Correct Then Detect: Zero-Shot FVMC Annotation for Child Language Sample Analysis

Shuwei Hou [1,**], *Wei Bo* [1], *Varun Shijo* [1], *Chuhui Liu* [1], *Manav Kanaganapalli* [1], *Ling-Yu Guo* [2], *Wenyao Xu* [1]

[1] Department of Computer Science and Engineering, University at Buffalo, USA
[2] Department of Communicative Disorders and Sciences, University at Buffalo, USA

shuweiho@buffalo.edu, weibo@buffalo.edu, varunshi@buffalo.edu, chuhuili@buffalo.edu, manavkan@buffalo.edu, lingyugu@buffalo.edu, wenyaoxu@buffalo.edu

Abstract

The Finite Verb Morphology Composite (FVMC) is a clinically validated metric for identifying Developmental Language Disorder (DLD), yet its computation requires labor-intensive manual annotation of obligatory tense contexts. Since no annotated corpora exist for training, large language models (LLMs) have been the only approach to automated FVMC annotation. Rather than prompting LLMs end-to-end, we present the first zero-shot framework that decomposes the task into grammar correction and obligatory context detection, followed by alignment to label each context as correct, incorrect, or omitted. Evaluated on the ENNI child narrative dataset, our best configurations achieve F1 scores of 97.05%, 60.15%, and 70.23% for correct, incorrect, and omitted contexts, improving by 1.99, 4.72, and 6.21 points over the strongest LLMs as of February 2026 (GPT 5.2 and Sonnet 4.6 Reasoning Models). The proposed paradigm enables scalable, cost-efficient automated language assessment.

Index Terms: Developmental Language Disorder, Automated Language Assessment, Grammatical Error Correction, Large Language Models

1. Introduction

Language sample analysis (LSA) is widely regarded as the gold standard for evaluating children's language development in clinical and research settings [1, 2]. Automating such clinical language measures could enable scalable assessment from speech data [3] and support AI-driven screening tools [4]. Among various metrics derived from LSA, the Finite Verb Morphology Composite (FVMC) has emerged as a particularly sensitive metric for identifying Developmental Language Disorder (DLD). FVMC calculates the accuracy of four tense morphemes, including third-person singular *–s*, past tense *–ed*, copula *be* and auxiliary *be* [5, 6]. Research has demonstrated that children with DLD are more likely to have difficulty with tense and agreement morphemes compared to their typically developing peers [7, 8], and FVMC has been validated as a clinically effective measure for differentiating children with DLD across a range of ages and narrative elicitation tasks [9, 5]. However, computing FVMC requires trained professionals to manually identify whether the target morpheme is correctly used, incorrectly used, or omitted, and aggregate these judgments across an entire language sample. This labor-intensive process has substantially limited both the clinical adoption of FVMC and its large-scale validation in research. To our knowledge, no prior work has provided a practical automatic pipeline for FVMC labeling and scoring.

Currently, automated methods for detecting grammatical errors have been explored in the Natural Language Processing (NLP) field, including rule-based approaches [10, 11, 12] and sequence labeling models [13, 14, 15]. However, directly applying these techniques to FVMC annotation presents several fundamental challenges. First, beyond identifying morphological errors on existing verbs, FVMC requires detecting omitted elements that are entirely absent from the utterance. Such omissions cannot be straightforwardly captured by sequence labeling models that operate over observed tokens. Second, the limited availability of FVMC-annotated corpora precludes the supervised training of task-specific models.

Recent advances in large language models (LLMs) offer another potential avenue, whereby FVMC annotation rules could be provided as prompt to guide the model in labeling FVMC directly. However, direct FVMC labeling requires multi-step reasoning over child utterances that often contain grammatical errors and disfluencies, which can destabilize inference and increase hallucination even in advanced reasoning models. The extent to which even state-of-the-art LLMs can reliably perform this task remains an open question. This makes reliable direct annotation difficult without additional task structuring.

