

Workers' Compensation Psychosocial Recovery Journeys: A Comparison of Large Language Model Insights

John McMahon, Ashley Craig, Ian Douglas Cameron

Submitted to: JMIR AI
on: September 20, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 19

 Multimedia Appendixes..... 20

 Multimedia Appendix 1..... 20

 Multimedia Appendix 2..... 20

 Multimedia Appendix 3..... 20

Workers' Compensation Psychosocial Recovery Journeys: A Comparison of Large Language Model Insights

John McMahon¹ BPSYCH(HON), DPSYCH (CLINICAL); Ashley Craig¹ BSC(HON), PhD; Ian Douglas Cameron¹ MBBS, PhD

¹John Walsh Centre for Rehabilitation Research Kolling Institute of Medical Research The University of Sydney St Leonards AU

Corresponding Author:

John McMahon BPSYCH(HON), DPSYCH (CLINICAL)
John Walsh Centre for Rehabilitation Research
Kolling Institute of Medical Research
The University of Sydney
10 Westbourne Street
St Leonards
AU

Abstract

Background: Artificial intelligence is a potentially useful tool to derive insights from large data sets of both structured and unstructured data. The psychosocial recovery journey of injured workers in compensable environments is complex and Large Language Models (LLMs) are tools that can illuminate relationships otherwise indiscernible.

Objective: The aim of this study was to model factors important in recovery from injury in a compensable environment, compare two service delivery modes for a support service, and use 3 LLMs to derive insights from a data set comprising of psychometric test data and free form notes from telephone coaching interviews of injured workers by generating recovery personas and personas for feigning spectrum behaviour (FSB).

Methods: We used Claude 3.5 Sonnet, ChatGPT-4-Turbo and Gemini 1.5 Pro to derive insights from a deidentified data set of 7472 cases who were coached and treated by the Navigator Support Program between 17th June 2020 and 1st May 2024. Each LLM was used to generate natural recovery personas, personas for each month of the recovery course, and personas for different treatment coaching conditions being: "routine", "escalated", "intervention" with therapy, and also "feigning" personas. Two versions of the program were compared: "support only" and "support and therapy".

Results: Claude 3.5 generated 7 natural personas, ChatGPT-4-Turbo generated 6 natural personas, and Gemini 1.5 Pro generated 7 natural persons. The models used different features and were "sentiment" based, "functional recovery" based, and "adaptation and support" based respectively, with only moderate overlap between ChatGPT-4-Turbo and Gemini 1.5 Pro (0.3994). Forcing the LLMs to generate personas for coaching conditions month on month showed high similarity (0.991 – 0.979) in persona generation. Five sub-personas were identified for FSB showing the diversity of this behaviour.

Conclusions: Systematic modeling with LLMs show the models utilize different features and derive significantly different insights. There were slight differences in complexity of cases in service delivery modes, which likely reflected the differences in the insurer portfolios.

(JMIR Preprints 20/09/2025:84481)

DOI: <https://doi.org/10.2196/preprints.84481>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.



Original Manuscript

Original Paper

Workers' Compensation Psychosocial Recovery Journeys: A Comparison of Large Language Model Insights

John E McMahon¹, BPSY(HON) DPSY; Ashley Craig¹, BSC(HON) PHD; Ian D Cameron¹, MBBS PHD

¹ John Walsh Centre for Rehabilitation Research, Kolling Institute, School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, St. Leonards, NSW 2065, Australia.

Corresponding Author:

John E McMahon, BPSY(HON), DPSY (CLINICAL)
John Walsh Centre for Rehabilitation Research, Kolling
Institute of Medical Research, School of Health Sciences,
Faculty of Medicine and Health, The University of Sydney,
St. Leonards,
NSW 2065 Australia
Phone: +61428460664
Email: drjohnmcmahon@icloud.com

Abstract

Background: Artificial intelligence is a potentially useful tool to derive insights from large data sets of both structured and unstructured data. The psychosocial recovery journey of injured workers in compensable environments is complex and Large Language Models (LLMs) are tools that can illuminate relationships otherwise indiscernible.

Objective: The aim of this study was to model factors important in recovery from injury in a compensable environment, compare two service delivery modes for a support service, and use 3 LLMs to derive insights from a data set comprising of psychometric test data and free form notes from telephone coaching interviews of injured workers by generating recovery personas and personas for feigning spectrum behaviour (FSB).

Methods: We used Claude 3.5 Sonnet, ChatGPT-4-Turbo and Gemini 1.5 Pro to derive insights from a deidentified data set of 7472 cases who were coached and treated by the Navigator Support Program between 17th June 2020 and 1st May 2024. Each LLM was used to generate natural recovery personas, personas for each month of the recovery course, and personas for different treatment coaching conditions being: "routine", "escalated", "intervention" with therapy, and also "feigning" personas. Two versions of the program were compared: "support only" and "support and therapy".

Results: Claude 3.5 generated 7 natural personas, ChatGPT-4-Turbo generated 6 natural personas, and Gemini 1.5 Pro generated 7 natural persons. The models used different features and were "sentiment" based, "functional recovery" based, and "adaptation and support" based respectively, with only moderate overlap between ChatGPT-4-Turbo and Gemini 1.5 Pro (0.3994). Forcing the LLMs to generate personas for coaching conditions month on month showed high similarity (0.991 – 0.979) in persona generation. Five sub-personas were identified for FSB showing the diversity of this behaviour.

Conclusions: Systematic modeling with LLMs show the models utilize different features and derive significantly different insights. There were slight differences in complexity of cases in service delivery modes, which likely reflected the differences in the insurer portfolios.

JMIR AI 2025;4:e76056; doi:

Keywords: artificial intelligence; large language models; psycho-social factors; insured injury; workers compensation; feigning

Introduction

Evans (1994) found after the first month that 785 of people with insured injuries found medical symptoms the greatest stressor and at month 24 insurers' failure to honor financial obligations was the greatest stressor [1]. People with compensable injuries find they are ensconced in a complex multistakeholder environment, which renders them subject to stereotypes, expectations about recovery, "nocebo" effect where adverse reactions are expected from benign events, and where there is confusion between injustice and entitlement which can produce symptom magnification [2]. We previously generated cut scores for identifying feigning spectrum behaviour (FSB) in injured workers [3] on the Orebro Musculoskeletal Pain Screening Questionnaire-Short Form (OMSPQ-SF) [4]. Large scale studies aggregating data across multiple medicolegal and forensic assessment providers show that FSB occurs in 31% of chronic pain assessments and 29% of personal injury cases [5]. Yet no large-scale study has attempted to examine systematically the possible motivations and circumstances that FSB occurs beyond its base rate.

