



Limitations in cultural context: systematic biases of LLMs in Chinese sentiment analysis

Ruili Geng, Jiwen Zhang, Ruixian Yang, Mingzhe Quan, Xiang Zheng *, and Yishuai Xu

DOI: <https://doi.org/10.47989/ir31iConf64278>

Abstract

Introduction. Large language models (LLMs) are increasingly applied in Chinese sentiment analysis; however, their ability to interpret sentiment in culturally specific contexts remains uncertain. This study systematically evaluates the potential biases of LLM-based Chinese sentiment analysis, providing an empirical basis for model optimisation and responsible application.

Method. A human-AI comparison experiment was conducted at both word and sentence levels. Two hundred Chinese words from four categories and 225 sentences were used as materials. Judgment data on sentiment polarity, intensity, valence, and arousal were collected from native Chinese speakers and representative Chinese and international LLMs. Data analysis was performed using Chi-square tests and Kruskal-Wallis tests.

Results. In sentiment polarity judgment, LLMs are highly consistent with humans and outperform the traditional sentiment lexicon. For continuous dimensions (intensity, valence, arousal), LLMs generally show an exaggeration bias, especially for sensory-perceptual words and ironic sentences. The bias stems from the disembodied cognitive nature of models, exaggerated linguistic patterns in training data, and technical characteristics such as the attention mechanism.

Conclusion. This study reveals a systematic limitation of LLMs in Chinese sentiment analysis, characterised by accurate classification but exaggerated quantitative evaluations. Consequently, in practical applications, outputs related to sentiment intensity should be interpreted with caution.

Introduction

With the rapid development of large language models (LLMs), these systems are increasingly employed in both academic research and practical applications for sentiment analysis, demonstrating potential across specialised domains such as psychological therapy (Qiu et al., 2024) and financial investment (Mohsin, 2025). Against this backdrop, systematically evaluating the sentiment analysis capabilities of LLMs is crucial, as such capabilities directly influence user decision-making and the effectiveness of human-AI interaction.

Unlike human sentiment judgments, the sentiment analysis capabilities of LLMs rely primarily on large-scale training data and are consequently constrained by factors such as data quality and training strategies (Lin, 2025). Furthermore, due to differences in cultural context, LLMs may carry biases that reflect predominantly Western cognitive and cultural patterns (Qi et al., 2025; Dai, 2025). As a result, when processing languages like Chinese, where sentiment expression is relatively subtle and rich in cultural metaphors, LLMs' sentiment judgments may diverge from those of native speakers.

In light of this, we selected Chinese words and sentences representing various semantic types as experimental materials and conducted a systematic comparative evaluation between Chinese speakers and representative LLMs. The study aims to identify biases in LLM-based Chinese sentiment analysis and to offer insights for both technological refinement and responsible deployment.

Related works

LLMs offer a new technical pathway for sentiment analysis. Compared with sentiment lexicons and traditional machine learning methods, LLMs leverage large-scale pre-trained data to enable sentiment annotation and inference for unseen texts (Krugmann & Hartmann, 2024). Currently, models such as GPT-3.5 and GPT-4 have been applied in fields including social media (Mabokela et al., 2025) and financial news (Kirtac & Germano, 2024; Liu et al., 2024), demonstrating their effectiveness in sentiment analysis tasks.

However, existing studies have shown that the sentiment perception of LLMs may exhibit systematic biases compared to human cognition. Some experiments have found that GPT models tend to deviate from a neutral stance when describing mathematical concepts, tending to associate them with negative perceptions (Abramski et al., 2023). Based on social media texts related to English scientific papers, Pontes et al. found that models such as GPT and Gemini exhibit a polarisation tendency, showing a strong bias toward negative labeling, whereas human annotations are relatively evenly distributed among positive, neutral, and negative categories (Pontes & Maricato, 2025). Liu et al. designed a scenario in which participants encountered unfair reward distribution after a collaborative task, and the results indicated that LLMs displayed more intense negative reactions compared to humans (Liu et al., 2025). Furthermore, LLMs remain deficient in cross-cultural reasoning capabilities, demonstrating flexibility and depth in understanding ancient Chinese poetry that are inferior to human capabilities (Gao et al., 2025).

As one of the most widely spoken languages in the world, Chinese is rooted in a unique cultural tradition and often conveys sentiment through indirect forms such as metaphor and irony, differing in certain respects from English expression. However, existing research has largely relied on English corpora and focused primarily on models like GPT and Gemini. Models widely used in China, such as DeepSeek and Doubao, have large user bases but have yet to be systematically evaluated with respect to sentiment perception. Therefore, this study aimed to comparatively assess the sentiment analysis performance of Chinese and international LLMs in the Chinese context.

Methodology

Procedure and sample selection

We designed experiments at two levels: words and sentences (see Figure 1). The word-level experiment aimed to examine models' default sentiment tendencies at the minimal semantic unit, while the sentence-level experiment simulated more contextually grounded sentiment judgments closer to real-world applications. Drawing on the three-dimensional structure of sentiment and the valence-arousal model in affective research, we quantified sentiment judgments using polarity (positive/negative/neutral), intensity (1–5 scale), and valence and arousal (1–9 scale).

At the word level, 312 participants were recruited (160 females and 152 males); at the sentence level, 338 participants were involved (173 females and 165 males). All participants were native Chinese speakers, familiar with both mainstream and online cultures, and suitable for the experimental tasks. They completed sentiment polarity classification and intensity ratings via online questionnaires.

For the LLM sample, six representative models from China and abroad were selected, covering diverse technical and cultural backgrounds. All models were evaluated using the default parameters and online mode on their web interfaces, with each test conducted in an independent session without incorporating multi-turn-dialogue or historical conversational context. Uniform prompts were input, and the output of each model was recorded. Data collection was conducted between June 1 and July 31, 2025.

