electronics

MDPI

Article

Enhancing Interprofessional Communication in Healthcare Using Large Language Models: Study on Similarity Measurement Methods with Weighted Noun Embeddings

Ji-Young Yeo [1], Sungkwan Youm [2,*] and Kwang-Seong Shin [3,*]

[1] College of Nursing, Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea; shine73@hanyang.ac.kr
[2] Department of Information and Communication Engineering, Wonkwang University, Iksan 54538, Republic of Korea
[3] Department of Computer Engineering, Sunchon National University, 255, Jungang-ro, Suncheon-si 57922, Republic of Korea
* Correspondence: skyoum@gmail.com (S.Y.); waver@scnu.ac.kr (K.-S.S.)

Abstract: Large language models (LLMs) are increasingly applied to specialized domains like medical education, necessitating tailored approaches to evaluate structured responses such as SBAR (Situation, Background, Assessment, Recommendation). This study developed an evaluation tool for nursing student responses using LLMs, focusing on word-based learning and assessment methods to align automated scoring with expert evaluations. We propose a three-stage biasing approach: (1) integrating reference answers into the training corpus; (2) incorporating high-scoring student responses; (3) applying domain-critical token weighting through Weighted Noun Embeddings to enhance similarity measurements. By assigning higher weights to critical medical nouns and lower weights to less relevant terms, the embeddings prioritize domain-specific terminology. Employing Word2Vec and FastText models trained on general conversation, medical, and reference answer corpora alongside Sentence-BERT for comparison, our results demonstrate that biasing with reference answers, high-scoring responses, and weighted embeddings improves alignment with human evaluations. Word-based models, particularly after biasing, effectively distinguish high-performing responses from lower ones, as evidenced by increased cosine similarity differences. These findings validate that the proposed methodology enhances the precision and objectivity of evaluating descriptive answers, offering a practical solution for educational settings where fairness and consistency are paramount.

Keywords: corpus; fast text; LLM; SBAR; Word2Vec

Academic Editors: Aleksandra Świetlicka, Aleksandra Kawala-Sterniuk and Dariusz Mikołajewski

Received: 6 February 2025
Revised: 17 May 2025
Accepted: 23 May 2025
Published: 30 May 2025

Citation: Yeo, J.-Y.; Youm, S.; Shin, K.-S. Enhancing Interprofessional Communication in Healthcare Using Large Language Models: Study on Similarity Measurement Methods with Weighted Noun Embeddings. *Electronics* **2025**, *14*, 2240. https://doi.org/10.3390/electronics14112240

1. Introduction

The Situation, Background, Assessment, Recommendation (SBAR) framework is widely utilized in medical communication training to enhance clarity and reduce errors in clinical settings [1,2]. However, traditional SBAR assessments rely heavily on subjective human evaluation, leading to inconsistencies in grading and difficulties in maintaining standardized assessment criteria [3].

Word and sentence similarity measurement plays a crucial role in various natural language processing (NLP) applications, such as automated grading systems, information retrieval, and document summarization [4,5]. By quantifying how closely two text inputs resemble each other, similarity measurement enables systems to perform tasks such as automatic feedback generation and text-based assessments with improved accuracy.

In medical communication training, accurately assessing student responses within the SBAR framework requires precise similarity measurement. Given that responses may vary in structure while conveying the same essential meaning, a robust similarity metric must account for both lexical and semantic equivalence [6]. Conventional evaluation methods often fail to capture subtle linguistic nuances, leading to inconsistencies in grading. Therefore, the development of reliable similarity assessment methods tailored to SBAR responses is essential.

Recent advancements in NLP, including unsupervised learning [7], few-shot learning [8], and GPT-based models [9], have expanded similarity measurement techniques [10–12], with cosine similarity widely applied in essay assessments [13–15]. Various NLP-based approaches have been developed for measuring text similarity, each with its own strengths and applications. One of the most widely used methods is cosine similarity, which calculates the cosine of the angle between two word or sentence vectors, making it useful in vector space models for tasks such as information retrieval and text classification [16]. Another widely adopted approach is Word2Vec and FastText, which generate word embeddings based on co-occurrence statistics within large text corpora. FastText, in particular, is advantageous for domain-specific tasks as it represents words as subword units, enhancing its effectiveness in handling medical terminology [17].

Despite the effectiveness of these techniques, general NLP models often struggle with domain-specific language, particularly in medical and educational contexts. Standard embedding models trained on general corpora may not adequately capture the structured nature of SBAR communication. As a result, fine-tuning models on specialized datasets is necessary to achieve higher accuracy in SBAR response evaluation [18].

While LLMs have demonstrated remarkable performance in various NLP tasks, their application in medical training assessments, particularly structured evaluations such as SBAR, remains an underexplored area. Existing LLM-based evaluation systems primarily focus on general text similarity measurement, often utilizing BERT-based models [19,20]. However, these models struggle to effectively capture the structured and domain-specific nature of SBAR responses.

Recent studies further underscore the potential of LLMs in healthcare and educational contexts. For instance, Hang et al. [21] developed an LLM-driven system for generating multiple-choice questions to support personalized learning, demonstrating how LLMs can adapt to educational tasks through prompt engineering and retrieval-augmented generation. Similarly, Burisch et al. [22] proposed a protocol to evaluate ChatGPT-4's performance in German continuing medical education, exploring its utility in structured healthcare assessments. These works highlight the growing application of LLMs in domain-specific training, yet they also reveal a gap in tailored approaches for structured tasks like SBAR, where fine-tuning on small, specialized datasets remains underexplored. Our study builds on this foundation by addressing this gap with a focused biasing methodology.

In addition, previous studies on automated medical response evaluation often rely on simple similarity scores without incorporating contextual adaptation [23]. This limitation results in inadequate performance when evaluating structured medical assessments, where key domain terms and response format significantly impact evaluation accuracy. Furthermore, LLMs trained on general corpora may fail to recognize the importance of medical-specific phrasing, leading to inconsistencies in automated scoring. These challenges highlight the need for a more domain-adaptive approach to similarity measurement in medical training.

To address these challenges, we propose a three-stage biasing approach for LLM-based similarity measurement. First, Reference Answer Integration involves training the model with expert-curated reference answers to establish a baseline for accurate similarity com-

parisons. Second, High-Scoring Student Response Incorporation integrates top-performing student responses into the training corpus, ensuring that the model aligns with real-world variations in high-quality answers. Finally, Domain-Critical Token Weighting applies Weighted Noun Embeddings to prioritize domain-specific terminology and key medical phrases, assigning higher weights to critical medical nouns and lower weights to less relevant terms. This approach enhances the alignment between automated scoring and human evaluations, leading to greater accuracy in SBAR assessments.