There have been many attempts to examine adjustment to injury and psychosocial factors in recovery [6-8]. Previous research has been hampered by small sample sizes or were systematic literature reviews of small sample size studies, and none have systematically considered the effects of FSB and accommodated the effects in their findings. A recent paper by Collie et al determined the prevalence and predictors of psychological distress in a large sample of 3755 workplace injured people [9] consisting of 3160 with a musculoskeletal injury and 595 with a primary psychological condition. They found 27.5% suffered moderate distress and 14% experienced severe distress as measured by the Kessler-6 scale. Multivariate ordinal logistic regression, a statistical technique, showed absence from work, malaise, poor work efficacy, financial duress, difficult interactions with health professionals, and having a diagnosis were features contributing to psychological distress. Yet only 27.2% reported accessing a mental health service. This demonstrates the need for the systematic assessment of psychological distress for people suffering injuries within the worker's compensation scheme. Notably, Collie et al identified the need for support to navigate the claims process, that conflict and contact with the insurer contributed to stress, and that interactions with healthcare providers could be stressful for injured people.

In 2017, the first author co-founded Navigator Group and developed the Navigator Support Program. Described as an "independent support service for people injured in the workplace", the program was successfully piloted and refined to systematically address psychosocial barriers to recovery of the type identified by Collie et al [9]. Trust is associated with improved employee

performance [10], job satisfaction [11-12], and general economic factors like investment [13] and savings within and between organizations [14]. The philosophy of the program was a series of "trust" transactions: insurers/scheme agents trust the program and did not require reporting apart from aggregated data and the worker's assumed that their personal information would not be disclosed in ways that may compromise their claim. At the time of the data collection period the program was Australia wide and utilized by several of the main insurers/scheme agents in Australia [15]. The standardized assessment and coaching process facilitated systematic data collection which was enhanced by use of the Salesforce Health Cloud environment by the telephone operatives.

AI is being increasingly used in mental health, including to improve client engagement and user experience, through use of tools such as chatbots, and predictive modeling techniques [16]. Amirian et al used LLMs to research the prevalence of psychosocial disability and rehabilitation through computational text mining [17]. Personas are short natural language descriptions of properties and behaviour generated from data [18]. The objective of our use of AI was to systematically apply the LLMs Claude 3.5 Sonnet [19], ChatGPT-4-Turbo [20] and Gemini 1.5 Pro [21] to the data set to identify personas that inform on the recovery experience of injured workers, excluding workers with FSB and developing insights into it.

Methods

Overview

The Navigator Support Program applies psychometric measures known to predict maladjustment to injury, the Orebro Musculoskeletal Pain Screening Questionnaire-Short Form (OMPSQ-SF), and the Perceived Stress Scale-4 (PSS-4). Measures were embedded periodically at the end of a semi-structured interview which allowed the telephone operative to code information into Sales Force Health Cloud as open text (OT); 3-point rating scale (3PRS) of "unhelpful", "ok" and "good"; date; or a rating out of 10 (x/10) and which assessed: the persons location at the time of the call (OT), how their day has been (OT), are they getting better (OT), what was their injury (OT), what medication they were taking (OT), what treatment they were having (OT), how do they find their doctor (3PRS), when did they last see their doctor (date), do they have any other health problems (OT), when will they see their doctor next (date), how do they find their case manager (3PRS), have they spoken with their employer (OT), have they spoken

with coworkers (OT), were they seeing a therapist(OT), do they have social/home support (OT), how they were going financially (x/10) and an open prompt “is there anything else you would like to talk about today?”(OT) and a general “case notes” field. The OMPSQ-SF is a 10-item test for features that are known to predict maladjustment to injury recording answers as x/10. The PSS-4 is a four-item test that assesses stress over the last month and rates the frequency from “never” to “very often”. The aim was that the enquiries would be administered in a conversational style to engender trust and engagement. A key performance indicator (KPI) for the operatives was that the OMPSQ-SF and PSS would be administered at least once every 60-days.

The telephone operatives were qualified counsellors and intern psychologists, most of whom had worked on support lines previously. There was a structured induction and training program. The operatives had team and individual supervision to ensure consistent quality of service delivery. In the “support” function, when issues were identified, the telephone operatives coached the injured worker on assertiveness skills [22], structured problem solving [23], worry management techniques[24], goal setting [23], activity planning [24], and used motivational interviewing techniques to address barriers to recovery and substance use issues [25], and role-played conversations with doctors, case managers, employers and social supports [22]. There were three levels of support “routine” (R) in which the injured person as contacted monthly, “escalated” (ESC) in which the injured person was contacted again within 3-days due to psychosocial challenge, and “intervention” (INT) where the person was formally referred for treatment by a psychologist or counsellor; generically referred to as “therapy”. In the “therapy” function, a different telephone operative conducted structured and directed cognitive-behavioural therapy. This included psychoeducation about the survival cascade and stress [26], injury adjustment involving cognitive restructuring [27], relaxation training [24], grounding techniques [28], imaginal exposure for trauma [24], and graded exposure

therapy [24]. The “therapy” was conducted by external service providers in the “support only” (SO) version of the support program and “in house” in the “support and therapy” (S+T) version of the program. Different insurers opted for different versions of the program.

The data was recorded from 17th June 2020 to 1st May 2024 when all people and information systems were working optimally. The manager of the Navigator Support Program exported the cases for that epoch and a de-identified data set was supplied to the researchers. Preliminary statistical comparisons were made between the SO and S+T program variants. The cut score of OMPSQ-SF > 70 [3] was applied to separate out instances of feigning for initial persona generation using Claude 3.5, ChatGPT-4-Turbo, and Gemini 1.5 Pro. Each

LLM was prompted with “Using other scores and unstructured data, see if there are different personas for people scoring above 70 on the OMPSQ-SF”. Various features of these personas were calculated such as employer engagement, trust metrics and rating of doctor and case manager. These FSB Personas were then used to further clean the remaining data of probable feigners. The LLMs were then used to generate recovery Personas on the genuine responding data set. The prompt used was “Examine this data and see if there are different recovery personas”. This generated naive personas (NPs), generated purely by the LLMs. These personas were compared for sentiment using Valence Aware Dictionary and sEntiment Reasoner (VADER), with semantic similarity and clustering methods, intra-model cohesion, pairwise cosine similarity among personas, and a global silhouette score using cosine distance, and lexical overlap via Jaccard indices [28-34]. Personas were then generated monthly for R, ESC, INT and also for Discharge (DISC) for each month of the program. Using the same method as for the NPs month-to-month personas were compared for sentiment etc. Short 60-word summary versions of the personas were generated for publication within supplementary material.

Ethical Considerations

The University of Sydney Research Integrity & Ethics Administration Human Research Ethics Committee 3 approved this study under cover letter dated 28th June 2019 (HREC3, 2019/136). Oral informed consent was obtained for the use of responses to the telephone interview of injured persons for research purposes with the informed consent statement: “We are researching our support service. We do not use any personally identifying information in the research, just the ratings from the questions we ask. You can decline and it does not affect the service we will provide to support you, your inclusion in the study is voluntary. May we have your permission to use your responses to questions to improve the quality of our calls, train our staff, and for research purposes?”