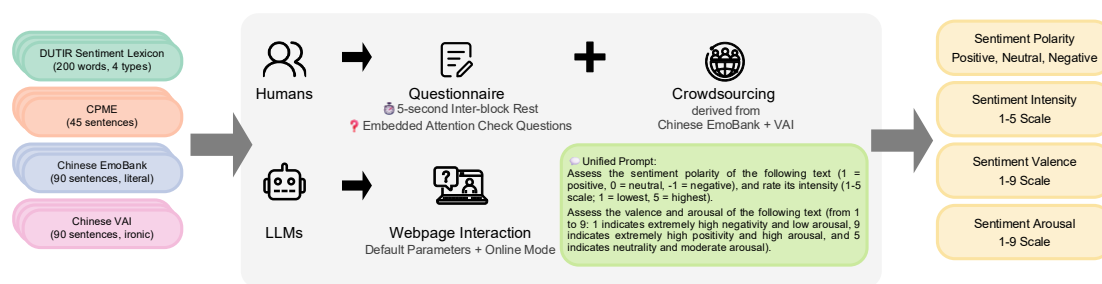


Figure 1. Experimental flowchart.

Materials

The words were selected from the DUTIR sentiment lexicon (Xu et al., 2008), while the sentences were drawn from the CPME (Zhou et al., 2024), Chinese VAI (Xie et al., 2021), and Chinese EmoBank (Lee et al., 2022) datasets (for detailed descriptions, see the Appendix), all of which are annotated for sentiment and can serve as reference frameworks. Based on the experimental scale used in studies of LLM sentiment recognition (Wang et al., 2023) and cultural comprehension (Li et al., 2024), we selected 200 words and 225 sentences. Prior to the formal experiment, the familiarity and comprehensibility of the experimental materials were screened by multiple authors and evaluators. The final selection of materials exhibited no significant unfamiliarity or interpretative ambiguity. The words were categorised into four types: culture-loaded words (e.g., ‘Dragon Boat Festival’ [端午], ‘Golden Hoop Spell’ [紧箍咒]), whose sentiment is rooted in specific historical and cultural contexts; sensory-perceptual words (e.g., ‘itchy’ [刺痒], ‘refreshing’ [爽口]), linked to embodied physiological experiences; internet slang (e.g., ‘like’ [稀饭], ‘embarrassed’ [糗大了]), with rapidly evolving semantics and sentiment in online contexts; and general basic words (e.g., ‘happy’ [开心], ‘dislike’ [讨厌]), featuring relatively stable and widely shared sentiment associations. The sentence materials comprised both literal and ironic sentences, covering straightforward as well as sarcastic expressions commonly used in Chinese.

Results

Reliability assessment

To assess the reliability of human annotations, we calculated Fleiss' Kappa. The coefficient for word-level sentiment polarity judgments was 0.671 ($p < 0.001$), and 0.613 ($p < 0.001$) at the sentence level, indicating strong consensus among participants and clear sentiment direction in the materials. For word-level sentiment intensity ratings, Kendall's W was 0.246 ($p < 0.001$), consistent with established patterns of sentiment perception. While intensity assessments are more susceptible to individual subjective experience, the overall agreement was within an acceptable range.

Categorical dimensions of sentiment

Figures 2 and 3 show the distribution of sentiment polarity and inter-group differences at the word and sentence levels for each group. Overall, LLMs and humans demonstrated a higher proportion of positive and negative sentiments compared to the DUTIR lexicon and CPME dataset, indicating a relative sentiment polarisation against these reference annotations. For different word types, LLMs and humans exhibited high consistency in sensory-perceptual words and general basic words, both significantly deviating from the lexicon with lower neutral proportions. For culture-loaded words and Internet slang, inter-group consistency decreased but the deviation from the lexicon was relatively mild.

At the word level, chi-square tests revealed significant differences in sentiment polarity distributions among LLMs, humans, and the lexicon for culture-loaded words, sensory-perceptual words, and general basic words. Post-hoc comparisons indicated that only Doubao 1.6 differed significantly from the lexicon in culture-loaded words; LLMs except Gemini 2.5 Flash showed significant differences from the lexicon in sensory-perceptual words; and all LLMs were significantly distinct from the lexicon in general basic words. Notably, there were no significant differences in sentiment polarity among LLMs or between LLMs and humans, with similar overall performance. At the sentence level, the chi-square test indicated that there were no significant differences in the distribution of sentiment polarity among LLMs, humans, and the dataset.

