

## Task-Adapted AI Training for L2 Pragmatics Assessment

Traditional assessments of L2 pragmatic competence rely heavily on human raters, whose judgments of appropriateness and politeness are inherently variable. Differences in rater severity and self-consistency frequently result in reliability concerns (Alemi & Khanlarzadeh, 2017; Li et al., 2019, 2023). Advances in natural language processing (NLP), particularly the emergence of large language models (LLMs), have motivated increased interest in automated language assessment tasks such as writing evaluation and short-answer scoring (Grévisse et al., 2024; Liu et al., 2025). However, the general-purpose architecture, high computational cost, and limited interpretability of LLMs limit their suitability for fine-grained, domain-specific assessment.

We address this limitation by fine-tuning a small open-weight transformer-based language model (Qwen-3B) for automated L2 speech act assessment. Qwen-3B is a decoder-only architecture with approximately 3 billion parameters, which allows efficient task adaptation while maintaining strong representational capacity. Using supervised instruction-style fine-tuning, the model was trained to predict holistic appropriateness ratings from transcribed learner responses and contextualized discourse prompts. Training data consisted of 489 responses (39.4%) that received unanimous agreement from three human raters, ensuring high-quality supervision. The remaining 751 responses (60.6%) were reserved for held-out evaluation.

L2 pragmatics data were elicited from 62 American learners of Chinese using a computerized oral discourse completion task comprising 20 request-making scenarios, yielding 1,240 transcribed responses. Model performance was evaluated against human ratings using correlation analysis and many-facet Rasch measurement. Results indicate that the trained model achieves strong alignment with human judgments ( $r = .65$ ,  $p < .001$ ) and demonstrates rating self-consistency comparable to that of human raters ( $MnSq = 0.87$ ), within the acceptable range (0.85–1.23).

These findings suggest that task-adapted training of small transformer-based language models with explicit rating objectives enables effective modeling of L2 pragmatic assessment, offering a computationally efficient and more interpretable alternative to large, general-purpose LLMs.

## References

Alemi, M., & Khanlarzadeh, N. (2017). Rater variability and bias in pragmatic assessment: A many-facet Rasch analysis. *Language Testing*, 34(4), 555–575.  
<https://doi.org/10.1177/0265532216682991>

Grévisse, C., Dufour, R., & Lemaire, B. (2024). Large language models for automated short-answer grading: Opportunities and limitations. *Computational Linguistics*, 50(1), 1–28.  
[https://doi.org/10.1162/coli\\_a\\_00487](https://doi.org/10.1162/coli_a_00487)

Li, S., Fan, X., & Ellis, R. (2019). The effects of rater background and rating experience on the assessment of pragmatic performance. *Language Testing*, 36(3), 353–375.  
<https://doi.org/10.1177/0265532218805360>

Li, S., Fan, X., & Ellis, R. (2023). Consistency and severity in rating L2 pragmatic performance: A longitudinal many-facet Rasch study. *Language Assessment Quarterly*, 20(2), 123–145.  
<https://doi.org/10.1080/15434303.2022.2146321>

Liu, Y., Zhang, H., & Wang, Y. (2025). Evaluating second language writing with large language models. *Natural Language Engineering*, 31(1), 45–67.  
<https://doi.org/10.1017/S1351324924000123>