Results

Only consenting calls were exported and then number of non-consenting cases is unknown. There were 7472 cases in the data set generating 1,048,795 rows of data, with 4572 cases in SO and 2900 in S+T. Overall, there was 7.88% data completion, but much of the missing data included instances in which psychometric screening was not conducted and redundant administrative data. Little’s Missing Completely at Random [25] was rejected ($P < .001$). However, imputation was not conducted due to the significance of the use of OMPSQ-SF, unreliability of imputation at such scale, and use of LLMs. Each case received an average of 3.75 sessions (min = 1, max = 48). After removing side of injury and consolidating overlapping terms and cleaning idiosyncratic responses, the most common injuries can be seen in Table 1 along with

demographic differences, differences in pre-existing mental health issues and mean OMPSQ-SF and PSS-4 scores. Pre-existing health issues that were recorded at the first session included high blood pressure (1%), diabetes (0.9%), asthma (0.5%) and arthritis (0.4%). 17.1% of people were explicitly recorded as not having a pre-existing health problem. In total 32.7% of the sample had a record about their previous health status. The S+T group was slightly more complex in presentation, reflecting the more demanding portfolios of the insurers utilizing this program. There were no significant differences in occupation between S+T and SO, both having top occupations of truck driver, laborer and cleaner, with slightly different order of proportions of other occupations including: support worker, carpenter, electrician, disability support worker, machine operator and process worker.

Table 1. Demographic, Injury, Pre-existing Mental Health and Psychometric Differences between Support Only (SO) and Support and Therapy (S+T) Groups.

Feature	SO	S+T
Age	42.7 (14.24) [17-79]	44.6 (13.24) [17-85]**
Percent Female	30.5%	32.9%*
Back (n=1505)	20.10%	22.97%*
Knee (n = 864)	11.61%	10.12%*
Shoulder (n=854)	11.41%	11.37%
Ankle (n=539)	7.22%	5.72%*
Finger (n=468)	6.33%	6.30%
Wrist	6.00%	4.97%*

(n=449)		
Hand	4.80%	6.15%*
(n=358)		
Pre-existing Mental Health Issue	2.1%	3%*
Initial OMPSQ-SF Mean Score	41.2 (3.0)	44.3 (3.1)**
Initial PSS-4 Mean Score	5.56 (0.34)	5.89 (0.51)**

(Standard Deviation), [Range], *p<0.05, **p<0.001

Over successive three-month periods from onboarding the retention rate across the program declined rapidly with 14.2%, 5.3%, and 0.7% of the onboarded sample requiring the program respectively. This shows that the majority of people recovered and were discharged (DISC), with more complex cases persisting. This is reflected in the relative stability of the month-on-month DISC rate (26%-33%) and increasing ESC rate (21%-39.8%) over time. Figure 1 shows the reported proportion of clinical interventions over time, showing conservative treatments were the most common treatment type proportionately. Figure 2 shows proportion of satisfaction ratings for doctor and case manager over time, and engagement with employer over time. There is a downward trend in satisfaction in relationships with stakeholders over time, although the INT shows fluctuations in later months reflecting the small number of cases with return to work as the focus of intervention. Despite higher baseline pain, S+T had higher DISC rates (33.9%-51.9%) compared to SO (20.5%-33.9%) showing S+T generated better long term-outcomes. However, for clients that did not DISC OMPSQ-SF were comparably high (SO: $x=60.6, SD=7.3$, S+T: $x=61.9, SD=7.5$).

Figure 1 Proportion of treatment types over time.

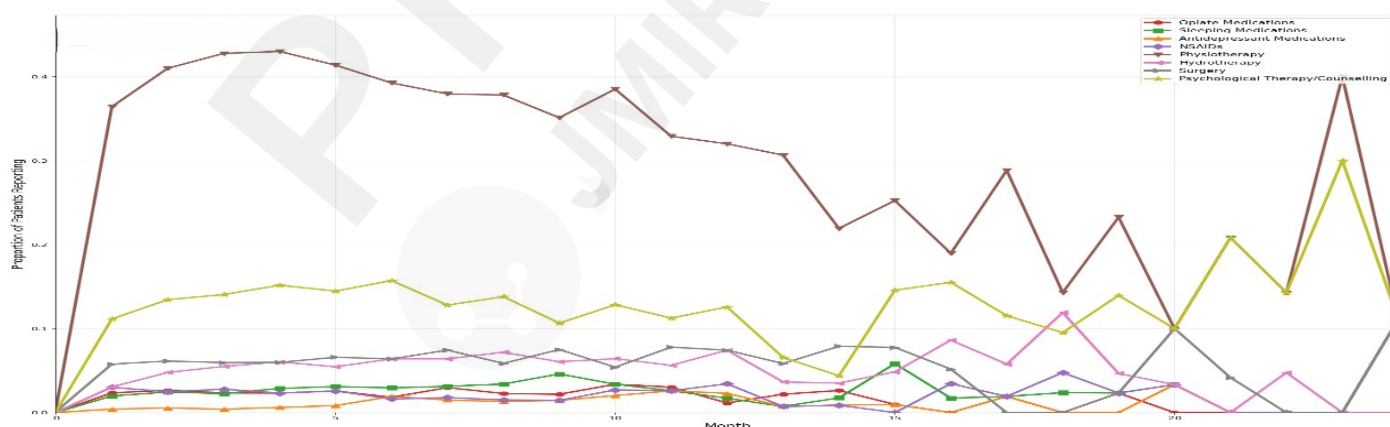


Figure 2 Proportion of satisfaction over time with doctor and case manager, and contact with employer



Table 2. FSB Personas generated by Claude 3.5, ChatGPT-4-Turbo and Gemini 1.5 Pro.

LLM	Persona	Prevalence within FSB	OMPSQ- % SF	Age Range (Years)	LLM Suggested Intervention
Claude 3.5 Sonnet	High-Severity Functional Decline	16.1	98-98	Variable	Multidisciplinary pain management
	Moderate-Severe Adaptive Copers	26.9	76-85	Variable	CBT & Graduated Activity
	Early-Stage High-Risk	24.9	71-78	Variable	Early intensive rehab
	Chronic Pain Syndrome	14.5	80-92	Variable	Chronic pain management
	Occupational Injury Complex	12.4	74-86	Variable	Legal liaison
	Rural/Access-Limited	5.2	72-89	Variable	Telehealth services
ChatGPT-4-Turbo	Severe Pain Overwhelmed	19.7	85-98	35-60	Intensive pain management
	Middle-Aged Work Warriors	30.1	75-84	40-55	Work-focused rehab
	Young Adult Strugglers	21.8	71-80	20-35	Early intervention
	Pre-Retirement Complex	18.1	76-88	55+	Geriatric assessment
	System Disengaged	10.4	78-94	Variable	Motivational interviewing
Gemini 1.5 Pro	Severe Multi-Dimensional Impairment	24.4	85-98	45±12	Intensive case management
	Moderate-High Adaptive Potential	40.4	74-84	42±10	Structured rehab
	Early-Intervention Responsive	26.9	71-79	35±8	Rapid access programs
	Complex Psychosocial Barriers	8.3	76-92	Variable	Motivational enhancement
	Severe Multi-Dimensional Impairment	24.4	85-98	45±12	Intensive case management
Combined Models	Severe Functional Impairment	15-20	85-94	40-55	Intensive Multidisciplinary Management
	Moderate-High Adaptive Copers	40-50	75-84	Variable	Work-Focused CBT
	Early-Stage High Risk	25-30	71-74	20-35	Early Prevention
	Chronic Complex Cases	10-15	70+	55+	Specialized Chronic Pain Program
	System Disengaged/Access-Limited	5-10	85-98	Variable	Improve Access to Treatment