In this work, we propose a zero-shot framework for FVMC labeling that requires no task-specific training. Our approach introduces a novel task decomposition paradigm: rather than attempting to label FVMC contexts directly from raw text that contain possible errors or omissions, we first apply a grammar error correction (GEC) model constrained to verb-related edits, and then detect obligatory contexts from the corrected output using a morphosyntactic parser. An alignment between the original and corrected utterances then determines whether each FVMC context was correctly produced, incorrectly produced, or omitted. This two-stage design reduces the complexity required of any single model while avoiding the confounding effects of grammatical errors. Experimental results on the ENNI dataset [16] (a narrative language corpus for children) demonstrate that our method substantially outperforms direct LLM-based annotation, achieving higher performance than the current reasoning models.

2. FVMC Rules & Calculation

FVMC measures the accuracy of four tense morphemes in obligatory contexts: third-person singular *-s*, past tense *-ed*, and copula and auxiliary *be* forms (e.g., *am, are, is, was, were*). It specifically indexes grammatical tense marking ability rather

** indicates the corresponding author.

than general language proficiency. And an obligatory context is defined as an instance in which a particular tense morpheme is required for the communication unit (C-unit) to be grammatical. For example, in the C-unit "The rabbit walking on the street", the auxiliary *be* (e.g., "is") is required for grammaticality but has been omitted, creating an obligatory context for auxiliary be. Only contexts in which the target morpheme is required for grammaticality are included in FVMC scoring.

Several exclusion criteria constrain the set of obligatory contexts. First, C-units containing verb forms but lacking an explicit subject are excluded because they do not establish obligatory contexts for tense marking. Second, overgeneralizations of third-person singular *-s* (e.g., "He haves an airplane") or regular past tense *-ed* (e.g., "She just standed there") are excluded, as these verbs do not, by definition, require the past tense or third-person singular morphemes in those contexts. Third, the infinitive of *be* (e.g., "The rabbit will be sick"), present participle *being* (e.g., "She is being funny"), past participle *been* (e.g., "He has been trying to get the ball"), and gerund (e.g., "Being happy is easy") are all excluded because these forms do not themselves mark tense; instead, tense is expressed by another element in the clause.

For each identified obligatory context, the target morpheme is coded as one of three outcomes: correctly used, omitted, or incorrectly used. FVMC is computed by dividing the total number of correctly used target morphemes by the total number of obligatory contexts across all four morpheme categories in the sample:

$$\text{FVMC} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{incorrect}} + N_{\text{omitted}}}, \quad (1)$$

where N_{correct}, $N_{\text{incorrect}}$, and N_{omitted} denote the number of correctly produced, incorrectly produced, and omitted obligatory contexts, respectively, across all four categories.

3. Methodology

When human annotators identify FVMC labels, the typical workflow involves first locating obligatory contexts based on surrounding linguistic cues and contextual information and then judging whether the target morpheme is correctly used, incorrectly used, or omitted. However, the grammatical errors inherent in the language samples of children with language disorders frequently mislead annotators and even state-of-the-art large language reasoning models. We propose a novel paradigm that circumvents this difficulty by decomposing the FVMC annotation task into two simpler subtasks: grammar error correction followed by obligatory context detection. As illustrated in Figure 1, the pipeline consists of three stages described below.

3.1. Stage 1: Grammar Error Correction

The first stage transforms each input utterance into a grammatically correct version using a Grammar Error Correction (GEC) model. Crucially, the scope of correction must be restricted to verb-related morphological errors, specifically correcting verb tense and form errors and inserting omitted copula, auxiliary, or main verbs. Other modifications such as noun number changes, article insertion, or word reordering are prohibited. This restriction is essential because the subsequent alignment stage relies on minimal and interpretable edits; additional modifications would introduce unnecessary differences between the original and corrected sequences and propagate errors into FVMC label recovery.