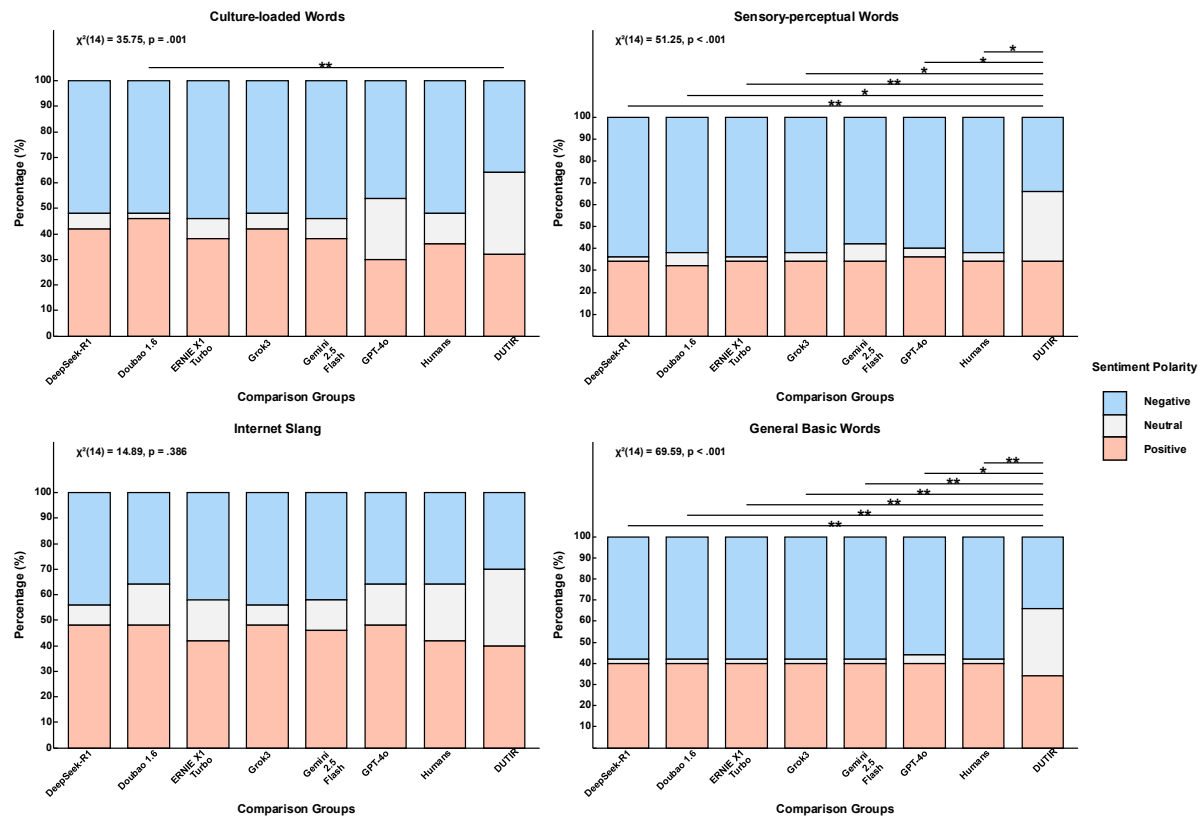


Figure 2. Polarity distribution and significant differences at the word level.

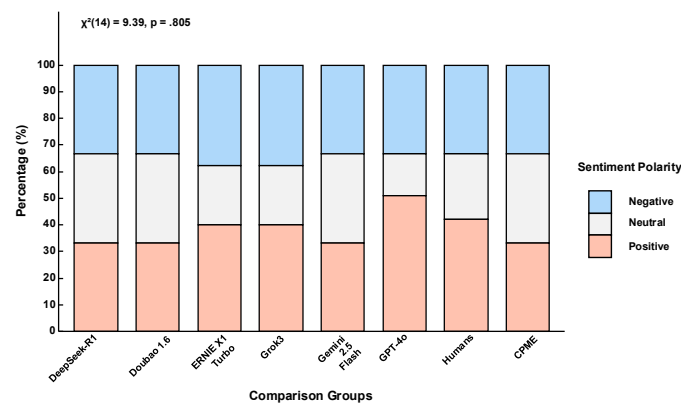


Figure 3. Polarity distribution and significant differences at the sentence level.

Continuous dimensions of sentiment

Figure 4 shows the distribution of sentiment intensity, mean values, and inter-group differences at the word level for each group. LLMs' sentiment intensity primarily clustered between levels 3 and 5 with mean values generally close to or higher than those of humans and the lexicon. Kruskal-Wallis tests revealed significant differences in sentiment intensity ratings across all four-word types among LLMs, humans, and the sentiment lexicon. Post-hoc comparisons indicated that significant differences in culture-loaded words occur between humans and DeepSeek-R1, and between the lexicon and DeepSeek-R1/Doubao 1.6; in sensory-perceptual words, differences existed between humans and all LLMs, and between the lexicon and DeepSeek-R1/Gemini 2.5

Flash; in Internet slang, significant differences are mainly observed between humans and DeepSeek-R1/GPT-4o; in general basic words, differences existed between humans and multiple LLMs, and between the lexicon and DeepSeek-R1/ERNIE X1 Turbo.

No statistically significant differences were found between humans and the lexicon, but their scatter distribution patterns differed—human ratings were more concentrated at level 3 while the lexicon’s values were relatively evenly distributed.