There were 1119 classifications of FSB on the basis OMPSQ-SF >70. Table 2 contains the LLM personas generated. The classification of FSB showed an upward trend over time in both SO ($r=0.626, P<.001$) and S+T ($r=0.556, P=.006$) with detection rates of 5.4%/9.3%, 16.3%/21.9%, and 25%/30.2%, respectively over 0-5months, 6-11months and month and 12-months epochs. For a detailed account of each FSB persona please see LLM FSB PERSONA ANALYSIS REPORT in the Supplementary Materials. In summary: (1) Claude 3.5 Sonnet focused on evidence-based clinical reasoning with a biopsychosocial framework. (2) ChatGPT-4-Turbo focused on behavioural patterns, demographics and work-related features. (3) Gemini 1.5 Pro focused on machine learning clustering with feature optimization.

Comparing sentiment for session notes immediately prior (VADER Compound = -0.036 ± 0.533) to the feigning session (VADER Compound = -0.336 ± 0.537) and subsequent session (VADER Compound = -0.090 ± 0.565) showed significant

deterioration in sentiment at the feigning session compared to the prior session ($t\text{-test}=9.020, P<.001$) and some recovery of sentiment in the subsequent session ($t\text{-test}=-7.912, P<.001$). Complaints peaked during the feigning period (0.71 per note vs 0.51 previous), while mentions of interventions remain relatively stable.

The most common terms during feigning include "surgery," "pain," "stress," "anxiety," and "frustrated." Comparing the personas between the LLMs using Jaccard Similarity Scores showed different approaches by the LLMs: Claude 3.5 v ChatGPT-4-Turbo (0.192), Claude 3.5 v Gemini 1.5 Pro (0.032), and ChatGPT-4-Turbo v Gemini 1.5 Pro (0.031). Using only data other than OMPSQ-SF Claude 3.5 Personas performed best at identifying 25% of FSB, ChatGPT-4-Turbo identified 13.8% of FSB, and Gemini only 1.1% of FSB.

Using the three LLMs to synthesis the 16 FSB personas generated on the basis of the descriptions and using Claude 3.5 Sonnet to synthesize terms to maximise semantic similarity, five combined personas were created

as reported in Table 2 as “Combined Models”. See the LLM PERSONA ANALYSIS REPORT in the supplementary materials for a detailed account of these personas. These combined personas were used to see if there was a relationship between FSB and “trust” inferred from relationship “goodness” ratings. Figure 3 shows the relationships between these FSB combined models and ratings of doctor, case manager and engagement with employer as indicators of trust. The System Disengaged/Access-Limited combined subtype are characterized by very low levels of trust as indicated by proportion of goodness of positive relationship, although 16.7% report engagement with their employer. A critical insight is that the LLM personas identify cases with negative sentiment, however, some FSB reported positive sentiment.

Only 4-cases were supported beyond 24-months. Two cases had ongoing complex injuries, one case had a complex cervical spine injury with a collateral diagnosis of post-traumatic stress disorder, and one long-term shoulder injury with extended treatment. These were excluded from subsequent analyses as were the 1192 cases engaging in FSB. This left a cohort of 6954 genuine responders for further analysis.

Systematic recovery persona generation with the LLMs resulted in 20-personas (Claude 3.5 Sonnet = 7, ChatGPT-4-Turbo = 6, Gemini 1.5 Pro = 7). Using multiple similarity methods Gemini 1.5 Pro showed highest internal similarity (average Cosine Similarity = .699), ChatGPT-4-Turbo personas demonstrated strong clustering (average Silhouette coefficient = .647), Claude 3.5 Sonnet personas showed moderate internal cohesion (average Within Cluster Sum of Squares = .529). Cross-model overlap analysis of 190 pairs of personas showed they only had modest common ground and remain more different than alike (best average Cosine Similarity=0.399). Jaccard Similarity was highest within Gemini 1.5 Pro personas but this was only moderate (.522) showing a balance of overlap and uniqueness. Term Frequency-Inverse Document Frequency cosine (TF-IDF) showed strongest clustering was again within Gemini 1.5 Pro personas (.661), but otherwise TF-IDF within LLMs was low ($x=.251$) and cross-LLM comparisons were very low (0.147-0.158) suggesting the personas have strong discrimination between language constructs and the LLMs use distinct vocabularies and terminology. Jensen-Shannon comparison across all pairs yielded moderate overlap (average = .4778) showing balanced discrimination, with models being more varied in content than similar and having moderate overlap. The within-model similarity suggests each LLM captures distinct aspects of recovery, and inspection showed there were common temporal progression patterns that are consistent between the models. Ultimately, the models are most likely complementary rather than overlapping. The models used different approaches, with the personas generated by Claude 3.5 Sonnet personas showing sentiment-based

recovery patterns, ChatGPT-4-Turbo personas emphasising functional recovery patterns, and Gemini 1.5 Pro personas focusing on adaptation and support patterns. Gemini 1.5 Pro did, however, fail to classify 28.2% of the cases, instead, referring to them as “other” and not providing characteristics. Table 2 contains the LLM Generated Recovery Personas and see the LLM Recovery Personas in the supplementary materials for detailed persona narratives. Figure 3 contains doctor and case manager ratings and employer contact as indicators of trust across LLM models. Claude 3.5 Sonnet and Gemini 1.5 Pro models show differentiation on these ratings.

327 comparisons were made comparing each intervention level month-on-month to see if there were differences between the themes resulting in a change of support level. The LLMs generated very high similarity across the comparisons month-on-month of 98.4% (average Cosine Similarity=.984, range=0.957-0.996). By comparing recovery trajectory within DISC, R, ESC, and INT over time. R showed significantly but only slightly less pain (Slope=-0.066, $r^2=.393$, $P<.01$), anxiety (Slope=-0.048, $r^2=.383$, $P<.05$), and improvement rating (Slope=-0.026, $r^2=.913$, $P<.001$) month-on-month. INT showed significantly but slightly reduced improvement rating (Slope=-0.015, $r^2=.764$, $P<.001$). Comparing LLMs there was only one significant difference between narratives in utilization of anxiety level in ESC personas (Difference = -.040, Cohen’s $d = -0.984$, $P<.05$). These differences were considered minor, and inspection of month-on-month personas showed that apart from averages changing for age, gender proportion and recovery duration, there were negligible differences.