Experimental results demonstrate that models fine-tuned with reference answers and high-scoring student responses achieve significantly higher correlations with expert ratings. Among the models tested, FastText exhibited a higher correlation in handling domain-specific vocabulary, making it a strong candidate for practical implementation in medical education [24].

The remainder of this paper is organized as follows: Section 2 details the similarity measurement methods and corpora used in this study. Section 3 presents our experimental setup and results, and Section 4 discusses the implications of our findings. Finally, Section 5 concludes the study and suggests future research directions.

2. Materials and Methods

2.1. Similarity Analysis Methods

To evaluate the similarity between student SBAR responses and reference answers, we employed three metrics: Cosine Similarity, Euclidean Distance, and Manhattan Distance. These metrics provide complementary perspectives on text similarity, capturing directional alignment, magnitude differences, and coordinate-wise disparities, respectively.

2.1.1. Cosine Similarity

Cosine Similarity is a widely adopted metric for measuring text similarity, as it focuses on the angular difference between two vectors, normalizing for magnitude to mitigate the influence of response length [25]. This makes it particularly suitable for SBAR responses, which vary in length but share common domain-specific tokens. For two vectors v_s (student response) and v_r (reference answer), Cosine Similarity is defined as follows:

$$\text{Cos}(v_s, v_r) = \frac{v_s \cdot v_r}{\|v_s\| \|v_r\|}, \tag{1}$$

where $\cdot$ denotes the dot product, and $\|\cdot\|$ represents the Euclidean norm. The similarity score is scaled to [0, 100] for consistency. Vectors are generated by averaging noun embeddings (extracted using the Mecab morphological analyzer [26]) weighted by domain-critical tokens, as described in Section 2.3.

2.1.2. Word2Vec

Word2Vec generates word embeddings by training a shallow neural network to predict word contexts, capturing semantic relationships. We used a pre-trained model on a medical corpus, fine-tuned with reference answers and high-scoring student responses. Vectors are 100-dimensional, and sentence embeddings are computed as weighted averages of noun embeddings.

2.1.3. FastText

FastText extends Word2Vec by representing words as bags of character n-grams, enhancing robustness for morphologically complex medical terms [27]. Like Word2Vec, it was pre-trained on a medical corpus and fine-tuned, producing 100-dimensional vectors aggregated into sentence embeddings.

2.1.4. Sentence-BERT

Sentence-BERT (S-BERT) generates 768-dimensional sentence embeddings via a Siamese BERT architecture optimized for semantic similarity tasks [28–30]. We fine-tuned a pre-trained Ko-SRoBERTa model using Contrastive Learning to align high-scoring responses with references.

2.2. Corpus Utilization

The study utilized three distinct corpora to provide diverse linguistic contexts for the evaluation of SBAR responses. The Conversational Corpus, comprising 2.9 million tokens of general dialogue, served as a source of everyday language patterns. The Medical Corpus, with 42 million tokens of healthcare-related texts, offered domain-specific terminology and context relevant to medical communication. Finally, the Reference Answer Corpus, containing 732 tokens of expert-crafted SBAR responses, enabled targeted fine-tuning to enhance the precision of SBAR evaluation. These corpora collectively supported the development and refinement of the similarity measurement models used in the study.

2.3. Weighted Noun Embeddings

To prioritize domain-critical tokens, we assign weights to nouns based on their membership in predefined groups. The weight $w(n)$ for a noun n is defined as follows:

$$w(n) = \begin{cases} w_k & \text{if } n \in \mathcal{G}_k, \quad k \in \{1, 2, \ldots, K\}, \\ w_{\text{default}} & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{G}_k$ represents the k-th group of nouns, w_k is the corresponding weight, and w_{default} is the default weight for nouns not in any group. In this study, we use $K = 4$ groups:

- $\mathcal{G}_1$: High-scoring nouns, $w_1 = 1.5$,
- $\mathcal{G}_2$: Positive domain terms, $w_2 = 1.0$,
- $\mathcal{G}_3$: Mid-scoring nouns, $w_3 = 0.1$,
- $\mathcal{G}_4$: Low-scoring nouns, $w_4 = 0.001$,

With $w_{\text{default}} = 0.3$. Sentence embeddings are computed as the weighted average of noun embeddings, enhancing the influence of critical medical terms.

2.3.1. Word2Vec and FastText Fine-Tuning

For Word2Vec and FastText, we updated the pre-trained models using a dataset combining reference answers and high-scoring student responses. Let $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$ denote the set of reference answers for M sections (e.g., $M = 4$ for SBAR), and $\mathcal{S}_{\text{high}} = \{s_1, s_2, \ldots, s_N\}$ represent the set of high-scoring students. For each section $i \in \{1, 2, \ldots, M\}$, the training data are defined as follows:

$$\mathcal{T}_i = \bigcup_{k=1}^{R} \{r_i\} \cup \{\mathcal{H}_{j,i} \mid s_j \in \mathcal{S}_{\text{high}}\}, \tag{3}$$

where $r_i \in \mathcal{R}$ is the reference answer for section i, $\mathcal{H}_{j,i}$ is the response of student s_j for section i, and R is the number of repetitions of the reference answer. In this study, we set $R = 30$ to amplify the influence of reference answers and used $N = 3$ high-scoring students. The models were trained for 50 epochs with negative sampling (15 negative samples), updating the vocabulary and embeddings to prioritize domain-critical tokens.

2.3.2. Sentence-BERT Fine-Tuning

S-BERT was fine-tuned using Contrastive Learning to align high-scoring responses with references while distinguishing mid- and low-scoring ones. The training dataset comprised positive pairs $\{(r_i, h_{j,i}) \mid r_i \in \mathcal{R}, h_{j,i} \in \mathcal{H}\}$ labeled 1.0 and negative pairs $\{(r_i, m_{k,i}) \mid r_i \in \mathcal{R}, m_{k,i} \in \mathcal{M}\}$ labeled 0.0, where $\mathcal{M}$ includes mid- and low-scoring responses. The loss function is as follows:

$$\mathcal{L} = \sum_{(x_1, x_2, y) \in \mathcal{D}} \max(0, 1 - y \cdot \mathrm{Cos}(v_{x_1}, v_{x_2}) + \epsilon), \tag{4}$$

where $\mathcal{D}$ is the set of training pairs, $y \in \{0, 1\}$ is the label, v_{x_1}, v_{x_2} are sentence embeddings, and ϵ is a margin (set to 1). The training was conducted for 10 epochs with a batch size of 4 and 5 warmup steps.