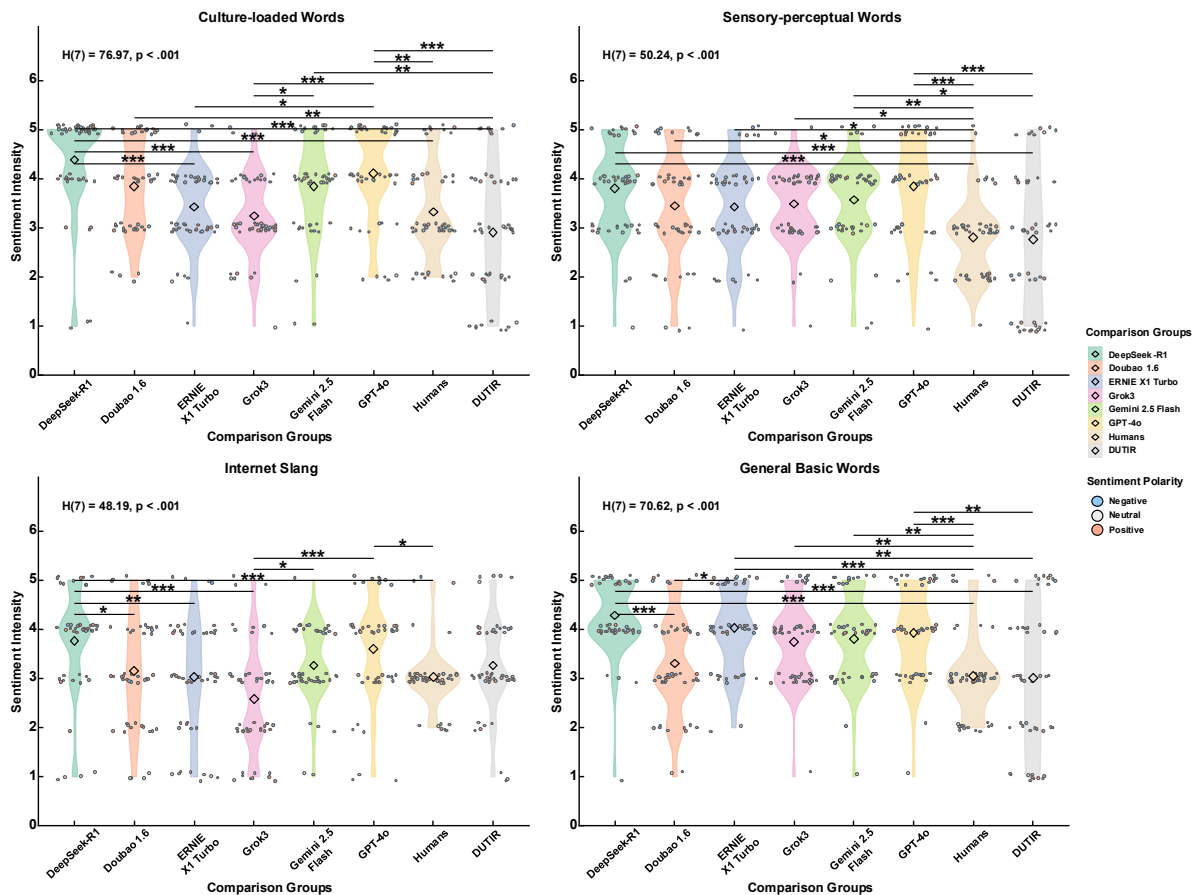


Figure 4. Intensity distribution and significant differences at the word level.

Figure 5 presents the distribution, mean, and intergroup differences of valence for each model and humans on literal and ironic sentences. For literal sentences, LLMs’ valence distributions are similar to those of humans with mean values close to human annotations. For ironic sentences, LLMs exhibited pronounced deviations from humans with mean values generally lower than human annotations, tending to overestimate the degree of negativity. Kruskal-Wallis tests revealed no significant differences in valence ratings among LLMs or between LLMs and human annotations for literal sentences while significant differences were observed for ironic sentences. Post-hoc tests indicated that all LLMs exhibited significant differences from humans in valence ratings for ironic sentences.

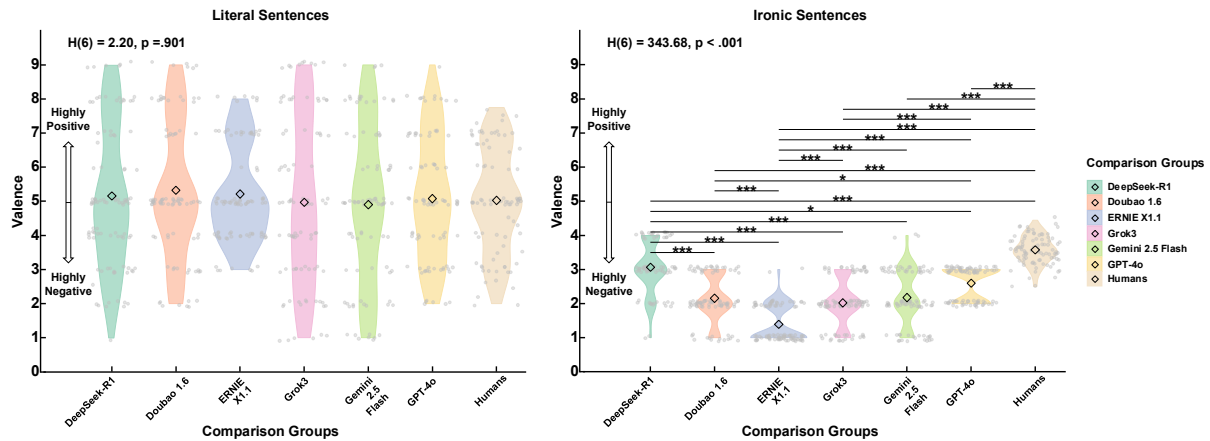


Figure 5. Valence distribution and significant differences at the sentence level.

Figure 6 shows the distribution, mean, and intergroup differences of arousal for each model and humans on literal and ironic sentences. For literal sentences, LLMs' arousal ratings cluster between 5 and 6 with mean values generally higher than human annotations. For ironic sentences, LLMs exhibited greater deviations from human annotations and all models except GPT-4o have mean values higher than human annotations. This indicates that LLMs generally exhibit an arousal bias, tending to exaggerate the sentiment activation level triggered by texts. In addition, for ironic sentences, arousal ratings show greater variability among LLMs, reflecting significant uncertainty in models' understanding of arousal in irony.

Kruskal-Wallis tests revealed significant overall differences in arousal ratings among LLMs and between LLMs and human annotations for both literal and ironic sentences. Post-hoc comparisons indicated that for literal sentences, significant differences existed between humans and Doubao 1.6/ERNIE X1.1/Gemini 2.5 Flash; for ironic sentences, all models except GPT-4o exhibited significant differences from human annotations and widespread differences existed among LLMs.

