Narrative Text Anxiety Prediction Based on Large Language Models: A Comparison of Different Models and Prompt Styles

Authors: Yutong Zhai, Wang Wenqian, Ren Cheng, Luo Min, Ting-Shao Zhu, Ting-Shao Zhu

Date: 2026-03-09T13:48:29+00:00

Abstract

Anxiety primarily encompasses subjective anxiety experiences, cognitive alertness, and physiological and somatic symptoms. With the development of social media, identifying anxiety from individual textual data to assist clinical treatment has broad application prospects. In the past, traditional machine learning methods were mostly employed to identify anxiety in text, but their accuracy was generally suboptimal. Large language models developed in recent years have provided new potential pathways for textual anxiety recognition.

This study explores the effectiveness of large language models in identifying anxiety within self-reported narrative texts under different models and various prompt styles. In the study, four experts pre-scored 50 texts, demonstrating good inter-rater consistency, followed by expert scoring of the final 252 texts. Under the guidance of four prompting strategies (zero-shot, few-shot, chain of thought, and few-shot + chain of thought), Qwen-max and Deepseek-reasoner were utilized for large model scoring.

The results indicated that the consistency between Qwen-max and the experts reached approximately 0.8, which was significantly superior to Deepseek-reasoner (0.6-0.7). The prompting strategy with both examples and a chain of thought (few-shot + chain of thought) yielded the best evaluation performance, while the strategy without examples or a chain of thought (zero-shot) performed the worst. Compared to Deepseek-reasoner, Qwen-max was less sensitive to different prompting strategies. Furthermore, for Deepseek-reasoner, the inclusion of examples in the prompts (compared to the chain of thought) was more helpful in improving its evaluation. This study provides a preliminary technical solution for the future assessment and screening of anxiety in text.

Full Text

Preamble

Narrative Text Anxiety Prediction Based on Large Language Models: A Comparison of Different Models and Prompting Styles

Zhai Yutong[1,2], **Wang Wenqian**[1,2], **Ren Cheng**[1,2], **Luo Min**[1,2], **Zhu Tingshao**[1,2*] [1](Department of Psychology, University of Chinese Academy of Sciences, Beijing 101408) [2](Institute of Psychology, Chinese Academy of Sciences, Beijing 100101)

Abstract

Anxiety mainly encompasses subjective anxious experiences, cognitive vigilance, and physiological or somatic symptoms. With the rapid development of social media, identifying anxiety from individual textual data to assist clinical treatment has demonstrated broad application prospects. Historically, traditional machine learning methods have been the primary tool for identifying anxiety in text, yet their accuracy has consistently remained suboptimal. The recent emergence of Large Language Models (LLMs) offers a promising new pathway for textual anxiety recognition. This study investigates the effectiveness of LLMs in identifying anxiety within self-reported narrative texts across different models and prompting styles. In this research, four experts performed preliminary scoring on 50 texts, achieving high inter-rater reliability, followed by expert scoring of the final 252 texts. We utilized Qwen-max and DeepSeek-reasoner to perform model-based scoring guided by four prompting strategies: zero-shot, few-shot, Chain of Thought (CoT), and few-shot + Chain of Thought (FS+CoT). The results indicate that Qwen-max achieved a consistency of approximately 0.8 with expert ratings, significantly outperforming DeepSeek-reasoner (0.6–0.7). The prompting strategy incorporating both examples and reasoning steps (few-shot + CoT) yielded the best evaluative performance, while the strategy lacking both (zero-shot) performed the worst. Furthermore, Qwen-max demonstrated less sensitivity to different prompting strategies compared to DeepSeek-reasoner. For DeepSeek-reasoner specifically, the inclusion of examples in the prompt was more effective at improving evaluation performance than the inclusion of a chain of thought alone. This study provides a preliminary technical framework for the future assessment and screening of anxiety in textual data.

Keywords: Anxiety; Large Language Models; Prompting Style; Self-reported Narrative Text

---

1. Introduction

1.1 Operational Definition of Anxiety

Anxiety, as a core emotional experience, is ubiquitous in human life. It is primarily rooted in the cognitive appraisal system and manifests as excessive worry about future threats and avoidant behaviors [**?**, **?**, **?**].

Distinct from the immediate fear response to real-world threats, anxiety emphasizes the prospective processing of uncertainty and potential risks. Research indicates that anxious individuals not only exhibit heightened responses to explicit threat cues but also show significant fear and vigilance toward non-threatening cues, previously threat-related cues, and ambiguous stimuli. This is accompanied by high reactivity to aversive stimuli, attentional bias toward threat-related information, and a cognitive tendency to interpret ambiguous stimuli as threatening [**?**, **?**]. Consequently, anxiety manifests not only as a subjective emotional experience but also involves systemic changes in cognitive processing patterns and physiological arousal levels.