2.4. Experimental Setup

The models were implemented using Python's gensim for Word2Vec and FastText and sentence-transformers for S-BERT. Fine-tuning was performed on a standard CPU with 16 GB RAM, leveraging the lightweight nature of the Reference Answer Corpus (732 tokens). The Mecab analyzer processed texts to extract nouns, ensuring consistency across models.

2.5. Experiment Setup

We employed Python's gensim library (version 4.3.3) to train Word2Vec and FastText models on each of the three corpora. We then applied additional fine-tuning with the Reference Answer Corpus to assess whether it would improve agreement with human evaluators. During training, we used extended epochs when corpora were small, ensuring that the model fully captured the domain-specific vocabulary.

The fine-tuning process for FastText and Word2Vec began with pre-trained models initially developed using a medical corpus with noun weighting adjustments, as outlined earlier. For additional training, we constructed a dataset by combining tokenized reference answers and high-scoring student responses. The reference answers, consisting of expert-crafted SBAR responses, were sourced from a text file and tokenized using the Mecab morphological analyzer, isolating the response text from each entry. High-scoring student responses were extracted from a CSV file containing SBAR answers from 13 students, with the top three performers identified based on human evaluations. For each high-scoring student, responses across the four SBAR sections (Situation, Background, Assessment, Recommendation) were concatenated into a single sentence and tokenized with Mecab. To prioritize these exemplary responses, they were repeated five times in the training data and combined with the reference answers.

The training was conducted for an additional 5000 epochs using the `gensim` library's training functionality, with the total number of sentences in the combined dataset defining the training sample size. Other hyperparameters, such as vector size, window size, and learning rate, were inherited from the pre-trained models, as the objective was to bias the embeddings toward the reference and high-scoring data rather than retrain from scratch. This process was performed on a standard personal computer equipped with a single CPU and approximately 16 GB of RAM within a Python environment managed by Anaconda. Given the small dataset size and the lightweight nature of FastText and Word2Vec, no GPU resources were necessary. The fine-tuned models were saved for subsequent analysis, enabling reproducible evaluation.

3. Results

3.1. SBAR Scenario and Data Collection

An assessment involving 13 participants was conducted to evaluate students' SBAR communication skills in a pediatric scenario. Each student was presented with a situation involving a pediatric patient and asked to respond using SBAR (Situation, Background, Assessment, Recommendation). All participants voluntarily participated in the study after providing informed consent. Table 1 shows the reference answer used in our experiments. Students' textual responses were segmented according to the SBAR structure, and each response was given a cumulative score based on the cosine similarity to the reference answer components. A similarity score above a certain threshold (e.g., 60) was considered a success.

Table 1. Reference answer in a pediatric nursing scenario.

SBAR	Contents
Situation	Hello, I am Kim Jiwoo, a nurse in the emergency room. Are you Dr. Choi Junsu? I am contacting you regarding a 6-year-old boy, Kim Rian, who has a history of asthma and has been admitted to the emergency room with difficulty breathing, coughing, and fever symptoms.
Background	The symptoms started a week ago, and he was treated with medication at a local clinic, but there has been no improvement. He was brought in today due to a fever and difficulty breathing.
Assessment	Vital signs measurements show a pulse rate of 92 beats per minute and a respiratory rate of 28. He took an antipyretic two hours ago, but he is still showing symptoms of fever and difficulty breathing. His SpO_2 is checked at 94%.
Recommendation	The child is in a lot of distress, and the guardian wishes to see the primary physician. Please come quickly to assess the patient's condition and prescribe medication and oxygen as necessary.

3.2. Corpus Statistics

Table 2 summarizes the main characteristics of the three corpora: Conversational, Medical, and Reference Answer. The Conversational Corpus has 2.9 million tokens, while the Medical Corpus is much larger at 42 million tokens. The Reference Answer Corpus is small but has a high vocabulary diversity relative to its size.

Table 2. Training corpus characteristics.

Characteristic	Conversation Corpus	Medical Corpus	Reference Answers
Total Tokens	2,922,486	42,093,425	732
Vocabulary Diversity	0.0184	0.0057	0.4249
Number of Sentences	65,117	1,106,104	49
Average Sentence Length	44.88	38.06	14.94
File Size	18 MB	433 MB	8 KB

The experiment's sample of 13 students, while sufficient for this proof-of-concept, limits conclusiveness on a larger scale, and future work should incorporate a broader external test set to validate results beyond this initial dataset.

3.3. Token Frequency and Embedding Analysis

Figure 1 shows the normalized frequency distribution of nouns in both medical and general corpora, which were analyzed by two different morphological analyzers (Okt and

Mecab). The 2 × 2 grid comparison reveals distinct noun frequency patterns across different domains and analyzers.