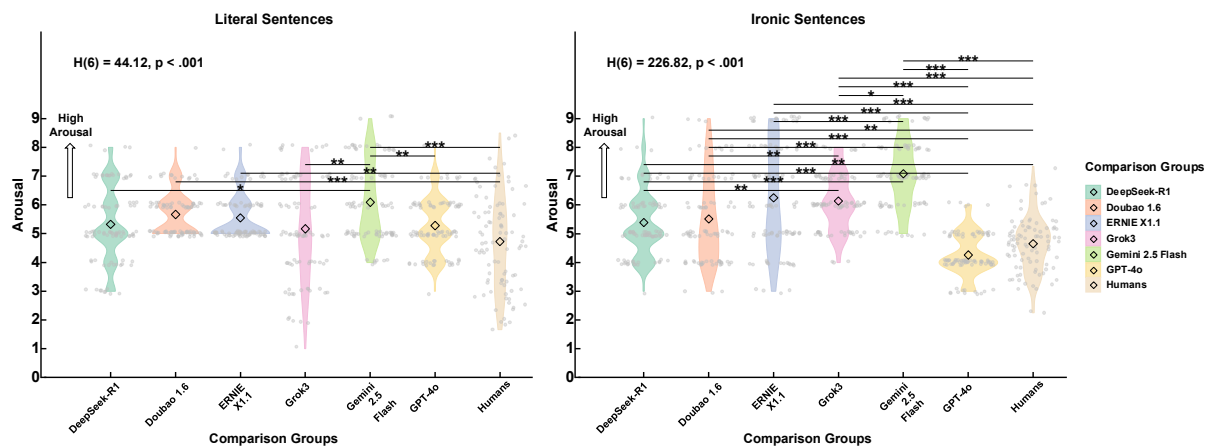


Figure 6. Arousal distribution and significant differences at the sentence level.

Findings

In sentiment polarity judgment, LLMs demonstrated high inter-model consistency and strong alignment with human consensus, tending to classify lexicon-labeled neutral words as positive or negative. This aligns with existing research—LLMs' sentiment polarity classification is closer to human cognition and outperforms traditional sentiment lexicon methods (Rathje et al., 2024). This superior performance likely stems from the fact that traditional sentiment lexicons, which are

based on relatively limited expert annotations, may be sensitive to annotator-specific characteristics such as gender, age, and cultural background (Ding et al., 2022), and may struggle to keep pace with the evolving nature of language use (Fu et al., 2024). In contrast, LLMs are trained on large-scale corpora and aligned with human preferences through supervised fine-tuning and reinforcement learning from human feedback (Bai et al., 2022).

In sentiment valence judgment, models converge with human annotators for literal sentences while showing significant deviations for ironic sentences—their valence ratings are significantly lower than human annotations, overestimating negative sentiments. For sentiment intensity, models' mean ratings are generally higher than those of humans and the lexicon. They also overestimate negative valence and arousal when processing ironic sentences. This indicates that models' sentiment overestimation is not merely random error but a systematic bias that tends to map sentiments to higher intensity levels. Existing research also confirms that LLMs exhibit stronger sentiment responses to positive and negative evaluations than humans (Bai et al., 2025).

This systematic bias toward sentiment exaggeration can be plausibly attributed to distinct technical and cognitive factors. Technically, LLMs are primarily trained on internet text, which often contains deliberately intensified sentiment expressions (Caspi & Etgar, 2023). After internalising this language style, models may exhibit reduced sensitivity to subtle sentiment cues. In addition, the training data lack fine-grained annotations of sentiment intensity based on real physiological experience and often underrepresent implicit intent in Chinese text (Mindner et al., 2023; Rimban, 2023). These limitations may cause a discrepancy between the statistical patterns learned by the model and actual sentiment semantics, resulting in overestimation of sentiment intensity in the output. Additionally, the Transformer architecture adopted by most models relies on the attention mechanism (Vaswani et al., 2017), which tends to predispose models to disproportionately attend to salient sentiment signals. Coupled with their probabilistic prediction-based generation mechanism, models are more inclined to select the most common and prominent sentiment response patterns in training data, resulting in a statistically significant tendency toward stronger sentiments (Wang et al., 2025).

From a theoretical perspective, Embodied Cognition Theory posits that humans' understanding of sensory-perceptual words such as 'itchy' [刺痒] and 'sweet' [甘甜] stems from direct bodily experiences (Barsalou, 2008; Wilson, 2002), whereas models derive the meanings of such words solely from textual statistical correlations. Affective Construction Theory, by contrast, emphasises that sentiments are dynamically constructed within specific cultural contexts (Barrett, 2017). Models exhibit significant divergence and biases in sentiment judgments of culture-loaded words (e.g., 'Dragon Boat Festival' [端午], 'Golden Hoop Spell' [紧箍咒]) and ironic sentences, reflecting the discrepancy between the cultural conceptual system internalised by models and the dynamically constructed sentiment cognitive system of native speakers. While these mechanisms are hypothesised based on current findings, they provide a theoretical framework that warrants further empirical testing.

LLMs' sentiment biases vary across different word and sentence types. For word types, intensity overestimation is most pronounced in sensory-perceptual words, culture-loaded words, and general basic words. This may be because the sentiment meanings of such words rely heavily on direct bodily experience or shared social conventions, which LLMs lack. As a result, models struggle to fully capture culture-specific contexts and dynamic usage, leading to relatively larger sentiment biases. For sentence types, models exhibit significant valence and arousal biases in ironic sentences, in stark contrast to their relative accuracy in literal sentences. This reveals that models still lack sufficient understanding of context and intent.

Further analysis reveals that inter-model consistency varies with textual semantic complexity. For texts with clear sentiment cues (e.g., general basic words, sensory-perceptual words, literal

sentences), the divergence among models' sentiment judgments is relatively small, likely because these texts possess cross-cultural commonality, allowing models to learn a high degree of agreement with human annotations. In contrast, complex expressions involving metaphors, homophones, or intent reversal—such as culture-loaded words, Internet slang, ironic sentences—exhibit significant inter-model divergence (see Appendix for examples). These texts lack a unified, definitive human annotation and require specific cultural knowledge and contextual inference, which may amplify differences in models' sentiment reasoning. This suggests that LLMs still face limitations in handling culturally specific expressions and context-dependent meanings.