Conventional sequence-to-sequence GEC systems [17, 18, 19] based on Neural Machine Translation (NMT) approaches are designed to freely rewrite sentences and cannot reliably constrain the categories of corrections they apply. Therefore, we adopt two alternative GEC approaches that allow fine-grained control over the correction scope.

3.1.1. GECToR.

GECToR [20] is a Transformer encoder based model that formulates GEC as a sequence tagging problem rather than a sequence-to-sequence generation task. Instead of generating corrected text autoregressively, it predicts token-level transformations that are then applied to the source tokens. The model defines 5,000 transformations, including basic operations such as *$KEEP* (retain the token), *$DELETE* (remove the token), *$APPEND_t* (insert token *t* after the current position), and *$REPLACE_t* (substitute the current token with *t*), as well as 29 grammatical-transformations organized into five categories: *CASE*, *MERGE*, *SPLIT*, *NOUN_NUMBER*, and *VERB_FORM*. To restrict corrections to FVMC-relevant edits, we retain only a subset of these transformations: *REPLACE* and *APPEND* operations involving the copula and auxiliary forms *am, is, are, was,* and *were*, together with 20 *VERB_FORM* grammatical-transformations that permit conversions among five verb inflections: base form (VB), third-person singular present (VBZ), past participle (VBN), regular past tense (VBD), and present participle (VBG).

3.1.2. LLM-based GEC.

An alternative approach leverages large language models by combining task-specific prompts with the input utterance to guide the model to produce a grammatically corrected version of the utterance. The prompt explicitly instructs the model to correct only verb morphology errors and to insert any omitted verbs, auxiliary or copula *be* forms, while strictly prohibiting all other modifications. The model is further instructed to return only the corrected utterance without explanations or additional text.

3.2. Stage 2: Obligatory Context Labeling

The second stage identifies FVMC obligatory contexts from the grammatically corrected utterance. We employ Universal Dependencies (UD) models from the Stanza NLP toolkit [21] to extract the part-of-speech (POS) tags and morphological features of each word. Using these linguistic annotations, we apply rule-based filters to identify four categories of obligatory contexts: third-person singular present tense verbs, regular past tense verbs, and copula and auxiliary be forms. Consistent with the definitions in Section 2, clauses without explicit subjects, infinitive *be*, present participle *being*, past participle *been*, and gerund are excluded. Words involving irregular verb morphology and overgeneralization errors are likewise excluded to ensure strict adherence to the FVMC criteria.

Because the input to this stage is a grammatically correct utterance, the POS tagger and morphological analyzer operate under ideal conditions, free from the confounding effects of grammatical errors that would otherwise degrade tagging accuracy. This stage assigns each word in the corrected utterance a binary label: obligatory context or non-obligatory context.