Table 2 LLM Generated Recovery Personas

LLM	Persona	Prevalence %	Age (SD)	%Female	Average Months to DISC (SD)	Important Model Features
Claude 3.5 Sonnet	Early Positive Responder	76.4	42.3 (14.3)	32.7	1.8 (1.7)	Sentiment of notes = 0.269 (SD =.110)
	Early Neutral Responder	15.0	42.5 (13.6)	33.8	1.6 (1.8)	Sentiment of notes = 0.035 (SD =.049)
	Mid Positive Responder	5.6	48.4 (13.5)	35.2	8.7 (1.6)	Sentiment of notes = 0.258 (SD =.096)
	Mid Neutral Responder	1.1	48.0 (14.9)	39.2	8.8 (1.7)	Sentiment of notes = 0.037 (SD =.055)
	Late Positive Responder	1.0	50.2 (11.8)	41.1	16.1 (2.9)	Sentiment of notes = 0.238 (SD =.105)
	Late Neutral Responder	0.3	53.3 (12.5)	36.4	17.8 (3.3)	Sentiment of notes = 0.019 (SD =.065)
	Baseline Neutral Responder	0.3	38.5 (12.3)	35.0	0.0	Sentiment if notes = -0.151 (SD = .031)
ChatGPT-4-Turbo	Early Functional Responder	64.6	42.4 (14.1)	32.8	1.8 (1.7)	100.0% of communications mention work/employment.
	Early Stable Responder	27.3	42.2 (14.3)	33.1	1.6 (1.8)	Maintaining stable work capacity while managing ongoing recovery needs.
	Mid Functional Responder	4.6	48.4 (13.1)	33.9	8.7 (1.6)	Sustained work rehabilitation, demonstrating ongoing occupational adaptation
	Mid Stable Responder	2.1	48.2 (15/0)	40.3	8.8 (1.7)	Sustained work rehabilitation, demonstrating ongoing occupational adaptation
	Late Functional Responder	1.0	49.7 (12.9)	37.7	16.2 (3.0)	In long-term occupational management showing established work-life balance strategies.
	Late Stable Responder	0.4	53.0 (10.3)	48.1	16.9 (3.3)	In long-term occupational management showing established work-life balance strategies.
Gemini 1.5 Pro	Early Supported Responder	38.2	42.0 (14.4)	33.7	1.7 (1.7)	100.0% of communications mention support/family/help. 11.0% of communications include adaptation/coping terms
	Other	28.2				
	Early Engaged Responder	16.0	42.6 (13.7)	35.7	1.9 (1.8)	82.3% of notes mention support/family/help. 18.8% of notes include adaptation/coping terms
	Baseline Brief Responder	12.6	41.5 (14.0)	29.9	0.0 (0.0)	9.7% of notes mention support/family/help. 7.9% notes include adaptation/coping terms
	Mid Supported Responder	2.6	49.0 (13.1)	38.8	8.8 (1.7)	100% of notes mention support/family/help. 10.4% notes include adaptation/coping terms
	Mid Engaged Responder	1.6	47.4 (13.5)	34.5	8.6 (1.5)	81.8% of notes mention support/family/help. 21.8% notes include adaptation/coping terms
	Late Supported Responder	0.5	48.1 (14.2)	39.5	16.2 (3.4)	100% of notes mention support/family/help. 23.7% notes include adaptation/coping terms
	Late Engaged Responder	0.3	51.9 (12.5)	50.0	15.8 (2.4)	94.4% of notes mention support/family/help. 27.8% notes include adaptation/coping terms

Therefore, the persona generation process was conducted by each LLM for Recovery Status. 12-recovery personas were concluded and compared using Jaccard Similarity metrics. Claude 3.5 Sonnet and ChatGPT-4-Turbo personas had 40.4% similarity, Claude 3.5 Sonnet and Gemini 1.5 Pro personas had 64.6% similarity and ChatGPT-4-Turbo and Gemini 1.5 Pro personas had 29.6% similarity, suggesting that the LLMs employed complementary approaches. Feature analysis showed “trust”, the combined goodness of doctor and case manager

ratings, emerged as a significant feature in 11 of the 12 personas. ChatGPT-4-Turbo had a unique focus on emotional support. R, INT, and DISC all show LLM agreement on trust and ESC had least convergence between LLMs. Across all Navigation Status OMPSQ-SF and PSS-4 were significantly and moderately correlated ($r = .3682-.506$, $P < .001$). Higher stress and pain were associated with lower trust ratings, and while statistically significant ($P < .001$) this explained little of the variance ($r^2 = .022-.001$). Across all Navigation Status personas

relationship with doctors were rated as better than with case managers with small effect size for R, ESC, and INT (Cohen's $d=0.412-0.493$) and medium effect

size at DISC (Cohen's $d=0.541$). The INT had lowest trust, highest stress and pain, and moderate social support with a high degree of variability.

