From Narratives to Numbers: Evaluating a Large Language Model (LLM) for Transforming Workplace Interview Narratives into Numerical Wellbeing Indicators

Full research paper

Riitta Hekkala
School of Business
Aalto University
Espoo, Finland
Email: riitta.hekkala@aalto.fi

Riko Nyberg
Adalyon Ltd,
Espoo, Finland
Email: riko.nyberg@adalyon.com

* The authors have contributed equally to the paper

Abstract

This study utilizes GPT-4o, an LLM, to interpret semi-structured interview narratives into structured numerical indicators of wellbeing. Key findings of this study are: firstly, different LLM instances produce consistent numerical wellbeing results. Secondly, we observe close alignment of these quantitative wellbeing results with the qualitative wellbeing interpretations. Thirdly, the strong correlation between the LLM's PHQ-9 and GAD-7 scores suggests that the model treats these constructs as a single general wellbeing measure when assessing workplace narratives. The study is based on 18 interviews (21 hours of speech) conducted between 2017 and 2021 with four IT professionals. Our study presents the LLM as a methodological tool for Information Systems (IS) researchers to scale mixed-method analysis. The study contributes (1) a demonstration of how an LLM can be applied to workplace narratives and (2) a validation and limitations when utilizing an LLM in producing numerical results for wellbeing questionnaires.

Keywords LLM, GPT-4, wellbeing, PHQ-9, GAD-7

1 Introduction

Stress, depression, and burnout are the major challenges in today's workplaces, and the information technology (IT) field is no exception. Moreover, high levels of stress and burnout among IT professionals have prevailed over the decades (Ahuja et al., 2007; Armstrong et al., 2015; Pawlowski et al., 2007) and continue to be a significant challenge for those working in this field. Scholars in the IT field have identified multiple causes of stress related to the IT work environment, including social aspects such as role ambiguity, lack of autonomy/ control, lack of rewards (Moore, 2000), management style, and corporate culture (Pawlowski et al., 2007). These stress-causing factors represent a fundamental threat to work performance, levels of employee motivation, and wellbeing. Previous research has widely recognized specific consequences of stress and factors that may lead to burnout (e.g., Ahuja et al., 2007; Armstrong et al., 2015; Chuang et al., 2025). However, there is a scarcity of literature investigating how IT professionals cope with the causes of stress in order to preserve their wellbeing. Furthermore, due to constant technological advancements, digital technologies are significantly transforming how healthcare services are planned, delivered, and investigated (cf., Rosenman et al., 2024).

There still remains, however, a remarkable gap in information technology (IT)/ information systems (IS) literature on how artificial intelligence (AI), large language models (LLMs), and advanced analytics in personal health management have been studied and used in research. Moreover, there have been calls for studies on machine learning and other smart analytics of individual healthcare, among other solutions in other fields, too, such as psychology (Rosenman et al., 2024) and psychiatry (Mazur et al., 2025). Recent studies in other fields have pointed out that LLMs could be used to transform how healthcare services are delivered (Mazur et al., 2025; Rosenman et al., 2024; Stade et al., 2024). For example, Stade et al. (2024) argued that different AI models have huge potential to support, augment, or even automate psychotherapy. The rapid development of LLMs has opened new methodological possibilities for IS scholars by enabling the transformation of interview narratives into structured numerical indicators of wellbeing. This paper evaluates the reliability and limitations of using an LLM to complete standardized wellbeing questionnaires from workplace narratives; thereby exploring how LLMs can serve as methodological tools to augment narrative-driven and data-driven approaches within IS research.

The workplace interviews for this study were not initially focused on issues related to wellbeing, such as depression and anxiety. Rather, the aim was to understand the lived experiences of IS development (ISD) work, and the narratives of the research participants became the data collection method (e.g., Polkinghorne, 2006). The qualitative results on exhaustion and anxiety, as well as on extended sick leaves among participants, generated the idea of testing how an LLM could be used with different questionnaires to recognize and develop deeper insights into issues related to wellbeing. However, although questionnaires were utilized that measure depression and anxiety, we do not claim that our informants would have suffered from these, nor do we diagnose them.

This study instead focuses on how specific questionnaires (if any) support or find results that align with the qualitative research undertaken here. Hence, this approach is exploratory and methodological in nature rather than inferential. With a small number of participants, the findings are not generalisable but instead demonstrate the potential of LLMs as methodological instruments for IS researchers. Our study is guided by the research question:

How reliably can an LLM transform workplace narrative interviews into standardized wellbeing indicators (PHQ-9; GAD-7), and what methodological implications does this hold for IS research?

This framing allows examination of both the technical reliability of LLM outputs (e.g., internal consistency, alignment with qualitative analysis) as well as the research implications and limitations for IS scholars who seek to integrate a combination of narrative-driven and data-driven approaches. By focusing on methodological evaluation rather than clinical diagnosis, we position the LLM as a tool that can augment IS research practices, particularly in contexts where retrospective, unobtrusive, and scalable wellbeing assessments are desirable.

The remainder of the paper is organized as follows: Section 2 presents the theoretical background on topics that are relevant to this investigation. The paper then outlines the methodology adopted for this study, followed by the findings, discussion, conclusions, and references.

2 Theoretical Background

This section presents, first, earlier research on wellbeing in the IT sector; second, important information regarding the Patient Health Questionnaire (PHQ-9) and Generalized Anxiety Disorder Scale-7 (GAD-7) for diagnosing health, examples of which are found in Appendix 1; and finally, recent studies using LLMs to predict wellbeing.

2.1 Research on Wellbeing in the IT Sector

Wellbeing is something of an ambiguous concept for which various interpretations have emerged over the years. Earlier literature in the IS field has outlined the importance of understanding the diverse factors that may cause harmful stress symptoms or even more serious consequences (e.g., Ahuja et al., 2007; Armstrong et al., 2015; Moore, 2000; Pawlowski et al., 2007).

Earlier research has also explained how factors such as workload, lack of autonomy, role ambiguity (Moore, 2000), coworkers, performance evaluations, management (Pawlowski et al., 2007), and conflicts (Khosravi et al., 2020) may cause stress and have negative consequences on the wellbeing of individuals. It has been recognized earlier that some of these factors may lead to even more severe problems, such as burnout (Pawlowski et al., 2007), and employee turnover (Ahuja et al., 2007). Pawlowski et al. (2007) also indicate three different burnout-induced outcomes: (1) declining job performance, (2) leaving a job or profession, and (3) reduced physical wellbeing. Moreover, studies in other fields have investigated the connections between personality and stress, among other things (Bowling and Jex, 2013).