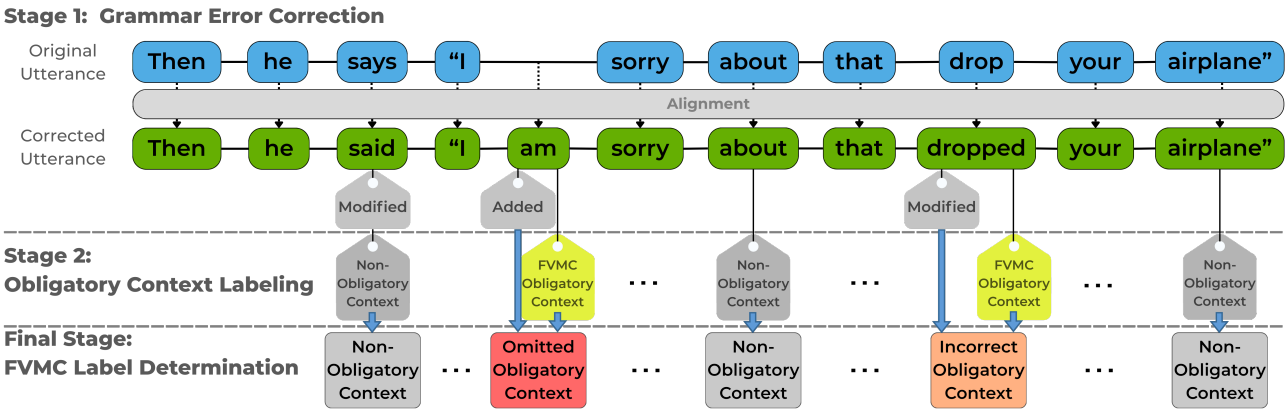

Figure 1: *Overview of the proposed FVMC annotation framework: The corrected utterance is aligned with the original utterance to obtain two types of modification labels (Modified or Added). Obligatory context labels are then detected from the corrected utterance, yielding binary labels (Obligatory Context or Non-obligatory Context). Based on the labels obtained from these two stages, the final four FVMC labels are determined: Non-obligatory Context, Correct Obligatory Context, Incorrect Obligatory Context, and Omitted Obligatory Context.*

3.3. Final Stage: Alignment, Determination, and Calculation

The final stage maps the obligatory context labels from the corrected utterance back onto the original utterance and determines the final FVMC labels based on the edit operations. We first align the original utterance $X = [x_1, x_2, \ldots, x_n]$ with the corrected utterance $Y = [y_1, y_2, \ldots, y_m]$ using the minimum edit distance algorithm [22]. By tracing the edit operations in the alignment path, we assign each original word one of four FVMC labels according to the following rules:

1. **Non-obligatory context.** If a corrected word y_j is labeled as non-obligatory context and y_j corresponds to an original word x_i (whether modified or unchanged), then x_i is labeled as non-obligatory context.
2. **Correct obligatory context.** If y_j is labeled as obligatory context and y_j is identical to x_i (i.e., no correction was applied), then x_i is labeled as correct obligatory context.
3. **Incorrect obligatory context.** If y_j is labeled as obligatory context and y_j was produced by modifying x_i (i.e., a substitution edit), then x_i is labeled as incorrect obligatory context.
4. **Omitted obligatory context.** If y_j is labeled as obligatory context and y_j was inserted after x_i (i.e., an insertion edit), then an omitted obligatory context label is appended following the label of x_i.

After applying these rules across all words in correct utterances, we obtain the complete FVMC label sequence for the original utterances. The FVMC score is then computed by dividing the total number of correct obligatory contexts by the total number of obligatory contexts (correct, incorrect, and omitted combined).

4. Experiments & Results

4.1. Dataset

We evaluate our method on the Edmonton Narrative Norms Instrument (ENNI) dataset [16], which contains narrative language samples from 377 children, including both typically developing (TD) children and children with Developmental Language Disorder (DLD). Each child narrated three independent stories based on pictures, yielding a story-telling language sample. The transcripts are publicly available through TalkBank [23], and all FVMC annotations were performed by certified speech-language pathologists (SLPs), ensuring professional-grade ground truth labels. The dataset contains 17,163 correct obligatory context, 507 incorrect obligatory context, and 272 omitted obligatory context labels. We segment the input at the story level rather than the sentence level, so that models have access to sufficient discourse context when inferring the intended tense of each utterance.

4.2. Evaluation Metrics

Since the FVMC score is entirely determined by the accuracy of its labels, we evaluate performance using standard sequence labeling metrics. Specifically, we report the precision, recall, and F1 score for each of the four label types: non-obligatory context, correct obligatory context, incorrect obligatory context, and omitted obligatory context. Because the non-obligatory context class dominates the label distribution and is trivially easy to identify, we focus our analysis on the remaining three obligatory context categories, which are critical to FVMC computation.

