved distinguishing Type-1 narcolepsy (T1N) patients from non-T1N individuals (including other hypersomnia patients and healthy controls), an arguably more challenging task than distinguishing general sleep-disorder patients from healthy controls. The XGBoost classifier trained on hypnogram and hypnodensity features achieved balanced accuracies of 84.39% and 85.16%, respectively. Although the BaseNet model performed reasonably well in diagnosing T1N patients, achieving a balanced accuracy of 79.69%, it struggled with sensitivity, identifying only 67.07% of T1N subjects. The HTN model

trained from scratch achieved balanced accuracy of 85.49%, sensitivity of 78.05%, and specificity of 92.92%, matching the performance of the XGBoost classifier trained on hypnodensity features. When fine-tuned with pretrained SleepGPT parameters, the HTN model further boosted performance, achieving a balanced accuracy of 92.81%, a sensitivity of 91.46%, and a specificity of 94.15%. This improvement underscores the benefits of the SleepGPT model's self-supervised pretraining on large-scale sleep stage datasets. Importantly, the pretrained SleepGPT model demonstrated the potential for fine-tuning on smaller datasets, which is crucial in sleep medicine, where data scarcity often presents a challenge.

Testing generalization of SleepGPT-powered diagnosis models

We further assessed the out-of-sample generalizability of SleepGPT-powered models in diagnosing abnormal sleep. We evaluated the prediction performance by training models using the CAP dataset, and directly applied them to the ISRUC, MNC, and HANG7 datasets. The ISRUC sleep dataset contains whole-night PSG recordings of obstructive sleep apnea (OSA) patients (N=76) and healthy controls (N=10). On this external dataset, the pretrained HTN model achieved a balanced accuracy of 90.00%, with a sensitivity of 100% and a specificity of 80.00% when distinguishing OSA patients from healthy controls (Figure 4 and Table **S4**). When applied to the MNC dataset, the same model achieved a balanced accuracy of 83.72%, with a sensitivity of 88.00% and a specificity of 79.43% in discriminating healthy controls (N=282) from patients with T1N or other hypersomnia (N=125). On the HANG7 dataset, which contains 51 patients with narcolepsy and 33 healthy controls, the model achieved a balanced accuracy of 84.05%, with a sensitivity of 86.27% and a specificity of 81.82%.

For the more challenging task of differentiating T1N patients from other hypersomnia patients and healthy controls, the HTN model trained on the MNC dataset was externally validated using the HANG7 dataset. As shown in Figure 5 (and in Table **S5**), the HTN model fine-tuned with SleepGPT parameters achieved a balanced accuracy of 84.18%, with a sensitivity of 92.31% and a specificity of 76.06%, outperforming the compared methods. In all settings, SleepGPT-based models outperformed conventional baselines, demonstrating robust generalization across datasets and diagnostic categories.

Interpretation of the SleepGPT-powered diagnosis models

To better understand feature separability, t-SNE visualizations of the SleepGPT-powered HTN model for sleep disorder diagnosis are provided in Figure 5A and Figure 5C. The features extracted by the HTN model with pretrained SleepGPT parameters show a more distinct separation between normal/T1N and abnormal/non-T1N sleep patterns compared to the features extracted by the HTN model trained from scratch. This indicates that the pretrained SleepGPT model effectively captures discriminative features relevant to sleep disorder diagnosis, enhancing the model's ability to differentiate between normal/T1N and abnormal/non-T1N sleep patterns.

Moreover, to provide insights into the global feature extractor of the HTN model, a visualization of the attention weights from the global feature extractor on the CAP

dataset is presented in Figure 5B for one healthy subject and five subjects with distinct sleep disorders—insomnia, narcolepsy, nocturnal frontal lobe epilepsy (NFLE), periodic leg movements (PLMs), and REM behavior disorder (RBD). By focusing on the attention weights assigned to the *CLS* token (a special token that outputs a global representation of the entire sequence, detailed in the Methods section) within the final transformer layer, we can observe how each model identifies and prioritizes salient features in the data. The results demonstrate the effectiveness of the global feature extractor in capturing atypical sleep patterns associated with sleep disorders. For instance, in the case of the insomnia subject, characterized by small N3 and REM sleep ratios and direct transitions from REM sleep to wakefulness, the global feature extractor allocates more attention to the segments marked by these transitions (Figure 5B(2)). Likewise, the global transformer encoder of HTN highlights segments with short wake-to-REM sleep latency or abrupt transitions from wakefulness to REM sleep for the narcolepsy subject (Figure 5B(3)). Furthermore, the self-attention module identifies segments featuring abnormal stage transitions, short N3 and REM sleep durations, direct shifts from REM sleep to wakefulness, and frequent toggling between deep (N3) or REM sleep and wakefulness (Figure 5B(4-6)). Conversely, for subjects without pathologies, the attention weights are more evenly dispersed across the hypnogram segments (Figure 5B(1)).

We also show the attention weights obtained by the global feature extractor of the HTN model on the MNC data set for a healthy control, a patient with hypersomnia, and a patient with T1N (Figure 5D). For the healthy control, attention weights are more evenly distributed across the hypnogram segments. While for the narcolepsy patient, there is increased attention to segments featuring short wake-to-REM latency or direct transitions from wakefulness to REM sleep. Since we classify T1N patients from other hypersomnia patients and healthy controls, the attention weights are more focused on segments similar to those of the narcolepsy patient. These segments include short wake-to-REM latency, direct transitions from wakefulness to REM sleep, and dissociated REM sleep. These results suggest that the HTN's attention mechanism captures meaningful macrostructural patterns associated with specific sleep disorders. This interpretability not only enhances model trustworthiness but also points to potential biomarkers for clinical diagnosis.

Discussion

This study introduces SleepGPT, a novel sleep language model adapted from the GPT architecture and trained in a self-supervised manner on a comprehensive sleep stage dataset. SleepGPT offers several advantages for efficient sleep assessment: it consistently enhances existing sleep staging methods, effectively captures sleep stage transition dynamics, and integrates as a feature extractor within hierarchical transformer networks to improve sleep disorder diagnosis. Additionally, it identifies interpretable abnormal sleep patterns, potentially providing mechanistic insights into sleep disorders. Taken together, these findings highlight SleepGPT's potential as a scalable, clinically translatable artificial intelligence (AI)-powered solution for automated sleep assessment.

In sleep staging, SleepGPT's attention mechanisms focus on critical stage transitions, identifying inconsistent patterns, such as abrupt shifts from deep sleep (N3) directly to wakefulness, that may signal staging errors. The benefits of SleepGPT are particularly pronounced in blinded evaluations on independent datasets, where data heterogeneity often hinders the generalizability of sleep staging models. By effectively capturing the intrinsic sequential patterns of sleep stages, SleepGPT enhances model generalizability across diverse datasets.

In contrast to traditional sleep disorder diagnosis approaches that rely on handcrafted features derived from hypnogram analysis (e.g., stage latencies, durations, and transitions), our proposed HTN directly models and categorizes sleep stage sequences in an end-to-end fashion. This hierarchical architecture allows the model to capture both local and global contextual information within sleep hypnograms, enabling more accurate classification of sleep stage sequences, especially those with varying lengths, compared to the baseline model. Notably, the HTN effectively addresses the challenge of subject-level sleep disorder diagnosis in weakly supervised learning scenarios, where only session-level labels are available and sleep disorder symptoms may not be consistently present throughout the entire sleep session [30]. As such, the HTN achieves accurate subject-level classification and identifies potential biomarkers indicative of specific sleep disorders by leveraging the self-attention mechanism in its global transformer module. These insights hold promise for developing novel diagnostic methodologies and designing precision therapies. Moreover, the HTN's adaptability allows for its potential application to other sleep-related brain disorders, such as depression, anxiety, dementia, and Parkinson's disease [1, 2], e.g., by replacing sleep disorder labels with corresponding condition labels. This flexibility is particularly valuable given the frequent comorbidity of sleep disorders with other neurological or psychiatric conditions.

