Can large language models build psychological theories automatically? Take rest intolerance as an example

Authors: Wang, Fei, Song,Haoran, Meng,Xiaoxuan, Wang, Ting, Ji,Xuanjing, Ji,Yongkang, Jiang,Can, Zhao,Nan, Zhu,Tingshao, Zhu, Tingshao

Date: 2025-12-22T13:18:43+00:00

Abstract

This study aimed to investigate whether large language models (LLMs) can independently and automatically construct psychological theories based on grounded theory and to examine whether the constructed theories reach expert-level quality. To this end, we developed a technical approach integrating prompt engineering with an embedding vector similarity merging mechanism. Utilizing the Qwen3-Max and Qwen3-embedding-0.6B models, we performed a complete grounded theory coding process, including open coding, preliminary merging, axial coding, and selective coding, on open-ended responses regarding the causes of rest intolerance collected from 469 university students. Concurrently, three human experts independently performed the same coding process. Evaluations of the internal consistency, content validity, and similarity between the LLM-derived and expert-derived coding results demonstrated that the LLM coding exhibited good reliability and validity. Furthermore, a comparison of theoretical narratives rated by 130 participants, along with structural equation modeling based on questionnaire data from 392 participants, showed that the "Performance Internalization Model of Rest Intolerance" constructed by the LLM did not significantly differ in narrative quality from that produced by human experts, and its theoretical structure was supported by empirical data. This study not only validates the feasibility and effectiveness of using LLMs for psychological theory construction but also provides the first systematic theoretical explanation for the causes of rest intolerance.

Full Text

Preamble

Can Large Language Models Build Psychological Theories Automatically?

Taking Rest Intolerance as an Example

Fei Wang[1,2], Haoran Song[3], Xiaoxuan Meng[4], Ting Wang[5], Xuanjing Ji[6], Yongkang Ji[7], Can Jiang[8], Nan Zhao[1,2], *Tingshao Zhu*[1,2]

[1] State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, 100101
[8] School of Economics and Management, Yanbian University, Hunchun, Jilin, China, 133300

Corresponding authors:
Nan Zhao, 16 Lincui Road, Chaoyang District, Beijing 100101, China. E-mail: zhaonan@psych.ac.cn
Tingshao Zhu, 16 Lincui Road, Chaoyang District, Beijing 100101, China. E-mail: tszhu@psych.ac.cn

Abstract

This study investigated whether large language models (LLMs) can independently and automatically construct psychological theories using grounded theory methodology, and whether such theories achieve expert-level quality. We developed a technical approach integrating prompt engineering with an embedding vector similarity merging mechanism. Using the Qwen3-Max and Qwen3-embedding-0.6B models, we performed a complete grounded theory coding process—including open coding, preliminary merging, axial coding, and selective coding—on open-ended responses about the causes of rest intolerance collected from 469 university students. Concurrently, three human experts independently performed the same coding process. Evaluations of internal consistency, content validity, and similarity between LLM-derived and expert-derived coding results demonstrated that LLM coding exhibited good reliability and validity. Furthermore, comparisons of theoretical narratives rated by 130 participants, along with structural equation modeling based on questionnaire data from 392 participants, showed that the "Performance Internalization Model of Rest Intolerance" constructed by the LLM did not differ significantly in narrative quality from that produced by human experts, and its theoretical structure was supported by empirical data. This study not only validates the feasibility and effectiveness of using LLMs for psychological theory construction but also provides the first systematic theoretical explanation for the causes of rest intolerance.

Keywords: Large Language Models; embedding; rest intolerance; Performance Internalization Model

1 Introduction

Unlike physics, which seeks universal laws applicable throughout the cosmos, psychology focuses on the complex and multidimensional inner world of human beings. Given this inherent heterogeneity and dynamic nature, constructing universal "laws of mental phenomena" faces fundamental challenges (Green, 2015). This disciplinary essence has led psychology to develop numerous subfields, where scholars typically adopt domain-specific research strategies aimed at constructing dedicated theoretical models for particular psychological phenomena to achieve precise description, explanation, prediction, and control (van Rooij & Baggio, 2021).

Against this backdrop, when research is in its early stages or confronts a novel and complex phenomenon, qualitative research becomes a crucial pathway for constructing preliminary theoretical frameworks. Among these approaches, grounded theory, as a systematic qualitative research method, plays an indispensable role (J. Corbin & Strauss, 2025). This methodology emphasizes in-depth coding and analysis of textual data to extract recurrent core categories, inherent contradictions, and key themes from the bottom up, thereby providing a solid empirical foundation for initial theoretical framework establishment (Cepellos & Tonelli, 2020; Walker & Myrick, 2006). However, a significant practical challenge is that the rigorous process of conducting grounded theory research typically demands substantial human and time resources. Furthermore, its stringent coding procedures require high theoretical sophistication and analytical skills from researchers, which limits its widespread application and the efficiency of theoretical output (Pope et al., 2000).

The rise of large language models (LLMs) has injected new vitality into grounded theory-based theoretical construction, as researchers can now leverage LLMs to assist in the coding process, significantly enhancing efficiency and scalability. Previous empirical studies utilizing ChatGPT for qualitative analysis have found that it can efficiently generate analytical outcomes comparable to manual coding (De Paoli, 2024; Jalali & Akhavan, 2024; Kabir et al., 2025; Morgan, 2023; Siiman et al., 2023). This indicates that LLMs hold broad application prospects in qualitative research. As a crucial methodology in qualitative research, grounded theory is increasingly being integrated with LLMs to explore their potential in performing grounded theory coding (Schroeder et al., 2025; Sinha et al., 2024).

Specifically, a recent study employed ChatGPT-4 to conduct a complete three-stage coding process on research abstracts from a specific domain, ultimately generating a conceptual model. Evaluation revealed that while the LLM could efficiently complete coding tasks and produce a theoretical framework with considerable plausibility, its analysis still had limitations in depth and originality, necessitating integration with the deep insights of human researchers (Sammon et al., 2024). Building on this, Yue et al. have further provided a tutorial on using LLMs to assist in grounded theory coding (Yue et al., 2025). These studies

collectively demonstrate the effectiveness of leveraging LLMs to support human experts in conducting grounded theory coding.