4.3. Baselines

We compare our correction-then-detection approach against a direct LLM baseline. In this baseline, the LLM is tasked with performing FVMC annotation end-to-end: the prompt explicitly specifies the complete FVMC annotation rules described in Section 2, and instructs the model to output the positions and categories of correct, incorrect, and omitted obligatory contexts in a structured output format. This use of LLMs is fundamentally different from their role in our proposed pipeline. In our method, the LLM performs only grammar error correction, a well-defined and constrained generation task. In contrast, the baseline requires the LLM to simultaneously infer obligatory contexts, evaluate morpheme correctness, and detect omissions within a single inference step, substantially increasing reasoning complexity and susceptibility to errors when processing grammatically noisy child language.

Table 1: *Precision (%), Recall (%), and F1 (%) for three obligatory context categories across all methods. GPT 5.2-R denotes GPT 5.2 Reasoning models, where (L), (M), and (H) indicate Low, Medium, and High reasoning levels, respectively. Sonnet 4.6-AT denotes Sonnet 4.6 with Adaptive Thinking.*

		Correct Obligatory Context			**Incorrect Obligatory Context**			**Omitted Obligatory Context**		
Type	**Method**	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LLM-only	Sonnet 4.5	43.97	58.13	50.07	20.88	43.25	28.17	38.20	54.60	44.95
LLM-only	Sonnet 4.6-AT	91.43	94.30	92.84	46.85	67.86	55.43	56.28	74.23	64.02
LLM-only	GPT 5.2	15.08	23.60	18.40	3.76	5.56	4.49	2.53	16.56	4.39
LLM-only	GPT 5.2-R (L)	91.72	93.63	92.67	52.97	38.89	44.85	33.62	73.01	46.03
LLM-only	GPT 5.2-R (M)	93.28	95.73	94.49	**55.38**	40.87	47.03	33.76	80.98	47.65
LLM-only	GPT 5.2-R (H)	93.99	96.16	95.06	43.80	47.62	45.63	37.60	**84.66**	52.08
Ours	GECToR + Stanza	96.18	**97.36**	96.77	43.79	50.40	46.86	**90.32**	17.18	28.87
Ours	GPT 5.2 + Stanza	96.87	97.23	**97.05**	44.00	78.57	56.41	77.21	64.42	**70.23**
Ours	Sonnet 4.5 + Stanza	**96.92**	96.79	96.85	48.43	**79.37**	**60.15**	89.11	55.21	68.18

4.4. Experimental Setup

For the baseline approach, we evaluate six LLMs via their APIs: ChatGPT 5.2 and ChatGPT 5.2 Reasoning (Low, Medium, and High) from OpenAI, and Claude Sonnet 4.5 and Claude Sonnet 4.6 Adaptive Thinking from Anthropic. All LLM API calls used default decoding parameters *(temperature = 1.0, top_p = 1.0)* without explicit configuration. For our proposed pipeline, we test three GEC configurations in Stage 1: the GECToR model with a RoBERTa [24] encoder, and Claude Sonnet 4.5 prompted for grammar correction, as well as ChatGPT 5.2 prompted for grammar correction. Stage 2 uses the Stanza UD model for obligatory context labeling in all configurations. All experiments use identical input data segmented at the story level.

4.5. Results

Table 1 presents the performance of all evaluated methods across the three obligatory context categories. Several observations emerge from these results. First, non-reasoning LLMs are largely incapable of performing FVMC annotation when used directly. Both Sonnet 4.5 and GPT 5.2, despite being frontier-class models, achieve poor F1 scores across all three categories in the direct annotation setting. GPT 5.2 performs particularly poorly, with F1 scores below 5% for both incorrect and omitted obligatory contexts, suggesting that the model struggles to follow the complex, multi-step annotation rules specified in the prompt.

Second, enabling reasoning capabilities yields substantial improvements across all models. GPT 5.2 Reasoning models dramatically outperform their non-reasoning counterpart, and Claude Sonnet 4.6 with Adaptive Thinking similarly surpasses Sonnet 4.5, achieving the best performance among LLM-only baselines with F1 scores of 92.84%, 55.43%, and 64.02% on the three categories. This confirms that FVMC annotation requires multi-step reasoning over linguistic context, which standard autoregressive generation alone cannot reliably perform.