On the CAP dataset, the pretrained HTN model effectively identifies nearly all abnormal subjects, with only 1.09% of abnormal subjects being erroneously classified as normal. This high sensitivity is crucial in sleep disorder diagnosis, where false positives (misclassifying normal as abnormal) are generally preferable to false negatives, as they lead to further medical investigation, while undetected abnormalities could have serious consequences. Moreover, previous literature indicates that sensitivity and specificity for T1N are 75-90% and 90-98%, respectively [25]. The performance achieved by the pretrained HTN model on the MNC dataset is comparable to or exceeds these reported values. These results suggest that analyzing a single-night PSG can be as effective as the PSG-MSLT gold standard, which involves a 24-hour procedure and is expensive. Consequently, our model offers a reliable and cost-effective alternative as a screening tool for T1N, streamlining the pathway for patients requiring further diagnostic evaluation.

The strong performance of SleepGPT-based models in sleep staging with wearable EEG and disorder identification opens exciting possibilities for real-time, at-home sleep monitoring. Wearable devices powered by SleepGPT could provide immediate feedback on sleep quality, identify potential sleep disturbances as they occur, and offer personalized advice for improving sleep. Such real-time monitoring could be invaluable for individuals with chronic sleep disorders, allowing for timely interventions and

adjustments to treatment plans. Furthermore, continuous sleep monitoring could facilitate early detection of sleep problems, potentially preventing them from escalating into more serious health issues.

While our experiments demonstrate the effectiveness of SleepGPT, it is important to acknowledge certain limitations. The SHHS dataset used for pretraining SleepGPT is relatively small compared to the massive text corpora used to train traditional language models. Expanding the pretraining dataset could further enhance model performance. Future research should also explore fine-grained sleep disorder classification by leveraging larger and more diverse datasets to gain deeper insights into specific sleep stage patterns and transitions associated with particular disorders. Furthermore, while hypnograms provide valuable information on sleep macrostructure, they have limited ability to capture microstructural events like sleep spindles and K-complexes. Integrating additional data modalities, such as EEG or other physiological signals, may enhance the accuracy and granularity of sleep disorder diagnosis.

Methods

Datasets

This study utilizes three distinct data sets: sleep data sets for SleepGPT pretraining, sleep staging, and sleep disorder diagnosis. A total of eight publicly accessible sleep datasets are employed in the experiments. Table 1 provides a comprehensive summary of these datasets, and further details are provided below.

Datasets for SleepGPT Pretraining

Training a transformer-based language model typically requires an extensive text corpus, often encompassing millions or even billions of web pages or documents. However, a sleep dataset of comparable scale, complete with sleep stage annotations, is not available. Fortunately, the increasing advancements in sleep medicine research, coupled with the research community's open data policy, have produced several publicly accessible sleep datasets with sleep stage annotations. The Sleep Heart Health Study (SHHS) database, a multi-center cohort study examining the cardiovascular and other consequences of sleep-disordered breathing [19, 31], is a noteworthy example. This database comprises two rounds of PSG records: Visit 1 (SHHS-1) and Visit 2 (SHHS-2). In this work, we use the SHHS-1 cohort, which encompasses 5,793 subjects aged between 39 and 90 years, to train the SleepGPT model. Notably, to ensure alignment with the AASM scoring standard [33], we merge the N3 and N4 stages into the N3 stage while discarding the MOVEMENT and UNKNOWN epochs, as the SHHS-1 database was manually scored following the R&K guidelines [34].

Datasets for Sleep Staging

The proposed SleepGPT model's performance on sleep staging tasks is evaluated with two widely used sleep datasets, namely, the Sleep-EDF and the Montreal Archive of Sleep Studies (MASS). The Physio2018 dataset serves as a benchmark to evaluate the

Table 1: Overview of the involved sleep datasets. **BMI** = body mass index, **AHI** = apnea-hypopnea index.

Dataset	Subjects / Sessions	Recording Duration	Age	Sex (% Male)	BMI	AHI	Health Conditions
SHHS [19, 31]	5793 / 5793	Overnight	63.1 ± 11.2	47.6	28.2 ± 5.1	17.9 ± 16.1	Sleep-disordered breathing, heart diseases, and others
SleepEDF [20, 32]	78 / 153	Around 9h	59 ± 22.1	46.4	-	-	Healthy subjects
MASS [21]	200 / 200	Overnight	40.6 ± 19.4	48.5	-	≤ 20	Healthy subjects
Physio2018 [22, 32]	994 / 994	7.7h	55 ± 14.3	67.0	33 ± 7.8	19 ± 14.6	Sleep disorders, healthy subjects
BOAS [23]	128 / 128	Overnight	42.2 ± 19.0	40.6	23.8 ± 3.2	-	Healthy subjects
CAP [24, 32]	108 / 108	8–10h	45.2 ± 19.7	61.1	-	-	Sleep disorders (N=92), healthy subjects (N=16)
ISRUC [26]	86 / 86	Overnight	49.7 ± 15.7	59.4	-	-	Sleep apnea (N=76), healthy subjects (N=10)
MNC-CNC [25, 31]	77 / 77	Overnight	28.5 ± 16.9	51.3	23.2 ± 11.5	5.34 ± 1.51	Type-1 narcolepsy (N=54), healthy subjects (N=23)
MNC-DHC [25, 31]	79 / 79	Overnight	33.4 ± 14.8	50.0	24.8 ± 4.9	-	Type-1 narcolepsy (N=21), hypersomnia (N=38), healthy subjects (N=20)
MNC-SSC [25, 31]	251 / 251	Overnight	45.4 ± 13.8	59.4	23.9 ± 6.5	13.7 ± 0.7	Type-1 narcolepsy (N=7), hypersomnia (N=5), healthy subjects (N=239)
HANG7 [27]	84 / 84	8h	24.5 ± 9.6	47.6	22.72 ± 3.65	-	Type-1 narcolepsy (N=13), other narcolepsy (N=38), healthy subjects (N=33)

generalization performance of SleepGPT in enhancing sleep staging. Table 2 provides a detailed summary of these datasets.

Table 2: Number of subjects, EEG channels, and sleep stage distribution of the sleep staging datasets.

Datasets	Subjects	EEG channel	W	N1	N2	N3	REM	Total
SleepEDF	78	Fpz-Cz	69824	21522	69132	13039	25835	199352
MASS	200	C4-A1/C3-A2	31184	19359	107930	30383	40184	229040
Physio2018	994	C3-A2	157945	136978	377870	102592	116877	892262
BOAS	128	C4/AF7	19137	4462	72181	5225	18754	120095

SleepEDF Dataset: We utilize the 2018 version of the SleepEDF Expanded dataset [20, 32]. This collection comprises data from 78 healthy Caucasian subjects aged 25 to 101 years. Each subject contributed two consecutive day-night PSG recordings, except for subjects 13, 36, and 52, where one recording was lost due to device failure. Consequently, the dataset contains 153 overnight recordings. Sleep experts manually scored the epochs based on the R&K standard [34], assigning each 30-second PSG epoch to one of eight categories: {W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN}. To align with convention, the N3 and N4 stages were merged into the N3 stage, while the MOVEMENT and UNKNOWN epochs were excluded. Notably, the SleepEDF-20 dataset was not assessed because it is a subset of this particular version of SleepEDF.

MASS Dataset: Derived from different hospital-based sleep laboratories, the MASS database comprises whole-night recordings from 200 subjects (97 males and 103 females) aged 18 to 76 years [21]. The annotation process involved sleep experts adhering to either the AASM standard [33] (for the SS1 and SS3 subsets) or the R&K standard [34] (for the SS2, SS4, and SS5 subsets). In alignment with the previously mentioned datasets, we harmonized the R&K annotations with the five sleep stages {W, N1, N2, N3, REM} according to the AASM standard. Epochs initially spanning 20 seconds were extended to 30 seconds by incorporating the 5-second segments before and after them.

Physio2018 Dataset: The PhysioNet 2018 Challenge dataset, also known as the Physio2018 dataset, comprises 1,985 polysomnographic recordings provided by the Computational Clinical Neurophysiology Laboratory (CCNL) and the Clinical Data Animation Center (CDAC) at Massachusetts General Hospital (MGH). This dataset was used in the 2018 PhysioNet Challenge [22, 32] to detect sleep arousals. We used the training set for our experiments, which included 944 subjects aged 18 to 90. Sleep experts manually scored the recordings according to the American Academy of Sleep Medicine (AASM) guidelines [33], annotating five sleep stages: W, N1, N2, N3, and REM. This dataset is employed to blindly validate the sleep staging model derived from the MASS dataset.