A subsequent question arises: can large language models be utilized to construct psychological theories independently and autonomously? If feasible, this would imply a potential fundamental shift in the paradigm of psychological theory construction. In the future, the process of building theories from textual data might only require participants to provide data, with pre-programmed systems automatically handling all subsequent tasks, thereby significantly alleviating the burden on researchers. In the context of automated grounded theory development using LLMs, existing research such as AcademiaOS has preliminarily achieved automated coding of qualitative data and theory construction. A subsequent user study involving 19 participants indicated the system' s potential to assist humans in conducting qualitative research (Übellacker, 2024). However, this methodology still exhibits certain limitations, particularly the potential for redundancy and semantic overlap when processing large-scale, fine-grained coding. To enhance coding precision and the efficiency of theory construction, this study proposes an improved workflow incorporating an embedding vector similarity merging mechanism. Building upon the traditional three-stage coding of grounded theory, two additional steps of embedding and merging are introduced. This aims to effectively integrate highly similar conceptual codes, prevent performance degradation and resource waste caused by redundant inputs, and simultaneously strengthen the rigor and interpretability of theory construction. This approach not only optimizes the performance of LLMs in complex theory-generation tasks but also offers a more robust and cost-effective implementation pathway for automated qualitative research.

To validate the reliability and validity of our proposed technical pathway for psychological theory construction using LLMs with grounded theory, we selected an emerging research area lacking theoretical guidance—rest intolerance —for verification. The concept of rest intolerance was first proposed by Wang et al., defined as a psychological phenomenon characterized primarily by negative feelings arising from choosing rest over productive activities, accompanied by cognitive features such as obsessive thoughts, social comparisons, and cognitive biases (Wang et al., 2025). Rest intolerance has been found to be closely associated with negative outcomes including social media addiction, status anxiety, insomnia, emotional distress, stress, and anxiety (Avcı, 2025; Cheng et al., 2025; Zhao et al., 2025). As a concept widely discussed in China in recent years but still in its early research stages, the causes of rest intolerance within the Chinese cultural context remain unclear. Therefore, an additional objective of this study is to investigate the causes of rest intolerance in the Chinese cultural environment and develop a theoretical framework explaining its origins.

In summary, this study aims to develop an automated and independent technical method for constructing psychological theories through grounded theory using large language models, and on this basis, to develop and validate a causal theory of rest intolerance.

2 Methods

2.1 Participants

This study employed three rounds of data collection. Sample 1 was utilized to collect open-ended responses concerning the causes of rest intolerance. Sample 2 was used to gather participants' evaluations of the quality of theoretical narratives generated by human experts and LLMs. Sample 3 was used to administer a questionnaire based on theories generated by LLMs to validate their effectiveness.

Sample 1 consisted of 469 university students, from which 168 individuals with high levels of rest intolerance were selected. Their open-ended responses served as the primary data for grounded theory coding. This sample included 130 females, with a mean age of 20.73 (SD = 2.13). The composition was as follows: 139 undergraduate students, 3 associate degree students, 20 master' s students, and 6 doctoral students.

Sample 2 included 130 university students, from which 62 individuals with high rest intolerance were selected to evaluate the quality of theoretical narratives produced by LLMs and human experts. This sample comprised 54 females, with a mean age of 20.79 (SD = 2.62). The distribution was as follows: 51 undergraduate students, 8 master' s students, and 3 doctoral students.

Sample 3 consisted of 392 university students, from which 352 valid responses were included in the final data analysis. This sample included 259 females, with a mean age of 22.06 (SD = 3.77). The composition was as follows: 211 undergraduate students, 11 associate degree students, 83 master' s students, and 47 doctoral students. Regarding monthly household income per capita, 116 participants reported below 4,500 RMB, 122 reported between 4,501 and 9,000 RMB, and 114 reported above 9,000 RMB.

2.2 Measures

2.2.1 Rest Intolerance Scale We utilized the 8-item Rest Intolerance Scale developed by Wang et al., which comprises four dimensions: Negative Feelings, Cognitive Bias, Obsessive Thinking, and Social Comparison (Wang et al., 2025). Each dimension is measured by two items. The scale employs a 5-point Likert scoring system, with the total score being the sum of all eight items. A higher total score indicates a more severe level of rest intolerance. In this study, the Cronbach' s alpha coefficients for the four dimensions were 0.82, 0.69, 0.79, and 0.87, respectively. The overall internal consistency coefficient for the full scale was 0.91.

2.2.2 Open-ended Question on Causes of Rest Intolerance A self-designed open-ended question was used to elicit participants' perceived causes of rest intolerance. The question stated: "What do you think are the reasons that

cause rest intolerance?" Participants were required to provide a text response of no less than 30 Chinese words.

2.2.3 Rest Intolerance Theory Questionnaire Based on Large Language Models The axial coding derived from the large language model revealed six core categories: Internalization of Performative Values, Pressure from Competitive Social Comparison, Institutionalized Discipline and Cultural Inculcation, Task and Time Management Pressure, Deficits in Self-Perception and Emotional Regulation, and Dependence on External Evaluation and Validation. Based on the definitions of these six categories and their corresponding open codes, we developed four items for each category. Participants were asked to rate their agreement on a 5-point Likert scale ranging from "Strongly Disagree" (1 point) to "Strongly Agree" (5 points). The total score for each category ranged from 4 to 20. In this study, the Cronbach' s alpha coefficients for the six category-based questionnaires were 0.90, 0.94, 0.87, 0.90, 0.90, and 0.94, respectively. More detailed reliability and validity tests can be found in Supplementary Material 1.

2.3 Large Language Model Tools

2.3.1 Qwen3-Max This study utilized Qwen3-Max, the flagship large language model released by Alibaba Group in September 2025, as the analytical tool. This model is a large-scale language model trained on over 36 trillion tokens with more than a trillion parameters, demonstrating powerful capabilities in deep reasoning and long-context processing. The model was accessed via API through the Alibaba Cloud Bailian platform (https://bailian.console.aliyun.com/#/home).

2.3.2 Qwen3-Embedding-0.6B This study employed the Qwen3-Embedding-0.6B text embedding model developed by Alibaba' s Tongyi Qianwen team. Based on the Qwen3 architecture, the model specializes in text representation, retrieval, and reranking. It supports multilingual and long-text processing, and features user-customizable embeddings of up to 1024 dimensions with instruction-aware capabilities. The model was obtained through the ModelScope platform (https://www.modelscope.cn/models?page=1&tabKey=task) and deployed locally for access.

2.4 Study Procedure

After obtaining ethical approval, we conducted this study. First, we recruited 469 university students through convenience sampling to complete a questionnaire survey. The questionnaire included basic sociodemographic information, the Rest Intolerance Scale, and an open-ended question. Participants were explicitly instructed not to use large language models when responding to the open-ended question. Subsequently, 174 individuals with high rest intolerance were initially identified based on the cutoff score of 28 on the RIS-8. Six of

these participants were excluded due to issues such as repetitive responses, substitution of characters for meaningful text, or suspected AI-generated content. Ultimately, data from 168 participants with high rest intolerance were retained for analysis.

