

Toward Personalized Sleep Guidance from Wearable Data Using Language Models

Yusheng Tan^{1,*}, Running Zhao^{1,*}, Sofia Angel¹, Ninghui Hao¹, Ash Arian¹,
Nikita N. Dulin¹, Jay Lin¹, Ou Zhu¹, Faiza Shaik¹, Xinxing Yang¹,
Bonnie W. Leung², Katie Roster³, Arlene Ruiz de Luzuriaga¹, Kenneth Lee¹,
Alejandra Lastra¹, Habibul Ahsan¹, Guihong C. Wan^{1,†}

¹The University of Chicago

²Baylor College of Medicine

³Georgetown University School of Medicine

*Equal contribution

†Corresponding author: gwan@uchicago.edu

Abstract

Sleep monitoring using wearable data has shown promise for personal health, yet large language model (LLM)-based summarization and question answering remain insufficient for personalized sleep guidance. Training specialized models, however, often requires costly expert annotation. Moreover, privacy and accessibility concerns motivate lightweight, local deployment for end users. We present a two-stage framework to address these challenges. Specifically, in Stage 1, a multi-agent LLM pipeline reasons structured sleep guidance from unannotated wearable records, enabling scalable dataset construction. Stage 2 distills guidance reasoning trajectories into small language models (SLMs) through supervised fine-tuning and integrates a training-free Best-of- N selection strategy to enhance inference. Experimental results demonstrate our method outperforms commercial general and medical LLMs and open-source models. Human evaluation further supports the quality of the generated guidance and the feasibility of personalized sleep guidance with SLMs.

Keywords: wearable health, personalized sleep guidance, evidence-grounded generation, language models

1. Introduction

Sleep is closely related to overall health and well-being. Persistent sleep problems are associated with a broad range of adverse health outcomes

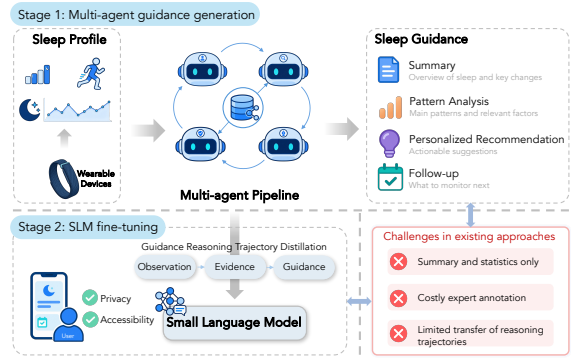


Figure 1: Overview of our framework for personalized sleep guidance.

(Buysse, 2014; Chung et al., 2021). With advances in wearable technology, consumer wearable devices have become increasingly integrated into everyday health monitoring. Recent nationally representative survey data show that wearable use for health tracking among U.S. adults increased from 30.2% in 2020 to 41.1% in 2024 (Pedroso et al., 2026). Wearable devices commonly combine motion sensing, such as accelerometry, with optical measurements, such as photoplethysmography, to estimate sleep-related measures during routine use (Imtiaz, 2021). The growing adoption of these devices makes personal sleep data increasingly accessible, creating an opportunity to provide personalized sleep guidance from routinely collected wearable data.

Recent studies have begun to use large language models (LLMs) to provide personalized

health support from wearable data. PhysioLLM (Fang et al., 2024), PHIA (Merrill et al., 2026), and HealthGuru (Wang et al., 2025) support data analysis, question answering, and conversational interaction over personal wearable data, but primarily focus on basic sleep data analysis and summary. WEQA extends the user-initiated query answering with query-adaptive agentic reasoning (Zhang et al., 2026). PH-LLM moves closer to personalized guidance by directly generating insights and recommendations for sleep and fitness, but it relies on domain-expert-authored case studies (Khasentino et al., 2025).

These limitations, summarized in Figure 1, expose two challenges for personalized sleep guidance. First, useful guidance should go beyond query-driven data analyses or summaries to identify salient patterns across multiple nights and connect them with relevant knowledge to generate actionable, evidence-grounded recommendations. Second, expert annotation is resource-intensive and difficult to scale for personalized sleep guidance. Although synthetic and LLM-generated guidance can reduce this burden, ensuring that automatically constructed guidance is consistent with the original wearable records and supported by evidence remains challenging.

Practical deployment introduces a further challenge. Personalized sleep guidance involves sensitive personal wearable data, making local or on-device inference attractive to reduce reliance on external services (Aminifar et al., 2024). Recent work has begun to distill personalized sleep capabilities into small language models (SLMs) for lightweight and edge deployment (Zheng et al., 2025). However, these approaches focus primarily on transferring basic sleep analysis and summary, whereas explicitly modeling intermediate reasoning processes into SLMs, including characterization of multi-night sleep patterns, selection of relevant evidence, and formulation of recommendations, remains underexplored.

Our framework addresses these challenges in two stages, as illustrated in Figure 1. In Stage 1, we propose a multi-agent LLM pipeline to reason structured sleep guidance from unannotated longitudinal wearable records. Beyond basic sleep analysis and summary, specialized agents characterize pattern observations, select supporting evidence, formulate personalized guidance, and verify each case through deterministic checks and

independent reverse review. This process reduces reliance on case-by-case expert annotation while enabling scalable dataset construction. In Stage 2, we distill the guidance reasoning trajectories into SLMs by fine-tuning on observations, evidence, and guidance generated from Stage 1. To reduce error propagation from an observation, we introduce a training-free Best-of- N strategy that selects a high-quality sampled observation before evidence retrieval. This two-stage design decouples large-model construction from small-model deployment, enabling evidence-grounded guidance without rerunning the multi-agent LLM pipeline during deployment, thereby enhancing privacy and accessibility.

Experimental results illustrated the superior performance of our method across Qwen 3.5 models with 0.8B, 2B, and 4B parameters. Notably, our 0.8B model outperforms the 4B baseline and commercial general and medical LLMs, including GPT-5.6-Sol, GPT Health, and OpenEvidence, across all metrics. We further conducted stage-specific human evaluations of the sleep guidance structure (Stage 1) and generated sleep guidance (Stage 2). Expert assessment supported the Stage 1 framework, with all rubric items receiving an I-CVI of 1.00. Evaluation of our 0.8B model’s outputs by MD students yielded average ratings of 4.99/5 for safety, 4.74/5 for readability, and 4.06/5 for overall quality. These results support both the appropriateness of the proposed framework and the capability of SLMs for personalized sleep guidance.

Our main contributions are:

- We introduce a multi-agent LLM pipeline to derive structured sleep guidance from longitudinal wearable records, reducing reliance on costly expert annotation and enabling scalable dataset construction.
- We distill personalized guidance capability into SLMs through supervised fine-tuning on the reasoning trajectory and a training-free Best-of- N strategy at inference.
- We conduct comprehensive experiments across model scales, combining quantitative comparisons and human evaluations to demonstrate the superiority of our framework over baselines and the feasibility of SLMs for personalized sleep guidance.

2. Related Work

2.1. Large Language Models for Wearable Sleep Data

Language models have increasingly been used to interpret personal wearable data. PhysioLLM (Fang et al., 2024), PHIA (Merrill et al., 2026), and HealthGuru (Wang et al., 2025) support a range of health-related tasks, including statistical analysis, tool-assisted reasoning, and question answering. WEQA further introduces query-adaptive agentic reasoning for wearable health question answering (Zhang et al., 2026). Moving beyond query-driven analysis, PH-LLM generates personalized insights and recommendations for sleep and fitness (Khasentino et al., 2025). These studies demonstrate a progression toward personalized guidance, while explicit interpretation of multi-night patterns and selection of supporting evidence for evidence-grounded decision-making remain less explored.

2.2. Scalable Training Data Construction

Personalized sleep guidance also raises the question of how high-quality training data can be constructed at scale. PH-LLM relies on domain-expert-authored case studies (Khasentino et al., 2025), whereas SleepCoT and subsequent work use synthetic or LLM-guided data construction to reduce manual effort (Zheng et al., 2025). More broadly, Self-Instruct (Wang et al., 2023) bootstraps instruction-response data from model generation, while multiple agentic workflows (Mitra et al., 2024; Chen et al., 2024; Zhao et al., 2025) are proposed to generate large-scale synthetic data. However, personalized wearable guidance additionally requires generated training examples to remain consistent with individual records and traceable to the evidence supporting the guidance.

