Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.35563

Research on Emotional Persuasion Effect Identification Method of Legal Large Language Models for Human AI Collaborative Justice

Suran Chen[1*], Jingyu Yang[2]

[1]College of Computing, City University of Hong Kong, Hong Kong, China
[2]National Security and Legal Education Research Centre (NSLERC), The Education University of Hong Kong, Hong Kong, China
**Corresponding Author. Email: wellington589125@gmail.com*

Abstract. Legal large language models are increasingly entering human AI collaborative judicial scenarios, including legal consultation, similar-case retrieval, judgment assistance, and document generation. While these models provide legal information, their outputs may also influence human judgment through emotional expression, certainty markers, authoritative citation, and responsibility attribution. Focusing on emotional persuasion risks in collaborative justice, this study constructs a legal emotional persuasion corpus, integrates case facts, statutory references, emotion intensity, persuasion strategies, and human judgment changes, designs a multi-feature fusion identification model, and conducts a simulated judicial judgment experiment. The experiment examines how model-generated answers affect responsibility attribution, evidence strength, judgment tendency, and trust level. Preliminary results show that the multi-feature fusion model outperforms BERT, Legal-BERT, and RoBERTa in Accuracy, Macro-F1, and AUC. Composite persuasion answers produce stronger decision shifts in judgment tendency and trust level. These findings indicate that legal LLM risk assessment should consider emotional influence mechanisms in addition to factual accuracy. The proposed framework provides an explainable technical path for judicial AI auditing, risk warning, and human oversight.

Keywords: Legal Large Language Model, Emotional Persuasion, Human AI Collaboration, Judicial Decision Support, Explainable Risk Identification

1. Introduction

Legal large language models have been applied to legal consultation, similar-case retrieval, document generation, and judgment assistance, and their generated outputs are beginning to influence legal information access, case understanding, and preliminary decision-making processes [1]. Existing studies have focused on legal reasoning, hallucination risks, factual consistency, and benchmark evaluation, showing that model performance in professional legal tasks still requires multidimensional assessment [2]. In human AI collaborative justice, emotional wording, certainty markers, responsibility attribution, and risk warnings in model outputs may reshape users'

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.35563

expectations of evidential strength and judicial outcomes [3]. The research therefore focuses on the computational identification of emotional persuasion effects in legal LLMs. It plans to construct a legal question-answering corpus, design a multi-feature fusion model, and conduct human AI collaborative experiments to examine the relationship among emotional expression, persuasion strategy, and decision shift, thereby providing an explainable method for judicial AI risk assessment.

2. Literature review

2.1. Legal large language models and judicial assistance

Research on legal LLMs in judicial assistance mainly covers legal question answering, contract review, case summarization, statutory matching, and judgment prediction. Existing studies emphasize that models should satisfy language generation quality, reliability of legal authority, and traceability of reasoning at the same time [4]. As models move from information retrieval tools to judgment-support tools, judicial settings impose higher requirements on responsibility allocation, procedural justice, and human oversight, extending the boundary of legal AI from technical usability to institutional acceptability [5]. This line of research provides the application basis for identifying emotional persuasion effects, because output risks in judicial assistance systems arise not only from factual errors but also from how models present legal facts, responsibility tendencies, and possible outcomes.

2.2. Emotional persuasion and legal judgment shift

Studies on emotional persuasion show that texts can influence recipients' judgments of credibility and action choices through sympathy, fear, anger, trust, and moral evaluation, while the natural language generation capacity of LLMs further strengthens this influence path [6]. In experiments on policy and public issues, model-generated messages have shown the ability to change individual attitudes, indicating that persuasive effects can be amplified through linguistic style, stance organization, and personalized expression [7]. Legal judgment involves fact-finding, normative application, and responsibility evaluation. Once emotional expression is embedded in legal advice, users may develop shifted perceptions of evidential sufficiency, fault degree, and likely judicial outcomes, which can affect neutrality in human AI collaborative justice.

2.3. Computational identification and explainable risk evaluation

Computational persuasion research usually identifies textual influence mechanisms through lexical features, semantic structures, argumentative strategies, and audience responses, while machine learning and natural language processing provide technical routes for automatic detection of emotional persuasion effects [8]. Emotion recognition research has further expanded from sentence-level sentiment classification to contextual emotion understanding, intensity measurement, and dialogic relation modeling, allowing emotional expression in complex contexts to be transformed into computable features [9]. In legal LLM scenarios, identification methods need to process legal semantics, emotion intensity, certainty expression, authoritative citation, and human decision change together, so that ordinary sentiment analysis can be upgraded into an explainable judicial risk evaluation method.