Conclusion

This study systematically evaluated the performance of LLMs in Chinese sentiment analysis tasks. The study's primary findings indicate that for sentiment polarity judgment, the outputs of LLMs are highly consistent with the consensus among human participants and significantly outperform the traditional rule-based sentiment lexicon. Regarding continuous dimensions such as sentiment intensity, valence, and arousal, LLMs generally exhibited an overestimation bias, with mean ratings significantly exceeding those of humans and the lexicon. This bias is particularly pronounced when processing sensory-perceptual words that rely on embodied experiences and ironic sentences that require in-depth contextual integration.

These biases are closely associated with models' technical characteristics, such as training data distribution and attention mechanism, suggesting that current models excel at surface-level sentiment pattern recognition and classification while remaining limited in deeper forms of sentiment understanding that require the integration of bodily experience or complex situational reasoning.

Practical implications

Enhancing the cross-cultural adaptability of LLMs is imperative. This requires integrating multidisciplinary perspectives and combining fine-grained sentiment annotations with high-quality training data to reduce potential Western cognitive or sentiment annotation biases. In scenarios like public opinion monitoring and preliminary comment screening, LLMs can be used for rapid sentiment polarity identification. However, in decision-making scenarios relying on sentiment intensity or arousal, models should be calibrated with domain knowledge. Alternatively, they should serve as auxiliary references rather than the sole basis for decision-making, thereby improving application reliability.

Limitations and future work

The participant sample was relatively limited, and the generalisability of the findings to broader populations remains to be validated. The experiments focused on static texts and did not examine models' sentiment understanding in dynamic, multi-turn dialogues. Because sentiment judgment is inherently subjective, model outputs involve uncertainty, and the sentiment lexicons and annotated datasets used are not absolute standards, which may introduce bias. In addition, the study is situated within the Chinese cultural context and does not claim applicability across all languages or cultures. Furthermore, sentiment classification is restricted to three-way polarity, excluding finer-grained differences.

Future research could incorporate multimodal scenarios and integrate prompt engineering to mitigate sentiment biases in text-only models. Furthermore, future studies should aim to enhance robustness via repeated sampling and conduct cross-cultural comparisons to verify the cultural specificity of observed biases. In specific applications, adopting finer-grained sentiment categories offers a key expansion direction.

Acknowledgements

This paper is supported by the Innovative Talent Support Program for Philosophy and Social Sciences in Higher Education Institutions of Henan Province, China (NO. 2025-CXRC-24).

About the authors

Ruili Geng is an Associate Professor in the School of Information Management, Zhengzhou University, Zhengzhou, China. Her research interests include information behaviour and privacy management. She can be contacted at gengruili@zzu.edu.cn

Jiwen Zhang is a master's student in the School of Information Management, Zhengzhou University, Zhengzhou, China. Her research interest is information behaviour. She can be contacted at 2290101220@qq.com

Ruixian Yang is a Professor in the School of Information Management, Zhengzhou University, Zhengzhou, China. Her research interests include data governance, AI literacy, and AI governance. She can be contacted at yrx@zzu.edu.cn

Mingzhe Quan is an Assistant Professor in the School of Information Management, Zhengzhou University, Zhengzhou, China. His research interests include information behaviour and human-centered AI. He can be contacted at u3008560@connect.hku.hk

Xiang Zheng is a Lecturer in the School of Information Management, Zhengzhou University, Zhengzhou, China. Her research interests include cultural heritage intelligent computing, textual semantic mining, digital technologies, and the digital-intelligent transformation of documents. She is the corresponding author and can be contacted at zhengxiang@zzu.edu.cn

Yishuai Xu is a PhD student in the Department of Library and Information Science, Universiti Malaya, Kuala Lumpur, Malaysia. His research interests include machine learning, information retrieval and personalised recommendation. He can be contacted at yishuai@um.edu.my

References

- Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., & Stella, M. (2023). Cognitive network science reveals bias in GPT-3, GPT-3.5 Turbo, and GPT-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3), Article 124. <https://doi.org/10.3390/bdcc7030124>
- Bai, X., Chen, G., He, T., Zhou, C., & Guo, C. (2025). A holistic comparative study of large language models as emotional support dialogue systems. *Cognitive Computation*, 17(2), 71. <https://doi.org/10.1007/s12559-025-10429-x>
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., ... Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. <https://doi.org/10.48550/arXiv.2204.05862>
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617-645. <https://doi.org/10.1146/annurev.psych.59.103006.093639>