Third, our proposed correction-then-detection pipeline outperforms all LLM-only baselines. GPT 5.2 + Stanza achieves the highest F1 for correct (97.05%) and omitted (70.23%) obligatory contexts, while Sonnet 4.5 + Stanza achieves the highest F1 for incorrect contexts (60.15%) Compared to the best reasoning LLM baselines, our approach improves F1 by 1.99, 4.72, and 6.21 percentage points for the three categories, respectively. Notably, this superior performance is achieved using a non-reasoning model for grammar correction, which is substantially less computationally expensive than reasoning-mode inference.

Finally, the results validate the core insight of our approach: by decomposing the complex FVMC annotation task into two simpler, well-defined subtasks, each handled by a model operating within its strengths, we achieve performance that surpasses even expensive reasoning-augmented LLMs applied to the full task directly.

5. Discussion

This work presents the first automated pipeline for FVMC labeling and scoring, addressing a longstanding bottleneck in clinical language assessment.

Clinical and Research Impact. Manual FVMC annotation by certified SLPs is time-consuming and costly, limiting its clinical adoption and large-scale research validation. By automating this process, our pipeline could enable scalable DLD screening for telehealth [25] and school-based assessment, while allowing researchers to validate FVMC norms across diverse populations and elicitation contexts [26]. Combined with Automatic Speech Recognition (ASR), it could further serve as a component of end-to-end, AI-driven language assessment tools for broader real-world clinical and educational deployment.

Insights on LLM Utilization. Recent research shows that SLPs are increasingly exploring the use of LLMs to support LSA and decision making [27, 28]. However, our findings show that prompting LLMs to perform complex linguistic annotation directly remains unreliable. In contrast, decomposing the problem into well-scoped subtasks is more effective. In our pipeline, the LLM and GEC model handles only grammar correction, while structured linguistic analysis is delegated to specialized models that offer greater precision. This principle of combining flexible generation from LLMs with task-specific analytical tools may generalize to other complex annotation tasks in clinical NLP and beyond.

Limitations and Future Work. First, our evaluation is limited to a single narrative dataset (ENNI); generalization to conversational or other contexts remains to be validated. Second, although performance on correct obligatory contexts is near ceiling, the F1 scores for incorrect and omitted contexts still leave room for improvement. Future work will integrate ASR to enable a speech assessment pipeline and extend the framework to grammar-related metrics beyond FVMC.

6. Acknowledgments

This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, through Award # 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

7. Generative AI Use Disclosure

Generative AI tools were used only to polish the language of this manuscript to improve clarity and readability. Every AI-assisted editing was rigorously reviewed and verified by the authors, who take full responsibility for the content and integrity of this paper.

8. References

[1] J. J. Heilmann, "Myths and realities of language sample analysis," *Perspectives on Language Learning and Education*, vol. 17, no. 1, pp. 4–8, 2010.

[2] A. Costanza-Smith, "The clinical utility of language samples," *Perspectives on Language Learning and Education*, vol. 17, no. 1, pp. 9–15, 2010.

[3] S. Bhattacharjee and W. Xu, "Voicelens: A multi-view multi-class disease classification model through daily-life speech data," *Smart Health*, vol. 23, p. 100233, 2022.

[4] W. Bo, M. Rubino, and W. Xu, "A mispronunciation-based voiceomics representation framework for screening specific language impairments in children," in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2024, pp. 294–304.

[5] L.-Y. Guo, S. Eisenberg, P. Schneider, and L. Spencer, "Finite verb morphology composite between age 4 and age 9 for the edmonton narrative norms instrument: Reference data and psychometric properties," *Language, Speech, and Hearing Services in Schools*, vol. 51, no. 1, pp. 128–143, 2020.

[6] J. M. Rudolph, C. A. Dollaghan, and S. Crotteau, "The finite verb morphology composite: Values from a community sample," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 6, pp. 1813–1822, 2019.

