

Fine-tuning DeBERTa-v3 to Automate Spatial Language Classification

Sam Agnoli
Northwestern University

Qingzhou Shi
Northwestern University

David Uttal
Northwestern University

Abstract

To automate spatial language classification, we fine-tuned DeBERTa-v3, a transformer-based language model, using a 33,284-word dataset following a 70:15:15 train-validation-test split. The model performed well in a held-out test comparing performance to human coders (Cohen's $\kappa = .88$).

Keywords: Language classification; spatial language; fine-tuning; educational measurement; large language models

Structured Summary

1 Background

Children's spatial skills, including the ability to mentally visualize and manipulate objects within their environment (Uttal et al., 2013), are predictive of success in STEM fields later in life (Lauer & Lourenco, 2016; Frick, 2019). One way to index the degree of spatial abilities in children is through the analysis of their spatial language in conversations (Fox et al., 2025; Pruden et al., 2011). Spatial language includes words for location and direction (e.g., in, up), dimension and size (e.g., small, tall, shallow), shape (e.g., circle, curved), and orientation or transformation (e.g., turn, flip). Determining whether language is spatial is difficult because many of these words are used spatially only in specific contexts. For example, to can express a spatial relation in "go to the top." but is non-spatial in "I want to help."; similarly, feet refers to a spatial measurement in "You are four feet tall." but to a body part in "My feet hurt." Because of the contextual complexity of language, researchers historically could not automate coding and instead relied on human coders, which remains an expensive and time-consuming bottleneck for research. Transcribing and coding open-ended language can take eight times as long as the recording itself (d'Apice et

al., 2019; Margolin et al., 1998). Partly for this reason, spatial ability has more often been measured with non-language based spatial ability assessments, which are quick to administer and score despite being less ecologically valid (Harris, 2023).

Language models offer a scalable, cost-effective alternative to human coding. For example, recent work demonstrated that a multi-agent foundation model (i.e., GPT-4o with Administrator, Judge, and Critic agentic roles) performed similarly to human coders and identified novel spatial words (Shi & Uttal, 2025), which expanded the dictionary of possible spatial candidate words (Cannon et al., 2007). In the current study, we attempt to further improve the fidelity of language model coding by fine-tuning the DeBERTa-v3 base model on a large dataset of words previously coded by humans.

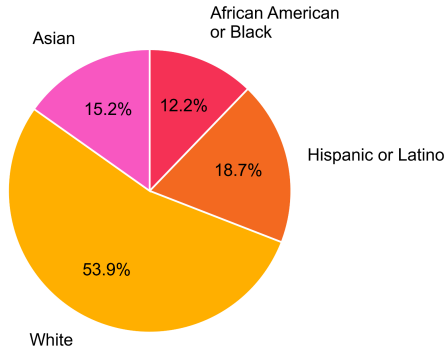
2 Aims

This study examined the validity of a fine-tuned DeBERTa-v3 model for spatial language coding (e.g., Yu et al., 2023). We compared a fine-tuned AI language model as a coder against traditional human coding, evaluated their performance, and explored implications for educational assessment.

3 Sample

Our sample consisted of recordings of 33,284 words across 239 parent-child conversations at a tinkering exhibit in the Chicago Children's Museum (see Polinsky et al., 2023). These counts reflect the final analytic sample after excluding eight sessions for which transcript and human-coding records could not be reliably matched due to inconsistencies between the transcript and coding files. Participating children ranged in age from 6 to 11 years, and families reflected the ethnic and socioeconomic diversity of Chicago (Figure 1). All families spoke fluent English.

Figure 1. Participant Demographics



The spatial dictionary (a predefined list of spatial candidate words that may have spatial meanings) developed by Cannon et al. (2007) was used to identify candidate words requiring contextual classification. The dictionary flagged 3,516 words as spatial candidates, representing 10.6% of all words and spanning 112 unique word types. Human coders subsequently judged whether each candidate was spatial in its utterance context. Approximately 41.4% of candidates were coded as spatial, corresponding to 4.4% of all words. The ten most frequent word types account for 61% of all spatial words.

4 Methods

Human coding followed the procedure detailed in Polinsky et al. (2023), in which a Bag-of-Words dictionary (Zhang et al., 2010) based on Cannon et al. (2007) flags candidate words that trained coders then judge in utterance context. This human-coded dataset serves as the criterion against which model performance is evaluated.

We used the open-source transformer-based language model DeBERTa-v3 base (He et al., 2021, 2023) for spatial language coding because of its efficiency in natural language understanding tasks. DeBERTa-v3 builds on DeBERTa’s disentangled attention, which separates a word’s meaning from its position, making it well-suited for binary word-level classification (spatial = 1, non-spatial = 0) with the full utterance provided as context. Spatial words were identified using Polinsky et al. (2023).