Subsequently, three human experts and a large language model (Qwen3-Max) performed grounded theory coding according to Corbin and Strauss' s approach (J. M. Corbin & Strauss, 1990) on these 168 text segments, with the aim of deriving a theoretical framework for the causes of rest intolerance. The three human experts possess extensive experience in grounded theory coding and had participated in the coding phase during the development of the Rest Intolerance Scale, making them the most suitable experts with dual expertise in both the concept of rest intolerance and grounded theory methodology. The selection of the Qwen3-Max model was primarily based on its demonstrated strengths in Chinese text processing and comprehension. As an ultra-large-scale language model launched by Alibaba, Qwen3-Max exceeds one trillion parameters, representing the largest and most powerful version in the current Qwen series. The model not only exhibits exceptional capabilities in text generation, code generation, and logical reasoning, but also employs a sparsely activated Mixture of Experts (MoE) architecture with support for context windows of up to 262K tokens. Furthermore, Qwen3-Max provides convenient API interface services that can be flexibly adapted to diverse application scenarios. Figure 1 [Figure 1: see original paper] shows a schematic diagram of the grounded theory coding process based on the large language model.

During the grounded theory coding phase conducted by human experts, the three experts followed a four-stage process: open coding, preliminary code integration, axial coding, and selective coding. (The four-stage design was implemented to facilitate comparison with the large language model' s performance, though it should be noted that preliminary code integration is conceptually part of axial coding in traditional grounded theory methodology.) Specifically, taking the open coding stage as an example, each expert independently performed open coding using the same instructions provided to the large language model. Subsequently, they convened to discuss and determine the final open coding results through a consensus-building process. The discussion followed a majority rule principle: if at least two experts proposed similar codes, these were directly incorporated; if all experts proposed different codes, they engaged in collective discussion to reach a final decision. This same procedure was applied throughout the preliminary code integration, axial coding, and selective coding stages. It is worth noting that the selective coding stage differed somewhat from the others, as it involved not only selecting core categories but also generating a theoretical narrative. Specifically, the three experts independently identified core categories, then jointly determined the final core categories through discussion, and collaboratively constructed a theoretical story based on these categories.

When examining the content validity of the large language model's grounded theory coding, we invited five additional experts with extensive experience in scale

development to evaluate the model' s integration process at each stage. These experts assessed the appropriateness of the coding integrations performed by the large language model using a 4-point scale ranging from 1 (highly inappropriate) to 4 (highly appropriate). The researchers calculated content validity indices based on these expert ratings.

Furthermore, to evaluate whether the theoretical narratives generated by the large language model were as reliable as those produced by human experts, we designed a comparative evaluation experiment to objectively compare the performance of the large language model and human experts in generating theoretical narratives. Prior to the experiment, we used the pwr R package to determine that a minimum sample size of 34 participants would be required for a paired samples t-test with a significance level of 0.05, statistical power of 0.8, and a medium effect size of 0.5. To recruit participants with sufficient understanding of rest intolerance, we specifically targeted individuals with high rest intolerance (i.e., total scores above 28 on the RIS-8), as they were expected to have greater insight into the construct than other participants.

The experiment recruited 130 university students from psychology or nursing programs, including 103 undergraduates, 19 master' s students, and 8 doctoral students, to ensure that raters possessed the necessary disciplinary background knowledge to make valid judgments. During the evaluation process, each participant rated two theoretical narratives—one generated by the large language model and one by human experts. To ensure objectivity and fairness, we implemented a single-blind design where the authorship of both narratives was completely concealed, preventing raters from knowing the source of each narrative. To control for potential order effects, we employed a counterbalanced design for material presentation. Specifically, we first collected 68 questionnaires (Group A) where the human expert' s narrative was presented first, followed by the model' s narrative. Subsequently, we collected 62 questionnaires (Group B) with the reverse order—the model' s narrative first, followed by the human expert' s. This approach allowed the potential influence of presentation order to be balanced out in the overall data.

Finally, to examine whether the theory generated by the large language model holds practical significance and to further validate its effectiveness, we developed a measurement questionnaire comprising four items for each category based on the results of the LLM-based axial coding. The questionnaire was designed by the first author of this study and three additional experts, using the grounded theory coding results derived from the large language model. The detailed measurement items corresponding to the six categories can be found in Supplementary Material 2. We then integrated the 24-item questionnaire with the 8-item Rest Intolerance Scale into a single survey. According to the empirical rule of 1:10 (items-to-responses ratio), a minimum of 320 valid responses was required (Wang, Tang, et al., 2024). Consequently, we distributed 392 questionnaires to university students online. After excluding 40 participants with completion times shorter than one minute, 352 valid responses were retained for analysis.

2.5 Data Analysis

All data, code, and programs associated with this research are publicly available and can be accessed at: https://github.com/kingfly51/RIS_{theory}.

During the grounded theory coding phase, the large language model completed the entire coding process in the following sequence: open coding, first embedding process, second embedding process, axial coding, and selective coding. Traditional grounded theory methodology comprises only three steps; the inclusion of two embedding processes was designed to address the issue that independent calls to Qwen3-Max for open coding of each participant generated numerous highly similar codes representing conceptually equivalent meanings. Failure to merge these codes could lead to several problems: excessive input content potentially causing model inertia and degraded performance, increased token consumption, and deviation from grounded theory requirements. Therefore, after performing open coding using the online Qwen3-Max model, we employed the locally deployed Qwen3-Embedding-0.6B model for word vector embedding and merged similar codes based on cosine similarity. Specifically, the first author of this study conducted open coding, axial coding, and selective coding through API calls combined with prompt engineering techniques using the online Qwen3-Max model. The two embedding processes were implemented locally by the first author using the Qwen3-Embedding-0.6B model for word vector embedding, calculating cosine similarity, and merging codes with high similarity scores. The decision to use the locally deployed Qwen3-Embedding-0.6B model was based on the recognition that the similarity merging phase represents a relatively straightforward task that can achieve satisfactory results without requiring additional online embedding services, making it the optimal choice for balancing cost efficiency and performance quality.