BOAS Dataset: The Bitbrain Open Access Sleep (BOAS) dataset serves to bridge the gap between gold-standard clinical sleep monitoring and emerging wearable EEG

technologies [23]. This dataset comprises data from 128 nights, during which healthy participants were simultaneously monitored using both a Brain Quick Plus Evolution PSG system by Micromed and a Bitbrain wearable EEG headband. The Micromed PSG system collected EEG signals from electrodes placed at F3, F4, C3, C4, O1, and O2, following the international 10-20 system. In contrast, the Bitbrain headband recorded EEG signals from the frontal AF7 and AF8 electrode sites. Both systems utilized a sampling rate of 256 Hz. Sleep staging was independently annotated by three expert scorers following the American Academy of Sleep Medicine (AASM) criteria [33], with a consensus label established by a fourth expert. The BOAS dataset enables the evaluation of the SleepGPT-based model's ability to achieve sleep staging accuracy comparable to PSG using wearable EEG data.

Datasets for Sleep Disorder Diagnosis

Two sleep-disorder-related datasets were identified for the sleep disorder diagnosis analysis, including the CAP Sleep Database and the Mignot Nature Communications (MNC) dataset, which offer both sleep stage annotations and sleep disorder labels.

CAP Dataset: The CAP Sleep Database is a collection of 108 polysomnographic recordings contributed by the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy [24, 32]. It contains data from seven groups of patients with distinct sleep disorders, as well as a healthy control group without any medical, neurological, or psychiatric conditions. The details of these groups are summarized in Table 3. Well-trained neurologists who are sleep experts manually scored the sleep recordings based on the R&K rules, categorizing the epochs into sleep stages 1-4, wake, REM sleep, and movement artifacts. To adhere to the AASM standard, we harmonized the R&K annotations, obtaining five sleep stages: {W, N1, N2, N3, REM}. Notably, the CAP Sleep Database has been widely used in numerous studies focusing on sleep disorder diagnosis.

Table 3: Sleep disorder groups in the CAP Sleep Database.

Sleep disorder	No. of subjects
Bruxism (BRUX)	2
Sleep-disordered breathing (SBD)	4
Insomnia (INS)	9
Narcolepsy (NARCO)	5
Nocturnal frontal lobe epilepsy (NFLE)	40
Periodic leg movements (PLMs)	10
REM behavior disorder (RBD)	22
No pathology (N)	16

Note that the CAP Sleep Database is inherently imbalanced. Several groups contain an extremely limited number of subjects. For instance, the BRUX and SBD groups consisted of only 2 and 3 subjects, respectively. To facilitate subject-level diagnoses, we conducted binary classification experiments on the CAP Sleep Database, specifically distinguishing abnormal from normal cases. Furthermore, after incorporating

the first-session data of 78 subjects (excluding subjects 36 and 52, who had only one session available) from the SleepEDF Expanded database into the normal group, the comprehensive dataset included 186 subjects, 94 of whom were labeled as normal and 92 of whom were labeled as abnormal.

ISRUC Dataset: The ISRUC-Sleep dataset [26] contains full-night PSG recordings, each approximately eight hours in duration, collected at the Sleep Medicine Centre of Coimbra University Hospital (CHUC) between 2009 and 2013. Data were acquired non-invasively using a SomnoStar Pro multi-channel system, with sensors placed according to the international 10-20 standard. The dataset includes recordings from both healthy adults and subjects with sleep disorders under medication, divided into three groups: 1) 100 subjects with one session each; 2) 8 subjects with two sessions for longitudinal studies; 3) 10 healthy subjects with one session, used for comparison with sleep disorder patients. Among the dataset, 76 obstructive sleep apnea (OSA) patients and 10 healthy controls are included to blindly validate the sleep disorder diagnosis model derived from the CAP dataset.

MNC Dataset: The Mignot Nature Communications (MNC) dataset comprises raw polysomnography data collected from an automated sleep staging project utilizing neural networks [25, 31]. It encompasses data from ten distinct cohorts recorded at twelve sleep centers across three continents: the patient-based Stanford Sleep Cohort (SSC), the population-based Wisconsin Sleep Cohort (WSC), the patient-based Inter-scorer Reliability Cohort (IS-RC), the Jazz Clinical Trial Sample (JCTS), the patient-based Korean Hypersomnia Cohort (KHC), the patient-based Austrian Hypersomnia Cohort (AHC), the patient-based Italian Hypersomnia Cohort (IHC), the patient-based Danish Hypersomnia Cohort (DHC), the patient-based French Hypersomnia Cohort (FHC), and the patient-based Chinese Narcolepsy Cohort (CNC). The study received approval from institutional review boards, and informed consent was obtained from all participants. Trained sleep-scoring technicians manually annotated all sleep studies according to the AASM Scoring Manual [33]. Additionally, the subjects were divided into the Type-1 narcolepsy (T1N; with either low CSF hypocretin-1 levels or clear cataplexy), other hypersomnia (OHS), or non-narcolepsy control (NNC) group based on multiple sleep latency test (MSLT) results, cataplexy symptoms, and human leukocyte antigen (HLA) results (if available). Further information about the MNC dataset can be found in [25].

However, as diagnostic results were only available for the SSC, DHC, and CNC cohorts, we exclusively utilized these three cohorts for the sleep disorder diagnosis task. The dataset compiled from these cohorts encompassed 407 subjects, among whom 82 were in the T1N group, 43 were in the OHS group, and 282 were in the NNC group (see Table 4). To facilitate comparison with previous research, we adhered to the methodology outlined in [25] and conducted binary classification experiments on the MNC dataset to discriminate T1N subjects from all other subjects, i.e., OHS and NNC subjects. This dataset is also employed to blindly assess the generalizability of the sleep disorder diagnosis model derived from the CAP dataset.

HANG7 Dataset: In the HANG7 dataset [27], 84 participants aged 11 to 57 years (mean age: 24.5 ± 9.6), including 44 females and 40 males, were recruited to collect polysomnography recordings at the Affiliated Mental Health Center & Hangzhou

Table 4: Sleep disorder groups in the MNC dataset.

Sleep disorder	No. of subjects			
	CNC	**DHC**	**SSC**	**Total**
Type-1 narcolepsy (T1N)	54	21	7	82
Other hypersomnia (OHS)	0	38	5	43
Non-narcolepsy control (NNC)	23	20	239	282
Total	77	79	251	407

Seventh People's Hospital, Zhejiang University School of Medicine. The study was conducted at Zhejiang University with Institutional Review Board approval, and written consent was obtained from all participants or their caregivers. PSG recordings were collected following the AASM sleep standards [33] and manually scored by experienced sleep technicians. Each participant's PSG recording covered one full night, from approximately 21:00 to 5:00 the next morning, totaling about 8 hours. Of the 84 participants, 13 were diagnosed with T1N (with clear cataplexy), 38 were other narcolepsy, and 33 were healthy controls. The dataset was used to blindly validate the generalizability of the sleep disorder diagnosis models derived from the CAP (abnormal vs. normal) and MNC (T1N vs. others) datasets.