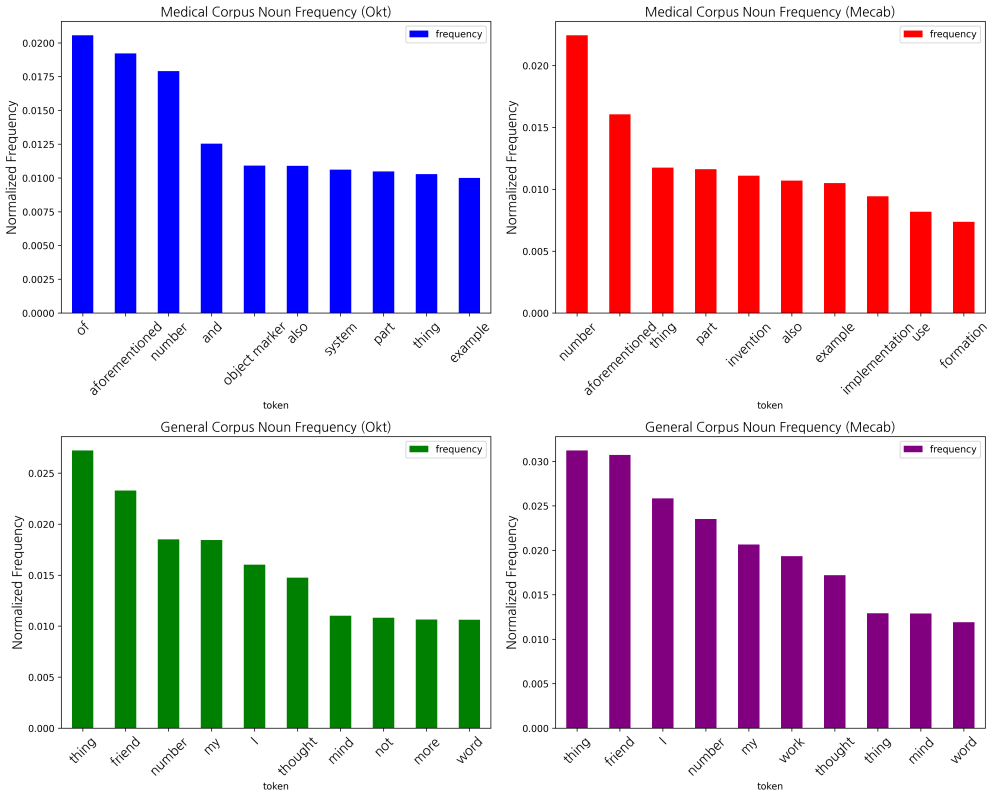

Figure 1. Comparison of noun frequency distributions across corpora and analyzers, with top 10 nouns translated from Korean to English (e.g., 'number', 'thing') for accessibility. (**Top left**) Medical corpus analyzed by Okt shows frequent functional terms. (**Top right**) Medical corpus analyzed by Mecab refines noun extraction. (**Bottom left**) General corpus analyzed by Okt highlights common terms. (**Bottom right**) General corpus analyzed by Mecab shows consistent patterns, supporting Mecab's role in training data preparation. Font sizes increased for readability.

The analysis demonstrates significant differences both between corpora and between analyzers. Processing times notably differed: Okt required 144.86 s for the medical corpus and 240.43 s for the general corpus, while Mecab completed the same analysis in 8.86 and 31.99 s, respectively. In the medical corpus, we observe domain-specific terminology dominating the frequency distribution, while the general corpus shows higher frequencies of everyday vocabulary. Mecab consistently demonstrated faster processing speeds while maintaining comparable accuracy in noun identification, particularly excelling in medical terminology analysis. This performance difference suggests Mecab's potential advantage for large-scale medical text processing applications.

To visualize how student-response vectors compare under different training conditions, we used dimension-reduction techniques to analyze the SBAR (Situation, Background, Assessment, Recommendation) sections separately. Figures 2 and 3 show the Word2Vec and FastText embeddings of student responses for each SBAR section.

To evaluate the effectiveness of our three-stage biasing approach—integrating reference answers, incorporating high-scoring student responses, and applying domain-critical token weighting—we analyzed the embeddings of student SBAR responses using Word2Vec, FastText, and Sentence-BERT models. These models were fine-tuned on a combined corpus comprising general conversational data, medical texts, and the Reference Answer Corpus, supplemented with high-scoring student responses repeated five times to emphasize exemplary patterns. The fine-tuning process aimed to shift the embedding space

such that high-performing student responses align more closely with the reference answers, enhancing the models' ability to distinguish between high- and low-scoring responses in a domain-specific context.

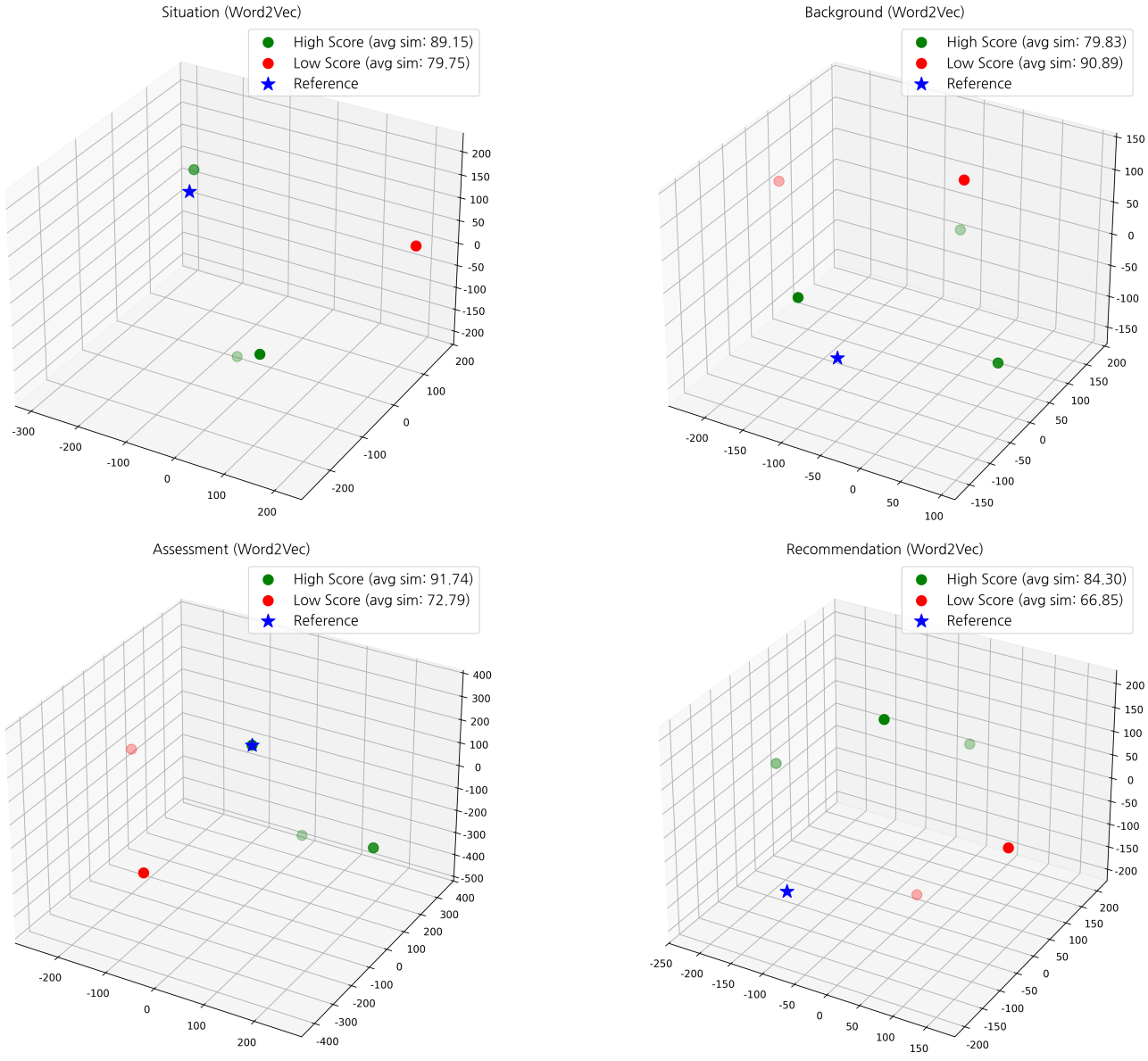

Figure 2. Word2Vec-based dimensionality reduction (t-SNE) of student responses across SBAR sections with noun count-weighted Cosine Similarity, visualizing clustering proximity to reference answers (blue star). Average similarities to references: Situation (High: 89.15, Low: 79.75, Diff: 9.40), Background (High: 79.83, Low: 90.89, Diff: −11.06), Assessment (High: 91.74, Low: 72.79, Diff: 18.95), Recommendation (High: 84.30, Low: 66.85, Diff: 17.45).