More recent research extends this understanding through the Job Demands–Resources (JD-R) model (Bakker and Demerouti, 2017; Bakker et al., 2023; Demerouti et al., 2001), which explains wellbeing as a dynamic balance between work demands (e.g., workload or emotional pressure) and available resources (e.g., autonomy or social support). Chuang et al. (2025) applied this model to account for technostress and artificial intelligence (AI) related demands that influence wellbeing and performance, demonstrating a dual impact: AI-related technostress increases exhaustion, whereas AI usefulness enhances engagement and productivity. Studies on technostress identify recurring pressures such as techno-overload, techno-complexity, techno-invasion, techno-uncertainty, and, techno-insecurity which can erode satisfaction, while supportive digital infrastructure and training mitigate these effects (Kumar, 2024).

Wellbeing is also a technology-sensitive construct that is shaped by the interplay of digital demands (e.g., technostress, AI demands), organisational or job resources (e.g., leadership, learning climate, digital support), and individual coping or buffer strategies, especially in AI-enabled workplaces (Chuang et al., 2025). In a nutshell, wellbeing is presented as a positive condition that holds significance for employees, contributing meaningfully to different results.

2.2 Self-Reported Questionnaires for Diagnosing Health

The PHQ-9 and the GAD-7 are standardized instruments for assessing symptoms of depression (PHQ-9) and anxiety (GAD-7). Both are brief self-report measures designed for clinical screening and research use, each asking respondents to rate the frequency of recent symptoms over the previous two weeks on four-point scales from 0 (“not at all”) to 3 (“nearly every day”). The total scores provide a level of symptom severity, with higher scores indicating greater depression or anxiety. The complete questionnaires are found in Appendix 1 for reference (Spitzer et al., 1999; Spitzer et al., 2006).

The PHQ-9 consists of nine items that assess the frequency of depressive symptoms with a total score of 0-27. A major depression can be diagnosed if at least five of the nine depressive symptom criteria have been present on more than half the days during the previous two-week period. Additionally, if two, three, or four depressive symptoms have at least been present on more than half the days during the previous two-week period, and one of the symptoms is depressed mood or anhedonia, another form of depression is diagnosed. The questionnaire also includes an item on suicidal ideation, which is considered on its own as being indicative of depression, regardless of duration. Before making any final diagnosis, the clinician is expected to rule out physical causes of depression, normal bereavement, and a history of manic episodes (Spitzer et al., 1999; Kroenke et al., 2001). The GAD-7 has seven questions assessing anxiety symptoms such as nervousness, excessive worry, restlessness, irritability, and difficulty relaxing during the previous two weeks, and each item is rated on the same 0–3 scale, totaling to a score of 0-21 (Spitzer et al., 2006).

2.3 Using LLMs to Predict Wellbeing

While LLMs have been a subject of interest to academics for years, the continuous development of those models has accelerated interest academically and practically with regard to this highly important sector. In the past two years alone, LLMs have already been researched in several fields such as psychology (Rosenman et al., 2024), healthcare (Lalk et al., 2025), IT education (Neumann et al., 2025), and management (Chuang et al., 2025).

There has also been a growing interest in exploring how different LLMs could be utilized in predicting health and wellbeing in different contexts (Lalk et al., 2025; Liu et al., 2025; Oparina et al., 2025; Rosenman et al., 2024; Stade et al., 2024). For instance, Stade et al. (2024) highlighted that LLMs, such as GPT-4 and Google's Gemini, have huge potential to support or even eventually automate psychotherapy. Moreover, recent literature (e.g., Oparina, 2025) has tested machine learning algorithms to contribute to a better understanding of people's self-reported wellbeing. They (Oparina et al., 2025) used extensive longitudinal and nationally representative surveys as data in their study. Furthermore, these authors argued that machine learning approaches predict wellbeing better than conventional linear models. This finding also supports Stade et al.'s (2025) research.

Additionally, recent studies are increasingly using LLMs together with different psychological questionnaires to test and predict wellbeing. For example, Liu et al. (2025) conducted a cross-sectional study to assess the possible applicability of ChatGPT-4 in evaluating different health problems (e.g., anxiety and depression) by comparing two questionnaires—PHQ-9 and GAD-7. Based on the results of these questionnaires, researchers entered the results into ChatGPT, and ChatGPT generated relevant questions based on the detailed information of the inputs. Participants (university students) answered specific ChatGPT-adapted questions via the ChatGPT webpage on their own computers, and then ChatGPT was used to automatically score each entry against the criteria of the PHQ-9 and GAD-7 scales. To ensure comparability of the assessment results, all participants also completed the traditional online questionnaires (Liu et al., 2025).

Rosenman et al. (2024) employed an LLM to transform unstructured psychological interviews into structured questionnaires with various psychiatric and personality domains. The LLM was prompted to answer the specific questionnaires by impersonating the interviewee. The answers obtained were coded as features, which were used to predict standardized psychiatric measures of depression (PHQ-8) and post-traumatic stress disorder (PTSD) (PCL-C) using a random forest regressor (RFR). An RFR is a type of supervised learning algorithm that can be used for classification and regression tasks. The approach used by Rosenman and co-authors has been shown to improve diagnostic accuracy compared to multiple baselines. Their study creates a connection between narrative-driven and data-driven approaches for mental health assessment (Rosenman et al., 2024).

A recent study by Lalk et al. (2025) also aimed to train an LLM for the perception of emotions in German speech. The motivation for focusing on emotions was that different emotions play a crucial role not only for symptom severity, but they also have an impact on the therapeutic relationship between patient and healthcare professional. The authors of the study applied a pre-trained LLM on psychotherapy transcripts to identify the most important emotions for predicting symptom severity. A public dataset of 28 different emotions was utilized, and the dataset contained 553 psychotherapy sessions of 124 patients. Another recent study by Mazur et al. (2025) examined free-form speech content. Their English-speaking sample contained almost 15,000 adults (recruited via social media) from the United States and Canada. However, the length of the free-form speech was only 25 seconds or less, and they did not utilize any written text or LLMs. The free speech voice results of these participants were compared with a self-reported PHQ-9 at a cut-off score of 10 (moderate to severe depression). The researchers highlighted that from 25 seconds of free-form speech, machine learning technology detected vocal characteristics consistent with PHQ-9 results (Mazur et al., 2025).

3 Methodology

The summarized process of this research and methodology are the following: (1) interview data was collected, (2) audio transcription, anonymization and qualitative wellbeing coding was done to the transcriptions, (3) the LLM (GPT-4o) prompt design was developed to (4) produce suitable output formatting for quantitative results, (5) validation of the LLM result consistency and comparisons to prior qualitative results and (6) from these outcomes we conducted the findings and limitations of the LLM's capabilities to understand and review the interviews. Figure 1 describes this summarized process, and the following subsections will describe this process in more detail.