Problem Formulation

Sleep Staging as Speech Recognition

A general speech recognition system consists of an acoustic model and a language model. The acoustic model $\mathcal{M}_1$ predicts the most likely word sequence given speech audio features X via

$$P(Y;\mathcal{M}_1) = P(Y|X;\mathcal{M}_1), \tag{1}$$

while the language model $\mathcal{M}_2$ is built based on a large-scale text corpus to capture the sequential characteristics of the word sequence via

$$P(Y;\mathcal{M}_2) = P(y_1, y_2, \cdots, y_T) = \prod_{t=1}^{T} P(y_t|y_1, y_2, \cdots, y_{t-1}). \tag{2}$$

The trained language model is then used to rectify the word sequence predicted by the acoustic model to improve the speech recognition performance as follows [35–37]:

$$\begin{aligned} Y^\star &= \arg\max_Y P(Y;\mathcal{M}_1,\mathcal{M}_2) \\ &= \arg\max_Y P(Y|X;\mathcal{M}_1)P(Y;\mathcal{M}_2). \end{aligned} \tag{3}$$

In the automated sleep staging process, features are extracted from a sleep data epoch, and a sleep staging model is then applied to predict the most likely sleep stage for that epoch. The model, whether it is a traditional machine learning model or a deep learning model, is similar to the acoustic model in a speech recognition system. It is designed and trained on a large-scale sleep dataset to achieve satisfactory prediction

accuracy on unseen sleep epochs. However, due to various factors such as architectural limitations, training inadequacies, and data scarcity, imperfections in the PSG-based sleep staging model may occur, leading to inaccurate predictions of sleep stages. The inter-scorer reliability for sleep stage scoring is reported to be 82.6% on average, which closely aligns with that achieved by machine learning-based automated staging systems [25]. Given the analogous nature of text and sleep stage sequences, emulating the speech recognition system paradigm may improve sleep staging performance. This involves training a sleep language model that can capture the inherent sequential characteristics of sleep stages to assume a role akin to that of a natural language model (see Figure 2B).

Sleep Disorder Diagnosis as Text Classification

It is believed that sleep architecture, i.e., the distribution of sleep stages, is strongly related to the quality of sleep. Certain sleep disorders exhibit disrupted sleep architecture and atypical sleep stage transitions. Examples include shortened deep sleep stages in cases of insomnia, and short REM sleep latency or immediate transitions from wakefulness to REM sleep in narcolepsy. An overnight sleep stage sequence has the form $\{\mathrm{W}, \cdots, \mathrm{N1}, \cdots, \mathrm{N2}, \cdots, \mathrm{N3}, \cdots, \mathrm{REM}, \cdots\}$. Such sequences are plotted as hypnograms in Figure 1 (top-left). They are akin to concise news pieces or textual documents with a vocabulary of only five words. In turn, the sleep stage sequence encodes the sequential attributes and transitional patterns characterizing the progression between sleep stages. The integration of a sequential model enables these characteristics to be identified and described.

Hence, a viable approach is to treat sleep disorder diagnosis as a long text classification task. Here, the trained sleep language model is harnessed to extract features from the sleep stage sequences (hypnograms), enabling the subsequent classification of these sequences as associated with a sleep disorder or not. This is an innovative method for sleep disorder diagnosis using only the sleep stages obtained from either human experts or automated sleep staging models.

The SleepGPT model

Recently, the success of transformer-based language models (LMs), including bidirectional encoder representations from transformers (BERT) and generative pretrained transformers (GPT), has been noted across various natural language processing (NLP) tasks, such as machine translation, question-answering, and text generation [28, 38, 39]. BERT, which functions as a bidirectional LM, is trained to predict masked words based on neighboring words and to perform next-sentence prediction. Conversely, GPT operates as a decoder-only transformer LM, autoregressively predicting the next character or word based on preceding ones. Both models undergo self-supervised training on extensive text corpora. The success of ChatGPT has illustrated its proficient ability to capture the inherent sequential attributes of natural language. This ability has been extensively utilized to improve performance in speech recognition and diverse NLP tasks, such as text classification.

Transformers consist of multiple transformer blocks, typically including a multi-head self-attention layer, a feed-forward layer, and residual and normalization layers [40]. GPT employs causal attention in its self-attention layer to ensure exclusive attention to preceding words. This concept has been applied in sequential sleep staging models that rely on prior sleep epochs for current sleep stage predictions. It has proven to be especially significant in online sleep staging systems, where only preceding epochs are accessible during the current prediction task. Hence, GPT is selected to capture the sequential traits of sleep stages. The SleepGPT model adopts the architecture of the GPT-2 language model developed by OpenAI [28, 39]. As illustrated in Figure 2A, the main part of the model is composed of a series of n transformer decoder blocks, marked by the masked multi-head self-attention layers. Token embedding and position embedding layers are employed to map the sleep stage tokens into a vector space and to infuse positional information into each token's sequence placement, respectively. The final transformer block's output is input to a linear layer to obtain the probability distribution of the next sleep stage.

The model undergoes self-supervised training through the autoregressive forecast of the most likely sleep stage based on preceding sleep stages. For each subject or sleep session, the overnight sleep stage annotations are structured as a long sequence $\mathcal{U} = \{u_1, ..., u_N\}$, where $u_i \in \{0, 1, 2, 3, 4\}$ represents the five sleep stages $\{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ as integers ranging from 0 to 4. Training samples are derived from the sequence using a sliding window of dimensions K, passing over the sequence with a stride of 1, where K signifies the sample sequence length. Consequently, the overnight stage sequence yields $N-K+1$ overlapping instances. The $(i-1)$th instance of the K sleep stages, denoted as $\mathbf{U}_{i-1} = \{u_{i-K}, ..., u_{i-1}\}$, constitutes the input and is fed into the model for predicting the target; the target is the ith instance $\mathbf{U}_i = \{u_{i-K+1}, ..., u_i\}$:

$$h_0 = \mathbf{U}_{i-1}\mathbf{W}_e + \mathbf{W}_p \tag{4}$$

$$h_l = \text{transformer_block}(h_{l-1}),\ \forall l \in [1, L] \tag{5}$$

$$P(\mathbf{U}_i) = \text{softmax}(h_n \mathbf{W}_s^T) \tag{6}$$

where L is the number of transformer layers, $\mathbf{W}_e$ is the token embedding matrix, $\mathbf{W}_p$ is the position embedding matrix, and $\mathbf{W}_s$ is the weight of the classification head. The model is trained with the objective of minimizing the cross-entropy loss, thereby reducing the disparity between the predicted sleep stage and the actual ground truth:

$$\mathcal{L}(\mathcal{U}) = -\sum_{i=1} \log P(\mathbf{U}_i | \mathbf{U}_{i-1}; \Theta), \tag{7}$$

where Θ is the parameter of the GPT sequential model.

Sleep Staging with SleepGPT

Automated sleep staging is an active research topic within the realm of sleep medicine; it aims to automate the prediction of sleep stages for individual sleep PSG data epochs,

typically spanning 30 seconds. Several excellent deep learning models have been proposed for this purpose. Most of them share a common architecture, comprising a CNN to extract intra-epoch features and an RNN to incorporate the contextual information in adjacent PSG data epochs[4–7, 12, 41]. Among cutting-edge sleep staging models, XSleepNet[12] employs two network streams to learn from multi-view inputs (e.g., both raw signals and time-frequency images) for sleep staging. By adapting the contributions of the two views on time to perform joint feature learning during training, XSleepNet outperforms the single-view baselines and multi-view baselines with a simple fusion strategy. However, while contextual information within PSG signals is incorporated into the above deep learning models, they neglect the inherent sequential traits and transition patterns within sleep stages.

To use the trained SleepGPT model to improve sleep staging performance, we treat the sleep staging task as a speech recognition task and follow the pipeline in Figure 2B. The sleep staging model (SSM) is used to predict the most likely sleep stage given a PSG data epoch or preceding epochs, while the SleepGPT model is used to rectify the predicted sleep stage given past stages. That is, the output logits of the SSM $P_{\mathrm{SSM}}(y)$ and those of the sleep language model (SLM) $P_{\mathrm{SLM}}(y)$ are weighted by a factor α to obtain the final sleep stage prediction:

$$P(y) = \alpha P_{\mathrm{SSM}}(y|\mathbf{x}) + (1-\alpha) P_{\mathrm{SLM}}(y|\mathbf{y}_{-}), \tag{8}$$

where $\mathbf{x}$ is the set of input data epochs, including the current and preceding epochs depending on whether a memory staging model is used, and $\mathbf{y}_{-}$ is the set of preceding sleep stages (context). Notably, the hyperparameter α governs the relative influences of these two models, ultimately steering the sleep stage prediction toward the highest probability outcome.