Figures 2–4 visualize the 3D t-SNE projections of student response embeddings across the four SBAR sections (Situation, Background, Assessment, Recommendation) for Word2Vec, FastText, and Sentence-BERT, respectively. Each figure depicts high-scoring responses (green), low-scoring responses (red), and the reference answer (blue star), with clustering reflecting Cosine Similarity weighted by noun counts. The t-SNE algorithm reduces high-dimensional embeddings (100D for Word2Vec and FastText, 768D for Sentence-BERT) into a 3D space, providing a spatial representation of semantic similarity.

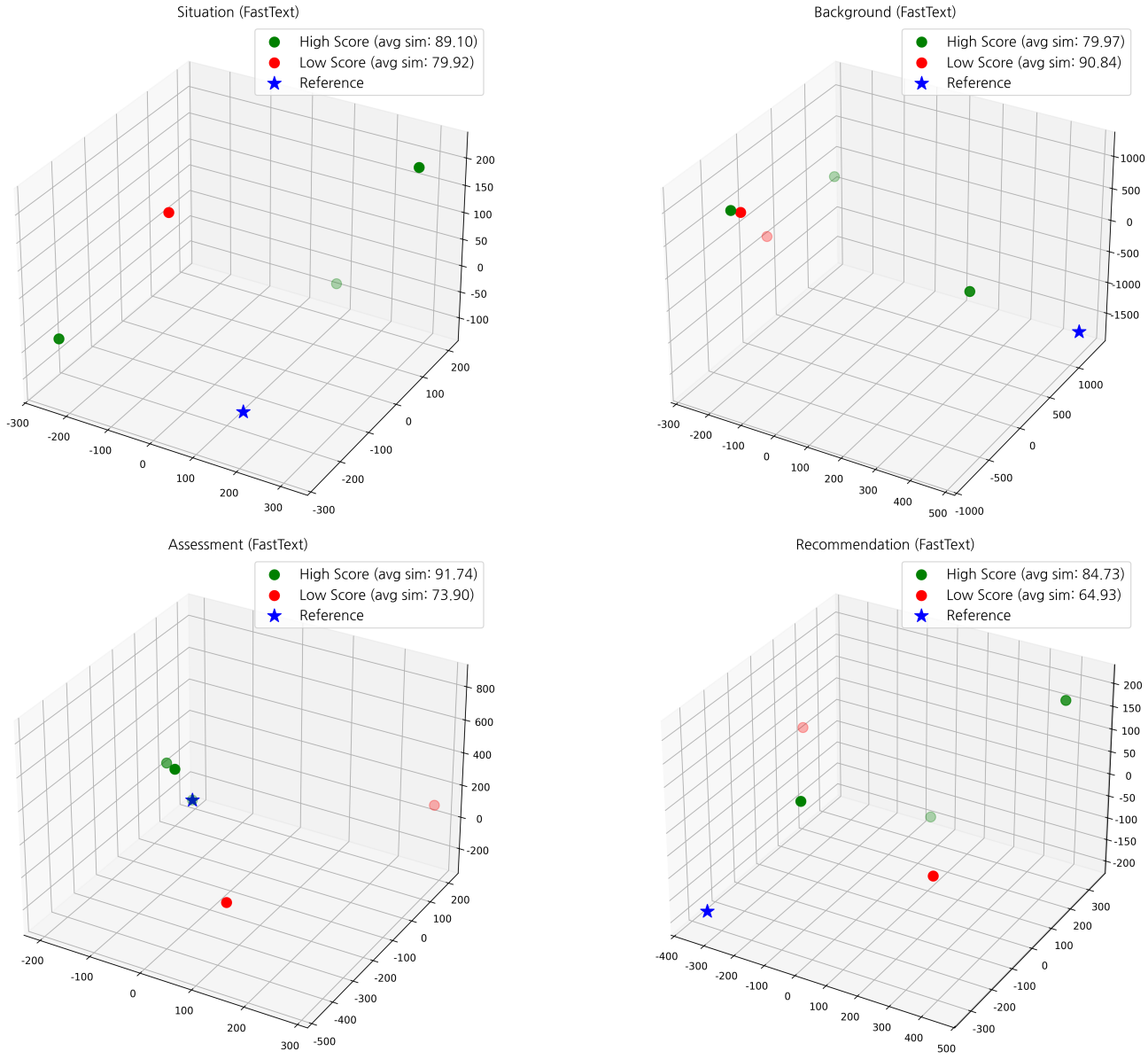

Figure 3. FastText-based dimensionality reduction (t-SNE) of student responses across SBAR sections with noun count-weighted Cosine Similarity, illustrating clustering proximity to reference answers (blue star). Average similarities: Situation (High: 89.10, Low: 79.92, Diff: 9.17), Background (High: 79.97, Low: 90.84, Diff: −10.87), Assessment (High: 91.74, Low: 73.90, Diff: 17.84), Recommendation (High: 84.73, Low: 64.93, Diff: 19.80).

In Figure 2 (Word2Vec), high-scoring responses generally cluster closer to the reference answer compared to low-scoring responses, particularly in the Assessment (High: 91.74, Low: 72.79, Diff: 18.95) and Recommendation (High: 84.30, Low: 66.85, Diff: 17.45) sections. This indicates that fine-tuning with reference answers and high-scoring responses effectively aligns the embedding space with expert expectations. However, the Background section shows an inverse trend (High: 79.83, Low: 90.89, Diff: −11.06), suggesting limitations in capturing context-specific nuances with Word2Vec's word-level embeddings, possibly due to its sensitivity to word co-occurrence patterns rather than structural coherence.

Figure 3 (FastText) exhibits similar trends, with high-scoring responses achieving higher similarities in most sections (e.g., Assessment: High 91.74, Low 73.90, Diff: 17.84; Recommendation: High 84.73, Low 64.93, Diff: 19.80). The positive differences highlight FastText's strength in leveraging subword information, enhancing its sensitivity to medical terminology and morphological variations in the SBAR context. Like Word2Vec, the Background section shows an inverse pattern (High: 79.97, Low: 90.84, Diff: −10.87), indicating challenges in modeling background context, potentially due to the diverse phrasing of student responses.