Figure 3 FSB Combined personas relationship ratings

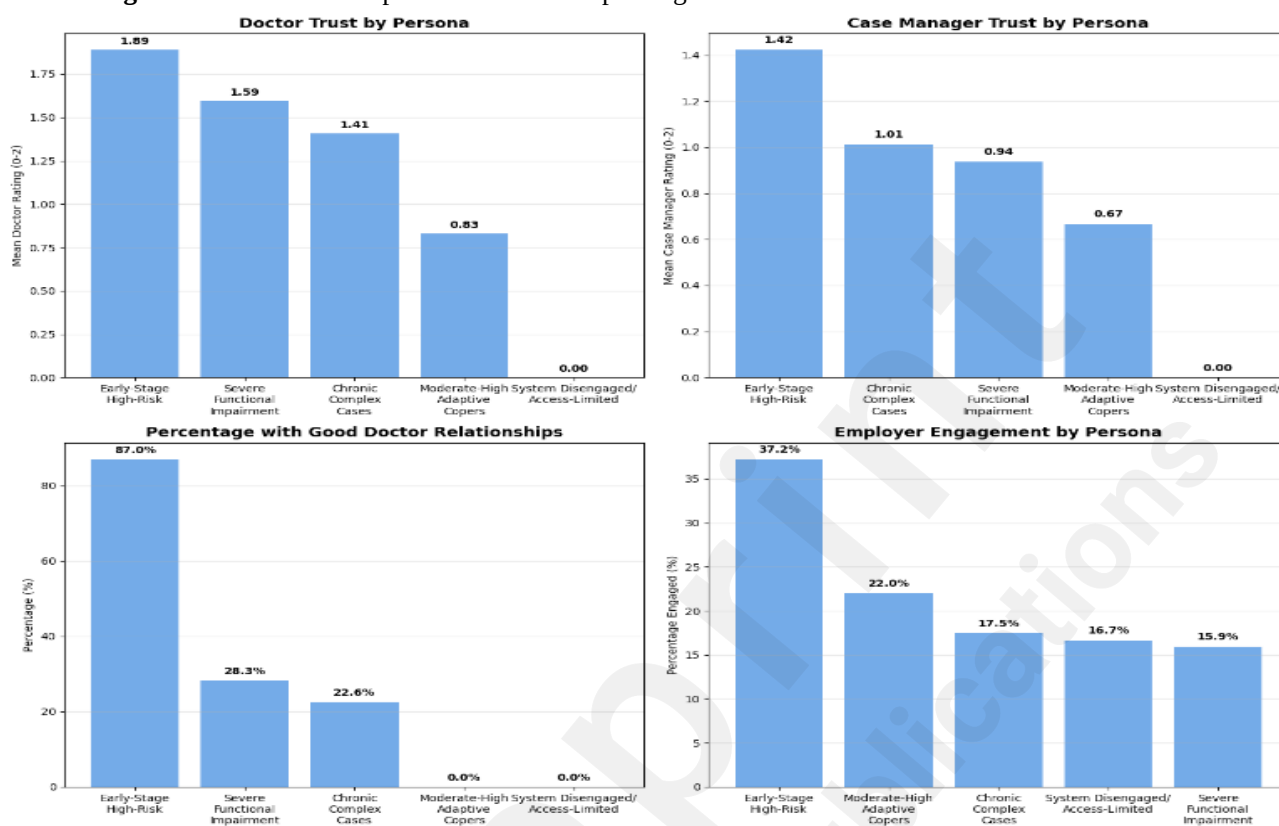


Figure 4 Doctor and case manager ratings and employer contact as indices of trust across LLM personas.

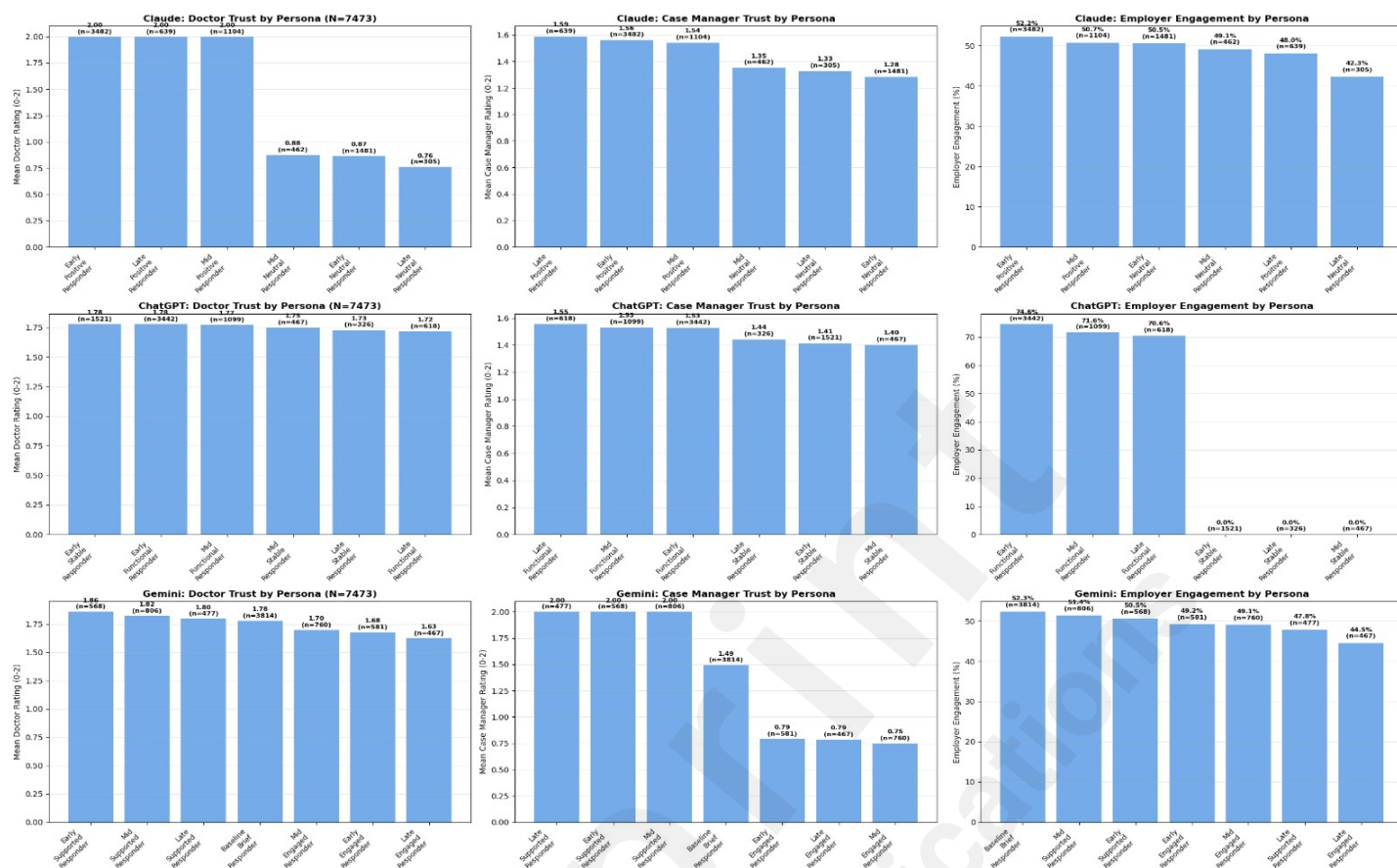


Table 3 LLM personas for Navigation Status

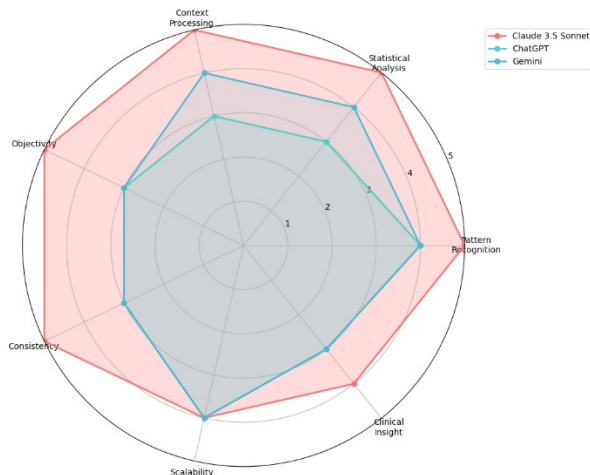
*Instances with meaningful notes about discharge for LLM processing.

Comparison was made between the LLMs on various performance metrics. Figure 5 shows a radar plot

comparing the LLMs on dimensions of context process, statistical analysis, pattern recognition,

clinical thought, scalability, consistency, and objectivity on a 5-point evaluation scale.