- Caspi, A., & Etgar, S. (2023). Exaggeration of emotional responses in online communication. *Computers in Human Behavior*, 146, 107818. <https://doi.org/10.1016/j.chb.2023.107818>
- Dai, X., Zhou, L., Wang, B., & Li, H. (2025). From word to world: Evaluate and mitigate culture bias in LLMs via word association test. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24521–24537. <https://doi.org/10.18653/v1/2025.emnlp-main.1246>
- Ding, Y., You, J., Machulla, T. K., Jacobs, J., Sen, P., & Hollerer, T. (2022). Impact of Annotator Demographics on Sentiment Dataset Labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 519. <https://doi.org/10.1145/3555632>
- Fu, Z., Hsu, Y. C., Chan, C. S., Lau, C. M., Liu, J., & Yip, P. S. F. (2024). Efficacy of ChatGPT in Cantonese sentiment analysis: Comparative study. *Journal of Medical Internet Research*, 26, Article e51069. <https://doi.org/10.2196/51069>
- Gao, C. H., Dang, B. B., Wang, B. J., & Wu, M. S. (2025). The linguistic strength and weakness of artificial intelligence: A comparison between large language model(s) and real students in the Chinese context. *Acta Psychologica Sinica*, 57(6), 27. <https://doi.org/10.3724/sp.J.1041.2025.0947>
- Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62, Article 105227. <https://doi.org/10.1016/j.frl.2024.105227>
- Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of generative AI. *Customer Needs and Solutions*, 11(1), 3. <https://doi.org/10.1007/s40547-024-00143-4>
- Lee, L. H., Li, J. H., & Yu, L. C. (2022). Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4), Article 65. <https://doi.org/10.1145/3489141>
- Li, J., Wang, J., Hu, J., & Jiang, M. (2024). How well do LLMs identify cultural unity in diversity. *arXiv preprint arXiv:2408.05102*. <https://doi.org/10.48550/arXiv.2408.05102>
- Lin, Z. C. (2025). Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3), Article 25152459251357566. <https://doi.org/10.1177/2515245925135756>
- Liu, C., Arulappan, A., Naha, R., Mahanti, A., Kamruzzaman, J., & Ra, I. H. (2024). Large language models and sentiment analysis in financial markets: A review, datasets, and case study. *IEEE Access*, 12, 134041–134061. <https://doi.org/10.1109/ACCESS.2024.3445413>
- Liu, H., Dai, Y., Tan, H., Lei, Y., Zhou, Y., & Wu, Z. (2025). Outraged AI: Large language models prioritise emotion over cost in fairness enforcement. *arXiv preprint arXiv:2510.17880*. <https://doi.org/10.48550/arXiv.2510.17880>
- Mabokela, K. R., Schlippe, T., & Wölfel, M. (2025). Large language models for sentiment analysis to detect social challenges: A use case with South African languages. *arXiv preprint arXiv:2511.17301*. <https://doi.org/10.48550/arXiv.2511.17301>
- Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of human- and AI-generated texts: Investigating features for ChatGPT. *Artificial Intelligence in Education Technologies: New Development and Innovative Practices. AIET 2023. Lecture Notes on Data Engineering and Communications Technologies*, 190, 152–170. https://doi.org/10.1007/978-981-99-7947-9_12

- Mohsin, T. M. (2025). Evaluating large language models (LLMs) in financial NLP: A comparative study on financial report analysis. arXiv preprint arXiv:2507.22936. <https://doi.org/10.48550/arXiv.2507.22936>
- Pontes, D. P. N., & Maricato, J. D. (2025). Using sentiment analysis to differentiate bots and humans in the dissemination of scientific publications on COVID-19 on social media X: A study with ChatGPT 3.5 and Gemini 1.5 Flash. *Biblios-Revista de Bibliotecologia y Ciencias de la Informacion*, Article e001. <https://doi.org/10.5195/biblios.2025.1297>
- Qi, Z., Perron, B. E., Wang, M., Fang, C., Chen, S. T., & Victor, B. G. (2025). AI and cultural context: An empirical investigation of large language models' performance on Chinese social work professional standards. arXiv preprint arXiv:2412.14971. <https://doi.org/10.48550/arXiv.2412.14971>
- Qiu, H., He, H., Zhang, S., Li, A., & Lan, Z. (2024). SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 615–636. <https://doi.org/10.18653/v1/2024.findings-emnlp.34>
- Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjeh, R., Robertson, C. E., & Bavel, J. J. V. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 121(34), Article e2308950121. <https://doi.org/10.1073/pnas.2308950121>
- Rimban, E. L. (2023). Challenges and limitations of ChatGPT and other large language models. *International Journal of Arts and Humanities*, 4(1), 147–152. <https://doi.org/10.25082/ijah.2023.01.003>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 3, 7. <https://doi.org/10.48550/arXiv.2402.01908>
- Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17, Article 18344909231213958. <https://doi.org/10.1177/18344909231213958>
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636. <https://doi.org/10.3758/bf03196322>
- Xie, H., Lin, W., Lin, S., Wang, J., & Yu, L.-C. (2021). A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*, 579, 832–844. <https://doi.org/10.1016/j.ins.2021.08.052>
- Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of the China Society for Scientific and Technical Information*, 27, 180–185.
- Zhou, J., Luo, S., & Chen, H. (2024). A Chinese multi-label affective computing dataset based on social media network users. arXiv preprint arXiv:2411.08347. <https://doi.org/10.48550/arXiv.2411.08347>

Appendix

Sources of experimental materials

The DUTIR sentiment lexicon, developed by the Information Retrieval Laboratory of Dalian University of Technology, contains 27,466 words, all annotated with sentiment category, polarity, and intensity. Sentiment polarity is divided into four categories: neutral (0), positive (1), negative (2), and both positive and negative (3); sentiment intensity is classified into five levels (1, 3, 5, 7, 9), with 9 representing the highest intensity and 1 the lowest. Widely used in sentiment analysis research, this lexicon demonstrates good reliability.

The CPME dataset comprises 80,000 text samples, generated by GPT based on real Weibo user posts to mimic users' expressive styles. Each text is annotated using the team's Chinese sentiment analysis toolkit, covering six basic sentiments: anger, fear, joy, neutrality, sadness, and surprise. With standardised annotations and reliable quality, the dataset is suitable for scientific research and model training. Empirical evaluations indicate that CPME performs excellently in both personality classification and sentiment recognition tasks.

The Chinese EmoBank dataset covers 5,512 words, 2,998 multi-word phrases, 2,582 single sentences, and 2,969 multi-sentence texts. Based on the Self-Assessment Manikin scale, it employs a crowdsourcing method where each text is independently rated on two dimensions: valence and arousal. Each text is evaluated by at least 10 different annotators. Subsequently, outlier scores and inappropriate texts are removed to enhance annotation quality, with the final sentiment annotation results being the average score for each dimension.