During the open coding phase, the researcher utilized the Qwen3-Max model (temperature=0.5, Top-p=0.5) to independently process each participant' s text sequentially, following the open coding prompt (see Supplementary Material 3). The Qwen3-Max model extracted concepts and their meanings from each sentence in accordance with open coding requirements. After processing, all results were locally integrated to form the initial open coding outcomes. The first embedding process involved converting each code and its definition from the open coding results into word vectors. Cosine similarity was then calculated between each pair of word vectors. All codes with cosine similarity exceeding the threshold of 0.85 were merged, and the code name and definition with the highest average similarity to all merged codes were selected as the final representation. The 0.85 threshold was empirically determined, as testing revealed it effectively merged codes conveying nearly identical meanings with only minor variations.

Following the first embedding round, the second embedding process converted the resulting code names, definitions, and the names of merged codes from the first round into word vectors. Cosine similarity was recalculated, and a second round of merging was performed using a threshold of 0.75. This lower threshold

was also empirically chosen, aligning with the objective of the second merge: to combine codes expressing semantically similar concepts. Subsequently, during the axial coding phase, the researcher employed the Qwen3-Max model (temperature=0.5, Top-p=0.5) with the axial coding prompt (see Supplementary Material 4) to process the codes resulting from the second embedding round. The model summarized higher-level concepts capable of encompassing all codes, adhering to axial coding requirements. In the final selective coding stage, the researcher invoked the Qwen3-Max model (temperature=0.3, Top-p=0.5) using the selective coding prompt (see Supplementary Material 5) to analyze the axial coding results. The model selected a core category from those identified during axial coding and constructed a theoretical narrative linking this core category to other categories and the phenomenon of rest intolerance, in line with selective coding specifications. It is noteworthy that the temperature parameter was reduced to 0.3 for the selective coding phase to enhance output stability, given the theoretical nature of this stage.

During the reliability analysis phase, to assess the internal consistency of the coding process based on the large language model, we first utilized the Qwen3-Embedding-0.6B model to convert all codes into vector representations. This model generated a 1024-dimensional embedding vector for each statement and code. Subsequently, we calculated Cronbach' s Alpha coefficients stage by stage to measure reliability. Since the codes generated during the open coding phase were initially unmerged, most contained only one original statement, making internal consistency calculation impossible. Therefore, we used the coding results after the open coding phase and the two subsequent embedding phases to compute internal consistency, thereby assessing the reliability of the open coding stage. Nevertheless, many codes still had a frequency count of 1 even after embedding. Thus, internal consistency was calculated only for codes with a frequency greater than 1. For each such code, we computed the Cronbach' s Alpha coefficient among the vectors of all the original statements it contained. This reflected the similarity of statements under the same code. The average of these coefficients was taken as the overall reliability indicator for this stage.

During the axial coding phase, we applied the same method, calculating the Cronbach' s Alpha coefficient among the code vectors contained within each of the 6 core categories and then averaging these values again. Finally, in the selective coding phase, we directly computed the Cronbach' s Alpha coefficient among the vectors of all 6 core categories, serving as the reliability indicator for the overall theoretical structure.

During the validity analysis phase, to evaluate the effectiveness of the coding process based on the large language model, we used the human experts' coding results as the criterion. We employed the Qwen3-Embedding-0.6B model to convert the code names generated by the large language model and the human experts at each stage into two sets of vector representations. Based on the cosine similarity between each code, we paired the codes with the highest cosine similarity. The average of all these paired similarity scores was calculated as

the mean similarity between the large language model' s and the human experts' coding processes (considering that the number of codes generated by the large language model and the human experts might differ, unpaired codes were directly discarded and did not participate in the average similarity calculation). Additionally, by concatenating all codes generated at each stage into a long text string and converting them into two overall vector representations using the Qwen3-Embedding-0.6B model, we computed the cosine similarity to obtain the overall similarity between the large language model' s and the human experts' coding processes. Specifically, we calculated the similarity between the coding data obtained by the large language model and that obtained by the human experts across four stages. The first two stages involved calculating both the mean and overall similarity between the code names generated by the large language model and those generated by the human experts. This approach was necessary because the human experts did not provide explicit definitions for each initial code during their coding process, making it feasible to compute similarity based only on code names. These two stages were: 1) the initial open coding stage, and 2) the first and second embedding integration stages. The third stage was the axial coding stage. Here, we further calculated both the mean and overall similarity between the code names and definitions generated by the large language model and those generated by the human experts. The fourth stage was the selective coding stage. As this stage involved forming a theoretical narrative based on the axial coding results, the outputs from both the large language model and the human experts were long text passages. Consequently, at this stage, only the overall similarity between the two theoretical narrative texts was computed. In summary, by calculating both the mean and overall similarity between the large language model' s and the human experts' coding results across these four stages, we assessed the criterion-related validity of the large language model' s performance in grounded theory-based coding.

In addition to criterion-related validity, we further assessed the content validity of the large language model's performance in the grounded theory coding process. The specific evaluation method was as follows: Five domain experts were invited to sequentially rate the alignment between the model' s results and the encoded content across four stages (namely, open coding, the embedding phase, axial coding, and selective coding) using a 4-point scale. Based on these ratings, we calculated both the item-level content validity index (I-CVI) and the scale-level content validity index (S-CVI). The item-level CVI was derived by computing the proportion of experts who endorsed each individual code, while the scale-level CVI was calculated as the overall CVI for all codes within the respective coding stage (Wang, Wu, et al., 2022). This approach systematically measured the applicability and representativeness of the model's coding results at different levels of granularity.

After collecting all 130 valid questionnaires, we conducted statistical analyses to evaluate whether the theoretical narratives generated by the large language model were as reliable as those produced by human experts. The core objective of this step was to systematically examine whether the ratings for the theoret-

ical narratives generated by the large language model and the human experts differed significantly across seven core evaluation dimensions. These seven key dimensions included: logical coherence, explanatory depth, comprehensiveness, persuasiveness, alignment with real-life experience, conceptual clarity, and inspirational value (heuristic). It is important to note that these seven evaluation dimensions were determined through a rigorous deliberative process. Prior to the experiment, we invited three senior human experts in the field of psychology, along with the paper' s authors, to hold multiple discussion sessions. Through in-depth exploration and analysis of the core qualities that theoretical narratives should possess, they collectively reviewed and finalized these seven dimensions, which comprehensively and multi-dimensionally measure the quality of theoretical narratives. Among the 130 participants, 62 were identified as individuals with high rest intolerance. Of these, 33 participants were from Group A and 29 from Group B. Using data from these 62 participants, we performed paired-sample t-tests for each dimension to compare the mean ratings given by evaluators for the "large language model group" and the "human expert group." This statistical method effectively detects whether there is a statistically significant difference between the two sets of data. We set the significance level at the conventional threshold of 0.05. Detailed definitions of the seven-dimensional evaluation system can be found in Supplementary Material 6.