The fine-tuned model was trained using a 70:15:15 train-validation-test split grouped by parent-child conversation to prevent data leakage (see Raschka, 2018). Given that spatial words are rare (4.4% of all words), we did not apply class

weighting or resampling. Instead, the spatial dictionary restricted scoring to candidate words at inference, and checkpoint selection used validation minority-class F1 (computed over all words) rather than accuracy. We trained the model over two epochs to limit overfitting (see Devlin et al., 2019) and because that F1 did not improve on runs with three to five epochs. We then evaluated the held-out test set against human coding on both overall and candidate-only classification. After fine-tuning, we calibrated the model’s decision confidence using temperature scaling (Guo et al., 2017), a standard method to reduce model overconfidence.

Performance of the fine-tuned model was evaluated using the following metrics: Cohen’s kappa, accuracy, sensitivity (recall), specificity, positive predictive value (PPV or precision), negative predictive value (NPV), and F1 score. All metrics range from 0 to 1 except for Cohen’s kappa, which ranges from -1 to 1. Higher values reflect stronger consistency with human coding. We also computed the error metrics expected calibration error (ECE; Guo et al., 2017) and Brier score (Brier, 1950) before and after temperature calibration. ECE summarizes the discrepancy between predicted confidence and observed accuracy across confidence bins, whereas the Brier score measures the squared difference between predicted probabilities and observed binary outcomes. For both ECE and Brier scores, lower values indicate better probabilistic predictions.

5 Results

We tested the fine-tuned DeBERTa-v3 model on a held-out dataset of 526 spatial-candidate words from a total of 4,470 words. The held-out dataset was created using a group-aware split by conversation, ensuring no conversation overlap between training and test data.

The performance metrics are presented in Table 1, with the positive class representing spatial words. The fine-tuned model achieved strong agreement with human coders across all words (i.e., overall evaluation set Cohen’s kappa = .88) and within spatial candidate words (Cohen’s kappa = .81). Because 95.3% of the test words are non-spatial, overall accuracy is inflated by class imbalance. The candidates-only evaluation is the more informative test, in which an accuracy of .91 compares against a .60 majority-class baseline. The fine-tuned model also produced well-calibrated probability estimates

Table 1. Performance Metrics Across Fine-tuned DeBERTa-v3

Accuracy		Sensitivity		Specificity		Precision (PPV)		NPV		F1	
FD	FD_s	FD	FD_s	FD	FD_s	FD	FD_s	FD	FD_s	FD	FD_s
0.99	0.91	0.92	0.92	0.99	0.90	0.84	0.86	1.00	0.95	0.88	0.89

Note. FD = Fine-tuned DeBERTa-v3; FD_s = Fine-tuned DeBERTa-v3 including spatial candidates only. The base rate of spatial words on the test set is 4.7% overall and 40.1% among spatial candidates. Accuracy should be interpreted against these prevalences.

Table 2. Error Metrics of Fine-Tuned DeBERTa-v3

Evaluation Set	Method	ECE	ECE	Brier	<i>n</i> (words)
		Uniform	Quantile		
Overall	Raw	0.008	0.004	0.010	4470
	Temp	0.005	0.003	0.009	
Spatial Candidates	Raw	0.059	0.044	0.074	526
	Temp	0.050	0.041	0.069	

Note. Raw = pre-calibration; Temp = post-hoc temperature calibration; ECE = expected calibration error (lower is better). ECE Uniform uses equal-width confidence bins; ECE Quantile uses equal-frequency bins. Brier = mean squared error of predicted probabilities (lower is better).

which further improved after post-hoc temperature calibration (see Table 2 and Figure 2). Taken together, these fine-tuning results are promising as a scalable alternative to human coding.

To provide calibrated confidence estimates for the model’s decisions, we adjusted how closely the model’s decision probabilities matched observed outcomes. Notably, calibration only changed what the model claimed to know, and left coding decisions untouched. The raw probabilities were overconfident within the spatial-candidate set (ECE = .059, Brier = .074). We fit a single temperature parameter on the candidate words of the validation split to address the model’s overconfidence, yielding $T = 1.46$ (see Figure 2). Applying that temperature to the held-out test set reduced expected calibration error by roughly 15% (.059 to .050 under uniform binning; .044 to .041 under quantile binning) and reduced the Brier score from .074 to .069 (see Table 2). Metrics computed across all words were already near ceiling and changed little (ECE = .008 to .005), which is expected given that 95.3% of words are trivially classified as non-spatial. Calibration allows researchers to more readily accept high-confidence judgments and manually review cases where the model is less confident.