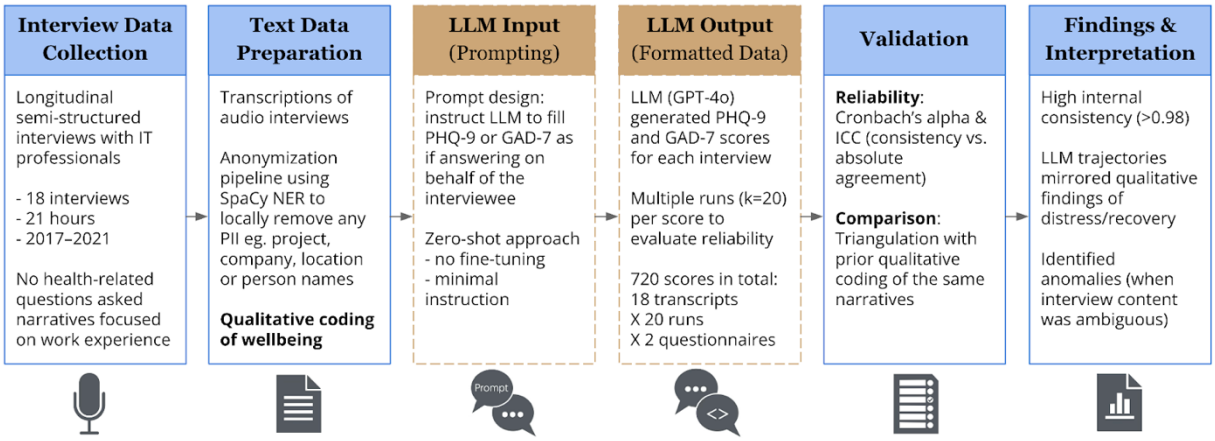

Figure 1. Process of Using an LLM to Transform Interviews into Numerical Wellbeing Indicators

3.1 Research Case, Design, and Data Collection

This study adopted a mixed-method longitudinal case study design, relying on 18 interviews with four IT professionals conducted yearly over a five-year period from 2017 to 2021 (see Table 1). Our approach was an interpretive case study (Walsham 2006), which offered deep insights into the meanings and thoughts of individuals in their ISD project. Prior to conducting the interviews, each participant was asked for permission to record the interviews to ease data collection and analysis, allowing us to focus fully on the conversation itself. Participation in the study was also voluntary, and we would redact any information from the interview upon the participant's request to encourage open and honest dialogue.

Interviewee	Number of interviews	Speech (total 1,264 min)
1	5	476 min
2	4	240 min
3	5	245 min
4	4	303 min

Table 1. The Interviewees, Number of Interviews, and Total Length of the Interviews.

Moreover, it is important to highlight that while the interview questions did not include health-related (e.g., depression, sick leaves) questions, the participants did talk about their health. To ensure data security and confidentiality, all interview recordings, transcripts, and notes were stored securely following the researchers' institutional data protection policies. Access to this data was restricted to the research team, and all identifiable information was removed or anonymized during transcription to protect the privacy of participants.

3.2 Data Analysis Methods

3.2.1 From an Interpretive Approach to Theory Development

The first phase of data analysis was our qualitative research part, which adopted an interpretive approach to grounded theory development (Glaser, 1992). We utilized an iterative theory-building process, which is based on constant comparison (Glaser, 1992). We consistently alternated between our data, memos, and the evolving categories. Throughout this process, we revisited earlier interviews as well as new ones alongside the emerging categories to either validate our findings or identify discrepancies between emerging concepts and our data. Our data analysis progressed through three stages: open coding, selective coding, and theoretical coding (Glaser, 1992). During the first stage of open coding, we already identified that employees in business and design roles stood out as extreme cases of 'victims' in current workplace cultures, enduring hardship for years at the cost of their wellbeing and eventually experiencing burnout. Moreover, our continued analysis revealed different responses and behaviors to their respective situations. Based on the results of our qualitative study on wellbeing and burnout, we decided to compare the transcribed interview data with standardized psychological and/or health questionnaires. This phase is explained in the following subsection.

3.2.2 Utilizing LLMs to Transform Interviews into Numerical Wellbeing Indicators

This study utilizes GPT-4o as the LLM to produce wellbeing indicators, because during the time of conducting this research, it is among the most advanced models. Furthermore, our study decided to use only one LLM API due to the large number of iterations we needed for each different prompt and wellbeing questionnaire. The GPT-4o is used through OpenAI's API without the data being used in any training for the provider. GPT-4o also had a suitable maximum context window for our use (128 000 tokens). The LLM runs had the following setup: OpenAI's gpt-4o-2024-11-20 model (accessed June 2, 2025) with a 128k-token context window, temperature, and top-p both set to 1, and 20 runs per transcript. More details can be found in Appendix 2.

This study's quantitative analysis focused on utilizing an LLM with a simple zero-shot prompt to transform transcribed workplace interview narratives into numerical results of standardized wellbeing questionnaires. An example prompt and response format can be found in Appendix 3. One goal was to simplify the application of this methodology in research and make it easily accessible and reproducible for other IS researchers, and hence, the choice of a simple zero-shot prompt was intentional. However, we recognise that this design choice may have encouraged construct overlap between depression and anxiety scores, as the model generated total scores rather than item-level responses.

We also examined several standardized wellbeing questionnaires, but ultimately focused on only two of them due to their stability and robustness when used with simple zero-shot prompts. Here is a list of the wellbeing questionnaires examined in the initial phase of this study: 1) PHQ-9: Assesses symptoms of depression, 2) GAD-7: Measures anxiety symptoms, 3) Depression, Anxiety and Stress Scale (DASS21): A quantitative measure of distress along the axes of depression, anxiety, and stress, 4) Bergen Burnout Indicator (BBI-15): Assesses Burnout, 5) Maslach Burnout Inventory – General Survey (MBI-GS): A widely used measure of burnout, and 6) Short Form Health Survey (SF-12): A general measure of health status.

From this list, we focus on the PHQ-9 and GAD-7, and complete versions of these questionnaires are provided in Appendix 1. Regarding the other questionnaires, problems were encountered when LLMs were transforming the transcripts into numerical results. These problems included results going beyond the scale of the questionnaire's maximum and LLMs not understanding the completion criteria or needing a detailed description of the questions. These problems could have been solved by more grounding and instructions for the LLM. However, we did not do this because the goal was to use as simple zero-shot prompts as possible. Hence, we relied upon established wellbeing questionnaires that would already be familiar to the LLMs.

To evaluate the internal consistency of LLMs for the given numerical wellbeing results, we ran the LLM analysis in batches, meaning that all interviews of each participant were included within a single LLM context window or inference. This approach was possible because each participant had up to five interviews that resulted in about ~100,000 tokens per LLM inference. Hence, the results have been separated into four different cases, one for each interviewee, where we consider the consistency of the quantitative results between runs, as well as how LLM assessments were conducted compared to the qualitative analysis.

