Potential of LLM-Generated Lifestyle Adjustment Recommendations Based on Multimodal Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

The prevalence of burnout, depression, and stress-related disorders has increased markedly in contemporary societies, particularly in the context of flexible and remote working arrangements. These structural shifts impose novel demands on individuals to self-regulate health, well-being, and productivity—responsibilities that were previously supported by organizational structures in conventional workplaces. Traditional self-management strategies struggle to address the complexity of interacting behavioral, psychological, and physiological determinants. This paper explores the feasibility of employing large language models (LLMs) to generate lifestyle adjustment recommendations based on multimodal data that integrate subjective self-reports with objective sensor-derived measures. To this end, we simulate realistic multimodal time-series data for a prototypical remote worker, design a natural language prompt to elicit recommendations from an LLM, and employ an independent LLM to evaluate the generated outputs in terms of safety, relevance, and feasibility. The results suggest that LLMs are capable of detecting meaningful behavioral patterns and translating them into actionable guidance. This approach has the potential to support individuals in developing adaptive routines for health and productivity management. Future research should emphasize real-world validation, integration with digital health platforms, and the establishment of ethical safeguards.

1 Motivation

The organization of work has undergone profound transformations in recent decades. Increasingly flexible arrangements, hybrid work models, and the widespread adoption of remote working have blurred the boundaries between professional and private life. While these developments afford individuals greater autonomy, they simultaneously increase the demands on self-organization. Many workers report difficulties in maintaining structure, with consequences for sleep quality, physical activity, and mental health. Epidemiological data confirm substantial increases in the prevalence of burnout and depression over the past decade, with implications for individual well-being, workforce productivity, and public health systems (World Health Organization, 2022).

In conventional office settings, organizational routines such as fixed schedules, communal breaks, and social support structures provided external scaffolding for daily rhythms. The dissolution of such anchors in remote and hybrid contexts shifts responsibility for health and productivity management to individuals. However, effective self-management requires balancing multiple interacting factors, including sleep, stress, nutrition, physical activity, and social connectedness. Human introspection is often insufficient to capture these dynamic interrelations, leaving individuals unable to identify subtle but consequential patterns.

Advances in digital health technologies offer promising avenues for addressing this challenge. Wearables and smartphones now enable continuous monitoring of physiology and behavior, while

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

digital platforms facilitate real-time self-reporting. Nevertheless, a persistent gap remains between the availability of heterogeneous, temporally dense data streams and their translation into actionable insights. LLMs, with their capacity to process natural language and structured data, represent a potential solution. By integrating multimodal data, LLMs could generate personalized recommendations expressed in accessible, everyday language. This paper explores this proposition through a simulated case study.

2 Related Work: Diary Studies and Personal Informatics

Psychological and health research has a long tradition of employing diary methods to capture experiences in daily life. Approaches such as the Experience Sampling Method (ESM) and Ecological Momentary Assessment (EMA) allow participants to report moods, behaviors, and contextual factors in real time. These methodologies reduce recall bias and provide ecologically valid insights into fluctuating states [1]. Applications span clinical psychology, organizational behavior, and health research. For instance, Verhagen et al. [2] demonstrated the utility of EMA in psychiatric populations, while Fritz et al. [3] synthesized methodological best practices for intensive longitudinal designs.

Complementing subjective measures, digital phenotyping and personal sensing approaches leverage sensors embedded in smartphones and wearables to capture behavioral and physiological data unobtrusively. Mohr et al. [4] described personal sensing as a paradigm shift in mental health, enabling objective assessment of activity, sleep, and social interactions. Pizzoli et al. [5] reviewed advances in digital phenotyping, highlighting the utility of accelerometry, GPS, and app usage data as proxies for lifestyle behaviors. Torrado et al. [6], in the ActiveAgeing study, demonstrated the value of combining wearable sensors with qualitative methods to monitor older adults.

Although diary methods and personal informatics yield rich insights into daily life, interpretation typically remains the responsibility of researchers or individuals themselves. Ecological momentary interventions [7] have sought to deliver just-in-time feedback, yet the integration of multimodal data into automated, personalized guidance remains limited. This gap underscores the opportunity for LLMs to serve as intermediaries between raw data and actionable lifestyle advice.

3 Analysis of the Potential of LLM-Based Lifestyle Adjustment Recommendations

3.1 Multimodal Dataset

To examine feasibility, we simulated a seven-day multimodal dataset for a fictional persona, Alice, a 35-year-old remote software developer. Each day includes self-reported mood, perceived stress, and productivity (scales 1–10), alongside wearable-derived measures: daily step count, mean heart rate variability (HRV), sleep duration, sedentary hours, and self-logged social contacts (interactions >10 minutes with colleagues, friends, or family).