6 Conclusion

This study contributes to the field of AI in educational measurement by supplying an open-access

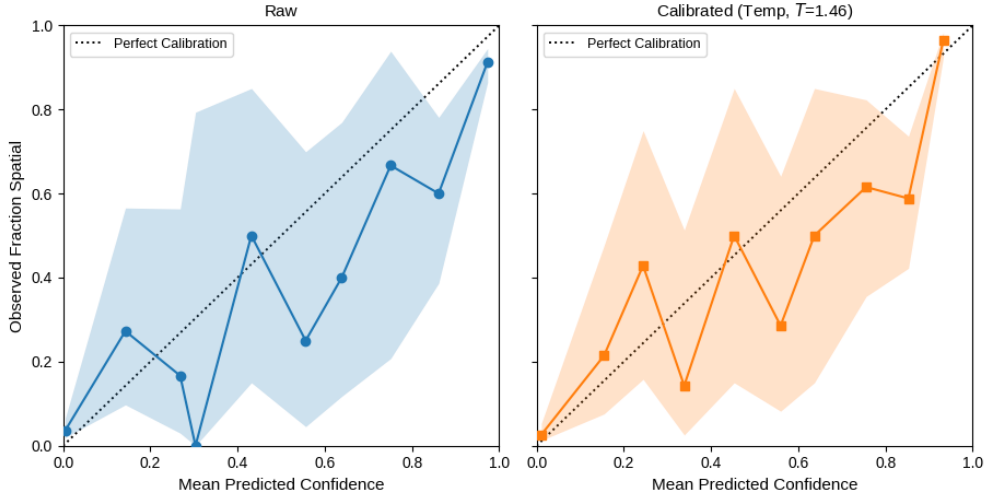
(<https://github.com/SamAgnoli/spatial-language-classifier/tree/main>), scalable model for spatial coding. Our findings provide empirical evidence that a fine-tuned DeBERTa-v3 model can achieve strong agreement with human coders.

We suggest that researchers can deploy this model as a supplement to the costly process of human coding for spatial language. However, we caution against using this model as a replacement for human coding until the generalizability concerns outlined below are addressed. As an example, the current model may be sufficient for identifying relative differences or rank-ordering spatial ability within a homogenous sample. However, the model does not capture absolute values in spatial ability and should not be used as a diagnostic tool. We note that humans also make errors in spatial coding, and that the model may fall within an acceptable range of human variance for researchers’ or educators’ specific goals.

More generally, the value of this research may lie in what automated coding supports rather than the specific performance of a given model. Spatial ability has historically been measured with tractable, easy to score laboratory tasks (Harris, 2023). These tasks have been criticized for failing to capture real-world behaviors and abilities. In contrast, natural language offers a more ecologically valid and perhaps powerful tool to measure spatial ability. Whereas previously natural language was too cost- and time-intensive to analyze without ample resources, researchers can now record students in any situation of interest and be rewarded with a rich dataset. Automated coding may eventually overtake lab tasks in both availability and ecological validity, albeit with potential confounds (e.g., socioeconomic status, Noble et al., 2007).

There are several limitations to consider. We note that our data are limited (i.e., a few tasks in a museum tinkering exhibit) among English-speaking parent-child dyads. Also, the fine-tuned model was limited by its spatial dictionary. Future

Figure 2. Calibrated Reliability Diagram for Spatial Candidate Subset



Note. Figure 2 presents the reliability diagram for the spatial-candidate subset. Both line graphs track the diagonal (i.e., perfect calibration) at the extremes of the confidence range. The temperature calibrated line-graph is closer to the diagonal than the raw line-graph across most of that range, demonstrating better calibration. Shaded bands are 95% Wilson score confidence intervals for each bin’s observed proportion, and are wide wherever a bin contains few words. Notably, 89.7% of candidate words fall in the lowest and highest confidence bins ($n = 267$ and $n = 169$, respectively, out of 526 candidates), so the volatility in the middle bins reflects greater noise due to the small sample.

research might develop a pipeline for expanding the dictionary for a given task following the multi-agentic approach outlined in Shi and Uttal (2025) and then fine-tuning DeBERTa-v3 (or a state-of-the-art model) provided researchers have sufficient data (i.e., spatial words within the utterance context). Additionally, human coding was based on the Cannon et al. (2007) and Polinsky et al. (2023) dictionary, whereas the model’s candidate gate might be expanded following testing with the expanded dictionary from Shi and Uttal (2025). Re-coding the expanded candidates would allow the model’s performance on novel spatial terms to be evaluated directly. Finally, to ensure generalizability, it is crucial to externally validate the model on independent datasets of varying demographics. Because the ten most frequent spatial words account for 61% of positive spatial language in the dataset and are domain-general (e.g., in, on, up, there), we suspect that the model’s classification abilities would transfer to other (i.e., non-museum) English-speaking contexts where middle-childhood participants are tasked with describing spatial objects and relations. Regarding spatial classification in other languages with unique spatial representations (see Bowerman & Choi, 2001), researchers could apply our methodology to develop their own trained classifier, presuming they have sufficient human-coded spatial data.