[7] M. L. Rice and K. Wexler, "Toward tense as a clinical marker of specific language impairment in english-speaking children," *Journal of speech, language, and hearing Research*, vol. 39, no. 6, pp. 1239–1257, 1996.

[8] L. B. Leonard, *Children with specific language impairment*. MIT press, 2017.

[9] P. A. Hadley and H. Short, "The onset of tense marking in children at risk for specific language impairment," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 6, pp. 1344–1362, 2005.

[10] D. Naber *et al.*, "A rule-based style and grammar checker," 2003.

[11] K. Gabani, M. Sherman, T. Solorio, Y. Liu, L. Bedore, and E. Pena, "A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children," in *Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the Association for Computational Linguistics*, 2009, pp. 46–55.

[12] M. Mozgovoy, "Dependency-based rules for grammar checking with languagetool," in *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2011, pp. 209–212.

[13] E. Volodina, C. Bryant, A. Caines, O. De Clercq, J.-C. Frey, E. Ershova, A. Rosen, and O. Vinogradova, "Multiged-2023 shared task at nlp4call: Multilingual grammatical error detection," in *Proceedings of the 12th workshop on NLP for computer assisted language learning*, 2023, pp. 1–16.

[14] S. Bell, H. Yannakoudakis, and M. Rei, "Context is key: Grammatical error detection with contextual word representations," in *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, 2019, pp. 103–115.

[15] S. Flachs, O. Lacroix, M. Rei, H. Yannakoudakis, and A. Søgaard, "A simple and robust approach to detecting subject-verb agreement errors," 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)*, 2019, pp. 2418–2427.

[16] L.-Y. Guo and P. Schneider, "Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 2, pp. 317–329, 2016.

[17] T. Ge, F. Wei, and M. Zhou, "Reaching human-level performance in automatic grammatical error correction: An empirical study," *arXiv preprint arXiv:1807.01270*, 2018.

[18] W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu, "Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data," 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)*, 2019, pp. 156–165.

[19] H. Zhou, Y. Liu, Z. Li, M. Zhang, B. Zhang, C. Li, J. Zhang, and F. Huang, "Improving seq2seq grammatical error correction via decoding interventions," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 7393–7405.

[20] K. Omelianchuk, V. Atrasevych, A. Chernodub, and O. Skurzhanskyi, "Gector–grammatical error correction: Tag, not rewrite," in *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, 2020, pp. 163–170.

[21] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," in *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, 2020, pp. 101–108.

[22] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM (JACM)*, vol. 21, no. 1, pp. 168–173, 1974.

[23] B. MacWhinney, "The talkbank project," in *Creating and digitizing language corpora: Volume 1: Synchronic databases*. Springer, 2007, pp. 163–180.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] W. Bo, S. S. Sullivan, X. Zhang, M. Gao, and W. Xu, "A telemedicine analytic framework for fully and semi-automatic alzheimer's disease screening using clock drawing test," *IEEE journal of biomedical and health informatics*, vol. 28, no. 12, pp. 7503–7516, 2024.

[26] B. Weiler, P. Schneider, and L.-Y. Guo, "The contribution of socioeconomic status to children's performance on three grammatical measures in the edmonton narrative norms instrument," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2776–2785, 2021.

[27] J. Austin, K. Benas, S. Caicedo, E. Imiolek, A. Piekutowski, and I. Ghanim, "Perceptions of artificial intelligence and chatgpt by speech-language pathologists and students," *American Journal of Speech-Language Pathology*, vol. 34, no. 1, pp. 174–200, 2025.

[28] N. Y. Birol, H. B. Çiftci, A. Yılmaz, A. Çağlayan, and F. Alkan, "Is there any room for chatgpt ai bot in speech-language pathology?" *European Archives of Oto-Rhino-Laryngology*, vol. 282, no. 6, pp. 3267–3280, 2025.