Sleep Disorder Diagnosis with SleepGPT

Sleep disorder diagnosis involves discerning the presence and specific type of sleep disorder in an individual. This task is typically conducted either by sleep specialists or by ML models that leverage PSG data for automated assessment. ML-based approaches often include PSG data feature extraction, followed by the classification of sleep disorders. However, the high dimensionality of PSG data, coupled with a lack of labeled instances, makes it challenging to train an ML model with robust generalizability. The situation is even worse in the case of deep learning models, which commonly demand an extensive volume of labeled data to achieve optimal model performance.

Pretrained language models, such as GPTs trained on expansive text corpora, can capture the intrinsic sequential characteristics of natural language. These pretrained sleep language models can function as feature extractors, processing text sequences to subsequently enable classification. Notably, these pretrained models can be further fine-tuned using limited-scale datasets, significantly enhancing the classification performance. This strategy, referred to as transfer learning, effectively addresses the challenge of limited sample size, and it has a proven track record of success across numerous NLP applications.

Drawing inspiration from this pretraining and fine-tuning paradigm, we leverage the pretrained SleepGPT model as a feature extractor. We replace the subsequent stage prediction layer with a classifier and perform comprehensive fine-tuning to facilitate sleep disorder classification. Nevertheless, a challenge arises for the SleepGPT model in handling long sequences: the sequence length limitation. A whole-night (8-hour) sleep stage sequence comprises 960 time steps, significantly exceeding the sequence length restriction of the SleepGPT model. We address this by segmenting the overnight sequence into shorter sections. Initially, the SleepGPT model is employed to extract local context features from these short sleep stage segments. Subsequently, another sequential model (a transformer encoder) is used to capture the global contextual features from the SleepGPT output [40]. Finally, the resulting global context features are fed into a classification head to predict sleep disorder labels. The configuration of the hierarchical transformer network (HTN) is depicted in Figure **S1**. This architecture, which is tailored for lengthy sequence classification, is borrowed from the hierarchical attention network (HAN) and hierarchical transformers utilized for long text classification [42, 43].

Additionally, to facilitate mini-batch training, we pad the short sleep stage sequences within a batch to achieve a uniform length. Subsequently, a mask matrix is employed to exclude the padded values during the computation of attention weights and the loss calculation. The loss function involves cross-entropy loss, which compares the predicted sleep disorder labels with the ground truth:

$$\mathcal{U} = \{\mathbf{U}_i\},\ \forall i \in [1, N] \tag{9}$$

$$\mathbf{z}_i = \text{SleepGPT}(\mathbf{U}_i) \tag{10}$$

$$\mathbf{V} = \text{transformer_encoder}(\mathbf{Z}) \tag{11}$$

$$P(\mathcal{U}) = \text{softmax}(\mathbf{W}_c \mathbf{v}_0 + b_c) \tag{12}$$

$$\mathcal{L}(\mathcal{U}) = -\sum_{i=1} \log P(\mathcal{U}; \Phi), \tag{13}$$

where $\mathbf{U}_i$ is the ith segment, and N represents the number of segments within a sleep stage sequence. $\mathbf{z}_i$ denotes the ith local feature vector generated by SleepGPT. Notably, $\mathbf{Z} = \{\mathbf{z}_0, \cdots, \mathbf{z}_N\}$ and $\mathbf{V} = \{\mathbf{v}_0, \cdots, \mathbf{v}_N\}$ denote the input and output sequences of the transformer encoder, respectively. Here, $\mathbf{z}_0$ corresponds to the CLS token, while $\mathbf{v}_0$ represents the global feature vector at the CLS token's output position. The CLS token in transformers is a special token added at the beginning of the input sequence, and its final output serves as a global representation of the entire sequence for tasks like classification [38, 44]. The parameters of the classification head are denoted as $\mathbf{W}_c$ and b_c, and Φ includes the trainable parameters of the entire model.

Conclusion

In summary, this study presents SleepGPT, a novel GPT-based sleep language model designed to advance automated sleep staging and sleep disorder diagnosis. Through extensive evaluation across multiple publicly available and independent datasets,

SleepGPT demonstrated consistent improvements in sleep staging accuracy and generalizability, outperforming existing state-of-the-art models. Its integration as a feature extractor within hierarchical transformer networks further enabled robust and scalable sleep disorder classification, including challenging tasks such as Type-1 narcolepsy identification. The model's ability to capture sequential sleep stage dynamics and highlight interpretable patterns underscores its potential for clinical translation and precision sleep medicine. Future work may focus on expanding pretraining datasets, refining model architectures, and integrating multimodal physiological signals to further enhance diagnostic capabilities and applicability in real-world settings.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The SHHS [19] and MNC [25] datasets are provided by the National Sleep Research Resource with appropriate deidentification. Permission and access for these datasets can be obtained via the online portal: https://www.sleepdata.org. The SleepEDF [20], Physio2018 [22], and CAP [24] datasets are available from PhysioNet at https://physionet.org/content/sleep-edfx/1.0.0/, https://physionet.org/content/challenge-2018/1.0.0/, and https://physionet.org/content/capslpdb/1.0.0/, respectively. The MASS [21] dataset is available at http://ceams-carsm.ca/mass/. The BOAS [23] dataset can be accessed at https://openneuro.org/datasets/ds005555/versions/1.0.0. The ISRUC [26] dataset can be accessed at https://sleeptight.isr.uc.pt/. Access to the HANG7 dataset is governed by data-use agreements, and it is therefore not publicly available.

Code availability

The models and source codes for reproducing the results reported in this paper can be accessed at https://github.com/yuty2009/sleepgpt.

Acknowledgements

This work was supported in part by STI2030-Major Projects under Grant 2022ZD0211700, AI Program of Shanghai Municipal Education Commission under Grant JWAIZD-4, the National Natural Science Foundation of China under Grant 62376098, 62276102, and U22A20293, and GuangDong Basic and Applied Basic Research Foundation 2024A1515011983.

Author contributions

T.Y. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. Z.G. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. Z.W., R.H., F.W., M.L., J.Y., Z.Y., J.Z., Y.X., H.J., W.L., G.D., Z.G.,

Y.W., J.L., Y.Z., and M.J. contributed to the analysis and interpretation of the data. Y.L. contributed to the analysis and interpretation of the data, and the drafting of the manuscript. J.X. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. W.W. contributed to the development of methods, the analysis and interpretation of the data, and the drafting of the manuscript. All authors reviewed the manuscript.

Competing interests

W.W. reports equity from Alto Neuroscience. None of the other authors has financial disclosures to report.

Additional information

Additional study and data information is contained in the supplementary material.

References

[1] Mahowald, M.W., Schenck, C.H.: Insights from studying human sleep disorders. Nature **437**(7063), 1279–85 (2005) https://doi.org/10.1038/nature04287

[2] Zheng, N.S., Annis, J., Master, H., Han, L., Gleichauf, K., Ching, J.H., Nasser, M., Coleman, P., Desine, S., Ruderfer, D.M., Hernandez, J., Schneider, L.D., Brittain, E.L.: Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the all of us research program. Nature Medicine **30**(9), 2648–2656 (2024) https://doi.org/10.1038/s41591-024-03155-8

[3] Rosenberg, R.S., Van Hout, S.: The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. Journal of Clinical Sleep Medicine **9**(1), 81–7 (2013) https://doi.org/10.5664/jcsm.2350

[4] Supratak, A., Dong, H., Wu, C., Guo, Y.: DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(11), 1998–2008 (2017) https://doi.org/10.1109/TNSRE.2017.2721116

[5] Supratak, A., Guo, Y.: TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In: The 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 641–644 (2020). https://doi.org/10.1109/EMBC44109.2020.9176741

[6] Phan, H., Andreotti, F., Cooray, N., Chen, O.Y., De Vos, M.: Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Transactions on Biomedical Engineering **66**(5), 1285–1296 (2019) https://doi.org/10.1109/TBME.2018.2872652

[7] Phan, H., Andreotti, F., Cooray, N., Chen, O.Y., De Vos, M.: SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering **27**(3), 400–410 (2019) https://doi.org/10.1109/TNSRE.2019.2896659

[8] Phan, H., Mikkelsen, K., Chén, O.Y., Koch, P., Mertins, A., De Vos, M.: SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Transactions on Biomedical Engineering **69**(8), 2456–2467 (2022) https://doi.org/10.1109/TBME.2022.3147187