Broadly, we underscore the importance of transparency in AI-related research as well as ensuring human alignment in AI-generated coding. This includes ensuring interpretable AI outputs, which we have attempted by using a spatial language dictionary and confidence calibration so that researchers can track which words or situations lead the model to be incorrect and/or overconfident. Such interpretability is essential for keeping humans aware and in control of model outputs. We also encourage other researchers to share their datasets either privately or in an open-source repository so that future researchers can use these datasets to train more advanced models as they become available.

Ultimately, this research advances the use of AI to support high-quality, context-aware educational measurement at scale.

7 Open Science and Code Availability

The fine-tuned model is publicly released under an MIT license at <https://huggingface.co/SamAgnoli/deberta-v3-base-spatial-language-detection>, and the full pipeline is available at <https://github.com/SamAgnoli/spatial-language-classifier>. The Hugging Face model card documents the input format, training hyperparameters, per-class evaluation metrics, and the temperature-scaling procedure, and the model can be loaded to classify

spatial language through a quick-start notebook accessible via the GitHub repository. In this repository, spatial words are identified using the dictionary developed by Polinsky et al. (2023), though we encourage future human coders to use the expanded dictionary by Shi and Uttal (2025). The repository also contains the annotated notebook covering dictionary construction and candidate matching, the conversation-grouped 70:15:15 split, fine-tuning, held-out evaluation on both the overall and candidates-only views, and post-hoc calibration.

Researchers applying the classifier should note that it is a sentence-pair classifier: it judges a target word together with its utterance. Performance outside English-language, parent-child tinkering speech has not been established, and we recommend the candidates-only metrics as the appropriate basis for judging model quality. Consistent with our recommendation that researchers share data where possible, we have provided a deidentified variant of our human-coded dataset within the repository under the data/example subfolder.

References

- Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. *Language acquisition and conceptual development*, 475-511.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Cannon, J., Levine, S., & Huttenlocher, J. (2007). A system for analyzing children and caregivers' language about space in structured and unstructured contexts. *Spatial intelligence and learning center (SILC) technical report*.
- d'Apice, K., Latham, R. M., & von Stumm, S. (2019). A naturalistic home observational approach to children's language, cognition, and behavior. *Developmental Psychology*, 55(7), 1414.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
- Fox, D. S., Ribner, A., Camarote, J., Vortuba-Drzal, E., Bachman, H. J., & Libertus, M. E. (2025). Parents' spatial talk predicts toddlers' spatial language gains. *Learning and Instruction*, 97, 102107. <https://doi.org/10.1016/j.learninstruc.2025.102107>
- Frick, A. (2019). Spatial transformation abilities and their relation to later mathematics performance. *Psychological research*, 83(7), 1465-1484.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321-1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- Harris, D. (2023). Spatial reasoning in context: bridging cognitive and educational perspectives of spatial-mathematics relations. *Frontiers in Education* (Vol. 8, p. 1302099).
- He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR 2021)*. <https://openreview.net/forum?id=XPZTaotutsD>
- He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. <https://openreview.net/forum?id=sE7-XhLxHA>
- Lauer, J. E., & Lourenco, S. F. (2016). Spatial processing in infancy predicts both spatial and mathematical aptitude in childhood. *Psychological science*, 27(10), 1291-1298.
- Margolin, G., Oliver, P. H., Gordis, E. B., O'hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical child and family psychology review*, 1(4), 195-213.
- Noble, K. G., McCandliss, B. D., & Farah, M. J. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental science*, 10(4), 464-480.
- Polinsky, N., Pagano, L. C., Acosta, D. I., Haden, C. A., & Uttal, D. H. (2023). Spatial language in families' conversational reflections about museum experiences. *Journal of Applied Developmental Psychology*, 86, 101539.
- Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? *Developmental Science*, 14(6), 1417-1430. <https://doi.org/10.1111/j.1467-7687.2011.01088.x>
- Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.
- Shi, Q., & Uttal, D. (2025, October 27-29). Evaluating AI methods for coding spatial language from children's naturalistic conversations [Paper presentation]. *Conference on Artificial Intelligence in Measurement and Education (AIME-Con)*, Pittsburgh, PA, United States.
- Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: a meta-analysis of training studies. *Psychological bulletin*, 139(2), 352.
- Yu, H., Yang, Z., Pelrine, K., Godbout, J. F., & Rabbany, R. (2023). Open, closed, or small language models for text classification?. *arXiv preprint arXiv:2308.10092*.
- Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1), 43-52.