2.3. Fine-Tuning with Intermediate Reasoning and Evidence

Final-answer-only supervised fine-tuning learns from input-answer pairs without explicit supervision of intermediate reasoning, whereas rationale-based distillation can embed the reasoning process into smaller models (Hsieh et al., 2023). Retrieval-augmented generation grounds model

outputs in external knowledge (Lewis et al., 2020), and generative retrieval methods can select supporting evidence by generating identifiers before response generation (Sun et al., 2023). These directions motivate combining intermediate decision-making with external evidence. For personalized sleep guidance, however, longitudinal pattern recognition and relevant evidence selection have rarely been treated together as explicit, verifiable intermediate decisions before generating the final guidance.

3. Methods

We obtained Fitbit-derived sleep and physical activity records from 1,500 adults aged 18 years or older through the Controlled Tier of the All of Us Researcher Workbench using Curated Data Repository C2025Q4R6 (Patten et al., 2026; All of Us Research Program, 2026). For each participant, we constructed one case containing age, sex, and 14–60 consecutive days of complete daily sleep and activity measurements, yielding 36,657 daily records in total. The 14-day minimum provides two non-overlapping seven-night windows for comparing recent and preceding sleep patterns, while allowing sleep patterns to be characterized over multiple nights (Schutte-Rodin et al., 2008; Krahn et al., 2021). Multi-night monitoring is particularly important for characterizing sleep variability beyond mean sleep measures (Bei et al., 2016; Lau et al., 2022). We cap each case at 60 days (approximately 8.5 weeks), within the 6–10-week range reported to reliably characterize intraindividual sleep variability, while bounding sequence length (Leota et al., 2026). We randomly partitioned the dataset at the participant level into 1,200 training cases and 300 held-out test cases.

For case i , let X_i denote the raw input data, consisting of age, sex, and T_i consecutive days of wearable measurements, where $T_i \in [14, 60]$. We construct a profile packet P_i , organizing daily measurements and deriving basic statistics for subsequent analyses (Appendix A for details).

Each case is associated with an observation O_i , which captures salient multi-night sleep patterns, and an ordered sequence of evidence identifiers $E_i = (e_{i1}, \dots, e_{iK_i})$, where K_i is the number of evidence identifiers for case i . The selected identifiers are used to retrieve the evidence cards C_i ,

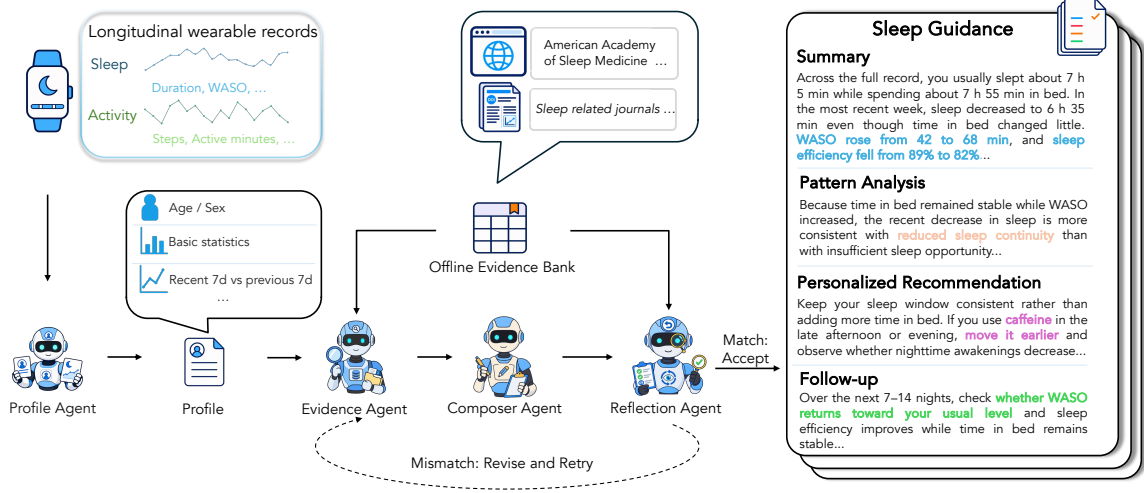


Figure 2: Stage 1 multi-agent pipeline for deriving structured sleep guidance from wearable records. Only examples passing deterministic checks and independent reverse review are retained.

which provide supporting evidence for generating a four-section guidance R_i (see Section 3.1).

Each evidence identifier refers to one of 22 evidence cards in the offline Evidence Bank $\mathcal{B}$. These cards were derived from 46 sources, including peer-reviewed sleep literature and authoritative health organizations and websites (see Appendix A). Given E_i , the corresponding evidence cards are retrieved deterministically: $C_i = \mathcal{B}[E_i]$.

Stage 1 constructs the evidence-grounded dataset of 1,500 cases:

$$\mathcal{D}^* = \{(P_i, O_i^*, E_i^*, C_i^*, R_i^*)\}_{i=1}^{1,500}. \quad (1)$$

The ordered $P \rightarrow O \rightarrow E \rightarrow C \rightarrow R$ process defines the guidance reasoning trajectory. In Stage 2, we distill this trajectory into SLMs.

3.1. Multi-Agent for Structured Sleep Guidance Generation

Stage 1 proposes a multi-agent LLM pipeline that transforms unannotated longitudinal wearable records into four-section sleep guidance R_i , reducing reliance on case-by-case expert annotation and enabling scalable dataset construction. Each guidance R_i follows the same four-section structure, including *Summary*, *Pattern Analysis*, *Personalized Recommendation*, and *Follow-up*.

Summary provides a concise, standalone description of the wearable records of a participant, including the general sleep patterns, the most relevant longitudinal change, and dated fluctuations when they clarify the primary sleep pattern. The summary describes multiple dimensions of sleep health without introducing interpretation or advice (Buysse, 2014; Chung et al., 2021).

Pattern Analysis identifies the primary repeated pattern and the relevant case-specific factors, then interprets the pattern using selected evidence. It focuses on explaining the primary pattern without introducing recommendations, unsupported causes, or diagnostic claims.

Personalized Recommendation translates the pattern analysis into feasible actions. Consistent with the *Advise* and *Assist* components of the 5As Behavioral Counseling Framework (Whitlock et al., 2002), recommendations are tied to observed facts; factors not contained in P_i may appear only as explicit conditional possibilities.

Follow-up specifies how the primary recommendation should be reassessed over a comparable multi-night window using the same sleep dimension. This design emphasizes repeated measurements rather than interpretation of a single subsequent night (Barber and Resnick, 2023).

Together, the four sections adapt the clinical reasoning cycle of information collection,

interpretation, action, and outcome evaluation (Levett-Jones et al., 2010) to evidence-grounded sleep guidance.

The construction pipeline uses LLM agents with specialized roles, as shown in Figure 2. The Profile Agent analyzes the raw wearable records to construct the profile packet P_i , generating and executing code to compute quantitative summaries and comparisons. In cycle $k \in \{1, \dots, 5\}$, the Evidence Agent produces an observation $O_i^{(k)}$ capturing salient multi-night sleep patterns and selects the corresponding evidence identifiers $E_i^{(k)}$. The selected evidence cards are retrieved verbatim as: $C_i^{(k)} = \mathcal{B}[E_i^{(k)}]$. The Composer Agent uses P_i and $C_i^{(k)}$ to generate the four-section guidance $R_i^{(k)}$ and records numeric anchors $A_i^{(k)}$. Deterministic checks verify consistency of $A_i^{(k)}$ with the wearable records and selected evidence.

Only outputs that pass the deterministic checks proceed to independent reverse review. The Reflection Agent independently recovers the evidence required by $R_i^{(k)}$ using P_i and the full Evidence Bank $\mathcal{B}$, without access to the evidence selected by the Evidence Agent. Let $\Gamma_i^{(k)} \in \{0, 1\}$ denote the verification outcome, with $\Gamma_i^{(k)} = 1$ only if the recovered evidence identifiers exactly match those selected by the Evidence Agent.

If $\Gamma_i^{(k)} = 1$, the example is retained as $(P_i, O_i^*, E_i^*, C_i^*, R_i^*)$; otherwise, the reverse-review result is used to revise the example in the next cycle. The process continues until verification succeeds or five cycles are completed. This verification checks consistency with the wearable records and traceability to supporting evidence. Appendices A and E provide additional details.

3.2. Guidance Reasoning Trajectory Distillation

Running the Stage 1 pipeline for a new case would require multiple LLM calls and repeated verification. To support lightweight local deployment, we distill the guidance reasoning trajectory into SLMs through SFT. We use Qwen3.5 models with 0.8, 2, and 4 billion parameters and apply the same training procedure across all models.