3.2.3 Statistical Validation for LLM Wellbeing Assessment Results

The objective of this step was to quantify the consistency and consensus among the independent LLM runs that produced the PHQ-9 and GAD-7 scores. Because our subsequent analysis uses averages across runs, establishing inter-rater reliability is essential before interpreting any patterns in the data. We assessed inter-rater reliability with intraclass correlation coefficients (ICCs).

For each questionnaire × interviewee, we formed a ratings matrix in which the targets were that interviewee's transcripts (4–5 per interviewee) and the raters were the $k = 20$ independent LLM runs (balanced design). ICCs were computed separately for PHQ-9 and GAD-7. Given our design, where a common set of raters evaluates the same set of targets within each interviewee, a two-way ICC framework is appropriate (Shrout and Fleiss, 1979). We report three complementary forms to address different generalization questions:

ICC(2,1) — two-way random, absolute agreement (single-measure). Treats both targets and raters as random samples. This answers: How well would a single score from a randomly chosen run agree, in absolute terms, with scores from any other run from the same population?

ICC(3,1) — two-way mixed, consistency (single-measure). Treats the specific set of runs as fixed. This answers: Are runs consistent in how they rank transcripts? Note that consistency ICC allows systematic

mean differences across raters; thus, it does not guarantee absolute interchangeability of single-run scores.

ICC(3,k) — two-way mixed, consistency (average-measure). Estimates the reliability of the mean score across the k runs used in this study (our primary downstream statistic). Under these assumptions, ICC(3,k) is mathematically equivalent to Cronbach's alpha; we therefore report ICC(3,k) as the reliability of the averaged LLM score used in later analyses.

For each ICC we report the point estimate, 95% confidence interval (CI), and a p-value for the null hypothesis ICC=0. CIs convey the precision of the reliability estimate, and it is emphasized because narrower CIs indicate higher precision. The p-values for the null hypothesis are reported for completeness, but statistical significance does not, by itself, imply strong reliability. We interpret ICC magnitudes using established guidelines (Koo and Li, 2016): < 0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, > 0.90 excellent reliability. Because our final analysis relies on the average across runs, ICC(3,k) is the most directly relevant indicator of whether the mean LLM score is stable and defensible for inference; ICC(2,1) and ICC(3,1) complement this by characterizing single-run agreement (absolute and consistency, respectively).

3.3 Ethical and Privacy Aspects of Utilizing LLMs

The privacy of the interview participants was protected by using a custom anonymization pipeline created with Python and the SpaCy library. This process was performed offline on a local computer to prevent any identifiable data from being transmitted externally. The pipeline utilized SpaCy's Named Entity Recognition to automatically identify named entities from the interview texts, such as personal names, company names, specific locations (such as cities), and project names. Each one of these entity names was consistently replaced with a unique alias (e.g., "[Person_A]" as "Kim" and another mention of "[Person_A]'s" as "Kim's"). We used consistent aliases to ensure that the meaning and context of the interviews were preserved, which was necessary for the later analysis using LLMs while still protecting the confidentiality of the interviewees. Ethically, this offline anonymization process was a key measure. Additionally, we ensured that the interview data, even after anonymization, would not be used to train or improve the LLM models.

4 Findings

4.1 Reliability of LLM Wellbeing Assessment Results

When an LLM was applied to assess wellbeing of interview transcripts, one significant finding was that across all four interviewees, the statistical validation revealed a high degree of internal consistency for the LLM-generated scores. The ICC for consistency, ICC(3,k), was exceptionally high, exceeding 0.98. These results indicate that, on average, different LLM runs produced nearly identical predictions when applied to the same transcript data (see Table 2).

ICC	Questionnaire	Int. 1	Int. 2	Int. 3	Int. 4	Avrg.
ICC(2,1)	PHQ-9	0.67	0.96	0.79	0.81	0.81
	GAD-7	0.63	0.94	0.86	0.74	0.79
ICC(3,1)	PHQ-9	0.76	0.97	0.80	0.80	0.83
	GAD-7	0.71	0.96	0.89	0.74	0.83
ICC(3,k) (Cronbach's alpha)	PHQ-9	0.99	0.99	0.98	0.98	0.99
	GAD-7	0.98	0.99	0.99	0.98	0.98

Table 2. Intraclass Correlation (ICC) for All the Interviewees ($p < 0.001$)

The measure of absolute agreement (ICC(2,1)), however, yielded more moderate results, ranging from 0.63 to 0.96. This fluctuation suggests that while the LLMs are consistent in their relative rankings, direct comparisons of scores from single, independent runs may be unreliable. To achieve more robust assessments of wellbeing, we recommend conducting multiple LLM runs and averaging the results. This approach mitigates the variability in absolute scores.

4.2 Alignment with Qualitative Analysis

To assess the validity of the quantitative results, the LLM-generated PHQ-9 and GAD-7 scores were compared with the themes identified in our qualitative coding of the interview narratives. Overall, the LLM assessments mirrored the observed wellbeing trajectories of the participants, particularly in periods of rising exhaustion, burnout, and recovery. This convergence suggests that the models were able to capture signals of wellbeing even though health was not the primary focus of the interviews.

Three cross-case patterns emerged in this analysis. First, rising LLM scores aligned with narratives of increasing strain between 2017–2019. Correspondingly, the LLM-based PHQ-9 and GAD-7 scores peaked between 2019-2020, indicating heightened symptoms of depression and anxiety. Second, peak scores coincided with burnout and sick leaves. Interviewee 2 described escalating interpersonal conflict and unrealistic demands, culminating in resignation after a period of sick leave. The LLM assessments reflected this trajectory, showing consistently elevated scores in the years leading up to the sick leave and resignation. Third, declining scores matched narratives of recovery and improved work conditions. Interviewees 1, 3, and 4 reported significant improvements following role changes and a reduction in pressure during 2020-2021 in the interviews held in 2021. Their PHQ-9 and GAD-7 outputs showed corresponding downward trends, suggesting that the LLMs were sensitive to improvements as well as deterioration.

Taken together, these patterns highlight the potential of LLMs to approximate the qualitative interpretation of workplace narratives in a structured and repeatable way. This strengthens the case for LLMs as methodological tools for bridging narrative-driven and data-driven assessments of wellbeing. Furthermore, the pipeline of automated anonymization and simple zero-shot prompting makes this exceptionally scalable for a data-driven approach to wellbeing assessments. These kinds of methodological tools and pipelines can immediately help accelerate IS research because they can be easily and retrospectively applied to any existing interview data.