Figure 5 “Radar Plot” of LLM Performance



Discussion

We had access to a significantly large and detailed dataset of structured and unstructured data, which was “natural” in that it reflects actual clinical behaviour from an organization. Such large datasets inflate the apparent meaningfulness of small differences between large clinical groups such as SO versus S+T comparisons if attention is only given the significance values [36], but due to the volume of cases, such a dataset allows meaningful comparisons of subgroups such as the FSB sub-personas which could not otherwise be examined. The use of LLMs generated insights is not likely to be easily determined otherwise, and this will be reviewed below. While somewhat obvious, the results also showed that if data is cut too finely, such as in the month-by-month by Navigation Status comparison, the LLMs do not produce differentiated models.

The FSB sub-personas indicate a variety of possible motivations for high scores on the OMPSQ-SF. There appear to be the chronic pain and somatic complexes that reflect previous findings such as those by Lees-Hayley on the Minnesota Multiphasic Personality Inventory-2 [37] or Sherman et al [38] “Invalid Somatic Symptom Presentation”. Interestingly the LLMs articulated some presentations anecdotally noted by experienced clinicians but rarely discussed including ChatGPT—Turbo’s “The Pre-Retirement Complex” and “System Disengaged”. The LLMs also identified other sources of injustice motivating over reporting of symptoms such as Claude 3.5 Sonnet’s “Rural/Access Limited” where the inference is that the psychometric expression of distress reflects resentment at lack of access to resources. Aggregating the FSB personas across LLMs produced a System Disengaged. This is

evident in the poor ratings of doctor and case manager by this cohort when aggregated with other System Disengaged/Access Limited groups who had poor ratings for both doctor and case manager reflecting low trust. A typical succession of case notes in an example of sessions around an FSB session as highlighted by Claude 3.5 Sonnet: *Previous session: IW (injured worker) went back for MRI and CT due to still having pain. No update as sees specialist in 1 month. Sees GP next week. Say its healing but not known where it will go. Feigning: Going to a few specialists who want to do a nerve block around his neck. Went for IME (independent medical examination) on Friday through insurer who says maybe a shoulder operation needed. Subsequent: IP hadn't been feeling well so missed my calls. Has had a lot of rain and was outside to help him feel better and get fresh air. [insurer] cancelled the nerve conduction study.*

The LLMs when asked to identify recovery personas with minimal other instruction, resulted in markedly different approaches with Claude 3.5 Sonnet focusing on the sentiment of notes, ChatGPT-4-Turbo focusing on functioning and outcomes, and Gemini Pro 1.5 focusing on support. Disappointingly, Gemini ignored 28% of cases. Interestingly, the Claude 3.5 Sonnet personas were based on the sentiment of notes made by the

telephone operatives, which demonstrates that the communication with the injured worker changed the emotional tone or sentiment of the contemporaneous notes such that this could be used to distinguish recovery patterns. This suggests that the dyadic interaction with the telephone operatives reflects the likely trajectory of recovery, and the potential to change recovery through the interaction. There appeared to be a relationship between sentiment and doctor rating, with neutral sentiment associated with lower ratings of goodness with doctor. The personas generated by Claude 3.5 Sonnet also provided feedback which would be valuable to coaching telephone operatives. The ChatGPT-4-Turbo personas focused on functioning, which is most meaningful to stakeholders within the insurance/scheme environment. By applying costing data, these personas may be useful to estimate resource consumption during case management. Gemini 1.5 Pro personas used support-based features, and those personas reflecting engagement rather than support showed lower goodness ratings in the relationship with case managers. This is a useful approach to improve the psychological wellbeing of injured workers and also reflects the impact of interpersonal traits and their impact on recovery. However, these personas did not contain much content useful for coaching telephone operatives or informing on interventions for injured workers.

The personas generated for R, ESC, INT, and DISC were not particularly informative, but rather confirmed that the intent behind these interventions were achieved. That is the description of the personas appeared to match the intent behind the level of intervention.

Interestingly ChatGPT-4-Turbo in the ESC persona identified a relationship between distress and report of financial duress; the only allusion to so called “compensation neurosis” [39]. But this did not shed light on this despite the inclusion of the financial wellbeing question to interrogate this possible presentation.

Limitations exist in that the data set involved. While well organized, there were many missing entries as the “support” function was the focus of telephone operatives rather than data collection. Consistent with the training processes of the LLMs there was the tendency for them to phrase things affirmatively, and to amplify minor differences that, though statistically significant, overstated the real-world value of the findings. Ideally, the narratives should be considered in a “dialectical” sense, that is, the value is found in questioning the propositions made by the LLMs in persona generation as much as in the insights derived from them. For example, the different features used by the LLMs and the similarity and differences between the recovery patterns helps to inform on recovery and interventions as much as the recommendations for intervention by the LLMs. The current research used text-

based data, and given the Claude 3.5 Sonnet’s focus on sentiment, it may be fruitful to use dynamic vocal sentiment as an input in future research. Another limitation is that there are a multitude of other LLMs with different structures and training, such as Large Language Model Meta AI and others that could be applied to such data sets and yield further insights. Comparing the LLMs on performance over the above tasks showed that for this kind of data Claude 3.5 performed better overall than ChatGPT-4-Turbo and Gemini 1.5 Pro, nevertheless, some of the features used by these LLMs could be considered more meaningful to industry. The ChatGPT-4-Turbo Personas appear most relevant to the business of insured injury on the face of it, however, the recommendations of the Claude 3.5 Sonnet for FSB Personas were the most realistic and relevant to systemic claims management of this subset. In conclusion, the current study showed the utility of LLM’s to organise and derive insights from a large naturalistic data set in insured injury. Despite limitations this represents a unique and first step to applying LLMs to research in insured injury.

Authors’ Contributions

JEM designed the program, study, carried out analyses and drafted the initial manuscript. AC and IDC critically reviewed and revised the manuscript. All authors approved the final manuscript and submitted and agree to be accountable for all aspects of the work.

Conflicts of Interest

JEM was the Director of Science of Navigator Group at the time of data collection and is a shareholder of Navigator Group. AC and IDC have no conflict of interest.

Data Availability Statement

Data is available on reasonable request, however, Navigator Group Pty Ltd reserves the right to refuse access to data for commercial, ethical and moral reasons.