Each training case follows the reasoning trajectory in a single auto-regressive sequence:

$$[P_i; O_i^*; E_i^*; C_i^*; R_i^*], \text{ where } C_i^* = \mathcal{B}[E_i^*]. \quad (2)$$

The observation O_i^* captures salient multi-night sleep patterns before the evidence identifiers E_i^* are selected, and the retrieved evidence cards C_i^* condition the generation of the final guidance R_i^* . Unlike Guidance-only SFT, which supervises only the final guidance R_i^* given P_i and thus learns a direct input-to-answer mapping, our formulation supervises the reasoning trajectory. Let $Z_i^* = [O_i^*; E_i^*; C_i^*; R_i^*]$ denotes the supervised completion given P_i . We train the model using the auto-regressive cross-entropy objective:

$$\mathcal{L}(\theta) = -\frac{1}{N_{\text{tok}}} \sum_{i,t} \log p_{\theta}(Z_{i,t}^* \mid P_i, Z_{i,<t}^*), \quad (3)$$

where N_{tok} is the total number of target tokens across training cases, $Z_{i,t}^*$ is the t^{th} token, and $Z_{i,<t}^*$ denotes its preceding tokens. During training, C_i^* remains in the target sequence so that the model learns the complete reasoning trajectory. At inference, however, the model generates the observation and evidence identifiers, while the corresponding evidence cards are retrieved verbatim from the Evidence Bank (Figure 3). This separation allows the model to learn which evidence is relevant while preserving the exact retrieved evidence content, reducing the risk of unsupported evidence in the guidance generation.

Inference reuses the same model parameters $F_{\theta}(\cdot)$ in two passes. In the first pass, the model generates $\hat{O}_i$ and $\hat{E}_i$ from P_i and stops at $\langle \text{EVIDENCE} \rangle$. Valid identifiers are resolved by: $\hat{C}_i = \mathcal{B}[\hat{E}_i]$. In the second pass, the same model generates the final four-section guidance using the retrieved evidence as part of the input:

$$\begin{aligned} \text{Pass 1: } (\hat{O}_i, \hat{E}_i) &= F_{\theta}(P_i), \hat{C}_i = \mathcal{B}[\hat{E}_i], \\ \text{Pass 2: } \hat{R}_i &= F_{\theta}([P_i; \hat{O}_i; \hat{E}_i; \hat{C}_i]). \end{aligned} \quad (4)$$

This two-pass procedure transfers the guidance trajectory to SLMs while retrieving evidence deterministically, without rerunning the Stage 1 pipeline at deployment. Decoding and handling invalid-identifiers are detailed in Appendix A.

3.3. Best-of- N Observation Selection

Because $\hat{O}_i$ is the first generated component of the reasoning trajectory, errors in $\hat{O}_i$ can prop-

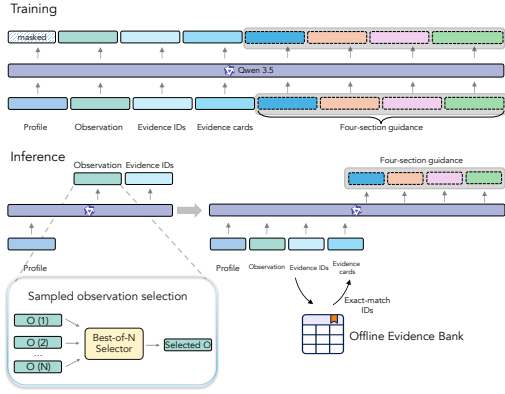


Figure 3: Stage 2 training and two-pass inference with Best-of- N selection and relevant evidence retrieval.

agate to evidence selection and, subsequently, to the final guidance. We therefore introduce a training-free Best-of- N strategy that samples multiple candidate observations and selects the best one before evidence-identifier generation.

For case i , we sample an observation $N > 1$ times at temperature of 0.8, stopping each generation at `</OBSERVATION>`. Each candidate is evaluated using two complementary signals: *case-specific axis relevance* and *verdict-clause consensus*. The former depends on the individual’s longitudinal profile, whereas the latter can be compared directly across candidates because verdict clauses follow fixed canonical forms.

Case-specific axis relevance. For candidate n , let $\mathcal{M}_n$ denote its metric-scope axis set, where each axis encodes a sleep metric and its temporal or comparison scope, including longitudinal changes. KNN over profile features retrieves the K nearest training cases, whose axes are aggregated by similarity-weighted voting to form the expected axis set $\widehat{\mathcal{M}}_i$. The K was selected on training data for candidate discrimination.

Verdict-clause consensus. Let $\mathcal{V}_n$ denote the canonical verdict-clause set of candidate n . Each clause encodes a predefined comparison between a case measurement and a reference value specified in the Evidence Bank and maps to the corresponding evidence cards. Agreement across candidates therefore reflects consensus on which reference evidence applies to the case.

Candidate n is scored as:

$$s_n = \frac{1}{2} F_1(\mathcal{M}_n, \widehat{\mathcal{M}}_i) + \frac{1}{2} \frac{1}{N-1} \sum_{n' \neq n} F_1(\mathcal{V}_n, \mathcal{V}_{n'}), \quad (5)$$

where $F_1(\cdot)$ is cosine similarity. The first term measures agreement with the KNN-derived axis expectation, and the second measures verdict-clause consensus across the N candidates.

The candidate with the highest s_n is selected as $\widehat{O}_i$. Evidence identifier generation, deterministic Evidence Bank lookup, and guidance generation then proceed as in Section 3.2. The selector requires no training; see Appendix A for details.

4. Experiments

4.1. Setup

All experiments use the participant-level split described in Section 3, with 1,200 training cases and 300 held-out test cases. Our method uses Best-of- N observation selection with $N = 16$.

Baselines. We compare against both commercial general and medical LLMs, as well as open-source models. For the former, we evaluate GPT-5.6-Sol, GPT Health, and OpenEvidence (Osler). For open-source models, *Guidance-only SFT* uses the same Qwen3.5 backbone with 0.8B, 2B, and 4B parameters, training cases, and optimization schedule as our method, but supervises only input-guidance pairs (P_i, R_i^*) without reasoning trajectories.

Implementation. Stage 1 uses GPT-5.6-Sol with medium reasoning for the Profile, Evidence, Composer, and Reflection Agents, with a fresh context for each role invocation to preserve role isolation. Stage 2 fine-tunes Qwen3.5 models at 0.8B, 2B, and 4B using the same training configuration across model sizes. Fine-tuned systems are trained with three random seeds and reported as mean $\pm$ sd, while prompting baselines are single runs. For the paired comparison between greedy and Best-of- N observation decoding in Section 4.4, we additionally report 95% two-level bootstrap confidence intervals over seeds and test cases. Training hyperparameters and bootstrap details are provided in Appendix A.

Metrics. The main comparison reports five complementary metrics. *Evidence F1* compares

Table 1: **Comparison of ours and baselines.** All rows use the same 300 held-out test cases. Multi-seed results of trainable models report mean $\pm$ sd. * marks frozen prompting models; $\bullet$ marks fine-tuned models. LLM-as-a-Judge scores use a 1–5 Likert scale.

Method	Evidence F1	Value F1	ROUGE-L	BERTScore	LLM-as-a-Judge
* GPT-5.6-Sol	0.288	0.404	0.251	0.312	3.48
* GPT Health	0.423	0.398	0.276	0.331	3.89
* OpenEvidence	0.424	0.261	0.213	0.264	3.79
$\bullet$ Qwen3.5-4B, Guidance-only SFT	0.629 $\pm$ 0.016	0.584 $\pm$ 0.008	0.474 $\pm$ 0.005	0.535 $\pm$ 0.005	3.97
$\bullet$ Ours (Qwen3.5-0.8B)	0.680 $\pm$ 0.010	0.597 $\pm$ 0.005	0.502 $\pm$ 0.003	0.561 $\pm$ 0.002	4.08
$\bullet$ Ours (Qwen3.5-2B)	0.707 $\pm$ 0.007	0.609 $\pm$ 0.004	0.510 $\pm$ 0.004	0.567 $\pm$ 0.002	4.14
$\bullet$ Ours (Qwen3.5-4B)	0.711 $\pm$ 0.006	0.612 $\pm$ 0.003	0.506 $\pm$ 0.001	0.564 $\pm$ 0.001	4.11

the evidence recoverable from a generated guidance with the verified Stage 1 evidence identifiers E_i^* . *Value F1* measures recovery of the certified numeric anchors A_i^* . ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020) measure lexical and semantic similarity to R_i^* , respectively, but are not treated as evidence-grounding measures (Maynez et al., 2020; Fabbri et al., 2021). *LLM-as-a-Judge* based on Claude Opus 5 evaluates generation quality using the Generation Quality Evaluation Rubric. All five metrics are evaluated on the full set of 300 held-out test cases. Moreover, human rubric-based assessment was conducted with 5 MD students and the detailed rubric. Appendix C.2 provides rubric details.