Finally, based on the data from the large language model theory questionnaire, we conducted an analysis using structural equation modeling. In terms of model selection, since the theoretical framework included interaction effects—and interaction models involving latent variables not only entail complex estimation processes but also require larger sample sizes—we opted for a path analysis based on observed variables to ensure model robustness and parsimony. To verify the reliability and validity of the developed questionnaire, we separately calculated Cronbach' s alpha coefficients, composite reliability, convergent validity, and the average factor loadings from factor analysis for the four items corresponding to each dimension. These results indicated that the six developed questionnaires exhibited good reliability and validity. We used the sum of the scores of all items within each dimension as the total score for that dimension and constructed the structural equation model according to the theoretical framework. Since the theoretical model included interaction terms, we constructed four interaction terms after centering the variables. It is worth noting that, given the relatively small sample size and the requirement of normality for parametric tests, non-parametric tests were considered more reliable in this context. Therefore, we primarily reported the results of non-parametric bootstrap tests, while the results of parametric tests are provided in Supplementary Material 7 (the significant results from non-parametric tests remained significant in parametric tests as well). During model evaluation, we mainly relied on absolute fit indices to assess whether the constructed model exhibited a good fit (Wang, Ge, et al., 2022). Furthermore, we used 5000 bootstrap resamples to test the mediation effects of indirect pathways in the model and conducted simple slope analyses to examine the specific direction of interactions for variables with significant

interaction effects.

3 Results

3.1 Grounded Theory-Based Coding Results from Large Language Models

The grounded theory-based coding process using the large language model was conducted in five main stages. During the initial open coding phase, utilizing prompt engineering techniques, the Qwen3-Max large language model generated a total of 364 independent codes. These 364 codes are provided in Supplementary Table 1 . Subsequently, considering that the large language model performed each round of open coding independently, the first embedding integration phase (step two) was conducted. We used the Qwen3-Embedding-0.6B model to convert the concatenated code names and definitions of the 364 codes into vector representations. By calculating the similarity between each vector representation and applying a threshold of 0.85, codes with similarity scores above 0.85 were merged into a single code. This process integrated codes that differed only slightly in wording—for example, "peer competition pressure," "peer pressure," "peer effort pressure," and "pressure from visibility of peer academic performance" were all categorized under "peer competition pressure." After this stage, the number of codes was reduced from 364 to 206. Detailed results are provided in Supplementary Table 2 .

In the third step, the second embedding integration phase, the Qwen3-Embedding-0.6B model was used to convert the concatenated code names, definitions, and names of merged codes from the 206 codes into vector representations. By calculating the similarity between each vector representation and applying a threshold of 0.75, codes with similarity scores above 0.75 were merged into a single code. This step aimed to integrate codes with semantically similar meanings—for example, "peer competition pressure," "competitive environmental pressure," "peer academic comparison," and "high-intensity academic competition environment" were all categorized under "peer competition pressure." After this stage, the number of codes was reduced from 206 to 82. Detailed results are provided in Supplementary Table 3 .

The fourth step constituted the axial coding phase, during which the Qwen3-Max large language model, employing prompt engineering techniques, systematically categorized the 82 codes into six core categories as presented in Table 1: "Internalization of Performative Values," "Pressure from Competitive Social Comparison," "Institutionalized Discipline and Cultural Inculcation," "Task and Time Management Pressure," "Deficits in Self-Perception and Emotional Regulation," and "Dependence on External Evaluation and Validation."

"Internalization of Performative Values" describes the process whereby individuals internalize performance-oriented values prevalent in mainstream society—such as "output equals value" and "effort equals morality" —as personal standards for self-evaluation, thereby experiencing rest as a betrayal of their own

value system. "Pressure from Competitive Social Comparison" refers to individuals in highly competitive environments consistently comparing their own resting state with others' perceived efforts through either direct observation or imagined monitoring, sensing a risk of being overtaken, which leads them to invalidate the legitimacy of rest. "Institutionalized Discipline and Cultural Inculcation" denotes the long-term conditioning of individuals through family, educational, and societal institutions with values such as "rest equals laziness or error" and "time must be devoted to productive activities," resulting in a conditioned reflexive aversion to rest. "Task and Time Management Pressure" arises when individuals, due to task overload, pressing deadlines, poor time control, or imbalanced planning, perceive rest as a threat to their responsibilities, efficiency, or goal progress, thus experiencing it as negligence or waste. "Deficits in Self-Perception and Emotional Regulation" characterize individuals who, owing to low self-efficacy, tendencies toward anxiety or depression, excessive rumination, or behavioral inertia, lack the capacity for positive cognitive and emotional regulation concerning rest, leading to intrusive negative thoughts during rest and difficulty achieving relaxation. Finally, "Dependence on External Evaluation and Validation" describes individuals heavily relying on assessments from family, peers, institutions, or society regarding their output and efforts to affirm their self-worth, where rest may induce shame due to the potential for negative judgment or being perceived as failing to meet expectations. More detailed results are available in Supplementary Table 4 .

The fifth step was the selective coding phase. During this stage, the Qwen3-Max large language model, guided by prompt engineering techniques, selected one of the six core categories as the central category. This central category represents the most direct psychological root of "rest intolerance" and is capable of integrating the other five categories. Based on this, a theoretical model of the causes of rest intolerance was further developed. The core category selected by the large language model was "Internalization of Performative Values." This category was chosen as the central category due to its high frequency in the data (58 occurrences) and its close logical connections with the other five axial categories: it acts as a "converter" that transforms external conditions, such as pressure from competitive social comparison, dependence on external evaluation, and task-related stress, into individual psychological responses; it is also the internalized outcome of long-term institutionalized discipline and cultural inculcation, while being amplified by deficits in self-perception. More importantly, it directly activates negative emotions such as shame and guilt during rest, constituting the most fundamental psychological mechanism of rest intolerance. It thus demonstrates a high degree of centrality, explanatory power, and integrative capacity. The theoretical model diagram constructed by the large language model is shown in Figure 2 [Figure 2: see original paper], and detailed results can be found in Supplementary Table 5 .