Interviewee Narratives: Early signals of distress were evident in 2018–2019, where participants reported feelings of despair, frustration, and uncertainty in their roles. By 2019–2020, all four interviewees experienced intensified difficulties, often leading to medical consultation or sick leave. Interviewee 1 described being "on overdrive" and unable to sleep, and Interviewee 2 recounted the collapse of a workplace friendship alongside escalating pressures that culminated in resignation after the 2020 interview. Similarly, Interviewee 3 reported severe team conflicts and took sick leave, and Interviewee 4 described heightened anxiety and ultimately also required leave. In each case, the LLM assessments peaked during these periods, aligning with the lived experiences of exhaustion and withdrawal from work for the participants. In 2021, the narratives reflect recovery and stabilization. The interviewees highlighted new rhythms of life, team changes, or reduced pressure that enabled improvements in wellbeing. The LLM-generated scores declined accordingly, indicating sensitivity not only to worsening conditions but also to positive change.

4.3 Sources of Variation and Anomalies

Although the LLM-generated assessments demonstrated strong overall reliability, some anomalies highlighted the method's limitations. A source of a notable anomaly was the overlap between depression and anxiety scoring. Across all interviewees, the correlation between PHQ-9 and GAD-7 outputs was extremely high ($r = 0.97–0.99$). This contrasts with what is typically reported in clinical studies, where PHQ-9 and GAD-7 demonstrate moderate to strong correlations ($r \approx 0.70–0.80$; Newman, 2022; Teymoori et al., 2020). Brattmyr et al. (2022) have proven that PHQ-9 and GAD-7 contain separable cognitive and somatic subdimensions. Showing that although depression and anxiety share variance, they represent distinct latent dimensions. Thus, the extremely high correlation in this study's results suggests that the zero-shot LLM approach likely collapses these distinct subdimensions into one general wellbeing factor.

Another significant anomaly emerged from Interviewee 3's interview in 2020. That transcription's wellbeing results had significantly wider confidence intervals compared to any other transcript (see Figure 2). Similarly, Interviewee 4's 2019 transcript also had above average confidence intervals, but not to the same extent as Interviewee 3's 2020. Closer examination revealed that these transcripts included mentions of both past and present experiences, or interviewees were describing simultaneous stress and recovery, which may have confused the models and hence made the emotional state difficult to classify. These cases suggest that when the transcription content itself is ambiguous, it would cause unstable predictions, rather than the LLM's incorrect interpretations. However, the level of transcription ambiguity and its effects on the LLM's wellbeing result stability can be measured by running the transcript multiple times through the LLM model and calculating the confidence interval of the results.

In such a way, the ambiguous transcripts and inconsistent results are visible and can be handled separately from the more reliable results.

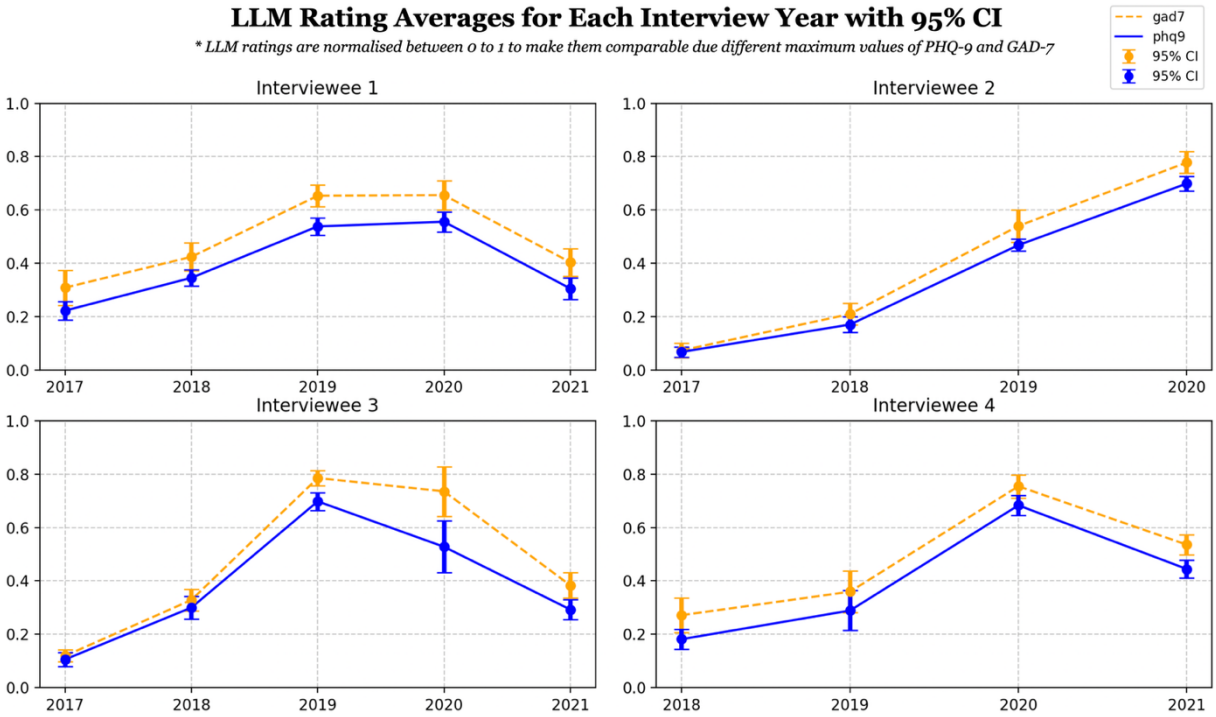

Figure 2. LLM Rating Averages and 95% Confidentiality Intervals for Each Interviewee

These anomalies illustrate that LLM-based wellbeing assessments are most reliable when transcripts contain clear and temporally consistent accounts of current emotional states, and least reliable when participants combine past, present, and future reflections or describe multiple states simultaneously. Recognizing these boundary conditions is essential for IS researchers seeking to apply LLMs to narrative data, as it indicates where additional grounding instructions or methodological safeguards are necessary.

4.4 Cross-Case Synthesis

Taken together, these findings suggest that an LLM is a reasonably reliable rater for a general numerical wellbeing score when using longitudinal interview narratives and simple zero-shot prompts. However, reliability can decrease due to ambiguous transcript content, which can be detected by examining the confidence interval results. The LLM was able to detect trajectories of distress leading up to burnout and improvements that resulted in recovery, and these results were aligned with the qualitative estimations across all four interviewees.

At the same time, the cross-case comparison reveals some limitations. The LLM generated nearly identical PHQ-9 and GAD-7 trajectories for all interviewees, suggesting a limited ability for the LLM to differentiate between depression and anxiety symptoms from workplace narratives. Hence, suggesting LLM results from PHQ-9 and GAD-7 are rather general wellbeing scores, than accurate item-level answers to the questions of each questionnaire. Furthermore, the most stable LLM predictions came from interview narratives that were clear, discussed current emotional states, and were weakest when participants mixed reflections on past difficulties and present improvements. In such cases, the models produced wider confidence intervals and inconsistent outputs.