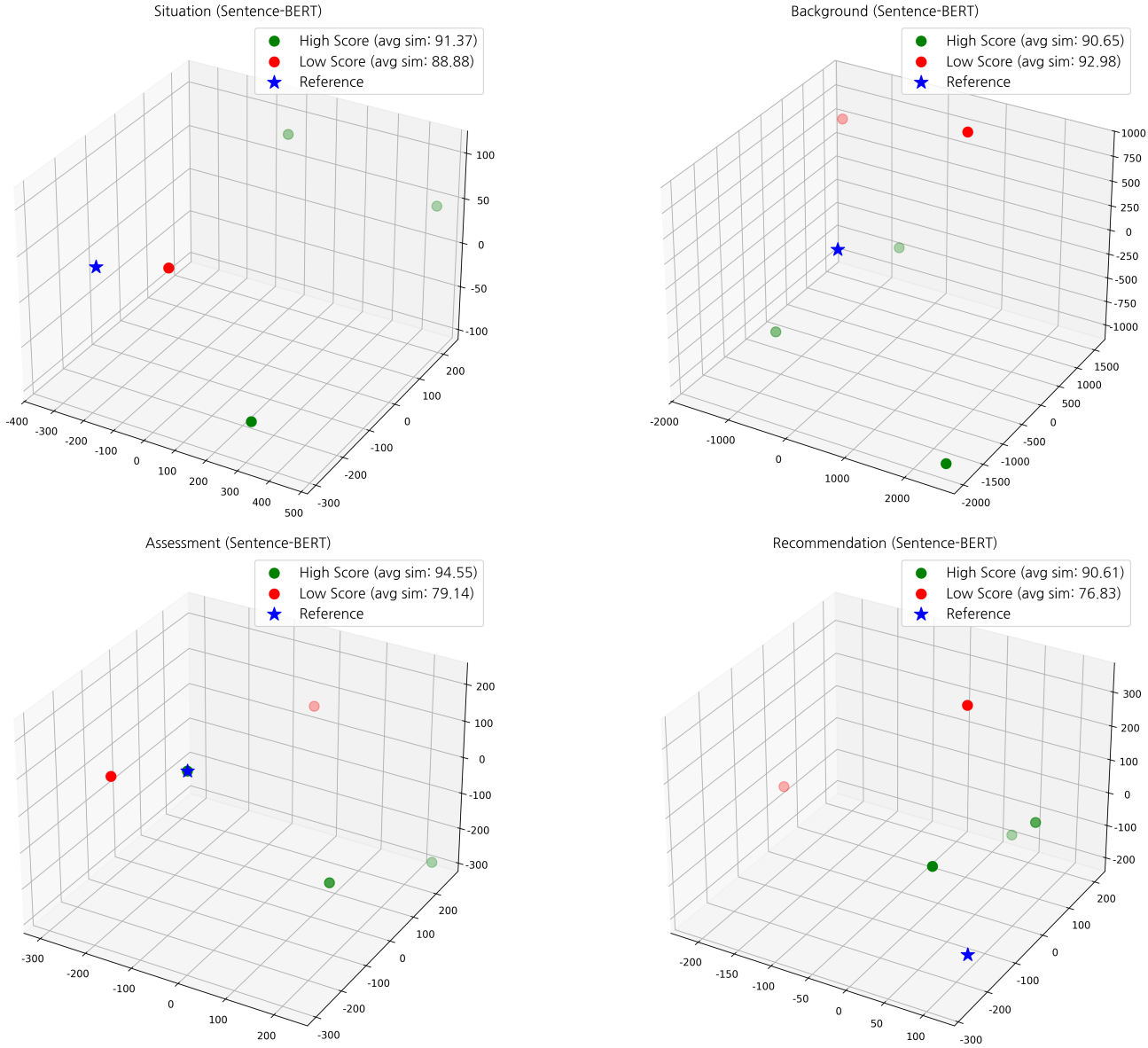

Figure 4. Sentence-BERT-based dimensionality reduction (t-SNE) of student responses across SBAR sections with noun count-weighted Cosine Similarity, illustrating clustering proximity to reference answers (blue star). Average similarities: Situation (High: 91.37, Low: 88.88, Diff: 2.50), Background (High: 90.65, Low: 92.98, Diff: −2.33), Assessment (High: 94.55, Low: 79.14, Diff: 15.41), Recommendation (High: 90.61, Low: 76.83, Diff: 13.78).

In contrast, Figure 4 (Sentence-BERT) shows more consistent clustering, with high-scoring responses achieving higher similarities in Assessment (High: 94.55, Low: 79.14, Diff: 15.41) and Recommendation (High: 90.61, Low: 76.83, Diff: 13.78). However, the differences are smaller than those observed with FastText and Word2Vec, particularly in Situation (Diff: 2.50) and Background (Diff: −2.33). The inverse trend in Background suggests that Sentence-BERT, despite its contextual embedding capabilities, may prioritize broader semantic patterns over the structured, domain-specific nuances of SBAR responses. This could stem from its training on general-purpose corpora, even after fine-tuning, limiting its ability to fully align with the specialized SBAR format.

Tables 3–5 quantify the impact of fine-tuning on FastText, Word2Vec, and Sentence-BERT embeddings by comparing cosine similarities before and after biasing for healthcare trainees' SBAR responses. Before fine-tuning, using baseline models trained on medical data, high-scoring responses exhibit greater similarity to the reference compared to low-scoring ones, though differentiation is moderate. For instance, in the FastText Situation section (Table 3), high-scoring responses score 86.71, low-scoring 72.97 (Diff: 13.74), with an average difference of 16.63 across sections. Word2Vec shows similar trends (Table 4), with Situation scores of High: 87.05, Low: 72.39 (Diff: 14.66), and an average difference of 17.15. Sentence-BERT (Table 5) yields smaller differences, e.g., Situation High: 92.56, Low: 91.54 (Diff: 1.02), with an average difference of 5.01, reflecting its broader semantic focus. After fine-tuning with reference answers, high-scoring responses, and weighted noun embeddings (emphasizing critical terms like "vital signs" via frequency adjustments), differentiation improves significantly. For FastText, Situation scores shift to High: 88.21, Low: 60.65 (Diff: 27.56), with an average difference of 27.38. Word2Vec shows Situation High: 88.26, Low: 58.11 (Diff: 30.15), with an average difference of 28.20. Sentence-BERT improves to Situation High: 90.56, Low: 80.78 (Diff: 9.79), with an average difference of 12.61. Notably, Assessment and Recommendation sections show consistent gains across models (e.g., Word2Vec Assessment Diff: 38.97 to 40.28; FastText Recommendation Diff: 25.86 to 40.70), indicating robust alignment of high-scoring responses with the reference. These results, scalable to n performance tiers, demonstrate that fine-tuning refines the models' semantic understanding of SBAR responses, though the small sample size (13 trainees) and single pediatric scenario limit generalizability, as discussed in Section 4.