The theory is named the Performance Internalization Model of Rest Intolerance. It proposes that within a highly performance-oriented sociocultural environment, individuals, through early familial upbringing, exam-oriented education, and

continuous socialization into societal norms, internalize beliefs such as "effort equals morality"and "output equals value"as core standards of self-identity (with institutionalized discipline and cultural inculcation serving as the contextual background). This internalization process leads individuals to automatically activate self-negating thoughts, such as "I am wasting my life" or "I am not good enough," during rest, directly triggering shame and guilt (with internalization of performative values acting as the core causal mechanism). Simultaneously, individuals immersed in a highly competitive environment constantly compare their own resting state with others' perceived efforts through direct observation or imagined monitoring, sensing a threat of "falling behind," thereby further constructing rest as a dangerous and immoral behavior (under pressure from competitive social comparison). Compounded by high-intensity task loads, a sense of time loss, and anxiety over unfinished tasks, rest is experienced as a betrayal of responsibilities and goals (under task and time management pressure). Furthermore, when individuals heavily rely on evaluations from family, peers, or society regarding their output to affirm their self-worth, rest is more likely to be interpreted as failing to meet expectations or inviting negative labels (under dependence on external evaluation and validation). Throughout this process, if individuals also exhibit low self-efficacy, ruminative thinking, or difficulties in emotional regulation (i.e., deficits in self-perception and emotional regulation), these traits significantly amplify the activation of shame by the aforementioned pressures, forming a vicious cycle wherein "the more one rests, the more shame one feels; the more shame one feels, the less one can truly rest." Thus, rest intolerance is not merely an emotional reaction but a systemic psychological dilemma dynamically generated through the internalization of performance values under specific social structures and psychological conditions.

The grounded theory coding results generated by human experts are provided in Supplementary Material 8. Detailed outcomes for the open coding, initial integration phase, axial coding, and selective coding can be found in Supplementary Tables 6-9.

3.3 Reliability Analysis

To evaluate the internal consistency of the large language model' s performance in grounded theory coding across open coding, axial coding, and selective coding stages, we calculated Cronbach' s alpha coefficients for the original statements contained within each of the 82 codes (42 codes were actually analyzed) resulting from open coding and two subsequent embedding phases. This served as the reliability indicator for the open coding and embedding processes. Similarly, we further computed Cronbach' s alpha coefficients for the codes contained within each of the six core categories derived from axial coding to assess the reliability of the model' s axial coding. Finally, we calculated the Cronbach' s alpha coefficient for the six core categories collectively as the reliability metric for selective coding. The results showed that the Cronbach' s alpha coefficients for the codes obtained after open coding and two embedding phases ranged

from 0.71 to 0.99, with a mean Cronbach' s alpha of 0.90. The Cronbach' s alpha coefficients for the six core categories derived from axial coding ranged from 0.97 to 0.99, with a mean value of 0.98. The Cronbach' s alpha coefficient corresponding to selective coding was 0.87.

3.4 Validity Analysis

3.4.1 Similarity Between Large Language Model and Human Expert in Grounded Theory Coding To evaluate the validity of the large language model (LLM) in performing grounded theory analysis, we calculated the similarity between the data generated by the LLM and that produced by human experts at each step of the grounded theory coding process. The mean similarity was derived by matching each code generated by the LLM with those from human experts, computing the cosine similarity for each pair, and then averaging these values. The overall similarity was obtained by concatenating all codes generated by the LLM and all codes generated by human experts into two respective text strings and calculating the cosine similarity between them. The results indicate that the grounded theory-based coding outcomes of the LLM exhibit a high degree of similarity to those of human experts. Specifically, the overall similarity scores for open coding, axial coding, and selective coding were 0.89, 0.81, and 0.85, respectively, as shown in Table 2.

Table 2 The similarity between large language models and human expert coding

Stage	Average similarity	Overall similarity
Open coding stage	0.73	0.89
The first and second embedding stages	0.71	0.88
Axial coding stage	0.84	0.81
Selective coding stage	-	0.85

3.4.2 Content Validity in the Large Language Model Coding Process We recruited five experts to evaluate the content validity of the large language model' s outputs at various stages: open coding, the two embedding phases, axial coding, and selective coding. Detailed information about the experts is provided in Supplementary Table 10 , and the rating forms for each stage can be found in Supplementary Tables 11-14. The results showed that among the 364 open codes, 361 had an I-CVI greater than 0.80 (>0.75), while only three codes had an I-CVI of 0.60. The S-CVI for the entire open coding phase was 0.98. During the embedding integration phase, 80 out of 82 codes achieved an I-CVI above 0.80 (>0.75), with only two codes scoring an I-CVI of 0.40. The S-CVI for the embedding phase was 0.96. In the axial coding phase, all six axial codes achieved an I-CVI of 1.00 (>0.75), resulting in an S-CVI of 1.00 (>0.75) for this phase. Similarly, in the selective coding phase, both the I-CVI and S-CVI for the identified core category were 1.00 (>0.75).

3.4.3 Differences Between Theoretical Narratives Generated by LLMs and Human Experts Paired-sample t-test results revealed no statistically significant differences in ratings between the theoretical narratives generated by the large language model and those produced by human experts across all seven dimensions: logical coherence (t=-0.20, df=61, p=0.84), explanatory depth (t=1.70, df=61, p=0.09), comprehensiveness (t=1.35, df=61, p=0.18), persuasiveness (t=0.00, df=61, p=1.00), alignment with real-world experience (t=0.59, df=61, p=0.56), conceptual clarity (t=1.96, df=61, p=0.06), and inspirational value (t=1.05, df=61, p=0.30). These results indicate that, under the experimental conditions, the theoretical narratives generated by the large language model were comparable in quality to those produced by human experts across these seven dimensions. Raters were unable to consistently distinguish between the two sources in terms of perceived quality. Please see Figure 3 [Figure 3: see original paper].

3.5 Quantitative Validation of the Theoretical Model Derived from Large Language Models

To examine whether the Performance Internalization Model of Rest Intolerance, generated by the large language model through qualitative analysis, can guide quantitative research and to further validate the effectiveness of this theoretical model, we developed four measurement items for each category derived from the axial coding phase to assess the corresponding constructs. We then collected a completely new sample to test the theory by constructing a structural equation model. The results from the 5000 bootstrap tests, as shown in Figure 4 [Figure 4: see original paper], indicate that the model demonstrates a good fit: CFI = 0.98, NFI = 0.97, GFI = 0.97, RFI = 0.93, TLI = 0.95, RMSEA = 0.074, SRMR = 0.0997.

By comparing the theoretical model diagram generated by the large language model with the final model diagram, we found that the majority of the predicted paths in the LLM-generated theory were supported by the data, with only some moderating effects being non-significant. We further tested whether the mediation effects in the model diagram were statistically significant. Using 5000 bootstrap samples for mediation analysis, the specific results presented in Table 3 show that all mediation effects in the diagram were significant ($p < 0.01$).