References

1. Evans RW. The effects of litigation on treatment outcome with personal injury patients. *Am J Forensic Psychol.* 1994;12(4):19-34.
2. Silver JM. Invalid Symptom reporting and performance: What are we missing? *NeuroRehab.* 2015;36:463-469.
3. McMahon JE, Craig A, Cameron ID. Development of Cut Scores for Feigning Spectrum Behaviour on the Orebro Musculoskeletal Pain Screening Questionnaire and the Perceived Stress Scale: A Simulation Study. *J Clin Med.* 2025;14,5504:1-16.
4. Lionton SJ, Nicholas, M. Macdonald S. Development of a Short Form of the Orebro Musculoskeletal Pain Screening Questionnaire: A Qualitative Study. *Spine.* 2011;36:1891-1895.
5. Mittenberg W, Patton C, Canyock EM, Condit DC. Base Rates of Malingering and Symptom Exaggeration. *J Clin Exp Neuropsychol.* 2002;24(8):1094-1102.
6. Laisne F, Lecomte C, Coriere M. Biopsychosocial determinants of work outcomes of workers with occupational injuries receiving compensation: A prospective study. *Work.* 2013;44(2):117-132.
7. Wall CL, Morrissey SA, Ogloff JRP. The worker’s compensation experience: A qualitative exploration of

workers' beliefs regarding the impact of the compensation system on their recovery and rehabilitation. *Int J Disab Manag*.2009;4(2):19-26

Eggert S. Psychosocial factors affecting employees abilities to return to work. *AAOHN J*.2010;58(2):51-55.

Collie A, Sheehan L, Lane TJ, Iles R. Psychological distress in workers' compensation claimants: Prevalence, predictors and mental health service use. *J Occup Rehab*. 2020;30:194-202.

Addison JT, Teixeira P. Trust and Workplace Performance. *J Ind Relat*.2020;58(4):874-903.

Zeffane R, Melhem SJB. Trust, job satisfaction, perceived organizational performance and turnover intention. *Emp Relat*.2017;39(7):1148-67.

Kramer RM. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annu rev psychol*.199;50:569-98

Zak PJ, Knack S. Trust and Growth. *Econom J*. 2001;111(470):295-321.

Phillip B, Harris, J. Trust, transaction cost economics and mechanisms. In Bachmann R and Zaheer A (Eds) *Handbook of Trust Research*. Edward Elgar N.Y.: 124-143.

McMahon JE, Roozegar R, Craig A, Cameron ID. Towards improving insurance processes: A Time Series Analysis of psychosocial recovery after workplace injury across legislative environments. *ICSOC Workshops 2023, LNCS 14518*.2024;27-38

Dehbozorgi R, Zangeneh S, Khooshab E, Nai DH, Hanil HR, Samian P, Yousefi M, Hashemi FH, Vakili M, Jamalimoghdam N, Lohrasebi F. The application of artificial intelligence in the field of mental health: a systematic review. *BMC Psychiatry*. 2025;25:132-152

Amirian S, Kekre A, Loganathan BJ, Chavan V, Kandula P, Littlefield N, Franco JR, Tafti AP, Ebuenyi ID. Advancing psychosocial disability and psychosocial rehabilitation research through large language models and computational text mining. *Cambridge Prisms: Global Mental Health*.2024;11,e123:1-13.

Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative Agents: Interactive Simulacra of Human Behaviour. *Proceedings of 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. 2023. doi:10.1145/3586183.3606763

Anthropic. Introducing the Claude 3 family. Anthropic; 2024. Available from: <https://www.anthropic.com/news/claude-3>

OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.09774.2023.

Google DeepMind. Gemini 1.5: Unlocking multimodal reasoning at scale. Deep Mind Blog. 2024. Available from: <https://deepmind.google/technologies/gemini>

Paterson, R. *The Assertiveness Workbook: How to express your ideas and stand up for yourself at work and in relationships*. New York: New Harbinger Publications; 2022

D'Zurilla TJ, Goldfried MR. Problem Solving and behaviour modification. *J Abnorm. Psychol*. 1971;8:107-26.

Andrews G, Dean K, Genderson M, Hunt C, Mitchell P, Sachdev P, Trollor J. *Management of Mental Disorders*:5th Edition. South Carolina: CreateSpace Independent Publishing Platform; 2014.

Miller WR, Rollnick S. *Motivational Interviewing: Helping People Change*: 3rd Edition. New York: The Guilford Press; 2013.

Schaur M, Elbert T. Dissociation following traumatic stress: Etiology and treatment. *J Psycholog*. 2010;218(2):109-127.

DeRubies RJ, Webb CA, Tang TZ, Beck AT. Cognitive Therapy. In: Dobson KS, ed. *Handbook of Cognitive Behavioral Therapies*. The Guilford Press; 2010: 277-316.

Boon S, Steele K, van der Hart O. *Coping with Trauma-Related Dissociation: Skills Training for Patients and Therapists*. New York: W.W. Norton & Company; 2011.

Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (3rd Ed). Hoboken;Wiley;2019.

Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of ICWSM*. 2014;8:216-25.

Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Theory Methods*. 1974;3(1):1-27.

Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;1(2):224-7.

Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*. 2019;3982-92.

Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. 2019;4171-86.

Lin CY. ROUGE: A package for automatic evaluation of summaries. *Proceedings of the ACL Workshop on Text Summarization Branches Out*. 2004;74-81.

Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *J Mach Learn Res*. 2003;3:993-1022.

37. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “ $p < 0.05$ ”. *Am Stat*. 2019;73(5):1-19.
38. Henry GK, Heilbronner RL, Mittenberg W, Enders C, Stanczak SR. Comparison of the Lees-Haley Fake Bad Scale, Henry-Heilbronner Index, and Restructured Clinical Scale 1 in Identifying Noncredible Symptom Reporting. *Clin Neuropsychol*. 2008;22:919-29.
39. Sherman EMS, Slick DJ, Iverson GL. Multidimensional Malingering Criteria for Neuropsychological Assessment: A 20-year Update of Malingered Neuropsychological Dysfunction Criteria. *Arch Clin Neuropsychol*. 2020;35(5):735-64.
40. Hall RCW, Hall RCW. Compensation Neurosis: A Too Quickly Forgotten Concept? *J Am Acad Psychiatry Law*. 2012;40:390-8.

Abbreviations

AI: artificial intelligence

ESC: Escalated Navigation

FSB: feigning spectrum behaviour

KPI: key performance indicator

LLM: large language model

NP: Naïve Persona

OMPSQ-SF: Orebro Musculoskeletal Pain Screening Questionnaire-Short Form

OT: Open Text

PSS-4: Perceived Stress Scale-4

R: Routine

SO: Support Only

S+T: Support Plus Therapy

VADER: Valence Aware Dictionary and sEntiment Reasoner

x/10: Rating from 0 to 10

3PRS: 3 Point Rating Scale

©. Originally published in JMIR AI (<https://ai.jmir.org>), 27.08.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Supplementary Files

Multimedia Appendixes

LLM Feigning spectrum behaviour persona analysis report.

URL: <http://asset.jmir.pub/assets/64607d1885e7bb7acc804b52ca80dfb6.docx>

Sixty-word recovery persona descriptions generated by LLMs.

URL: <http://asset.jmir.pub/assets/b027dde91852c64e6117c93eb7e2c2d6.docx>

Navigation status detailed LLM personas.

URL: <http://asset.jmir.pub/assets/a912d498434fa5b4ba84224d3624f276.docx>