Within the framework of emotional structure theories, the tripartite model proposed by Clark and Watson (1991) provides an essential foundation for understanding the distinction between anxiety and depression. This model posits that emotions are composed of three dimensions: negative affect, positive affect, and physiological hyperarousal. Specifically, anxiety is characterized primarily by high levels of physiological arousal and tension, whereas depression is more characterized by a lack of positive affect (anhedonia). Empirical research based on this model further suggests that anxiety is not a unidimensional construct but is composed of multiple components, including subjective experience, cognitive appraisal, and physiological responses [**?**, **?**].

At the measurement level, the operational definitions of anxiety across various scales reflect these multidimensional characteristics. The Depression Anxiety Stress Scales (DASS-21) define anxiety through physiological hyperarousal, subjective anxious experiences, and situational anxiety responses, excluding positive affect components [**?**, **?**, **?**]. The Beck Anxiety Inventory (BAI), the Generalized Anxiety Disorder 7-item scale (GAD-7), and the diagnostic criteria in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) focus primarily on subjective feelings of apprehension and physiological symptoms [**?**, **?**, **?**]. Furthermore, the Self-Rating Anxiety Scale (SAS) categorizes anxiety into four factors: subjective anxiety experience, cognitive and vigilance-related irritability, autonomic nervous system symptoms, and somatic symptoms. This systematically characterizes the comprehensive manifestation of anxiety across subjective, cognitive, and physiological domains.

Building upon emotional structure theories and clinical measurement frameworks, this study operationally defines anxiety and divides it into three interrelated core dimensions. First, the **Subjective Anxiety Experience** dimension reflects the emotional feelings individuals generate when facing potential threats,

including anxiety, fear, panic, feelings of losing control, and nightmares. Second, the **Cognitive and Vigilance Irritability** dimension reflects cognitive processing styles and sustained levels of alertness under anxious states, including premonitions of misfortune, fatigue, restlessness, and sleep disturbances. Third, the **Physiological Symptoms** dimension integrates autonomic nervous system and somatic responses, including dyspnea, urinary frequency, diaphoresis, tremors, somatic pain, palpitations, and dizziness. This three-dimensional operational definition structurally inherits the symptom classification logic of scales such as the SAS, BAI, and DASS-21, providing a clear theoretical foundation for text-based anxiety prediction.

1.2 Research Progress in AI-Based Textual Anxiety Recognition

With the widespread adoption of social media, text has increasingly become a vital medium for individuals to express their emotions. This shift provides a rich source of data for identifying anxiety using artificial intelligence. The primary advantage of these methods lies in their ability to automatically extract semantic features from large-scale unstructured text, enabling early identification and improving the accuracy, efficiency, and scalability of detection.

Existing research indicates that data sources for anxiety recognition primarily fall into two categories: pure text data (social media, blogs, journals, counseling transcripts) and multimodal data (speech, video, emojis). Traditional machine learning models, such as Support Vector Machines (SVM) and Naive Bayes, often rely on manually constructed features like word frequency and sentiment ratios. In contrast, deep learning methods, such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), automatically learn hierarchical representations of text.

Representative studies demonstrate progress: Yu et al. (2023) used an XGBoost model to predict anxiety levels from Weibo text with a reliability of approximately 72%. Kim et al. (2020) employed a CNN on Reddit data, achieving 77.81% accuracy. Tyshchenko et al. (2018) reached approximately 78% accuracy using SVM and CNN models on blog text. Despite these advancements, predictive performance has not yet reached the high-precision standards required for clinical application, and the ability of these models to understand deep semantic meanings remains limited.

1.3 Research Background and Problem Statement

To address these gaps, Large Language Models (LLMs) offer a promising new path. Unlike previous models, LLMs develop holistic representations of contextual relationships and implicit semantics through self-supervised learning on large-scale corpora. Existing research indicates that different LLMs exhibit significant variations in emotional judgment, and that prompt engineering styles (zero-shot, few-shot, and Chain-of-Thought) can substantially influence performance in mental health tasks [**?**, **?**].

However, systematic comparisons of various LLMs for anxiety text identification remain scarce. Most studies rely on structured datasets, leaving the applicability of these models to highly unstructured narrative texts from real-world social media under-investigated. Therefore, this study systematically compares the performance of different LLMs and prompting styles in anxiety prediction tasks using individual self-reported narrative texts.

---

2. Research Methodology

2.1 Data Collection and Processing

This study employs LLM technology to explore the effectiveness of psychological trait identification. The overall research design is illustrated in [Figure 1: see original paper].

(1) Data Sources This study selected two representative Chinese internet platforms and classroom instructional materials as primary data sources.

(2) Inclusion and Exclusion Criteria Inclusion criteria: (a) first-person self-expression; (b) explicit descriptions of emotional experiences or psychological states; (c) text length between 100–1000 characters. Exclusion criteria: advertising, purely objective factual statements, and garbled text.

(3) Data Cleaning and Sample Size Raw data were processed using Python to remove HTML tags, URLs, and special symbols. After manual review, a dataset of $N = 252$ high-quality text entries was constructed. All data were anonymized to ensure privacy compliance.