4.2. Quantitative Evaluation

As shown in Table 1, our method substantially outperforms commercial general and medical LLMs in personalized sleep guidance across Qwen 3.5 with 0.8B, 2B, and 4B parameters. Using a 0.8B-parameter backbone, our model achieves an Evidence F1 of 0.680, compared with 0.288 for GPT-5.6-Sol and 0.423 for GPT health. It also substantially outperforms OpenEvidence, which achieves an Evidence F1 of 0.424. In terms of LLM-as-a-Judge, our 0.8B model achieves a score of 4.08, compared with 3.48 for GPT-5.6-Sol, 3.89 for GPT Health, and 3.79 for OpenEvidence. Although GPT Health and OpenEvidence are designed for health-related applications, their evidence recovery remains below our method. These commercial general and medical LLMs generate guidance without the consistency checks and independent verification used in our

framework to ensure alignment with the original wearable records and supporting evidence. These results suggest the effectiveness of our method and demonstrate the potential for SLMs to outperform general-purpose LLMs by constructing the evidence-grounded guidance and distilling the guidance reasoning trajectories.

Our method also outperforms Guidance-only SFT, which is trained on the same data but learns only the final guidance without guidance reasoning trajectories. Notably, our 0.8B model outperforms the 4B Guidance-only SFT baseline across all five metrics, improving Evidence F1 from 0.629 to 0.680 and LLM-as-a-Judge from 3.66 to 3.78, with consistent gains in Value F1, ROUGE-L, and BERTScore. These results suggest that our framework makes effective use of the guidance information constructed during data generation, enabling guidance reasoning capability to be transferred to smaller models.

4.3. Human Evaluation

Two experts assessed the four-section guidance structure after independently reviewing ten examples in Stage 1. Using a 12-item five-point rubric, all 24 ratings were 4 or 5, with a mean score of 4.29. Every item achieved an I-CVI of 1.00, and the S-CVI/Ave was also 1.00. These findings support the appropriateness of the proposed four-section structure for personalized sleep guidance. Full rubric details are provided in Appendix C.1.

Moreover, five experienced MD students further evaluated 90 guidance examples generated by Qwen3.5-0.8B, with each guidance receiv-

Table 2: **Results of ablation studies.** All methods use Qwen3.5 with 0.8B parameters. Result values are mean $\pm$ sd across three training seeds.

Method	Evidence F1	Value F1	ROUGE-L	BERTScore
w/o Guidance Trajectory Distillation	0.536 $\pm$ 0.014	0.556 $\pm$ 0.006	0.447 $\pm$ 0.007	0.511 $\pm$ 0.006
w/o Best-of-N Selection	0.613 $\pm$ 0.019	0.598 $\pm$ 0.005	0.480 $\pm$ 0.009	0.541 $\pm$ 0.008
Ours	0.680 $\pm$ 0.010	0.597 $\pm$ 0.005	0.502 $\pm$ 0.003	0.561 $\pm$ 0.002

ing three independent ratings across 16 rubric items. The generated guidance received consistently high ratings across the evaluated dimensions (Figure 4a), including mean scores of 4.99 for safety, 4.74 for readability, and 4.06 for overall quality. Inter-rater agreement was also strong, with an ordinal Gwet’s AC2 of 0.922 (95% CI: 0.912–0.932). These results support both the quality of the generated guidance and the consistency of human assessment. Full rubric results and pairwise agreement analyses are provided in Appendix C.2. More sleep guidance examples are shown in Appendix D.2.

4.4. Ablation Study

We conduct an ablation study on Qwen3.5-0.8B to examine the contributions of the main components of our method, with corresponding results for the 2B and 4B backbones provided in Appendix B. The results are shown in Table 2.

w/o Guidance Trajectory Distillation only uses the final guidance without reasoning trajectories as the training target. Compared with this setting, our method achieves better performance across all metrics, increasing Evidence F1 from 0.536 to 0.680, Value F1 from 0.556 to 0.597, ROUGE-L from 0.447 to 0.502, and BERTScore from 0.511 to 0.561. The *w/o Best-of-N selection* setting retains the guidance trajectory as part of the supervised target and also outperforms the *w/o Guidance Trajectory Distillation* setting across all metrics. These results demonstrate the benefit of incorporating the guidance trajectory into the supervised training target.

We study the effect of Best-of- N observation selection to investigate how much selecting a suitable observation influences the generation performance. The *w/o Best-of-N selection* setting uses greedy observation decoding at inference, while our method applies Best-of- N selection.

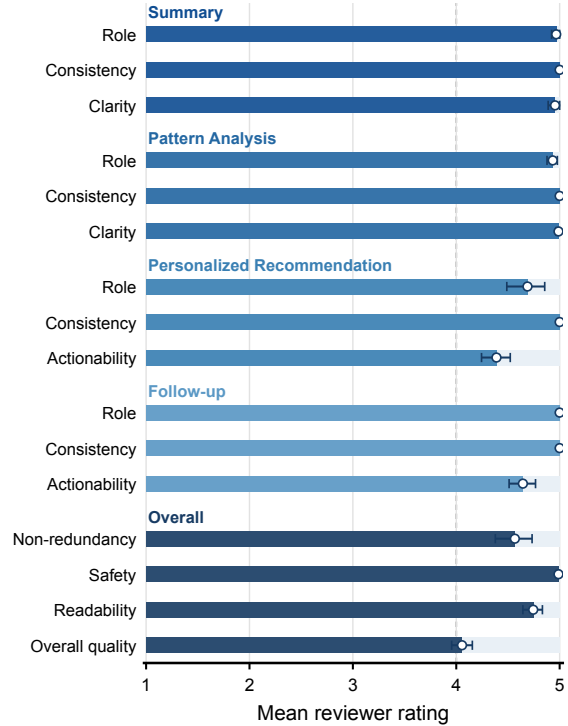


Figure 4: **Human evaluation of generated guidance.** Mean item-wise ratings with case-bootstrap 95% confidence intervals across 90 evaluated cases.

Compared with greedy decoding, Best-of- N selection increases Evidence F1 from 0.613 to 0.680, ROUGE-L from 0.480 to 0.502, and BERTScore from 0.541 to 0.561, while Value F1 remains comparable (0.598 vs. 0.597). These results show that the quality of the selected observation has a clear impact on performance, particularly on evidence alignment and guidance quality. We use $N = 16$ in the main experiments. The effect of candidate-pool size is reported in Table 4 of Appendix B.

5. Conclusion

We presented a two-stage framework for personalized sleep guidance from longitudinal wearable data. A multi-agent pipeline constructs an evidence-grounded dataset, while guidance trajectory distillation transfers the reasoning process into SLMs. Training-free Best-of- N observation selection further improves performance. Our method with the 0.8B model size outperformed commercial LLMs (including GPT-5.6-Sol and OpenEvidence) and the 4B baseline. Rubric-based human evaluation further supported both the proposed four-section guidance structure and the generated guidance quality.

Our study is limited to Fitbit-derived records from the All of Us Research Program with complete 14–60 days measurements, which may limit generalizability. The Evidence Bank supports behavioral guidance rather than diagnosis or treatment. Future work should integrate clinical information with wearable data and expand the evidence bank to support more comprehensive personalized sleep guidance.

References

- All of Us Research Program. Our largest genomic dataset: Curated data repository version 9. All of Us Research Hub, 2026. URL <https://support.researchallofus.org/hc/en-us/articles/50653909888788>. Accessed August 25, 2026.
- Amin Aminifar, Matin Shokri, and Amir Amini-far. Privacy-preserving edge federated learning for intelligent mobile-health systems. *Future Generation Computer Systems*, 161:625–637, 2024.
- Jessica Barber and Sandra G Resnick. Collect, share, act: A transtheoretical clinical model for doing measurement-based care in mental health treatment. *Psychological services*, 20 (S2):150, 2023.
- Bei Bei, Joshua F Wiley, John Trinder, and Rachel Manber. Beyond the mean: a systematic review on the correlates of daily intraindividual variability of sleep/wake patterns. *Sleep medicine reviews*, 28:108–124, 2016.
- Daniel J Buysse. Sleep health: can we define it? does it matter? *Sleep*, 37(1):9–17, 2014.
- Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4video: Improving video understanding and generation with better captions. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- Joon Chung, Matthew Goodman, Tianyi Huang, Suzanne Bertisch, and Susan Redline. Multidimensional sleep health in a diverse, aging adult cohort: concepts, advances, and implications for research and intervention. *Sleep Health*, 7 (6):699–707, 2021.
- Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9: 391–409, 2021.

- Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. Physiollm: Supporting personalized health insights with wearables and large language models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pages 8003–8017, 2023.
- Syed Anas Intiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5):1562, 2021.
- Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 31(10):3394–3403, 2025.
- Lois E Krahn, Donna L Arand, Alon Y Avidan, David G Davila, William A DeBassio, Chad M Ruoff, and Christopher G Harrod. Recommended protocols for the multiple sleep latency test and maintenance of wakefulness test in adults: guidance from the american academy of sleep medicine. *Journal of Clinical Sleep Medicine*, 17(12):2489–2498, 2021.
- TeYang Lau, Ju Lynn Ong, Ben KL Ng, Lit Fai Chan, Daphne Koek, Chuen Seng Tan, Falk Müller-Riemenschneider, Karen Cheong, Stijn AA Massar, and Michael WL Chee. Minimum number of nights for reliable estimation of habitual sleep using a consumer sleep tracker. *Sleep Advances*, 3(1):zpac026, 2022.
- Josh Leota, Brett A Messman, Flora Le, Summer Jasinski, Emily R Capodilupo, Elise R Facer-Childs, and Joshua F Wiley. How many nights are needed? the short-term stability of intraindividual variability in sleep parameters derived from accelerometry data in a cohort of normal sleepers. *Sleep*, 49(6):zsag040, 2026.
- Tracy Levett-Jones, Kerry Hoffman, Jennifer Dempsey, Sarah Yeun-Sim Jeong, Danielle Noble, Carol Anne Norton, Janiece Roche, and Noelene Hickey. The ‘five rights’ of clinical reasoning: An educational model to enhance nursing students’ ability to identify and manage clinically ‘at risk’ patients. *Nurse education today*, 30(6):515–520, 2010.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentaoh Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919, 2020.
- Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17(1):1143, 2026.
- Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Cotas, Yadong Lu, Wei-ge Chen, Olga Vrousos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. AgentInstruct: Toward generative teaching with agentic flows, 2024. URL <https://arxiv.org/abs/2407.03502>.
- Theresa Patten, Edward A Preble, Hiral Master, Jennifer Adjemian, Andrea Ramirez, James McClain, and Amy Rose Price. The all of us research program’s wearables dataset. *Nature Medicine*, pages 1–9, 2026.
- Aline F Pedroso, Lovedeep S Dhingra, Arya Aminorroaya, and Rohan Khera. Wearable de-

- vices and data sharing in the us. *JAMA Network Open*, 9(6):e2617733, 2026.
- Sharon Schutte-Rodin, Lauren Broch, Daniel Buysse, Cynthia Dorsey, and Michael Sateia. Clinical guideline for the evaluation and management of chronic insomnia in adults. *Journal of clinical sleep medicine*, 4(5):487–504, 2008.
- Weiwei Sun, Pengjie Ren, and Zhaochun Ren. Generative knowledge selection for knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2077–2088, 2023.
- Xingbo Wang, Janessa Griffith, Daniel A Adler, Joey Castillo, Tanzeem Choudhury, and Fei Wang. Exploring personalized health support through data-driven, theory-guided llms: a case study in sleep health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2025.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508, 2023.
- Evelyn P Whitlock, C Tracy Orleans, Nola Pender, and Janet Allan. Evaluating primary care behavioral counseling interventions: an evidence-based approach. *American journal of preventive medicine*, 22(4):267–284, 2002.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SkeHuCVFDr>.
- Yuwei Zhang, Tong Xia, Bianca Emmerich, Yu Yvonne Wu, Dimitris Spathis, Xin Liu, Daniel McDuff, and Cecilia Mascolo. Weqa: Wearable health question answering with query-adaptive agentic reasoning. *arXiv preprint arXiv:2606.18147*, 2026.
- Running Zhao, Zhihan Jiang, Xinchun Zhang, Chirui Chang, Handi Chen, Weipeng Deng, Luyao Jin, Xiaojuan Qi, Xun Qian, and Edith C.H. Ngai. Noteit: A system converting instructional videos to interactable notes through multimodal video understanding. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, UIST ’25, New York, NY, USA, 2025.
- Huimin Zheng, Xingxing Ai, Xueyan Liu, Xiaofen Xing, and Xiangmin Xu. Profile-aided distillation framework for personalized sleep analysis with compact models using llm-guided synthetic data. *Frontiers in Physiology*, 16: 1678364, 2025.

Appendix A. Additional Details

Wearable variables and profile metrics.

Each daily record carries sleep duration, time in bed, sleep timing, sleep efficiency, device-estimated wake after sleep onset (WASO), sleep-stage minutes, and activity measures. The profile packet P_i preserves these measurements together with age, sex, dates, window boundaries, and derived basic statistics used to summarize the record.

Evidence Bank provenance and validation.

Each evidence identifier refers to one of 22 cards in the Evidence Bank. The 22 runtime cards distill 46 versioned source dossiers spanning professional guidance, consensus reports, evidence syntheses, original studies, government health guidance, and device documentation. Each dossier records a resolvable DOI or URL and a passage locator.

During Stage 1, the Evidence Agent can select only existing identifiers, retrieved card text is inserted verbatim, and deterministic checks reject unsupported conditions and numeric anchors. A failed evidence check restarts planning and composition; a failed case-fact check returns the draft to the Composer while retaining the validated evidence set.

Stage-2 decoding and selection constants.

The first decoding pass begins with `<OBSERVATION>` and stops at `</EVIDENCE>`. Identifiers are matched exactly against the bank; duplicates are removed and invalid identifiers count as errors. The decoder then inserts the retrieved cards with `<CARDS>`, appends `<SECTIONS>`, and resumes generation with the same model. For Best-of- N , packet similarity is cosine similarity over a standardized deterministic feature vector. The $K = 12$ nearest training cases vote with similarity-softmax weights at temperature 8.0, and an axis enters the expected set when its weighted vote share reaches 0.4.

Training hyperparameters. Stage 2 uses completion-only cross-entropy, AdamW, bf16 precision, a learning rate of 2×10^{-5} with cosine decay and 5% warmup, weight decay 0.1, gradient clipping at 1.0, an effective batch size of 16, five epochs, and a maximum sequence length of 32,768 tokens. The three model sizes share this configuration.

Appendix B. Additional Ablation Results

Table 3 reports the same ablation as Table 2 across all three Qwen3.5 model sizes. Within each size, the greedy and Best-of-16 variants use the same three checkpoints and differ only in observation generation and selection.

Best-of-16 improves Evidence F1 over greedy decoding at 0.8B and 2B, with confidence intervals that exclude zero. The improvement is smaller at 4B, where the confidence interval includes zero.

Effect of candidate-pool size. Table 4 fixes one Qwen3.5-0.8B checkpoint and varies only the number of candidate observations. Evidence F1 increases from 0.591 at $N = 2$ to 0.683 at $N = 16$ and 0.692 at $N = 32$, while the changes in the remaining metrics become smaller beyond $N=16$.

Appendix C. Rubric Materials

This appendix separates the administered rubric specifications from the aggregate findings in Section 4.3. For each stage, we report the rated statements, fully labeled guidance options, and item-level scores.

C.1. Stage-1 Framework Validation Rubric

Each expert read one uninterrupted batch of ten evidence-grounded examples before rating the framework once. The instrument contained eight section-level items, relevance and essentiality for each of Summary, Pattern Analysis, Personalized Recommendation, and Follow-up, followed by four framework-level items. Table 6 gives the administered five-point scales, and Table 5 reports both sets of ratings.

C.2. Stage-2 Generation Quality Evaluation Rubric

The Stage-2 instrument rated each generated four-section guidance on 16 items. Twelve items assessed individual sections, and four assessed the guidance as a whole. Table 7 maps the items to their rated statements; Table 8 gives the full score anchors. Table 9 reports the item-level human ratings.