[9] Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P.M., Guo, Y.: Mixed neural network approach for temporal sleep stage classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **26**(2), 324–333 (2018) https://doi.org/10.1109/TNSRE.2017.2733220

[10] Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, 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) https://doi.org/10.1109/TNSRE.2021.3076234

[11] Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P.J., Igel, C.: U-Sleep: resilient high-frequency sleep staging. npj Digital Medicine **4**(1), 72 (2021) https://doi.org/10.1038/s41746-021-00440-5

[12] Phan, H., Chen, O.Y., Tran, M.C., Koch, P., Mertins, A., De Vos, M.: XSleepNet: Multi-view sequential model for automatic sleep staging. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5903–5915 (2022) https://doi.org/10.1109/TPAMI.2021.3070057

[13] Seo, H., Back, S., Lee, S., Park, D., Kim, T., Lee, K.: Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. Biomedical Signal Processing and Control **61**, 102037 (2020) https://doi.org/10.1016/j.bspc.2020.102037

[14] Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). American Psychiatric Publishing, Washington, DC, USA (2022)

[15] Thachayani, M., Loganayagi, M.: Artificial intelligence based classifier for sleep disorder detec-tion using eeg-bci data. International Journal of Computer Science Trends and Technology **9** (2021)

[16] Phyo, J., Ko, W., Jeon, E., Suk, H.-I.: TransSleep: Transitioning-aware attention-based deep neural network for sleep staging. IEEE Transactions on Cybernetics **53**(7), 4500–4510 (2023) https://doi.org/10.1109/TCYB.2022.3198997

[17] Born, J., Manica, M.: Regression transformer enables concurrent sequence

regression and generation for molecular language modelling. Nature Machine Intelligence **5**(4), 432–444 (2023) https://doi.org/10.1038/s42256-023-00639-z

[18] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023) https://doi.org/10.1126/science.ade2574

[19] Quan, S.F., Howard, B.V., Iber, C., Kiley, J.P., Nieto, F.J., O'Connor, G.T., Rapoport, D.M., Redline, S., Robbins, J., Samet, J.M., Wahl, P.W.: The sleep heart health study: Design, rationale, and methods. Sleep **20**(12), 1077–1085 (1997) https://doi.org/10.1093/sleep/20.12.1077

[20] Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A.C., Oberye, J.J.L.: Analysis of a sleep-dependent neuronal feedback loop: the slow-wave micro-continuity of the EEG. IEEE Transactions on Biomedical Engineering **47**(9), 1185–1194 (2000) https://doi.org/10.1109/10.867928

[21] O'Reilly, C., Gosselin, N., Carrier, J., Nielsen, T.: Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. Journal of Sleep Research **23**(6), 628–635 (2014) https://doi.org/10.1111/jsr.12169

[22] Ghassemi, M.M., Moody, B.E., Lehman, L.H., Song, C., Li, Q., Sun, H., Mark, R.G., Westover, M.B., Clifford, G.D.: You snooze, you win: the physionet/computing in cardiology challenge 2018. Computing in Cardiology (2010) **45** (2018) https://doi.org/10.22489/cinc.2018.049

[23] López-Larraz, E., Sierra-Torralba, M., Clemente, S., Fierro, G., Oriol, D., Minguez, J., Montesano, L., Klinzing, J.G.: Bitbrain open access sleep dataset (2024) https://doi.org/10.18112/openneuro.ds005555.v1.0.0

[24] Terzano, M.G., Parrino, L., Sherieri, A., Chervin, R., Chokroverty, S., Guilleminault, C., Hirshkowitz, M., Mahowald, M., Moldofsky, H., Rosa, A., Thomas, R., Walters, A.: Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Medicine **2**(6), 537–553 (2001) https://doi.org/10.1016/S1389-9457(01)00149-6

[25] Stephansen, J.B., Olesen, A.N., Olsen, M., Ambati, A., Leary, E.B., Moore, H.E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y.L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S.C., Kim, T.W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S.T., Schweitzer, P.K., Kushida, C., Peppard, P.E., Sorensen, H.B.D., Jennum, P., Mignot, E.: Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nature Communications **9**(1), 5229 (2018) https://doi.org/10.1038/s41467-018-07229-3

[26] Khalighi, S., Sousa, T., Santos, J.M., Nunes, U.: ISRUC-Sleep: A comprehensive public dataset for sleep researchers. Computer Methods and Programs in Biomedicine **124**, 180–192 (2016) https://doi.org/10.1016/j.cmpb.2015.10.013

[27] Wang, J., Zhao, S., Zhou, Y., Jiang, H., Yu, Z., Li, T., Li, S., Pan, G.: Narcolepsy diagnosis with sleep stage features using PSG recordings. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 3619–3629 (2023) https://doi.org/10.1109/TNSRE.2023.3312396

[28] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[29] Vallat, R., Walker, M.P.: An open-source, high-performance tool for automated sleep staging. eLife **10**, 70092 (2021) https://doi.org/10.7554/eLife.70092

[30] Xu, S., Faust, O., Seoni, S., Chakraborty, S., Barua, P.D., Loh, H.W., Elphick, H., Molinari, F., Acharya, U.R.: A review of automated sleep disorder detection. Computers in Biology and Medicine **150**, 106100 (2022) https://doi.org/10.1016/j.compbiomed.2022.106100

[31] Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., Redline, S.: The national sleep research resource: towards a sleep data commons. Journal of the American Medical Informatics Association **25**(10), 1351–1358 (2018) https://doi.org/10.1093/jamia/ocy064

[32] Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet. Circulation **101**(23), 215–220 (2000) https://doi.org/10.1161/01.CIR.101.23.e215

[33] Iber, C., Ancoli-Israel, S., Chesson, A.L., Quan, S.: The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. American Academy of Sleep Medicine (2007)

[34] Wolpert, E.A.: A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Archives of General Psychiatry **20**(2), 246–247 (1969) https://doi.org/10.1001/archpsyc.1969.01740140118016

[35] Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964 (2016). https://doi.org/10.1109/ICASSP.2016.7472621

[36] Rao, K., Sak, H., Prabhavalkar, R.: Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. In: 2017 IEEE

Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 193–199 (2017). https://doi.org/10.1109/ASRU.2017.8268935

[37] Karita, S., Soplin, N., Watanabe, S., Delcroix, M., Ogawa, A., Nakatani, T.: Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH **2019-September**, 1408–1412 (2019) https://doi.org/10.21437/Interspeech.2019-1938 . Publisher Copyright: Copyright © 2019 ISCA; 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019 ; Conference date: 15-09-2019 Through 19-09-2019

[38] Kenton, J.D.M.-W.C., Toutanova, L.K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)

[39] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning (2018)

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[41] Mousavi, S., Afghah, F., Acharya, U.R.: SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. PLoS One **14**(5), 0216456 (2019) https://doi.org/10.1371/journal.pone.0216456

[42] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-1174 . https://aclanthology.org/N16-1174

[43] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., Dehak, N.: Hierarchical transformers for long document classification. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 838–844 (2019). https://doi.org/10.1109/ASRU46091.2019.9003958