2.2 Expert Annotation and Consistency Testing

To establish a "gold standard,"this study followed an iterative expert annotation method [Figure 2: see original paper].

(1) Annotation Team and Standard Formulation The team consisted of four psychology master' s students. A coding scheme for anxiety (scale 0–3) was developed based on clinical diagnostic criteria.

(2) Iterative Annotation Process Fifty samples were used for training. Inter-rater reliability was calculated using the Intraclass Correlation Coefficient (ICC). After reaching $ICC > 0.7$, the remaining data were distributed. Team members worked in pairs to evaluate 126 texts each, resolving discrepancies through discussion to reach a final consensus for the 252 texts.

2.3 Large Language Model Prompt Engineering

Four distinct prompting styles were constructed [Figure 3: see original paper]:

1. **Zero-Shot Style**: Includes role setting, task description, operational definitions of the three assessment factors, and output requirements without examples.
2. **Few-Shot Style**: Adds 1–3 "anchor examples" (input text, expert ratings, and rationales) to the zero-shot prompt to calibrate the model' s understanding.
3. **Chain-of-Thought Style (CoT)**: Requires the model to explicitly generate its analytical process (identifying symptoms, checking dimensions, explaining the basis) before providing the final score.
4. **Few-Shot + CoT**: Combines both examples and reasoning steps.

2.4 Evaluation Metrics and Analysis Methods

1. **Consistency Analysis**: Pearson correlation and Kendall' s tau are used to measure the agreement between AI-generated scores and human expert ratings. A Kendall' s tau > 0.6 indicates high consistency.
2. **Model Comparison**: Two representative domestic LLMs, Qwen-max and DeepSeek-reasoner (V3), were selected for horizontal evaluation [Figure 4: see original paper].

2.5 Research Hypotheses

1. **Feasibility**: LLM ratings will achieve a consistency coefficient ≥ 0.6 with expert ratings.
2. **Model Performance**: High-performance general-purpose models (Qwen-max) will exhibit better adaptability than specific reasoning models for Chinese psychological texts.
3. **Prompting Strategies**: Prompts incorporating both examples and CoT will yield superior results [**?**].

---

3. Results

3.1 Expert Ratings

The initial scoring of 50 texts yielded an ICC of 0.863, 95% CI [.782, .917], indicating high reliability. The final consensus scores for the 252 texts served as the gold standard.

3.2 LLM Performance Under Different Prompts

Consistency results between LLMs and expert ratings are summarized in .

Consistency between machine evaluation and expert evaluation across different large language models and prompting styles.

Prompting Style	Qwen-max (Kendall' s τ)	DeepSeek-reasoner (Kendall' s τ)
Zero-shot	0.786*** [0.735, 0.837]	0.622*** [0.550, 0.695]
Few-shot	0.793*** [0.738, 0.848]	0.655*** [0.584, 0.726]
Zero-shot + CoT	0.798*** [0.745, 0.851]	0.627*** [0.558, 0.696]
Few-shot + CoT	0.810*** [0.763, 0.857]	0.664*** [0.601, 0.727]

Note: *** $p < 0.001$; CI denotes 95% Confidence Interval.

Both models met the consistency threshold (Kendall' s $\tau > 0.6$). Qwen-max significantly outperformed DeepSeek-reasoner across all categories. The "Few-shot + CoT" strategy yielded the highest performance, while "Zero-shot" performed the worst. Qwen-max showed lower sensitivity to prompt variations compared to DeepSeek-reasoner [Figure 5: see original paper].

---

4. Discussion

4.1 Reliability of Expert Ratings

The high ICC (0.863) confirms the stability of the expert assessment system [**?**, **?**]. This establishes a robust "gold standard" for validating LLM performance in transforming abstract anxiety levels into quantifiable metrics.

4.2 Overall Model Performance Differences

Both models achieved significant consistency with experts, but Qwen-max ($\approx$ 0.8) was notably superior to DeepSeek-reasoner (0.6–0.7). This likely reflects Qwen-max' s specialized optimization for emotional semantic understanding in Chinese corpora [**?**, **?**]. DeepSeek-reasoner, while strong in logic, may lack targeted optimization for fine-grained emotional recognition [**?**, **?**].

4.3 Impact of Prompting Styles

The "Few-shot + CoT" style' s success validates the synergy of providing benchmarks and structured reasoning [**?**, **?**, **?**]. Qwen-max's low sensitivity to prompts suggests robust foundational semantic understanding. Conversely, DeepSeek-reasoner' s higher sensitivity suggests that models with weaker baseline emotional understanding rely more heavily on few-shot examples to calibrate their performance [**?**, **?**].

---

5. Conclusion

This study demonstrates that LLMs can effectively identify anxiety in narrative texts. Qwen-max outperformed DeepSeek-reasoner, and the "Few-shot + CoT" strategy provided the most accurate results. These findings offer an efficient technical framework for automated anxiety screening and intelligent auxiliary diagnosis based on natural language.

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