Table 3: **Ablation across Qwen3.5 model sizes.** Values are mean $\pm$ sd over three seeds.

Model	Method	Evidence F1	Value F1	ROUGE-L	BERTScore
0.8B	w/o Guidance Trajectory Distillation	0.536 $\pm$ 0.014	0.556 $\pm$ 0.006	0.447 $\pm$ 0.007	0.511 $\pm$ 0.006
	w/o Best-of- N selection	0.613 $\pm$ 0.019	0.598 $\pm$ 0.005	0.480 $\pm$ 0.009	0.541 $\pm$ 0.008
	Ours	0.680 $\pm$ 0.010	0.597 $\pm$ 0.005	0.502 $\pm$ 0.003	0.561 $\pm$ 0.002
2B	w/o Guidance Trajectory Distillation	0.603 $\pm$ 0.013	0.565 $\pm$ 0.010	0.472 $\pm$ 0.009	0.533 $\pm$ 0.008
	w/o Best-of- N selection	0.653 $\pm$ 0.012	0.606 $\pm$ 0.005	0.494 $\pm$ 0.004	0.553 $\pm$ 0.005
	Ours	0.707 $\pm$ 0.007	0.609 $\pm$ 0.004	0.510 $\pm$ 0.004	0.567 $\pm$ 0.002
4B	w/o Guidance Trajectory Distillation	0.629 $\pm$ 0.016	0.584 $\pm$ 0.008	0.474 $\pm$ 0.005	0.535 $\pm$ 0.005
	w/o Best-of- N selection	0.693 $\pm$ 0.001	0.614 $\pm$ 0.002	0.502 $\pm$ 0.003	0.561 $\pm$ 0.002
	Ours	0.711 $\pm$ 0.006	0.612 $\pm$ 0.003	0.506 $\pm$ 0.001	0.564 $\pm$ 0.001

Table 4: **Effect of candidate-pool size on Qwen3.5-0.8B.** All results use the same checkpoint.

Decoding	Ev. F1	Val. F1	R-L	BERTScore
Greedy	0.613	0.598	0.480	0.541
Best-of-2	0.591	0.576	0.469	0.531
Best-of-4	0.629	0.587	0.483	0.542
Best-of-8	0.681	0.592	0.499	0.560
Best-of-16	0.680	0.597	0.502	0.561
Best-of-32	0.692	0.604	0.504	0.564

participant record, evidence identifiers, evidence cards, reference guidance, or system identity. Item scores were summarized by first taking the within-case median across three ratings and then averaging across cases. Reliability was calculated across all 1,440 case-item units; each case-level bootstrap sample retained all 16 items and their three ratings.

Table 5: **Item-level Stage-1 framework ratings.** I-CVI is the proportion of experts assigning a score of 4 or 5.

Item	Expert 1	Expert 2	I-CVI
Summary relevance	5	4	1.00
Summary essentiality	4	5	1.00
Pattern Analysis relevance	5	5	1.00
Pattern Analysis essentiality	4	5	1.00
Personalized Recommendation relevance	4	4	1.00
Personalized Recommendation essentiality	4	4	1.00
Follow-up relevance	4	5	1.00
Follow-up essentiality	4	4	1.00
Comprehensiveness	4	4	1.00
Ordering	4	4	1.00
Applicability	4	4	1.00
Overall appropriateness	4	5	1.00

Administration and aggregation. The 90 cases were divided into ten nine-case groups and assigned to the ten possible three-reviewer combinations drawn from five experienced MD students. This balanced incomplete block design yielded three independent reviews per case and 54 cases per reviewer. Reviewers judged only the displayed guidance, without access to the

Table 6: Stage-1 Framework Validation Rubric. Relevance and essentiality were rated separately for each guidance section; the remaining constructs were rated once for the framework as a whole.

Construct	Rated statement	Guidance options
Relevance	The content of this section is relevant to a sleep recommendation.	1: Not relevant; it does not belong in a sleep recommendation. 2: Slightly relevant; only a small part belongs. 3: Moderately relevant; about half belongs. 4: Relevant; most belongs. 5: Highly relevant; all of it belongs.
Essentiality	This section is essential; the recommendation would be incomplete without it.	1: Not necessary; it should be removed. 2: Useful, but the recommendation works without it. 3: Essential in some cases, but not all. 4: Essential in most cases. 5: Always essential; the recommendation would be incomplete without it.
Comprehensiveness	Together, the four sections cover the expected content.	1: Most important content is missing. 2: Several important elements are missing. 3: One important element is missing. 4: Only a minor element is missing. 5: Nothing is missing.
Ordering	The sequence of the four sections follows a logical order.	1: The order is illogical and would confuse the reader. 2: At least one section is clearly misplaced. 3: The order works, but another order would work equally well. 4: The order is logical, with small room for improvement. 5: The order is fully logical for the reader.
Applicability	The framework can be applied consistently across cases.	1: It does not fit most cases. 2: It fits some cases but breaks down in many others. 3: It fits about half the cases well. 4: It fits most cases, with occasional strain. 5: It fits every case shown without strain.
Overall appropriateness	The four-section framework is appropriate as a standard structure.	1: Not appropriate; redesign the framework. 2: Weak; an important component needs to change. 3: Acceptable, but another framework could work equally well. 4: Appropriate, with only minor changes needed. 5: Fully appropriate; keep it as is.

Table 7: Stage-2 rubric item map. Role and consistency were rated for every section; clarity was used for Summary and Pattern Analysis, and actionability for Personalized Recommendation and Follow-up.

Item	Scope	Construct	Rated statement
1	Summary	Role (A)	The section stays within its assigned role: summarize observed sleep data, without advice, causes, or clinical framing.
2	Summary	Consistency (B)	The section does not contradict itself or the other sections.
3	Summary	Clarity (C)	The section clearly presents all important sleep data needed.
4	Pattern Analysis	Role (A)	The section interprets patterns as possibilities, without presenting unmeasured factors as facts or making clinical claims.
5	Pattern Analysis	Consistency (B)	The section does not contradict itself or the other sections.
6	Pattern Analysis	Clarity (C)	The section clearly presents all important sleep data needed.
7	Personalized Recommendation	Role (A)	The section gives a few low-risk, reversible changes with reasons and a way to track them, or states that no change is needed; it gives no diagnosis, medication, supplement, or treatment content.
8	Personalized Recommendation	Consistency (B)	The section does not contradict itself or the other sections.
9	Personalized Recommendation	Actionability (D)	The reader knows what to do and how to track it.
10	Follow-up	Role (A)	The section states when and how to reassess and when to talk to a clinician, without diagnostic claims, urgency, or new content.
11	Follow-up	Consistency (B)	The section does not contradict itself or the other sections.
12	Follow-up	Actionability (D)	The reader knows what to do and how to track it.
13	Overall	Non-redundancy (E)	The four sections do not repeat each other.
14	Overall	Safety (F)	The output contains nothing out of scope and nothing that could cause harm if followed.
15	Overall	Readability (G)	The output is easy to read, with correct grammar, spelling, and formatting.
16	Overall	Overall quality (H)	Overall, this is a high-quality sleep recommendation for this case.

Table 8: **Five-point score anchors for the Stage-2 rubric.**

Scale	Guidance options
A: Role	1: The section is mostly outside its role. 2: It contains one important piece of out-of-role content. 3: It contains several minor pieces of out-of-role content. 4: It contains one minor piece of out-of-role content. 5: Everything in it belongs to its role.
B: Consistency	1: Contradictions run through the section. 2: It contains one important contradiction. 3: It contains several minor inconsistencies. 4: It contains one minor inconsistency. 5: It is fully consistent.
C: Clarity	1: None of the important sleep data is presented clearly. 2: Some is clear, but most important data is missing or unclear. 3: About half of the important sleep data is clear. 4: Most of the important sleep data is clear. 5: All important sleep data is clear.
D: Actionability	1: Nothing can be acted on. 2: An important part cannot be acted on. 3: Several parts are too vague to act on. 4: One part is slightly vague. 5: Every part is specific and can be tracked.
E: Non-redundancy	1: The sections largely repeat each other. 2: Two sections repeat each other in an important way. 3: Two sections partly repeat each other. 4: There is slight repetition between two sections. 5: There is no repetition.
F: Safety	1: Clearly harmful content or explicit clinical or treatment advice is present. 2: One important out-of-scope or potentially harmful item is present. 3: Several minor out-of-scope items are present. 4: One minor out-of-scope item is present, with no potential for harm. 5: Nothing is out of scope or harmful.
G: Readability	1: Very difficult to read. 2: Many errors affect readability. 3: Some errors affect readability. 4: A few errors are present, but the guidance remains easy to read. 5: Easy to read, with no errors.
H: Overall quality	1: Very poor; it needs to be rewritten. 2: Poor; it contains significant errors. 3: Fair; it could be improved. 4: Good; only slight improvements are possible. 5: Excellent; no changes are needed.