Table 3 Mediation Analysis of Indirect Paths in Models

Path	Estimate	95% CI lower	95% CI upper
IDCI→IPV→DEEV→RIS	0.12	0.06	0.19
IDCI→IPV→STMP→RIS	0.08	0.03	0.13
IDCI→IPV→PCSC→RIS	0.10	0.05	0.16
IDCI→IPV→RIS	0.15	0.07	0.24
IDCI→DEEV→RIS	0.11	0.05	0.18

Path	Estimate	95% CI lower	95% CI upper
IDCI→TTM→RIS	0.07	0.02	0.12
IDCI→PCSC→RIS	0.09	0.04	0.15

We further conducted simple effect analysis on the significant moderating effects. The results demonstrated that when deficits in self-perception and emotional regulation were at a low level, the predictive effect of competitive social comparison pressure on rest intolerance was relatively weak (β=0.40, t=4.11, p<0.001). When deficits in self-perception and emotional regulation were at an average level, the predictive effect of competitive social comparison pressure on rest intolerance became stronger (β=0.60, t=7.42, p<0.001). Notably, when deficits in self-perception and emotional regulation were at a high level, the predictive effect of competitive social comparison pressure on rest intolerance nearly doubled in magnitude (β=0.81, t=8.42, p<0.001), as illustrated in Figure 5 [Figure 5: see original paper].

4 Discussion

This study systematically designed and validated a technical pathway through which a large language model independently and automatically constructs psychological theories based on grounded theory. Leveraging prompt engineering combined with an embedding vector similarity merging mechanism can effectively support the large language model in performing theory construction in psychology grounded in empirical data. Notably, this research fills a theoretical gap in the field of rest intolerance, preliminarily revealing the theoretical basis and underlying psychological mechanisms of its formation.

The embedding vector similarity merging mechanism introduced in this study performs semantic-level clustering and deduplication of initial codes, effectively merging highly repetitive and redundant coding units. This approach significantly enhances the quality of codes input into the large language model. On one hand, it mitigates issues of delayed responses and performance degradation caused by excessive input volume (Liu et al., 2024). On the other hand, since the embedding process is handled locally, this strategy improves processing efficiency while reducing additional token consumption and computational costs associated with frequent calls to the large language model.

The coding processes of human experts and the large language model were conducted independently. Despite this, the coding results from both sides exhibited a high degree of similarity. In particular, the six categories derived from axial coding demonstrated substantial semantic consistency between the two sources, indicating the effectiveness of the large language model in performing grounded theory coding. It is worth noting, however, that certain differences emerged in the theoretical narratives generated by the model and the human experts. These divergences primarily revolved around the sequential relation-

ship between the "Internalization of Performative Values" and "Dependence on External Evaluation and Validation." The large language model tended to posit that the internalization of performative values leads to dependence on external evaluation and validation—that is, because an individual internally endorses the logic that "one' s worth must be demonstrated," they become heavily reliant on external judgments (from others or society) for confirmation, thereby alleviating inner uncertainty and insecurity. In contrast, human experts were inclined to argue that dependence on external evaluation fosters achievement-oriented self-worth, suggesting that an individual' s reliance on potential external evaluations shapes a stronger tendency to internalize achievement-oriented values. Both interpretations appear theoretically plausible, implying the possible existence of a bidirectional causal relationship between these two variables. The observed discrepancy in the theoretical narratives may be attributed to the fact that the prompt engineering process did not explicitly account for the potential for such bidirectional causality.

Furthermore, this study systematically verified that the large language model demonstrates high content validity in grounded theory-based coding. Moreover, individuals with high rest intolerance rated the quality of the theoretical narratives generated by the model as having no significant difference from those produced by human experts, indicating that the model has reached a level of theoretical completeness and acceptance comparable to that of human experts.

Further data analysis provided preliminary empirical support for the "Performance Internalization Model of Rest Intolerance" proposed by the model. Together, these findings suggest that large language models are capable of independently and automatically generating explanatory psychological theories at a level comparable to human experts. At least in the early stages of theory building, researchers can leverage such models to analyze qualitative texts, formulate insightful theoretical hypotheses, and design subsequent empirical studies for validation. This highlights the important methodological value of large language models in driving innovation in psychological theory and guiding quantitative research directions.

Another significant contribution of this study lies in the development of a preliminary theoretical framework for rest intolerance: the Performance Internalization Model of Rest Intolerance, which addresses a critical theoretical gap in this field. This model posits that the root cause of an individual' s rest intolerance lies in the Internalization of Performative Values, which directly explains why rest intolerance arises: rest is subjectively constructed as a deviation from the "ideal self," thereby triggering fundamental self-negating emotions and constituting the deepest psychological mechanism of rest intolerance. The internalization process of performative values systematically generates rest intolerance by shaping three interrelated dimensions of pressure. First, value internalization leads to tasks and time being alienated as the core metrics of self-worth. As a result, any form of rest is cognitively framed as an interruption of "value production," directly inducing existential task-time management pressure. Second, since in-

trinsic self-worth becomes conditionally dependent on external validation, individuals inevitably exhibit a high degree of reliance on external evaluation and recognition. Consequently, rest is experienced as a risky behavior that may invite negative judgment, due to its inability to generate positive feedback. Third, internalized performance standards compel individuals to engage in persistent social comparison. From this perspective, rest is constructed as a potential threat leading to loss of competitive advantage. Ultimately, at the convergence of these three pressures, rest is no longer perceived as necessary physiological or psychological recovery. Instead, it becomes subjectively constructed as a fundamental deviation from the "ideal performing self," thereby evoking deep-seated self-negating emotions and forming the core psychological mechanism of rest intolerance.

It is important to note that an individual' s internalization of performative values does not occur in a vacuum but is deeply embedded within their macrocultural context, reflecting and embodying the influence of specific socio-cultural structures. Specifically, the establishment of performance-oriented values at the individual level essentially results from the socialization process, through which individuals receive, identify with, and ultimately incorporate dominant cultural norms transmitted by social institutions such as family, school, and workplace into the core of their self-schema. When a society' s cultural core highly emphasizes competition, efficiency, and visible achievement, and imbues them with a sense of moral superiority, this logic of "performance supremacy" is deeply internalized by individuals through continuous discipline and indoctrination. This "culture-psychology" transmission mechanism provides a crucial theoretical benchmark for understanding cross-cultural variations in rest intolerance. For instance, in performance-driven societies such as those in East Asia (Wang, Zhu, et al., 2024), which are influenced by both Confucian work ethics and modern competitiveism, the cultural field glorifies "relentless diligence" while implicitly stigmatizing "leisure and comfort" (Wang et al., 2025). This combination creates a highly susceptible environment for the emergence of rest intolerance and constitutes the underlying socio-psychological dynamic behind phenomena such as the "China Speed" (Yang, 2025). Conversely, in societies where cultural traditions place higher value on personal well-being and life balance, the degree to which individuals internalize performance values and the resulting rest intolerance tend to be less pronounced. Therefore, a thorough examination of rest intolerance must go beyond individual psychological factors and critically examine the specific cultural contexts in which it is cultivated and reinforced.