The Chinese VAI dataset contains 1,465 sentences with a positive surface expression but implicit negative sentiments. Among these, 1,005 are sourced from Nanyang Technological University's irony corpus, and the remaining 460 from social media. Annotations for this dataset cover three dimensions—valence, arousal, and irony degree—all using a continuous 1-9 point scale: a score of 1 indicates strong negativity, low arousal, or low irony; 5 denotes neutrality, moderate arousal, or moderate irony; and 9 corresponds to strong positivity, high arousal, or high irony. The annotation process is conducted via the Google App Engine crowdsourcing platform, where annotators are required to rate the three dimensions of each text on a 1-9 scale based on the SAM. Each text is independently rated by at least 10 different annotators. After annotation, the corpus is cleaned and outliers are removed to optimise data quality, with the final sentiment annotation results being the average score for each dimension.

Partial sentiment judgment results

To facilitate readers' understanding of the specific manifestations of judgment discrepancies between models and humans, we provide several result examples (see Tables 1 and 2). However, it is important to note that sentiment judgment itself is context-dependent, and discrepancies in single-text judgments are insufficient to constitute direct evidence of sentiment biases in large language models. Instead, we find that models systematically exhibit a sentiment exaggeration tendency compared to humans across multiple texts.

Text	Supplementary Notes	DeepSeek-R1	Doubao 1.6	ERNIE X1 Turbo	Grok 3	Gemini 2.5 Flash	GPT-4o	Humans	DUTIR
the Great Wall (Wanli Changcheng 万里长城)	Cultural symbol	1	1	1	1	0	0	1	0
red envelope (Hongbao 红包)	Cultural-specific monetary	1	1	1	1	1	1	1	0

	gift								
traditional Chinese matchmaker (Hongniang 红娘)	Cultural-specific role	1	1	1	1	1	0	0	0
golden hoop spell (Jinguzhou 紧箍咒)	Literary allusion from Journey to the West	-1	-1	-1	-1	-1	-1	-1	0
Dragon Boat Festival (Duanwu 端午)	Traditional Chinese festival	1	1	0	1	0	0	0	0
act evasively (Shuahuatou 耍滑头)	Colloquial phrase	-1	-1	-1	-1	-1	-1	-1	-1
display arrogance (Baikuanr 摆款儿)	Colloquial phrase	-1	-1	-1	-1	-1	0	-1	-1
curry favor with (Liuxu Paima 溜须拍马)	Idiom	-1	-1	-1	-1	-1	-1	-1	-1
beat a retreat (Datuitanggu 打退堂鼓)	Colloquial phrase	-1	-1	-1	-1	-1	-1	-1	-1
itchy (Ciyang 刺痒)	Sensory adjective	-1	-1	-1	-1	-1	-1	-1	0
refreshing (Shuangkou 爽口)	Sensory adjective	1	1	1	1	1	1	1	1
sweet (Gantian 甘甜)	Sensory adjective	1	1	1	1	1	1	1	1
like (Xifan 稀饭; literal meaning: congee)	Chinese internet slang; homophonic pun	1	1	1	1	1	1	1	1
embarrassed (Qiudaile 糗大了)	Chinese internet slang; colloquial phrase	-1	-1	-1	-1	-1	-1	-1	0
happy (Kaixin 开心)	General basic word	1	1	1	1	1	1	1	1
dislike (Taoyan 讨厌)	General basic word	-1	-1	-1	-1	-1	-1	-1	-1

Table 1. Partial sentiment polarity results.

Text	Supplementary Notes	Deep Seek -R1	Doubao 1.6	ERNIE X1 Turbo	Grok 3	Gemini 2.5 Flash	GPT-4o	Humans	DUTIR
the Great Wall (Wanli Changcheng 万里长城)	Cultural symbol	5	5	4	3	1	4	2	1
red envelope (Hongbao 红包)	Cultural-specific monetary gift	5	4	3	3	4	4	5	1
traditional Chinese matchmaker (Hongniang 红娘)	Cultural-specific role	4	4	3	2	3	3	3	1
golden hoop spell (Jinguzhou 紧箍咒)	Literary allusion from Journey to the West	5	3	4	3	4	4	3	1
Dragon Boat Festival (Duanwu 端午)	Traditional Chinese festival	4	3	1	3	2	4	2	3
act evasively (Shuahuatou 耍滑头)	Colloquial phrase	5	3	3	3	3	4	2	1
display arrogance (Baikuanr 摆款儿)	Colloquial phrase	4	3	3	3	3	5	2	1
curry favor with (Liuxu Paima 溜须拍马)	Idiom	5	4	4	4	4	5	3	2
beat a retreat (Datuitanggu 打退堂鼓)	Colloquial phrase	4	3	3	3	4	4	3	2
itchy (Ciyang 刺痒)	Sensory adjective	3	3	3	3	3	4	2	2
refreshing (Shuangkou 爽口)	Sensory adjective	4	4	3	3	4	4	2	1
sweet (Gantian 甘甜)	Sensory adjective	4	5	4	4	4	4	3	3
like (Xifan 稀饭; literal meaning: congee)	Chinese internet slang; homophonic pun	4	3	2	2	4	3	3	4
embarrassed (Qiudaile 糗大了)	Chinese internet slang; colloquial phrase	4	4	4	2	4	4	3	4
happy (Kaixin 开心)	General basic word	5	3	5	4	4	4	3	3
dislike (Taoyan 讨厌)	General basic word	4	3	4	3	4	4	3	4

Table 2. Partial sentiment intensity results.