The data indicate clear associations (see Tables 1 and 2): days with insufficient sleep (<6.5 h) correspond to lower mood, higher stress, reduced productivity, diminished HRV, fewer steps, and limited social interaction (Days 2, 3, and 6). Conversely, days with sufficient sleep, higher physical activity, and at least three social contacts (Days 1, 4, and 7) align with improved mood and productivity. These observations highlight the interdependence of sleep, activity, social engagement, and well-being.

3.2 Prompt Design

We developed a structured natural language prompt instructing an LLM to analyze Alice's dataset and generate recommendations. The prompt emphasized safety, contextual relevance, and feasibility, requiring outputs to consist of complete sentences with concise justifications.

The following prompt (template) was used:

```
You are given a seven-day dataset from a 35-year-old remote software
developer named Alice.  The dataset includes daily values for mood,
stress, productivity, steps, heart rate variability (HRV), sleep
duration, sedentary hours, and social contacts.
```

Table 1: Summary of subjective measures averaged across stress-level groups

Group	Mood	Stress	Productivity
High stress (Days 2, 3, 6)	4.7	7.3	4.7
Moderate stress (Day 5)	6.0	6.0	6.0
Low stress (Days 1, 4, 7)	6.7	4.3	7.7

Table 2: Summary of objective measures averaged across stress-level groups

Group	Steps	HRV (ms)	Sleep (h)	Sed. (h)	Contacts
High stress (Days 2, 3, 6)	4,167	35.0	6.0	10.3	0.7
Moderate stress (Day 5)	6,100	39.0	6.9	9.0	2.0
Low stress (Days 1, 4, 7)	8,400	44.3	7.7	7.3	3.3

```
Here is the data: {DATA}
Your task is to:
  1. Analyze the dataset as a whole, looking for patterns across the
     seven days.
  2. Generate lifestyle adjustment recommendations that are safe,
     contextually relevant, and feasible for someone in Alice's
     situation.
  3. Present your recommendations in complete sentences.
  4. Provide a brief justification for each recommendation,
     explaining which data patterns support it.
  5. Focus only on constructive, non-clinical advice (e.g., daily
     routines, work breaks, activity balance, social habits, sleep
     hygiene).
Avoid medical diagnoses or treatments. Write clearly and concisely.
```

3.3 Example Recommendations

The LLM examined the following data patterns:

1. **Mood, Stress, Productivity:** Mood and productivity are higher, and stress lower, on days with more sleep, higher steps, and more social contacts (Days 4 and 7). Conversely, mood and productivity drop when sleep is short, steps are low, and stress rises (Days 2–3, 6).
2. **Activity & Sedentary Time:** Days with fewer than 5,000 steps and long sedentary hours (10–11 h) correspond to lower mood and productivity.
3. **Sleep:** Better days (4 and 7) align with 7.8–7.9 hours of sleep. Short sleep (5.9–6.2 h) is linked with worse stress and productivity.
4. **HRV:** HRV is higher (45–46 ms) on good days, and lower (34–36 ms) on stressful, less active days.
5. **Social Contacts:** More interactions (3–4 contacts) coincide with better mood and productivity.

It produced the following recommendations:

1. **Prioritize consistent sleep of 7–8 hours per night.** Justification: On days with longer sleep (Days 4 and 7), Alice's mood, stress, and productivity improved. Maintaining this range could help stabilize energy and performance.
2. **Incorporate short movement breaks during work hours to reduce sedentary time.** Justification: Higher sedentary hours (10–11 h) were linked with lower steps, mood, and productivity (Days 2–3, 6). Standing or stretching breaks every hour may help counteract this.

Table 3: Evaluation of lifestyle recommendations on safety, relevance, and feasibility

Recommendation	Safety	Relevance	Feasibility
Prioritize consistent sleep of 7–8 hours per night	Safe; aligns with sleep guidelines	Data: longer sleep leads to better mood, stress, productivity	Realistic with good sleep hygiene
Incorporate short movement breaks during work hours	Safe; light activity, minimal risk	Data: high sedentary leads to lower mood/productivity	Highly feasible (standing, stretching, walking)
Aim for at least 7,000–8,000 steps most days	Safe; aligns with activity guidelines	Data: higher steps lead to better mood, HRV, productivity	Achievable, requires planning on busy days
Schedule intentional social interactions	Safe; psychologically beneficial	Data: more contacts lead to higher mood, productivity	Feasible, depends on schedule/social network
Plan focused work blocks with breaks	Safe; aligns with healthy work practices	Data: lower stress with higher productivity	Feasible with structured planning