Overall, with our methodology, the LLM is a promising but conditional tool: it can effectively approximate qualitative wellbeing interpretations when narratives are clear and contextually consistent. However, their outputs require careful validation using confidence intervals, and the wellbeing scores appear general, rather than specific to a given questionnaire, such as PHQ-9 or GAD-7. For IS research, this suggests that LLMs are best positioned as methodological instruments that could augment, rather than replace, qualitative analysis by providing scalable and structured insights into longitudinal wellbeing patterns.

5 Discussion

This study investigated the potential use of an LLM (GPT-4o) for assessment and prediction of the wellbeing of IT professionals by analysing transcripts of longitudinal narrative data. The findings indicate a correlation between LLM-driven wellbeing assessments using the PHQ-9 and GAD-7 scales and the results of a qualitative analysis of the same data. This suggests that LLMs are tentatively capable of transforming interviews into structured, quantitative measures of wellbeing, with some limitations. With this methodology and results, our study contributes to the IS research in the following ways.

Firstly, this study demonstrates the methodological feasibility by showcasing how an LLM can be applied to complete standardized wellbeing questionnaires (PHQ-9, GAD-7) using simple zero-shot prompts and workplace narrative transcripts as the source data. This suggests that LLMs could act as general wellbeing coders of qualitative material, even without fine-tuning or domain-specific prompts. The bridging of qualitative and quantitative approaches in IS research allows IS scholars to scale longitudinal qualitative studies into formats that support statistical analysis, trend detection, and cross-case comparison. This enables researchers to use an LLM retrospectively to gain wellbeing insights from existing narrative data. This study provides a proof-of-concept demonstration that invites further exploration of LLMs as tools for broader IS research.

Secondly, using longitudinal interview data, this study assessed the reliability of LLM-generated results by comparing them with qualitative analysis. Also, the internal consistency of the LLM model was high (ICC(3,k) > 0.98), thus strengthening the case for an LLM's capabilities to consistently interpret wellbeing from interview transcripts. This study provides one benchmark for the reliability and consistency, but also the limitations, of LLMs in workplace wellbeing contexts. For example, sensitivity to ambiguous narratives and difficulty distinguishing between depression (PHQ-9) and anxiety (GAD-7) symptoms, suggesting LLM's tendency to give general wellbeing scores. This study also contributes to the growing body of research on the application of LLMs in wellbeing, offering a unique perspective that extends the findings of related research (e.g., Liu et al., 2025; Mazur et al., 2025).

5.1 Sources of Variation and Limitations

Despite this study's promising results, there are multiple challenges identified. Firstly, the unusually high correlation between PHQ-9 and GAD-7 (r = 0.97–0.99) indicates that the model may have collapsed these two questionnaires into a general wellbeing factor. This raises concerns about the validity of questionnaire responses and encourages future studies to make finer item-level distinctions between cognitive and somatic components to ensure discriminant validity, as stressed by Brattmyr et al. (2022). Secondly, the LLM exhibited higher variation in certain interviews, particularly those with Interviewee 3 in 2020 and Interviewee 4 in 2019, as reflected in larger 95% CIs and standard deviations for both PHQ-9 and GAD-7 scores. A potential explanation is the complexity of the interview content, where a participant discusses past burnout while also reporting current improvements, creating ambiguity for the model. This suggests that the source data itself, rather than the LLM, can be the cause of unstable predictions. This source data's effect on the reliability can be mitigated by calculating the confidence interval and separating the inconsistent results from the research.

Thirdly, the sample size of this research was limited to only four individuals and one LLM model (GPT-4o). This limits the generalizability of these results and calls for further research with bigger datasets and other LLM-based models to strengthen the validation of the potential use of LLMs for assessment and prediction of wellbeing by analysing the narrative data of transcripts. Fourthly, the LLM itself has practical model constraints, such as token limits and questionnaire knowledge gaps, that limit its capabilities to assess wellbeing from transcripts. The token limits set the maximum transcription length that can be given to the model during one iteration, limiting the lengths of the longitudinal interview transcripts. Furthermore, questionnaire knowledge gaps limit the use of simple zero-shot prompts because LLMs are missing the required knowledge to fill in questionnaires. For instance, more complex questionnaires, such as the DASS21 and BBI-15, required more explicit instructions and grounding to achieve reliable results. However, the advancement of LLMs will help with these limitations as the context windows are getting bigger, some already reaching one million tokens when GPT-4o had only 128 000 tokens, and their knowledge should deepen with increased training.

Finally, the rapid evolution of LLMs could also be a limitation for stable and reliable wellbeing assessment results. While this study demonstrates the capabilities of the current GPT-4o LLM model, the performance and potential biases of other LLM models may vary. Furthermore, the future updates to all LLM models might also affect these results, and hence, continuous re-evaluation will be necessary to ensure the ongoing reliability of this methodology.

5.2 Methodological Implications for IS Research

This study has multiple methodological implications for IS research. Firstly, by presenting how narrative data can be systematically transformed into standardized wellbeing indicators, bridging the gap between qualitative and quantitative analysis in IS research. Suggesting that simple wellbeing assessments, such as the PHQ-9 and GAD-7, are suitable for general wellbeing scoring with LLM-based analysis using a simple zero-shot approach. Secondly, the methodological implications of this study could have profound consequences for IS research because it would enable wellbeing assessment from archival qualitative datasets where questionnaires were not originally used. This means it could be used for all existing research interviews or transcription materials with minimal effort, creating vast amounts of new perspectives and research data. However, using archival qualitative datasets raises ethical considerations when using the data for purposes other than their originally intended use.

Thirdly, this methodological approach could also have an impact on existing and future research by reducing participant burden compared to repeated survey administration. This method can also offer an independent quantitative view, which might help to find interesting anomalies and add validity or raise mismatches to the claims from the qualitative analysis. Lastly, this is a demonstration of a pathway for IS scholars to integrate AI-driven tools into mixed-method research designs. It is important to present these new pathways of integrating AI-driven research tools because it is still in early stages for AI to be used in research, and these demonstrations help researchers to evaluate and understand the limitations and how AI could be effectively and ethically used in research. The AI-research toolset will expand in the following years, accelerating the pace of research and giving way to completely new approaches for analysing data and conducting research.

5.3 Directions for Future Research

The limitations of this study create several directions for future research. First, our research dataset was limited to only four individuals; a larger and more diverse samples are needed for generalization. Second, we tested one LLM with a simple zero-shot prompt. Thus, future studies should compare multiple models (e.g., GPT, Llama, Gemini, Claude) and prompting strategies to assess stability and performance. Third, other wellbeing questionnaires beyond PHQ-9 and GAD-7 should be explored. Fourthly, future research should directly pair transcripts with the real PHQ-9 responses of participants. This way, research could compare LLM-generated estimations against ground-truth questionnaire scores, thereby clarifying the model's true accuracy and potential biases. Finally, to address the extremely high correlation between PHQ-9 and GAD-7, future work should implement and examine item-level question prompting for both questionnaires to improve discriminant validity. Item-level prompting may encourage the LLM to develop a better understanding of each question, and thereby separate cognitive and somatic dimensions within the PHQ-9 and GAD-7, as highlighted by Brattmyr et al. (2022).