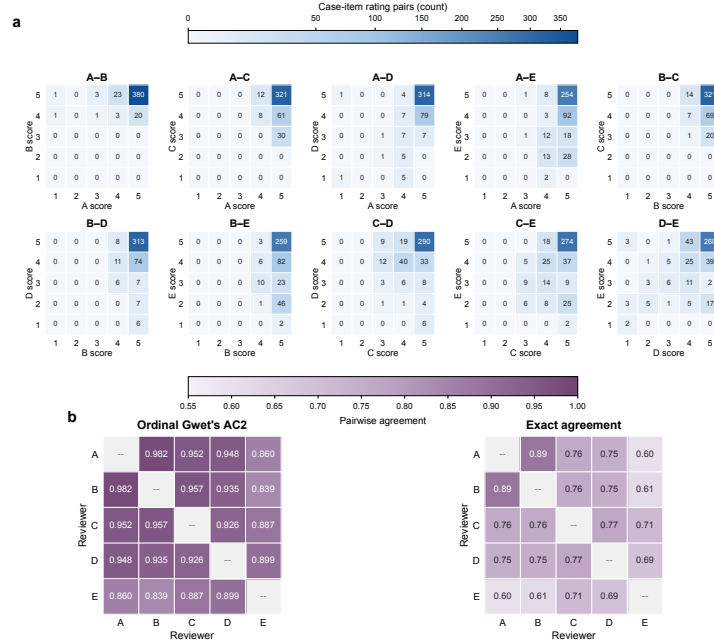


Figure 5: **Pairwise agreement among five experienced MD-student reviewers.** **a**, Pairwise 5×5 contingency matrices for all ten reviewer pairs. Columns correspond to the first named reviewer and rows to the second; cells report raw case-item counts. Each pair shared 27 cases, giving 432 case-item units. **b**, Full pairwise matrices of ordinal Gwet's AC2 and exact agreement. Diagonal cells are omitted because self-agreement is not an inter-rater comparison. Overall reliability estimates are reported in Table 10.

Table 9: **Item-level Stage-2 human rubric scores for the Qwen3.5-0.8B system.** Values are means of within-case median scores across 90 cases.

Item	Mean score
Summary role	4.97
Summary consistency	5.00
Summary clarity	4.96
Pattern Analysis role	4.93
Pattern Analysis consistency	5.00
Pattern Analysis clarity	4.99
Personalized Recommendation role	4.69
Personalized Recommendation consistency	5.00
Personalized Recommendation actionability	4.39
Follow-up role	5.00
Follow-up consistency	5.00
Follow-up actionability	4.64
Non-redundancy	4.57
Safety	4.99
Readability	4.74
Overall quality	4.06

Table 10: **Inter-rater agreement for the Stage-2 human evaluation.** Chance-corrected estimates include case-bootstrap 95% confidence intervals; observed agreement is reported descriptively.

Measure	Result
Ordinal Gwet’s AC2	0.922 95% CI: 0.912–0.932
Krippendorff’s ordinal α	0.235 95% CI: 0.186–0.284
Exact agreement	72.9%
Difference ≤ 1 point	92.4%

Appendix D. Qualitative Examples

This section presents one complete Stage 1 evidence-grounded guidance construction example and three representative Stage 2 input-output examples generated by the Qwen3.5-0.8B system.

D.1. Stage-1 Complete Evidence-Grounded Guidance Construction Example

Male, 61 years

Profile packet P (display-compacted). Thirty-one valid nights were recorded from 29 July to 28 August 2023, with 100% nightly coverage. Across the completed record, median sleep duration was 6 h 24 min. During the recent seven nights, median sleep duration was 6 h 8 min, compared with 6 h 24 min during the preceding seven nights. Recent median time in bed was 6 h 55 min. Median time awake after falling asleep rose from 48 min to 59 min, while median sleep efficiency decreased from 89.3% to 87.1%.

Observation O . Across the completed record, you usually slept about 6 h 24 min. Recently, you typically slept about 6 h 8 min, compared with about 6 h 24 min during the preceding period. You spent about 59 min awake after falling asleep recently, compared with about 48 min previously. Sleep efficiency, the percentage of time in bed spent asleep, was about 87.1% recently versus 89.3% previously. Recent time in bed was about 6 h 55 min.

Selected evidence identifiers E Across the completed record, you usually slept about 6 h 24 min. Recently, you typically slept about 6 h 8 min, compared with about 6 h 24 min during the preceding period. You spent about 59 min awake after falling asleep recently, compared with about 48 min previously. Sleep efficiency, the percentage of time in bed spent asleep, was about 87.1% recently versus 89.3% previously. Recent time in bed was about 6 h 55 min.

Evidence-card block C

Allow more time for sleep (allow-more-time-for-sleep; guidance). **Analysis.** When sleep is repeatedly short and the available time for sleep is also limited, making somewhat more time available can increase sleep duration, although the amount of change varies. **Recommendation.** When practical, make a little more time for sleep on the days when it is repeatedly short, for example by starting the sleep period a little earlier, without forcing a fixed bedtime. **Follow-up.** After making more time available, see whether sleep lasts longer across several nights. If it does, keep allowing the additional time. **Select when.** Sleep duration is repeatedly short recently and time in bed is also below 7 h in the same recent or weekday pattern. Do not infer limited time available merely because sleep is below 7 h. **Action.** Make more time for sleep. **Do not claim.** Do not promise a health, mood, or performance benefit; do not recommend an earlier sleep start when time in bed is already at least 7 h; and do not turn this into a fixed sleep schedule unless separate timing-consistency evidence is selected.

Adults aged 18–64 usually need at least 7 hours (adults-18-to-64-usually-need-at-least-7-hours; reference). **Analysis.** Adults generally need 7 h or more of sleep, although the amount an individual needs can vary. **Select when.** The person is aged 18–64 and the repeated recent sleep duration is clearly below 7 h in a way that changes the decision; interpret it alongside the earlier and whole-record pattern. **Reference.** Usual lower sleep-duration reference for adults aged 18–64: at least 420 min. **Do not claim.** Do not apply this card to adults aged 65 or older; do not use it for one short night or when the recommendation would be unchanged without the reference; and do not call 7 h an exact personal requirement.

Awake-time quality reference for adults aged 18–64 (waso-quality-reference-for-adults-18-to-64; reference). **Analysis.** Repeatedly spending longer awake after first falling asleep can indicate interrupted sleep. For adults, 20 min or less is commonly used as one general sign of good sleep quality, not as a diagnosis or an exact personal target. **Select when.** Use this card for an adult younger than 65 when time awake after falling asleep is repeatedly longer than 20 min across both the whole record and recent nights, and this helps explain an ongoing interruption pattern. **Reference.** Awake-time comparison for adults younger than 65: at most 20 min. **Do not claim.** Do not apply this card to adults aged 65 or older; do not diagnose insomnia or another sleep disorder; do not use one night or a small isolated increase as the trigger; and describe the measure in plain language as time awake after falling asleep.

Persistent sleep difficulty needs clinical context (persistent-sleep-difficulty-needs-clinical-context; boundary). **Analysis.** A repeated sleep pattern is more useful to discuss with a clinician when the person also has ongoing trouble sleeping, feels unusually sleepy during the day, or has trouble concentrating. The sleep record alone cannot establish a disorder. **Recommendation.** If you are also having ongoing sleep difficulty, unusual daytime sleepiness, or trouble concentrating, discuss the repeated pattern with a clinician. **Follow-up.** Before the discussion, note which of those problems you notice and how often they occur. **Select when.** A prominent repeated duration or continuity problem remains unresolved, no selected direct self-directed step addresses it, and clinical care remains explicitly conditional on unrecorded ongoing trouble sleeping, unusual daytime sleepiness, or trouble concentrating. **Action.** Discuss persistent sleep difficulty. **Only when.** If you are also having ongoing sleep difficulty, unusual daytime sleepiness, or trouble concentrating. **Do not claim.** Do not diagnose insomnia or infer daytime impairment; do not use professional care as a generic closing sentence; and do not recommend medication or a treatment protocol.