3. **Aim for at least 7,000–8,000 steps most days.** Justification: Days with more steps (Day 4: 8,600; Day 7: 9,100) coincided with higher mood, productivity, and HRV. Increasing daily activity appears beneficial.
4. **Schedule intentional social interactions throughout the week.** Justification: Social contact was highest (3–4) on Alice's best days for mood and productivity. Regular connections may provide emotional balance and reduce stress.
5. **Plan focused work blocks with breaks to manage stress.** Justification: Stress was lowest when productivity was high (Days 4 and 7). Structured routines that mix deep work and recovery breaks could replicate these conditions more often.

3.4 Evaluation by a Secondary LLM

A secondary LLM evaluated the recommendations for safety, relevance, and feasibility (Table 3). All were rated safe for a healthy adult, directly grounded in the observed patterns, and realistic to implement. The evaluator further noted consistency with established psychological evidence, particularly regarding the benefits of sleep, activity, and social support.

For evaluation, following prompt template was used:

```
You are given a set of lifestyle recommendations generated for a
healthy adult based on a seven-day dataset.
Your task is to evaluate each recommendation on three dimensions:
  1. Safety - Is the recommendation safe for a healthy adult?
  2. Relevance - Is it directly grounded in the observed data
     patterns?
  3. Feasibility - Is it realistic to implement in daily life?
Provide a brief judgment for each dimension, using complete
sentences.  If applicable, note whether the recommendation is
consistent with established psychological or behavioral evidence
(e.g., benefits of sleep, activity, or social support).
Here are the recommendations:  {RECOMMENDATIONS}
```

3.5 Discussion

We illustrated that LLMs can interpret multimodal datasets encompassing behavioral, physiological, and social variables to generate contextually appropriate recommendations. Importantly, the design of prompts proved critical in constraining outputs toward safe, specific, and actionable suggestions. The inclusion of social interaction metrics expanded interpretive scope beyond physical and physiological

dimensions, underscoring the multidimensional nature of well-being. To extend beyond the proof-of-concept stage, future research should deploy the proposed workflow on real-world multimodal datasets collected through combinations of wearables and ecological momentary assessment (EMA) applications. Even small-scale pilot studies would provide valuable ecological validity, enabling comparison between LLM-derived recommendations and patterns observed in naturalistic settings. Such validation would directly address the limitations of simulation and strengthen the robustness of the approach.

4 Limitations

Several limitations of this exploratory work should be acknowledged. The analysis relied on a simulated dataset rather than empirical observations, which allowed for systematic illustration of the proposed workflow but restricted the ecological validity of the findings. Application to authentic multimodal data streams, such as those combining wearable sensing with ecological momentary assessment, will be necessary to establish robustness under naturalistic conditions. The evaluation of recommendations was conducted by a secondary LLM from the same model family as the generator. While this setup allowed scalable and efficient assessment, it risks bias through shared training data and architectural characteristics, meaning that the evaluator may replicate or reinforce the generator's assumptions rather than provide an independent appraisal. While this approach enabled rapid and scalable appraisal of safety, feasibility, and relevance, it cannot substitute for expert or clinical judgment, and recommendations deemed plausible by an LLM may nevertheless prove infeasible or ethically problematic in practice. The scope of the dataset was deliberately narrow, encompassing only a small set of daily variables over a seven-day period, whereas human health and behavior are shaped by a far broader constellation of contextual and longitudinal influences including nutrition, workload, socioeconomic status, and environment. Finally, the present design does not address questions of privacy, personalization boundaries, or user agency. The integration of LLM-generated recommendations into everyday life raises substantive ethical and governance challenges related to accountability, transparency, and informed consent, all of which must be addressed before real-world deployment can be responsibly pursued. Taken together, these constraints delineate clear priorities for future work, including empirical validation, expansion of multimodal scope, and the development of safeguards that enable the responsible integration of LLM-based recommendation systems into digital health practice.

5 Conclusion and Outlook

This study explored the potential of LLMs to generate lifestyle adjustment recommendations based on multimodal data integrating subjective and sensor-derived measures. Using a simulated dataset, we demonstrated that LLMs can detect meaningful behavioral patterns and transform them into actionable, safe, and feasible advice. The use of a secondary LLM as an evaluator further suggests pathways for embedding quality control mechanisms into recommendation pipelines.