Finally, this study has several limitations. First, as it relies on a large language model for psychological theory construction, it inherits all the inherent limitations of such models. For instance, in terms of data privacy, the online invocation of the Qwen-Max model may pose potential data security concerns. Second, the empirical validation in this study is based on cross-sectional structural equation modeling, which does not allow for establishing causal relationships among the identified categories. Lastly, since the primary aim of this research was to

verify whether large language models can be used to automatically and independently construct psychological theories, rather than to compare performance across different models, no other large language models were invoked in this work.

5 Conclusions

This study developed an automated technical pathway for independently utilizing large language models to construct psychological theories based on grounded theory, and validated the effectiveness of this approach. More importantly, it proposed the Performance Internalization Model of Rest Intolerance, which identifies the internalization of performance-oriented values as the fundamental cause of rest intolerance, thereby establishing a theoretical foundation for future research in this field.

Reference

Avcı, M. (2025). Rest Intolerance, Emotional Distress, Insomnia, and Adaptive Coping Strategies: A Validation and Serial Mediation Analysis Study. *The Psychiatric Quarterly*. https://doi.org/10.1007/s11126-025-10176-0

Cepellos, V. M., & Tonelli, M. J. (2020). Grounded theory: The step-by-step and methodological issues in practice. *RAM. Revista de Administração Mackenzie*, 21, eRAMG200130.

Cheng, W., Cheng, L., Qian, Z., & Wang, H. (2025). When winning costs your peace: How does individualism Hijack relaxation capacity? Network analysis and vertical mediation models. *PLOS ONE*, 20(11), e0335851. https://doi.org/10.1371/journal.pone.0335851

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1), 3–21. https://doi.org/10.1007/BF00988593

Corbin, J., & Strauss, A. (2025). *Basics of Qualitative Research* (3rd ed.): Techniques and Procedures for Developing Grounded Theory. https://doi.org/10.4135/9781452230153

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4), 997–1019.

Green, C. D. (2015). Why psychology isn' t unified, and probably never will be. *Review of General Psychology*, 19(3), 207–214. https://doi.org/10.1037/gpr0000051

Jalali, M. S., & Akhavan, A. (2024). Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review*, 40(3), e1772.

Kabir, S. M. A., Ali, F., Ahmed, R. L., & Sulaiman-Hill, R. (2025). Exploring the Use of AI in Qualitative Data Analysis: Comparing Manual Processing with Avidnote for Theme Generation. *International Journal of Qualitative Methods*, 24, 16094069251336810. https://doi.org/10.1177/16094069251336810

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_{00638}

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, 22, 16094069231211248.

Pope, C., Ziebland, S., & Mays, N. (2000). Qualitative research in health care. Analysing qualitative data. *BMJ* (Clinical Research Ed.), 320(7227), 114. https://doi.org/10.1136/bmj.320.7227.114

Sammon, D., McCarthy, S., Thummadi, B. V., Wibisono, A., & Fitzgerald, B. (2024). Exploring the potential of large language models (LLMs) for grounded theorizing: A human-in-the-loop configuration.

Schroeder, H., Aubin Le Quéré, M., Randazzo, C., Mimno, D., & Schoenebeck, S. (2025). Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. 1–17.

Siiman, L. A., Rannastu-Avalos, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023). Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. 87–96.

Sinha, R., Solola, I., Nguyen, H., Swanson, H., & Lawrence, L. (2024). The role of generative AI in qualitative research: GPT-4' s contributions to a grounded theory analysis. 17–25.

Übellacker, T. (2024). AcademiaOS: Automating Grounded Theory Development in Qualitative Research with Large Language Models. *arXiv Preprint arXiv:2403.08844*.

van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 16(4). https://doi.org/10.1177/1745691620970604

Walker, D., & Myrick, F. (2006). Grounded theory: An exploration of process and procedure. *Qualitative Health Research*, 16(4), 547–559. https://doi.org/10.1177/1049732305285972

Wang, F., Ge, P., Li, D., Cai, L., Li, X., Sun, X., & Wu, Y. (2022). The impact of infectious disease prevention behavior on quality of life: A moderated mediation model. *Health Care Science*, 1(3), Article 3.

Wang, F., Song, H., Meng, X., Wang, T., Zhang, Q., Yu, Z., Fan, S., & Wu, Y. (2025). Development and validation of the long and short forms of the rest intolerance scale for college students. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2024.112869

Wang, F., Tang, J., Sun, X., Sun, X., Li, J., Meng, X., & Wu, Y. (2024). Design and Development of Scales in Primary Care: Practical Steps and Statistical Methods. *Chinese General Practice*, 27(13), 1573–1583.

Wang, F., Wu, Y., Sun, X., Wang, D., Ming, W.-K., Sun, X., & Wu, Y. (2022). Reliability and validity of the Chinese version of a short form of the family health scale. *BMC Primary Care*, 23(1). https://doi.org/10.1186/s12875-022-01702-1

Wang, F., Zhu, X., Pi, L., Xiao, X., & Zhang, J. (2024). Patterns of participation and performance at the class level in English online after-school education in China: A longitudinal cluster analysis of online education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12451-2

Yang, X. (2025). The Cinematic Construction of Social Acceleration: Nice View (2022) and the Other Side of 'China Speed' . *Asian Studies Review*, 1–16.

Yue, Y., Liu, D., Lv, Y., Hao, J., & Cui, P. (2025). A Practical Guide and Assessment on Using ChatGPT to Conduct Grounded Theory: Tutorial. *Journal of Medical Internet Research*, 27, e70122.

Zhao, J., Wu, Y., Xu, J., Du, K., Yang, Y., & Zang, S. (2025). Rest intolerance and associated factors among Chinese nursing students: A cross-sectional network analysis. *BMC Nursing*, 24(1), 793. https://doi.org/10.1186/s12912-025-03472-4

Note: Figure translations are in progress. See original paper for figures.