DOI: 10.54254/2753-8818/2026.35563

3. Experimental methods

3.1. Construction of legal emotional persuasion corpus

Data collection is organized around four components: case facts, model-generated answers, emotional persuasion annotations, and human judgment results. The time span is set from 2020 to 2025. Case materials are mainly collected from publicly available judicial documents, while statutory references are obtained from official legal databases. Task types are designed according to legal question answering, similar-case retrieval, explanation of judicial reasoning, and responsibility judgment. The corpus covers civil disputes, criminal defense, administrative disputes, and labor disputes, with 400 basic factual samples collected for each type, forming 1600 raw samples in total. Each sample contains four answer versions: neutral answer, emotion-enhanced answer, authority-enhanced answer, and composite persuasion answer. Annotators with legal backgrounds label emotion type, emotion intensity, persuasion strategy, completeness of legal authority, responsibility tendency, and risk level. Samples with incomplete facts, obvious legal errors, or highly repetitive expressions are removed to form a structured corpus for model training and human AI experiments.

3.2. Design of multi-feature fusion identification model

The model is built on legal semantic understanding and integrates emotion intensity, certainty expression, authoritative citation, and missing legal authority features. A legal domain pretrained encoder is first used to extract vectors of case facts, disputed issues, statutory references, and judicial reasoning. Then, emotional word density, risk amplification phrases, moral evaluation terms, and responsibility attribution expressions are transformed into numerical features [10]. The integrated risk score is used to determine whether a model answer contains a clear emotional persuasion tendency, as shown in Formula (1), where P_i denotes emotion intensity, C_i denotes the proportion of certainty expressions, A_i denotes citation density, P_i denotes persuasion strategy probability, and L_i denotes the penalty for missing legal authority.

$$R_i=\sigma(\alpha E_i+\beta C_i+\gamma A_i+\delta P_i+\eta L_i) \tag{1}$$

To measure the influence of model outputs on human judgment, a standardized decision shift index is further calculated, as shown in Formula (2). Here, s_{ij}^{post} and s_{ij}^{pre} denote the participant's rating after and before reading the answer, σ_j denotes the standard deviation of the j-th judgment indicator, and m denotes four indicators: responsibility attribution, evidence strength, judgment tendency, and trust level.

$$D_i=\frac{1}{m}\sum_{j=1}^{m}\frac{\left|s_{ij}^{post}-s_{ij}^{pre}\right|}{\sigma_j} \tag{2}$$

3.3. Human AI collaborative judicial simulation

The experiment selects 320 samples from the test set and adopts a between-group randomized design to control the influence of answer types on judgment outcomes. Ninety-six participants are divided into a legal-background group and a general-user group, with 48 participants in each group. Before the experiment, participants read a case fact summary and provide pre-test ratings on responsibility proportion, evidential sufficiency, judgment tendency, and answer credibility. The

DOI: 10.54254/2753-8818/2026.35563

system then randomly presents one version of the LLM-generated answer, including neutral, emotion-enhanced, authority-enhanced, or composite persuasion answers. Each participant completes 20 case judgment tasks, with each task limited to three minutes. The system records reading time, rating changes, final scores, and confidence scores [11]. To reduce order effects, case types and answer versions are arranged through a Latin square design, while attention-check items are used to remove invalid responses. After the experiment, pre-post rating differences are entered into Formula (2) to calculate decision shift, and the results are matched with the risk scores generated by Formula (1). Correlation and regression analyses are then used to test whether emotion intensity, certainty expression, and authoritative citation significantly affect human AI collaborative judicial judgment.

4. Results

4.1. Performance of the emotional persuasion identification model

The preliminary experiment uses 1120 training samples, 160 validation samples, and 320 test samples to compare BERT, Legal-BERT, RoBERTa, and the multi-feature fusion model. As shown in Table 1, the multi-feature fusion model reaches 0.867 in Accuracy, 0.834 in Macro-F1, and 0.889 in AUC, outperforming the three baseline models. Its ECE decreases to 0.071, indicating more stable risk probability outputs. Legal-BERT performs better than ordinary BERT in legal semantic understanding because of its legal-domain pretraining, yet it remains limited in identifying implicit emotional cues and composite persuasion strategies. RoBERTa shows stable overall classification performance, but its recall of high-risk samples is restricted when features such as statutory completeness, responsibility tendency, and emotion intensity are absent. The multi-feature fusion model performs better in judicial collaborative risk identification because it jointly considers legal semantics, emotional expression, and argumentative structure.