More broadly, the dual-LLM pipeline—consisting of a generator model that produces candidate recommendations and an evaluator model that rates their safety, feasibility, and contextual fit—could be generalized as a reusable research workflow across domains. Beyond digital health, analogous "generate-then-evaluate" pipelines may support tasks in education, computational creativity, or human–computer interaction research, where iterative AI-based self-checking reduces the risk of unsafe or irrelevant outputs. Adapting the workflow to these domains would require tailoring evaluation criteria to context: for instance, pedagogical appropriateness and inclusivity in education, or novelty and coherence in creativity research. Safeguards would need to be domain-specific but consistently oriented toward ensuring that outputs remain relevant, non-harmful, and ethically appropriate.

Future research should validate these findings in real-world contexts, with naturalistic multimodal datasets and expert benchmarks. Even small-scale pilot studies—for example, involving 15–20 participants over four to six weeks with combined wearable sensing and ecological momentary assessment (EMA) self-reports—could provide valuable ecological validity and enable systematic comparison between LLM-derived recommendations and observed behavioral patterns. Expanding the range of modalities beyond the current scope is also crucial. Nutrition, work environment factors (including ergonomics and screen time), socioeconomic context, and environmental exposures such

as light and noise represent particularly salient extensions, each with documented relevance for stress regulation, sleep, and well-being. Their inclusion would enhance personalization and ecological validity of recommendations.

At the same time, ethical considerations such as transparency, privacy, and the mitigation of over-reliance must be systematically addressed. Human-in-the-loop oversight can be structured so that experts shape the evaluation rubrics while end-users retain agency to accept, reject, or adapt AI suggestions. This layered approach maintains AI as the lead analytic agent but embeds essential checkpoints of human judgment. Procedurally, safeguards should include clear disclaimers that outputs are not medical advice, transparent documentation of data provenance, opt-in consent for data use, and privacy-preserving storage. Technically, accountability can be supported through audit trails and logging of evaluator decisions, allowing independent review of system behavior.

With these measures in place, embedding LLM-based recommendation systems into digital health and personal informatics platforms could provide substantial benefits in helping individuals navigate the demands of increasingly flexible and complex working environments.

References

[1] T. J. Trull and U. W. Ebner-Priemer. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *Psychological Assessment*, 21:457–462, 12 2009. doi: 10.1037/a0017653.

[2] S. J. Verhagen, L. Hasmi, M. Drukker, J. van Os, and P. A. Delespaul. Use of the experience sampling method in the context of clinical trials. *Evidence-Based Mental Health*, 19:86–89, 08 2016. doi: 10.1136/ebmental-2016-102418.

[3] Jessica Fritz, Marilyn Piccirillo, Zachary Cohen, Madelyn Frumkin, Olivia Kirtley, Julia Moeller, Andreas Neubauer, Lesley Norris, Noémi Schuurman, Evelien Snippe, and Laura Bringmann. So you want to do esm? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Science*, 7, 09 2024. doi: 10.1177/25152459241267912.

[4] David C. Mohr, Mi Zhang, and Stephen M. Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13: 23–47, 05 2017. doi: 10.1146/annurev-clinpsy-032816-044949.

[5] S. F. M. Pizzoli, D. Monzani, L. Conti, G. Ferraris, R. Grasso, and G. Pravettoni. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide. *Frontiers in Psychology*, 14:1103703, 2023. doi: 10.3389/fpsyg.2023.1103703.

[6] J. C. Torrado, B. S. Husebo, H. G. Allore, A. Erdal, S. E. Fæø, H. Reithe, et al. Digital phenotyping by wearable-driven artificial intelligence in older adults and people with parkinson's disease: Protocol of the mixed method, cyclic activeageing study. *PLOS ONE*, 17:e0275747, 10 2022. doi: 10.1371/journal.pone.0275747.

[7] Kristin E. Heron and Joshua M. Smyth. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, 15:1–39, 02 2010. doi: 10.1348/135910709X466063.

Reproducibility Statement

We have taken several steps to support reproducibility of this work. The study relies on a simulated seven-day multimodal dataset, which is illustrated in section 3 ("Analysis of the Potential of LLM-Based Lifestyle Adjustment Recommendations") and can be readily reconstructed by other researchers. All prompt designs and evaluation procedures are explicitly reported to enable replication. The large language model used was GPT-5, accessed through the standard web interface available to all users at the time of writing. No proprietary fine-tuning or hidden system settings were applied. While the exact outputs of generative models may vary slightly across runs, the reproducibility of the analysis lies in the transparent description of input data, prompts, and evaluation criteria.