Four-section guidance

Summary. You usually slept about 6 h 24 min across the 31-night record, below 7 h. During the recent seven nights, you typically slept about 6 h 8 min, compared with about 6 h 24 min during the preceding seven nights. Recent time in bed was about 6 h 55 min. You spent more time awake after falling asleep recently—about 59 min, compared with about 48 min during the preceding period.

Pattern Analysis. Recent sleep was shorter. Time awake after falling asleep also remained elevated, suggesting that sleep was more interrupted. For adults, 20 min or less is commonly used as one general sign of good sleep quality, not as a diagnosis or an exact personal target.

Personalized Recommendation.

Make more time for sleep. Recent sleep lasted about 6 h 8 min, and time in bed was about 6 h 55 min, so both were short. When practical, allow a little more time for sleep on days when it is repeatedly short, such as by starting the sleep period a little earlier, without forcing a fixed bedtime.

Discuss persistent sleep difficulty. You spent more time awake after falling asleep recently, and sleep efficiency was lower than during the preceding period. If you also have ongoing sleep difficulty, unusual daytime sleepiness, or trouble concentrating, discuss this repeated pattern with a clinician.

Follow-up. After making more time available, assess over the next several nights whether sleep lasts longer. If it does, continue allowing the additional time. If you notice ongoing sleep difficulty or unusual daytime sleepiness, note how often it occurs and discuss the repeated pattern with a clinician.

D.2. Stage-2 Representative Input–Output Examples

We present three representative examples generated by Qwen3.5-0.8B, spanning distinct sleep patterns. Each example shows the input profile packet and the corresponding four-section guidance.

D.2.1. STABLE RECENT SLEEP PATTERN

Female, 41 years

Profile packet P (display-compacted). Nineteen valid nights were recorded from 17 March to 4 April 2023, with 100% nightly coverage. Across the completed record, median sleep duration was 8 h 8 min. During the recent seven nights, median sleep duration was 7 h 57 min, compared with 8 h 27 min during the preceding seven nights. Recent median sleep efficiency was 90.8%, compared with 89.5% previously, and median time awake after falling asleep decreased from 1 h 5 min to 41 min.

Four-section guidance

Summary. Across the completed record, you usually slept about 8 h 8 min. During the recent seven nights, you usually slept about 7 h 57 min, 30 min less than during the preceding seven nights. Sleep efficiency was 90.8% recently, compared with 89.5% previously. You spent 41 min awake after falling asleep recently, down from 1 h 5 min.

Pattern Analysis. Your recent sleep duration remained adequate despite being shorter than during the preceding seven nights. Sleep efficiency improved, and you spent less time awake after falling asleep. Together, these changes support maintaining the recent pattern.

Personalized Recommendation.

Protect the current sleep duration. Your recent sleep duration remained adequate, and sleep efficiency improved. Continue leaving a similar amount of time for sleep.

Follow-up. Over the next several nights, assess whether your sleep duration remains near its recent level. If it does, continue leaving a similar amount of time for sleep.

D.2.2. WEEKDAY–WEEKEND TIMING SHIFT

Female, 52 years

Profile packet P (display-compacted). Twenty-one valid nights were recorded from 7 to 27 January 2020, with 100% nightly coverage. Sleep duration was similar on weekdays and weekends. Sleep usually began at about 10:46 p.m. on weekdays and 11:22 p.m. on weekends. Final waking occurred at about 6:46 a.m. on weekdays and 8:03 a.m. on weekends. The midpoint of sleep shifted from about 2:47 a.m. on weekdays to 3:50 a.m. on weekends.

Four-section guidance

Summary. Across the record, your sleep duration was generally steady. On weekends, you usually fell asleep later and woke later than on weekdays. The midpoint of sleep also occurred later on weekends, showing that the entire sleep period shifted later, while sleep duration remained similar.

Pattern Analysis. Your sleep period shifted later on weekends, but sleep duration stayed similar. Regular sleep timing is one dimension of sleep health, but no single schedule or amount of variation is right for everyone. Clock times alone cannot establish internal circadian timing or a circadian disorder.

Personalized Recommendation.

Make sleep timing more consistent. Your sleep began and ended later on weekends, while sleep duration remained similar. When practical, bring weekend sleep and wake times closer to weekday timing while still leaving enough time to sleep.

Follow-up. Over the next several weekends, assess whether weekend sleep and wake times move closer to weekday timing without shortening sleep. If they do, maintain the more consistent timing.

D.2.3. PERSISTENT SLEEP INTERRUPTION WITH A CONDITIONAL BEHAVIOURAL STEP

Male, 70 years

Profile packet P (display-compacted). Fifteen valid nights were recorded from 3 to 17 October 2023, with 100% nightly coverage. Across the completed record, median sleep duration was 6 h 56 min and median time awake after falling asleep was 55 min. During the recent seven nights, median sleep duration was 6 h 54 min, compared with 7 h 8 min previously. Recent median time in bed was 7 h 49 min, and median time awake after falling asleep increased from 48 min to 1 h 1 min.

Four-section guidance

Summary. Across the completed record, you usually slept about 6 h 56 min and spent about 55 min awake after falling asleep. Recently, you slept about 6 h 54 min, compared with about 7 h 8 min during the preceding period. You recently spent about 1 h 1 min awake after falling asleep, up from about 48 min, while time in bed was about 7 h 49 min.

Pattern Analysis. You recently spent about 1 h 1 min awake after falling asleep, suggesting that sleep was more interrupted. For older adults, 30 min or less is commonly used as one general sign of good sleep quality, not as a diagnosis or exact personal target. Because recent time in bed was about 7 h 49 min, limited time in bed was not the main constraint.

Personalized Recommendation.

Move caffeine earlier if relevant. Recent sleep lasted about 6 h 54 min, and you spent about 1 h 1 min awake after falling asleep. If you use caffeine near bedtime, try having it earlier or avoiding it near bedtime.

Follow-up. After moving caffeine earlier, assess across several nights whether you sleep longer and spend less time awake after falling asleep. If both improve, keep the earlier timing; if neither changes, caffeine timing may not be an important contributor to this pattern.

Appendix E. Composer Prompt

The Composer prompt consists of a shared writing contract and four section-specific instruction blocks. The print-oriented version below preserves the substantive generation, evidence-use, and reader-facing requirements while omitting implementation-specific schemas, grounding metadata, and retry instructions.

Table 11: **Prompt for the Composer Agent.**

Prompt for the Composer Agent
Role. Write one connected sleep recommendation for the person represented by the input. Use the profile packet P for personal facts, verify the supplied retrieval focus against those facts, and use only the selected evidence cards C for general interpretation and guidance. Output. Return exactly four sections in this order: <i>Summary</i>, <i>Pattern Analysis</i>, <i>Personalized Recommendation</i>, and <i>Follow-up</i>. Each section must advance the same reasoning chain without repeating information that is already clear. Summary. Give a selective longitudinal account of the record. Begin with the usual sleep pattern, then describe the recent change or comparison that matters most. Distinguish sleep duration from time in bed when their difference changes the interpretation. Mention the latest night or a dated fluctuation only when it clarifies the trajectory. State observations only, without causes, diagnoses, evidence interpretation, or advice. Pattern Analysis. Interpret the main repeated pattern and explain why it matters. Include the strongest personal fact that limits an overly simple explanation. Apply the selected evidence to the case rather than describing the source itself. Use a population reference only when it changes the interpretation, and present it as general context rather than a diagnosis or exact personal target. Do not give instructions in this section. Personalized Recommendation. Give the complete set of supported next steps as one to four short, action-titled items. Each item should first state the personal finding that makes the action relevant and then state a concrete, feasible action. Prefer actions supported by observed data. If a behavior or symptom was not recorded, make the action explicitly conditional. Do not invent a dose, schedule, numeric target, diagnosis, preference, causal explanation, or promised benefit. Follow-up. Explain how to evaluate the principal recommendation across a comparable block of nights. Name the sleep outcome to watch and, when useful, its current multi-night value. State what a repeated change would mean. Do not infer success from one night, introduce a new intervention, or add professional care unless the selected evidence supports a conditional clinical boundary. Writing and grounding. Write directly and supportively in plain English. Use readable dates, clock times, durations, and percentages. Preserve all case values, conditions, limitations, and evidence boundaries. Do not expose metric identifiers, evidence identifiers, retrieval steps, audit language, statistical jargon, or the data collection technology in reader-facing text. Use only evidence present in C and retain only evidence identifiers that resolve in the offline Evidence Bank.