[44] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YicbFdNTTy

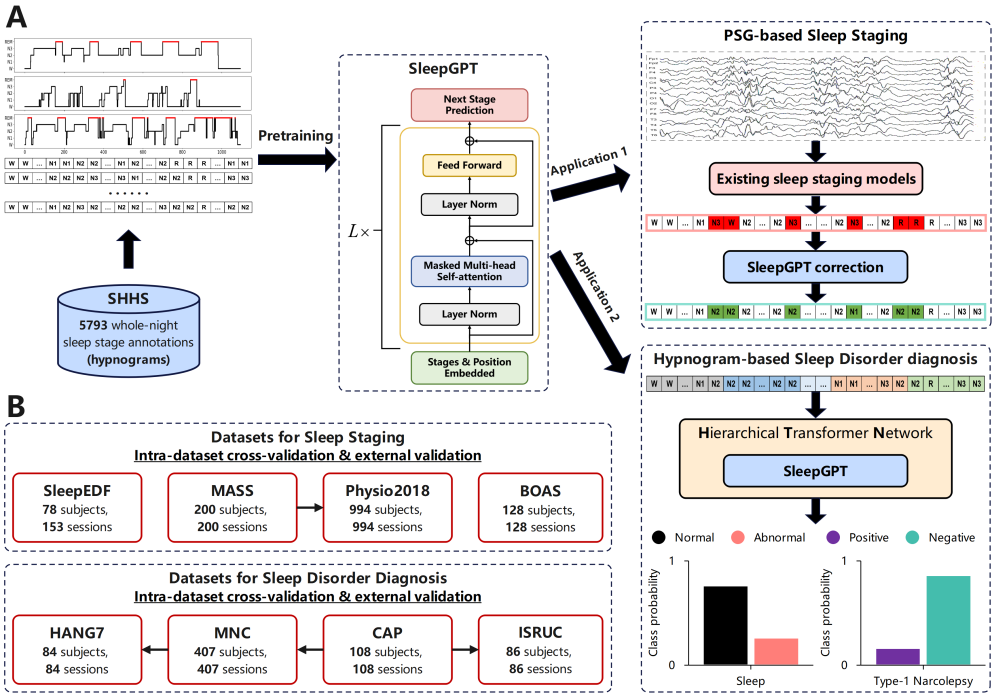

Fig. 1: Overview of the proposed SleepGPT model and applications to sleep staging enhancement and sleep disorder diagnosis. (**A**) The SleepGPT model is pretrained on a large sleep stage annotation dataset SHHS [19] and is used to correct the sleep stage predictions of existing sleep staging models. Moreover, a hierarchical transformer network (HTN) is employed for sleep disorder diagnosis, with SleepGPT acting as a local feature extractor. (**B**) Datasets for evaluating the proposed artificial intelligence (AI) models. For sleep staging, cross-validation of the AI models is performed on the SleepEDF [20] and MASS [21] datasets. The models trained from the MASS datasets are then externally validated on the Physio2018 [22] dataset for generalizability assessment. Furthermore, the translatability of the SleepGPT model on wearable EEG data is validated on the BOAS [23] dataset with simultaneously collected PSG and headband EEG data. For sleep disorder diagnosis, cross-validation of the AI models is performed on the CAP [24] and MNC [25] datasets. For generalizability assessment, models trained on the CAP dataset for distinguishing normal from abnormal sleep stage sequences are validated externally using the ISRUC [26], MNC, and HANG7 [27] datasets, while those trained on the MNC dataset for distinguishing Type-1 narcolepsy from other hypersomnia and healthy controls are externally validated on the HANG7 dataset.

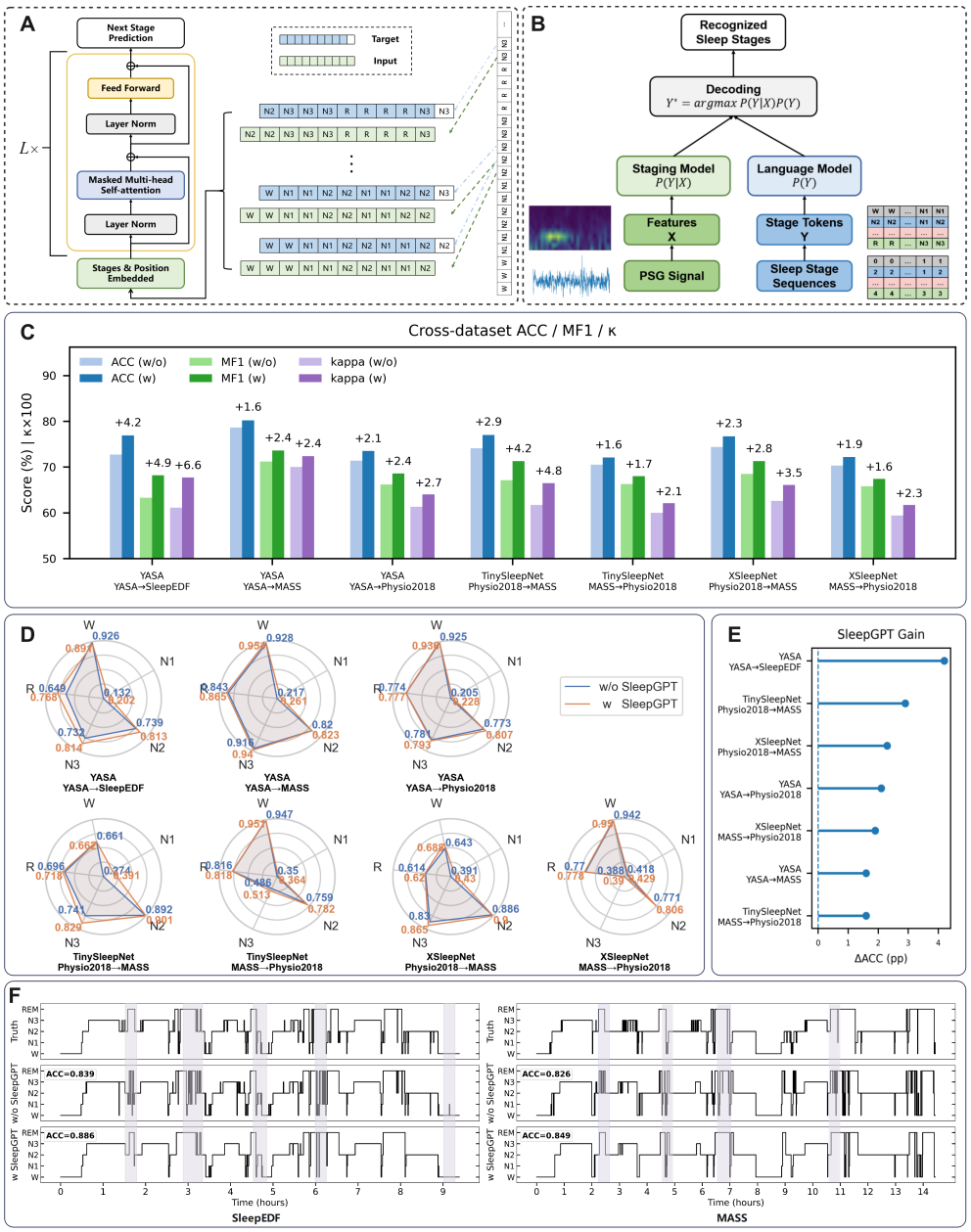

Fig. 2: Results of three state-of-the-art staging methods with and without SleepGPT when performing cross-dataset sleep staging on the SleepEDF, MASS, and Physio2018 datasets. The labels **Source** → **Target** indicate that the staging model is trained from the **Source** dataset and evaluated on the **Target** dataset. (**A**) Architecture and training procedure of the SleepGPT model. (**B**) The pipeline for enhancing existing sleep staging models with SleepGPT. (**C-E**) Performance comparison of YASA [29], TinySleepNet [5], and XSleepNet [12] with and without SleepGPT on the SleepEDF, MASS, and Physio2018 datasets. (**F**) Two examples of the corrections made by SleepGPT during cross-dataset validation. The shaded areas indicate where the SleepGPT model corrected the erroneous sleep stage predictions of the baseline staging models.