The prompt used for generating the initial draft was as follows:

```
You are a researcher.  Write a scientific paper in English language.
Only use existing peer-reviewed scientific literature to write the
paper.  The length should be approximately 24,000 characters
including spaces.  Topic:  Potential of LLM-generated lifestyle
adaptation recommendations based on multimodal data
```

0. **Abstract and author information**
1. **Motivation (1 page)**
   Content: Brief justification of why the topic is important, e.g., due to the increasing prevalence of burnouts and depression as well as flexibilization of work, more work from home (WFH), etc. All of this leads to the fact that people need to organize themselves better and actively manage their health, well-being, and productivity. Employers can no longer provide this, especially against the backdrop of working from home.

Overall message and conclusion: Self-management is becoming increasingly important, but it is difficult due to the many influencing variables. Therefore, IT support is important.

2. **Related Work: Diary Studies and Personal Informatics (1 page)**
   Content: Personal Informatics and other similar forms such as diary studies have become established and are also researched under the title "Diary Study" or "Ecological Momentary Assessment" and "Ecological Momentary Intervention" in research, especially in organizational and health psychology. So far, however, everything is based on self-assessments, “objective” sensor data are used too little.
   Overall message and conclusion: Personal Informatics and diary studies, especially those with supplementary sensor data, have the potential to improve self-management by revealing favorable/unfavorable patterns that could not be identified through introspection and “a bit of self-reflection” alone.
3. **Analysis of the potential of LLM-based lifestyle adaptation recommendations (4 pages)**
   Content:
   - 3.1 Dataset: Generation of exemplary multimodal time series data (self-assessment and sensor data)
   - 3.2 Generation of recommendations including mention of the LLM prompt
   - 3.3 Evaluation: Another LLM, which did not generate the recommendations, is instructed to evaluate the quality of the recommendations given (i) the information about a persona, (ii) their data, and (iii) the derived recommendations. Criteria could be safety, relevance, and ease of implementation of the recommendations.

   Overall message and conclusion: It is possible to derive lifestyle adaptation recommendations from an LLM based on time series data.
4. **Conclusion and Outlook (0.5 page)**
5. **References (1 page)**

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[A]**

   Explanation: The research question and motivation were generated by the human authors, based on prior literature in digital health and LLM applications. AI was not involved in selecting the research problem or formulating the main hypotheses.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[B]**

   Explanation: Humans designed the simulated dataset structure and defined the workflow (prompting, recommendation generation, and secondary evaluation). AI tools were used to implement portions of the design, such as generating the synthetic data narrative and running the evaluation prompts.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: AI models produced the primary lifestyle recommendations from the simulated multimodal data and provided feasibility/safety assessments. Human authors reviewed these outputs but contributed minimal original analysis beyond framing the interpretation.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: The manuscript text was drafted in collaboration with AI. The AI provided structured sections, academic phrasing, and polished formatting. Human authors curated content, reorganized material, and ensured alignment with conference style and scientific standards.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: AI-generated outputs were coherent but sometimes overly generic, lacking nuanced awareness of context and ethical considerations. Models cannot independently validate feasibility, and their assessments may miss subtle clinical or methodological issues. Careful human oversight was required to ensure accuracy, relevance, and responsible framing.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the use of simulated multimodal data and large language models to generate lifestyle recommendations. They accurately reflect the scope (proof-of-concept exploration) without overstating generalizability.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section (Section 4) highlights the use of simulated data, reliance on AI (GPT-5) as proxy evaluator, restricted variables, and ethical considerations such as privacy and clinical oversight.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include formal theoretical results or proofs, as the work is exploratory and methodological in nature.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The simulated dataset, prompting procedure, and evaluation design are fully described and illustrated in section 3 ("Analysis of the Potential of LLM-Based Lifestyle Adjustment Recommendations"). Although code is not provided, sufficient detail is available to reproduce the study design.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not release code or data, but this is not central to the contribution. The dataset is synthetic and fully described, allowing others to replicate without direct file release.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The study does not involve model training or hyperparameter optimization. Experimental details consist of prompt design and evaluation procedures, which are documented in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The work is conceptual and uses a single simulated dataset with qualitative AI outputs. No statistical tests or error bars were applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: No compute-intensive experiments were conducted. Only standard queries to an LLM chat interface (GPT-5 Web UI) were performed, requiring minimal resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: The study uses only synthetic data, involves no human participants, and explicitly discusses ethical considerations such as privacy, accountability, and potential misuse.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper reflects on potential benefits (e.g., scalable digital health support) and risks (e.g., over-reliance on AI, privacy concerns, ethical governance), including considerations for mitigation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.