The results in Table 6 reveal several key patterns in the similarity evaluation of student responses. Students rated as High by both evaluators (e.g., students 05, 06, 08) consistently exhibit higher similarity scores across all models, with student 05 achieving the highest scores (Word2Vec: 373.85, FastText: 375.45, Sentence-BERT: 394.40), indicating strong alignment with the reference answers. Conversely, students rated Low by at least one evaluator (e.g., students 03 and 09) show varied performance: student 03 scores 0.00 across all models, suggesting a complete mismatch with the reference, possibly due to missing or irrelevant responses, while student 09 maintains relatively high scores (e.g., Word2Vec: 348.02, FastText: 348.08), likely due to partial use of relevant medical terminology despite lower human ratings. Students rated Average generally fall within a moderate range (e.g., student 01: Word2Vec 368.05, FastText 367.85), reflecting typical performance aligned with expected proficiency.

Notably, Sentence-BERT produces higher absolute scores across all students (e.g., student 05: 394.40 vs. FastText: 375.45), likely due to its contextual embeddings capturing broader semantic relationships in full-sentence inputs, compared to the noun-focused, weighted embeddings of Word2Vec and FastText. Despite these absolute differences, the relative distinctions between high- and low-scoring students remain consistent across models, underscoring the robustness of our fine-tuning and domain-specific weighting approach. These scores form the foundation for correlation analysis with human evalua-

tions, demonstrating the models' ability to reflect expert judgments, with FastText showing particular promise due to its subword modeling capabilities, which enhance sensitivity to morphological variations in medical terminology.

Table 3. Comparison of FastText average cosine similarities (rounded to two decimal places) between high-scoring and low-scoring student responses before and after fine-tuning across SBAR sections, using noun-only embeddings with domain-specific weighting. Before uses the unbiased model for all responses, and After uses the fine-tuned model for high-scoring responses and is unbiased for low-scoring responses. Differences (High–Low) are included, complementing Figure 3.

SBAR Section	FastText—Before			FastText—After		
	High	Low	Diff.	High	Low	Diff.
Situation	86.71	72.97	13.74	88.21	60.65	27.56
Background	77.99	89.27	−11.28	76.80	74.29	2.51
Assessment	89.60	51.39	38.21	90.55	51.80	38.75
Recommendation	79.66	53.80	25.86	82.10	41.40	40.70
Avg. Difference			16.63			27.38

Table 4. Comparison of Word2Vec average cosine similarities (rounded to two decimal places) between high-scoring and low-scoring student responses before and after fine-tuning across SBAR sections, using noun-only embeddings with domain-specific weighting. Before uses the unbiased model for all responses, and After uses the fine-tuned model for high-scoring responses and is unbiased for low-scoring responses. Differences (High–Low) are included, complementing Figure 2.

SBAR Section	Word2Vec—Before			Word2Vec—After		
	High	Low	Diff.	High	Low	Diff.
Situation	87.05	72.39	14.66	88.26	58.11	30.15
Background	76.79	89.16	−12.37	75.58	74.63	0.95
Assessment	88.99	50.02	38.97	90.55	50.26	40.28
Recommendation	79.92	52.59	27.33	81.69	40.27	41.42
Avg. Difference			17.15			28.20

Table 5. Comparison of Sentence-BERT average cosine similarities (rounded to two decimal places) between high-scoring and low-scoring student responses before and after fine-tuning across SBAR sections, using noun-only embeddings with domain-specific weighting. Before uses the unbiased model for all responses, and After uses the fine-tuned model for high-scoring responses and is unbiased for low-scoring responses. Differences (High–Low) are included, complementing Figure 4.

SBAR Section	Sentence-BERT—Before			Sentence-BERT—After		
	High	Low	Diff.	High	Low	Diff.
Situation	92.56	91.54	1.02	90.56	80.78	9.79
Background	93.08	93.74	−0.66	89.85	84.84	5.01
Assessment	95.53	85.24	10.29	93.93	77.71	16.22
Recommendation	92.94	83.55	9.38	89.11	69.68	19.43
Avg. Difference			5.01			12.61

Table 6. Results of similarity evaluation for student responses using Word2Vec, FastText, and Sentence-BERT.

id	Eval.1	Eval.2	Word2Vec	FastText	Sentence-BERT
01	Average	Average	368.05	367.85	388.38
02	Average	Average	248.66	247.68	281.31
03	Low	Average	0.00	0.00	0.00
04	Average	Low	312.16	313.59	377.90
05	High	High	373.85	375.45	394.40
06	High	High	345.61	349.74	388.83
07	Average	Average	336.66	337.53	386.65
08	High	High	357.51	356.88	384.59
09	Low	Low	348.02	348.08	362.24
10	Average	Average	306.23	309.22	364.65
11	Average	Average	236.58	236.55	281.09
12	Average	Average	266.01	266.09	288.47
13	Average	Average	355.65	359.26	375.77

4. Discussion

Our study demonstrates that a compact Reference Answer Corpus (49 sentences, 732 tokens; Table 2) can effectively bias lightweight language models like Word2Vec and FastText, aligning high-performing healthcare trainees' SBAR responses with reference vectors, as shown in t-SNE visualizations (Figures 2 and 3). By applying Weighted Noun Embeddings, where critical medical nouns (e.g., "vital signs", "oxygen") were assigned higher weights (1.5) and less relevant terms (e.g., "measurement") lower weights (0.001), and adjusting word frequencies during fine-tuning to emphasize domain-specific terms, we enhanced model sensitivity to clinical vocabulary. This approach achieved strong correlations with human evaluations (Table 6), with Word2Vec and FastText showing robust alignment with Evaluator 1 ($r = 0.77$ for Reference Corpus), suggesting that minimal reference data can suffice for structured SBAR tasks in resource-constrained healthcare training settings.

The inclusion of the Reference Answer Corpus, combined with weighted fine-tuning, shifted embedding distributions, clustering high-performing trainees' responses closer to reference vectors, as evident in visualizations (Section 3.3). Post-biasing, cosine similarity differences between high- and low-scoring responses increased significantly (Word2Vec: 0.09 to 0.26; FastText: 0.09 to 0.23; Tables 3 and 4), reflecting refined differentiation across SBAR sections. Sentence-BERT, while producing higher absolute scores (Table 5), showed smaller differences, indicating its strength in broader semantic patterns but less precision in capturing SBAR's structured nuances. Notably, weighted fine-tuning, by amplifying critical nouns' frequency, modestly improved differentiation beyond the Medical Corpus baseline, highlighting the synergy of large-scale and curated datasets.