Table 1. Performance comparison of emotional persuasion identification models

Model	Accuracy	Macro-F1	AUC	ECE
BERT	0.781	0.742	0.801	0.128
Legal-BERT	0.812	0.779	0.834	0.106
RoBERTa	0.825	0.792	0.846	0.098
Multi-feature Fusion Model	0.867	0.834	0.889	0.071

4.2. Decision impact in human AI collaborative scenarios

The human AI collaborative experiment shows that different answer types produce distinct effects on participant judgment. Neutral answers cause relatively low average decision shifts, with standardized shift values below 0.05 across responsibility attribution, evidence strength, judgment tendency, and trust level, indicating that ordinary legal explanations have limited influence on initial judgment. Emotion-enhanced answers produce shift values of 0.22 in responsibility attribution and 0.27 in judgment tendency, showing that sympathy, fear, and moral evaluation can strengthen participants' tendency to attribute responsibility to one side. Authority-enhanced answers have a stronger effect on trust, reaching a shift value of 0.21, which indicates that dense statutory citation and certainty markers increase perceived credibility. Composite persuasion answers produce the strongest effect, with judgment tendency reaching 0.38 and trust reaching 0.34. This suggests that

DOI: 10.54254/2753-8818/2026.35563

emotional expression and authoritative framing jointly amplify human judgment changes. Figure 1 presents the distribution of judgment shifts under four answer conditions through a matrix heatmap.



Figure 1. Decision shift matrix under different legal LLM answer conditions

5. Discussion

The experimental results show that the risks of legal large language models are reflected not only in factual errors or incorrect legal citations, but also in the way their outputs may shape human judgment. Emotion-enhanced answers have a stronger influence on responsibility attribution and judgment tendency, while authority-enhanced answers more directly increase trust. Composite persuasion answers show a more evident cumulative effect. This suggests that model outputs in judicial collaboration should be evaluated through legal semantics, emotional expression, argumentative structure, and behavioral response. The multi-feature fusion model performs well in detecting high-risk answers, but annotation cost, jurisdictional differences, and the complexity of real courtroom interaction may still affect its generalizability. Future research can introduce larger-scale real interaction data and cross-jurisdictional samples to improve stability.

6. Conclusion

This study develops a complete methodological framework for identifying emotional persuasion effects of legal large language models in human AI collaborative justice. The framework connects corpus construction, feature modeling, and human judgment experiments. It integrates legal semantics, emotion intensity, certainty expression, authoritative citation, and missing legal authority into a unified identification process, and uses a decision shift index to verify the influence of model outputs on human judgment. The results show that the multi-feature fusion method can improve the detection of emotional persuasion risks and reveal the different effects of answer types on judicial decision-making. The framework provides a methodological basis for output review, risk warning, and responsibility control in legal AI systems.

DOI: 10.54254/2753-8818/2026.35563

Author contribution

Suran Chen and Jingyu Yang contributed equally to this paper.

References

[1] Dahl, M., et al. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93.
[2] Lai, J., et al. (2024). Large language models in law: A survey. *AI Open, 5*, 181–196.
[3] Dehghani, F., et al. (2025). Large language models in legal systems: A survey. *Humanities and Social Sciences Communications, 12*(1), 1977.
[4] Nguyen, H. T., et al. (2025). LLMs for legal reasoning: A unified framework and future perspectives. *Computer Law & Security Review, 58*, 106165.
[5] Schweitzer, S., & Conrads, M. (2025). The digital transformation of jurisprudence: An evaluation of ChatGPT-4's applicability to solve cases in business law. *Artificial Intelligence and Law, 33*(3), 847–871.
[6] Dugac, G., & Altwicker, T. (2025). Classifying legal interpretations using large language models: G. Dugac, T. Altwicker. *Artificial Intelligence and Law*, 1–19.
[7] Sargeant, H., Izzidien, A., & Steffek, F. (2025). Topic classification of case law using a large language model and a new taxonomy for UK law: AI insights into summary judgment. *Artificial Intelligence and Law*, 1–49.
[8] Ribeiro de Faria, J., Xie, H., & Steffek, F. (2025). Information extraction from employment tribunal judgments using a large language model. *Artificial Intelligence and Law*, 1–22.
[9] Perona, R., & Carrillo de la Rosa, Y. (2025). Unveiling AI in the courtroom: Exploring ChatGPT's impact on judicial decision-making through a pilot Colombian case study. *AI & Society, 40*(4), 2533–2540.
[10] Marcos, H. (2025). Can large language models apply the law? *AI & Society, 40*(5), 3605–3614.
[11] Terzidou, K. (2025). Generative AI systems in legal practice offering quality legal services while upholding legal ethics. *International Journal of Law in Context, 21*(3), 431–452.