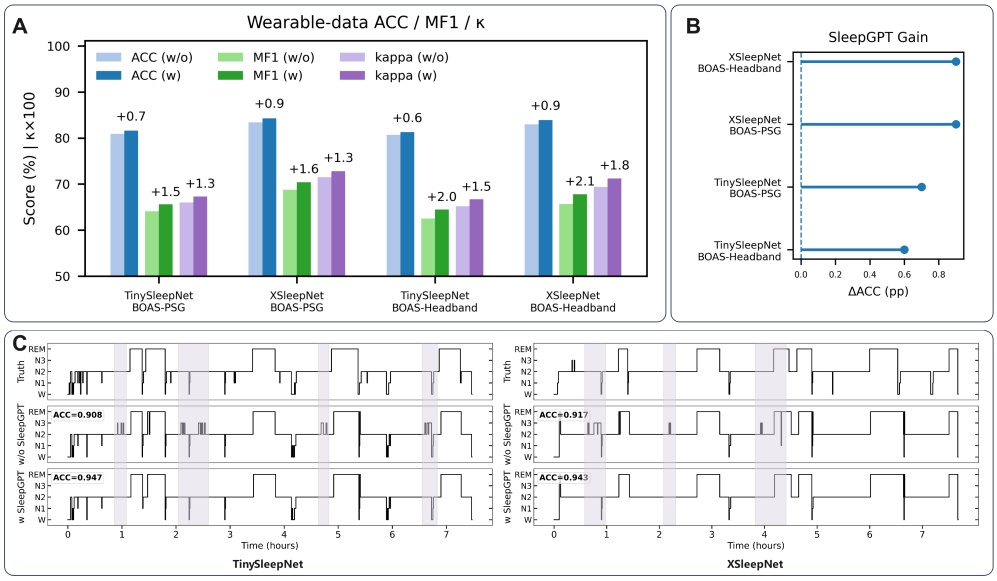

Fig. 3: Sleep staging results of the state-of-the-art staging methods with and without SleepGPT on the BOAS wearable dataset. (**A**) Performance comparison of TinySleepNet [5] and XSleepNet [12] with and without SleepGPT on the PSG and headband EEG data. (**B**) Performance gain obtained by integrating SleepGPT into the TinySleepNet and XSleepNet models on the PSG and headband EEG data. (**C**) Examples of the corrections made by SleepGPT on the headband EEG data. The shaded areas indicate where the SleepGPT model corrected the erroneous sleep stage predictions of the baseline staging models.

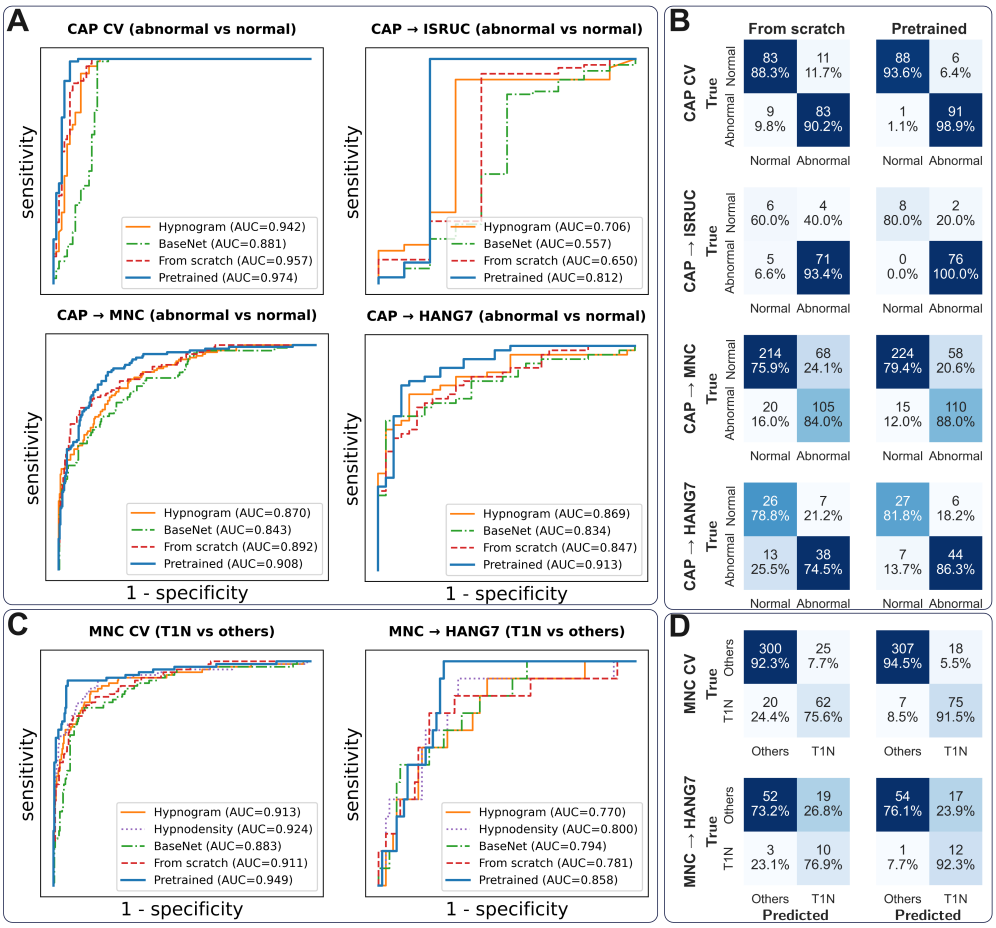

Fig. 4: Performance of the SleepGPT-powered hierarchical transformer network (HTN) and baseline methods for sleep disorder diagnosis, normal vs abnormal sleep on the CAP, ISRUC, MNC, and HANG7 datasets, and Type-1 narcolepsy (T1N) patients vs non-T1N individuals on the MNC and HANG7 datasets. (**A**) Receiver operating characteristic (ROC) curves for SleepGPT-powered and baseline models in abnormal sleep detection. Cross-validation (CV) results on CAP and external validation on ISRUC, MNC, and HANG7 are shown; the area under the ROC curve (AUC) is indicated in the legend. (**B**) Confusion matrices for HTN models trained from scratch and with pretrained SleepGPT parameters across four datasets. (**C**) ROC curves for SleepGPT-powered and baseline models in T1N detection, showing CV results on MNC and external validation on HANG7; AUC is indicated in the legend. (**D**) Confusion matrices for HTN models trained from scratch and with pretrained SleepGPT parameters on both datasets.

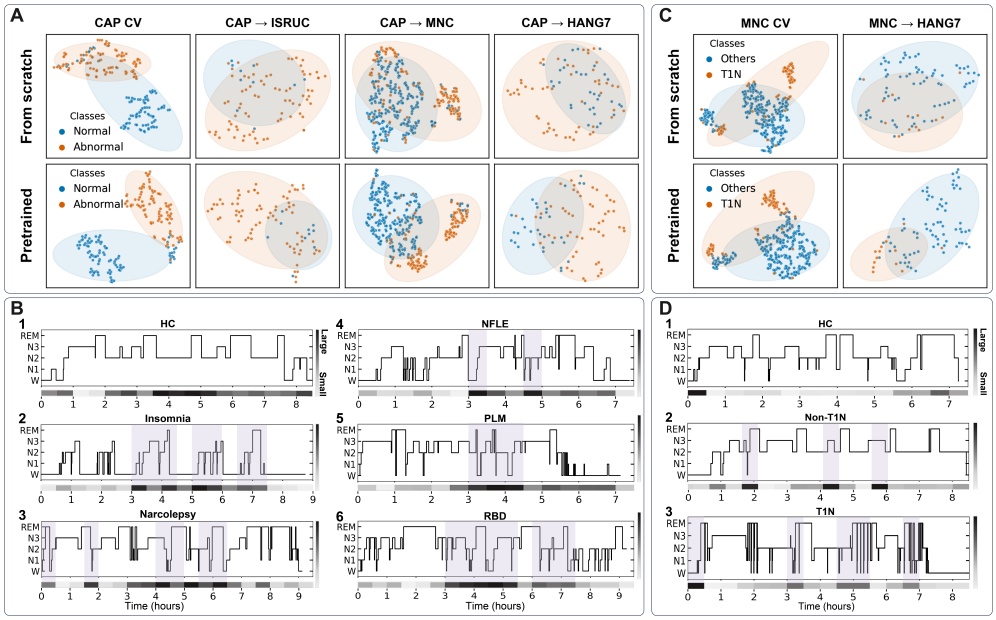

Fig. 5: Visualization and interpretation of the SleepGPT-powered hierarchical transformer network (HTN) for sleep disorder diagnosis. (**A**) t-SNE visualization of features extracted by HTN with and without SleepGPT pretraining on CAP dataset. (**B**) Visualization of global transformer encoder attention weights on CAP, highlighting healthy and representative subjects with insomnia, narcolepsy, NFLE, PLMs, and RBD. Gray-shaded areas indicate segments with abnormal sleep patterns. (**C**) t-SNE visualization of features extracted by HTN with and without SleepGPT pretraining on MNC dataset. (**D**) Visualization of global transformer encoder attention weights on MNC, highlighting representative subjects (healthy, T1N, other hypersomnia); gray-shaded areas indicate segments with discriminative sleep patterns.