6 Conclusion

This study evaluates the reliability and methodological implications of utilizing a large language model (LLM) to transform workplace narratives into numerical results of wellbeing questionnaires (PHQ-9; GAD-7). Using simple zero-shot prompts and a longitudinal interview dataset of IT professionals, we demonstrate how an LLM can generate PHQ-9 and GAD-7 scores that align closely with prior qualitative analysis and exhibit high internal consistency. However, we also identified important limitations to this methodology. These limitations include variation in results, the small sample size, and the LLM's inability to differentiate between depression (PHQ-9) and anxiety (GAD-7) symptoms, an issue also observed in clinical research when item-level distinctions are not modelled (Brattmyr et al., 2022). Future research should incorporate item-level prompting and triangulation with human-validated questionnaires to improve discriminant validity, following the methodological guidance of Johnson, Gray, and Sarker (2019).

By creating this methodological demonstration, our study is contributing to the field of Information Systems (IS) research. Firstly, by showing how LLM can bridge qualitative narratives and quantitative measures for IS scholars, enriching their methodological toolkit to analyse existing and future transcript narratives in a scalable way. However, using this toolkit retrospectively raises ethical considerations on using data for purposes other than their originally intended use. Secondly, validating the reliability and limitation of this methodology. To give benchmarks and guidance on what needs to be taken into consideration when using an LLM for wellbeing estimation. Future research on methodology should expand to using larger, more diverse datasets, evaluate different LLM models and prompting strategies (e.g., item-level prompting), and use LLMs with other wellbeing questionnaires.

7 References

Ahuja, M. K., Chudoba, K. M., Kacmar, C. J., McKnight, D. H., and George, J. F. 2007. "IT Road Warriors: Balancing Work–Family Conflict, Job Autonomy, and Work Overload to Mitigate Turnover Intentions," *MIS Quarterly* (31:1), pp. 1–17. (doi: 10.2307/25148778)

Armstrong, D. J., Brooks, N. G., and Riemenschneider, C. K. 2015. "Exhaustion from Information System Career Experience: Implications for Turn-Away Intention," *MIS Quarterly* (39:3), pp. 713–728.

Bakker, A. B., and Demerouti, E. 2017. "Job Demands–Resources Theory: Taking Stock and Looking Forward," *Journal of Occupational Health Psychology* (22:3), pp. 273–285. (doi: 10.1037/ocp0000056)

Bakker, A. B., Demerouti, E., and Sanz-Vergel, A. 2023. "Job Demands–Resources Theory: Ten Years Later," *Annual Review of Organizational Psychology and Organizational Behavior* (10:1), pp. 25–53. (doi: 10.1146/annurev-orgpsych-120920-053933)

Bowling, N. A., and Jex, S. M. 2013. "The Role of Personality in Occupational Stress: A Review and Future Research Agenda," in *Handbook of Personality at Work*, N. Christiansen and R. Tett (eds.), New York, NY: Routledge, pp. 692–717.

Brattmyr, M., Lindberg, M. S., Solem, S., Hjemdal, O., and Havnen, A. 2022. "Factor Structure, Measurement Invariance, and Concurrent Validity of the Patient Health Questionnaire-9 and the Generalized Anxiety Disorder Scale-7 in a Norwegian Psychiatric Outpatient Sample," *BioMed Central Psychiatry* (22:1), pp. 1–13. (doi: 10.1186/s12888-022-04101-z)

Chuang, Y.-T., Chiang, H.-L., and Lin, A.-P. 2025. "Insights from the Job Demands–Resources Model: AI's Dual Impact on Employees' Work and Life Well-Being," *International Journal of Information Management* (83), pp. 1–12. (doi: 10.1016/j.ijinfomgt.2025.102887)

Demerouti, E., Bakker, A. B., Nachreiner, F., and Schaufeli, W. B. 2001. "The Job Demands–Resources Model of Burnout," *Journal of Applied Psychology* (86:3), pp. 499–512. (doi: 10.1037/0021-9010.86.3.499)

Glaser, B. G. 1992. *Emergence vs. Forcing: Basics of Grounded Theory Analysis*, Mill Valley, CA: Sociology Press.

Johnson, S. L., Gray, P., and Sarker, S. 2019. "Revisiting IS Research Practice in the Era of Big Data," *Information and Organization* (29:1), pp. 41–56. (doi: 10.1016/j.infoandorg.2019.01.001)

Khosravi, P., Rezvani, A., and Ashkanasy, N. M. 2020. "Emotional Intelligence: A Preventive Strategy to Manage Destructive Influence of Conflict in Large-Scale Projects," *International Journal of Project Management* (38), pp. 36–46. (doi: 10.1016/j.ijproman.2019.11.001)

Koo, T. K., and Li, M. Y. 2016. "A Guideline for Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine* (15:2), pp. 155–163. (doi: 10.1016/j.jcm.2016.02.012)

Kroenke, K., Spitzer, R. L., and Williams, J. B. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure," *Journal of General Internal Medicine* (16:9), pp. 606–613. (doi: 10.1046/j.1525-1497.2001.016009606.x)

Kumar, P. S. 2024. "Technostress: A Comprehensive Literature Review on Dimensions, Impacts, and Management Strategies," *Computers in Human Behavior Reports* (16), pp. 1–15. (doi: 10.1016/j.chbr.2024.100475)

Lalk, C., Targan, K., Steinbrenner, T., Schaffrath, J., Eberhardt, S., Schwartz, B., Vehlen, A., Lutz, W., and Rubel, J. 2025. "Employing Large Language Models for Emotion Detection in Psychotherapy Transcripts," *Frontiers in Psychiatry* (16), pp. 1–12. (doi: 10.3389/fpsyt.2025.1504306)

Liu, J., Gu, J., Tong, M., Zhang, Y., Peng, K., Wu, B., and Chen, W. Y. 2025. "Evaluating the Agreement Between ChatGPT-4 and Validated Questionnaires in Screening for Anxiety and Depression in College Students: A Cross-Sectional Study," *BioMed Central Psychiatry* (25:359), pp. 1–9. (doi: 10.1186/s12888-025-06798-0)

Mazur, A., Constantino, H., Tom, P., Wilson, M. P., and Thompson, R. G. 2025. "Evaluation of an AI-Based Voice Biomarker Tool to Detect Signals Consistent with Moderate to Severe Depression," *Annals of Family Medicine* (23), pp. 60–65. (doi: 10.1370/afm.240091)

Moore, J. E. 2000. "One Road to Turnover: An Examination of Work Exhaustion in Technology Professionals," *MIS Quarterly* (24:1), pp. 141–168. (doi: 10.2307/3250982)

Neumann, A. T., Yin, Y., Sowe, S., Decker, S., and Jarke, M. 2025. "An LLM-Driven Chatbot in Higher Education for Databases and Information Systems," *IEEE Transactions on Education* (68:1), pp. 103–116. (doi: 10.1109/TE.2024.3467912)

Newman, M. W. 2022. "Value Added? A Pragmatic Analysis of the Routine Use of PHQ-9 and GAD-7 Scales in Primary Care," *General Hospital Psychiatry* (79), pp. 15–18. (doi: 10.1016/j.genhosppsych.2022.09.005)

Oparina, E., Kaiser, C., Gentile, N., Tkatchenko, A., Clark, A. E., De Neve, J. E., and D'Ambrosio, C. 2025. "Machine Learning in the Prediction of Human Wellbeing," *Scientific Reports* (15), pp. 1–11. (doi: 10.1038/s41598-024-84137-1)

Pawlowski, S. D., Kaganer, E. A., and Cater, J. J. 2007. "Focusing the Research Agenda on Burnout in IT: Social Representations of Burnout in the Profession," *European Journal of Information Systems* (16:5), pp. 612–627. (doi: 10.1057/palgrave.ejis.3000699)

Polkinghorne, D. E. 2006. "An Agenda for the Second Generation of Qualitative Studies," *International Journal of Qualitative Studies on Health and Well-Being* (1:2), pp. 68–77. (doi: 10.1080/17482620500539248)

Rosenman, G., Wolf, L., and Hendler, T. 2024. "LLM Questionnaire Completion for Automatic Psychiatric Assessment," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 403–415. Y, Al-Onaizan., M, Bansal., and Y-N, Chen. (Editors). Publisher: Association for Computational Linguistics.

Shrout, P. E., and Fleiss, J. L. 1979. "Intraclass Correlation: Uses in Assessing Rater Reliability," *Psychological Bulletin* (86:2), pp. 420–428. (doi: 10.1037/0033-2909.86.2.420)

Spitzer, R. L., Kroenke, K., and Williams, J. B. 1999. "Validation and Utility of a Self-Report Version of PRIME-MD: The PHQ Primary Care Study," *Journal of the American Medical Association* (282:18), pp. 1737–1744. (doi: 10.1001/jama.282.18.1737)

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. 2006. "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Archives of Internal Medicine* (166:10), pp. 1092–1097. (doi: 10.1001/archinte.166.10.1092)

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., and Eichstaedt, J. C. 2024. "Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation," *Mental Health Research* (3:12), pp. 1–10. (doi: 10.1038/s44184-024-00056-z)

Teymoori, A., Gorbunova, A., Haghish, F. E., Real, R., Zeldovich, M., Wu, Y.-J., Polinder, S., Asendorf, T., Menon, D., and Steinbüchel, N. 2020. "Factorial Structure and Validity of Depression (PHQ-9) and Anxiety (GAD-7) Scales After Traumatic Brain Injury," *Journal of Clinical Medicine* (9:3), pp. 1–15. (doi: 10.3390/jcm9030873)

Walsham, G. 2006. "Doing Interpretive Research," *European Journal of Information Systems* (15:3), pp. 320–330. (doi: 10.1057/palgrave.ejis.3000589)

Appendix 1

Patient Health Questionnaire and General Anxiety Disorder
(PHQ-9 and GAD-7)

Date_____ Patient Name:_____ Date of Birth: _____

Over the last 2 weeks, how often have you been bothered by any of the following problems?
Please circle your answers.

PHQ-9	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
1. Little interest or pleasure in doing things.	0	1	2	3
2. Feeling down, depressed, or hopeless.	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much.	0	1	2	3
4. Feeling tired or having little energy.	0	1	2	3
5. Poor appetite or overeating.	0	1	2	3
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down.	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television.	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual.	0	1	2	3
9. Thoughts that you would be better off dead, or of hurting yourself in some way.	0	1	2	3
Add the score for each column				

Total Score (add your column scores): _____

If you checked off any problems, how difficult have these made it for you to do your work, take care of things at home, or get along with other people? (Circle one)

Not difficult at all **Somewhat difficult** **Very Difficult** **Extremely Difficult**

Over the last 2 weeks, how often have you been bothered by any of the following problems?
Please circle your answers.

GAD-7	**Not at all sure**	**Several days**	**Over half the days**	**Nearly every day**
1. Feeling nervous, anxious, or on edge.	0	1	2	3
2. Not being able to stop or control worrying.	0	1	2	3
3. Worrying too much about different things.	0	1	2	3
4. Trouble relaxing.	0	1	2	3
5. Being so restless that it's hard to sit still.	0	1	2	3
6. Becoming easily annoyed or irritable.	0	1	2	3
7. Feeling afraid as if something awful might happen.	0	1	2	3
Add the score for each column				

Total Score (add your column scores): _____

If you checked off any problems, how difficult have these made it for you to do your work, take care of things at home, or get along with other people? (Circle one)

Not difficult at all **Somewhat difficult** **Very Difficult** **Extremely Difficult**

UHS Rev 4/2020

Developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues, with an educational grant from Pfizer Inc.

Appendix 2

Field	Value
LLM Provider	OpenAI
Exact model name & version	gpt-4o-2024-11-20
Access dates (UTC)	2025-05-28 to 2025-06-02
Context window (tokens)	128 000
Tokenization	tiktoken/bpe
Inference parameters	temperature=1, top_p=1, presence_penalty=0.0, frequency_penalty=0.0, seed=not supported
Prompt template IDs	PHQ-9:[PHQv1], GAD-7:[GADv1] (see Appendix 3)
Replicated runs per transcript	k = 20; independent calls per model with fixed params

Appendix 3

Select either PHQ-9 or GAD-7 questionare and remove the other one from this prompt template:

Can you fill out the PHQ-9/GAD-7 form (Patient Health Questionnaire-9)/(Generalized Anxiety Disorder 7) for each of these interviews and give the scores in the following json format?

```
{
  "questionnaire": "PHQ-9/GAD-7",
  "items": [
    {
      "interview_date": "YYYY-MM-DD",
      "total_score": integer, # this is the sum of questionnaire scores
      "rationale": "1–2 sentence justification grounded in evidence"
    }
  ]
}
```

Interview transcripts:

<<<BEGIN_TRANSCRIPT

{{ANONYMIZED_TRANSCRIPT_TEXTS}}

END_TRANSCRIPT>>>

Output strictly as JSON with the schema provided.

Acknowledgements

We are very thankful for our informants for their willingness to tell their story over time. This work was supported by Adalyon Ltd, which provided funding for Riko Nyberg's research activities.

Copyright