This proof-of-concept validates biasing lightweight LLMs with small, weighted datasets for healthcare training, achieving strong alignment with human evaluations across three performance tiers (High, Average, Low; Table 6). The methodology's scalability to n tiers enhances its flexibility for diverse grading schemes, supporting broader educational applications. However, the study's scope—relying on a single pediatric scenario with 13 trainee responses (Section 3.1) and a cosine-based metric—limits generalizability. The small Reference Answer Corpus, while practical, restricts validation across varied contexts, which is a trade-off prioritizing feasibility in data-scarce settings.

Future research should expand the Reference Answer Corpus beyond 49 sentences and increase the trainee sample to test scalability across healthcare training domains (e.g., nursing, allied health). Incorporating external test sets and diverse scenarios would

validate robustness. Exploring alternative metrics (e.g., Spearman's rank) and larger or more complex transformer-based models could further enhance applicability, though our focus on lightweight models suits resource-limited environments. This study provides a practical SBAR grading solution with the potential for broader impact through expanded evaluation, a direction we aim to pursue.

5. Conclusions

This study developed an LLM-based tool to assess SBAR responses from healthcare trainees using a three-stage biasing approach: integrating reference answers, incorporating high-scoring responses, and applying domain-critical token weighting. Weighted Noun Embeddings assigned higher weights (1.5) to key medical nouns and lower weights to irrelevant terms. During fine-tuning, word frequencies were adjusted based on these weights to emphasize domain-specific terms, enhancing clinical relevance. Results show that Word2Vec and FastText, fine-tuned with a compact Reference Answer Corpus (49 sentences, 732 tokens), effectively aligned automated cosine similarity scores with human evaluations, improving differentiation across three performance tiers (High, Average, Low) for 13 trainees. The Medical Corpus (42 million tokens) ensured domain coverage, while targeted fine-tuning refined alignment. This approach, scalable to n performance tiers, offers a practical, objective solution for SBAR grading in resource-limited settings. Future work should expand the corpus and sample size and explore contextual models for broader healthcare training applications.

Author Contributions: Conceptualization, J.-Y.Y.; methodology, J.-Y.Y.; software, S.Y.; validation, J.-Y.Y.; formal analysis, S.Y.; investigation, J.-Y.Y. and S.Y.; writing—original draft preparation, S.Y.; writing—review and editing, J.-Y.Y.; visualization, K.-S.S.; supervision, K.-S.S.; project administration, J.-Y.Y.; funding acquisition, K.-S.S.; resources, S.Y.; data curation, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon reasonable request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors thank the anonymous reviewers and editors for their insightful comments and suggestions.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Burisch, C.; Bellary, A.; Breuckmann, F.; Ehlers, J.; Thal, S.; Sellmann, T.; Gödde, D. ChatGPT-4 Performance on German Continuing Medical Education—Friend or Foe (Trick or Treat)? Protocol for a Randomized Controlled Trial. *JMIR Res. Protoc.* **2025**, *14*, e63887. [CrossRef] [PubMed]
2. Müller, M.; Jürgens, J.; Redaèlli, M.; Klingberg, K.; Hautz, W.E.; Stock, S. Impact of the communication and patient hand-off tool SBAR on patient safety: A systematic review. *BMJ Open* **2018**, *8*, e022202. [CrossRef]
3. Hang, C.N.; Tan, C.W.; Yu, P.-D. MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access* **2024**, *12*, 102261–102273. [CrossRef]
4. Reference Answer Corpus. Available online: https://github.com/skyoum00/SBAR_Assessment_Tool/blob/main/referenceAnswerCorpus (accessed on 7 March 2025).
5. Wang, J.; Dong, Y. Measurement of Text Similarity: A Survey. *Information* **2020**, *11*, 421. [CrossRef]
6. Medical Corpus. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71487 (accessed on 7 March 2025).

7. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
8. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *NeurIPS* **2020**, *33*, 1877–1901.
9. Yenduri, G.; Garg, D.; Oak, R.; Hooda, D.; Aggarwal, B.; Kumar, V.; Singh, A.; Lal, S. Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *arXiv* **2023**, arXiv:2305.10435. [CrossRef]
10. Song, G.; Ye, Y.; Du, X.; Huang, X.; Bie, S. Short Text Classification: A Survey. *J. Med. Microbiol.* **2014**, *9*, 635–643. [CrossRef]
11. Yi, E.; Koenig, J.-P.; Roland, D. Semantic similarity to high-frequency verbs affects syntactic frame selection. *Cogn. Linguist.* **2019**, *30*, 601–628. [CrossRef]
12. Zhou, Y.; Li, C.; Huang, G.; Guo, Q.; Li, H.; Wei, X. A Short-Text Similarity Model Combining Semantic and Syntactic Information. *Electronics* **2023**, *12*, 3126. [CrossRef]
13. Lahitani, A.R.; Permanasari, A.E.; Setiawan, N.A. Cosine similarity to determine similarity measure: Study case in online essay assessment. In Proceedings of the 2016 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 26–27 April 2016; pp. 1–6.
14. Nurfadila, P.D.; Wibawa, A.P.; Zaeni, I.A.E.; Nafalski, A. Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stopword Removal. *Int. J. Artif. Intell. Res.* **2019**, *3*, 41–47. [CrossRef]
15. Ristanti, P.Y.; Wibawa, A.P.; Pujianto, U. Cosine Similarity for Title and Abstract of Economic Journal Classification. In Proceedings of the 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 23–24 October 2019; pp. 123–127.
16. Conversational Corpus. Available online: https://raw.githubusercontent.com/byungjooyoo/Dataset/main/corpus.txt (accessed on 7 March 2025).
17. Steinberger, J.; Ježek, K. Text Summarization and Singular Value Decomposition. In Proceedings of the 7th International Conference on Advances in Information Systems (ADVIS), Istanbul, Turkey, 20–22 October 2004; pp. 245–254.
18. Li, X.; Yao, C.; Fan, F.; Yu, X. A Text Similarity Measurement Method Based on Singular Value Decomposition and Semantic Relevance. *J. Inf. Process. Syst.* **2017**, *13*, 863–875. [CrossRef]
19. Cheng, C.-H.; Chen, H.-H. Sentimental text mining based on an additional features method for text classification. *PLoS ONE* **2019**, *14*, e0217591. [CrossRef] [PubMed]
20. Li, Y.; Bandar, Z.; McLean, D.; O'Shea, J. A Method for Measuring Sentence Similarity and its Application to Conversational Agents. In Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, Miami Beach, FL, USA, 17–19 May 2004.
21. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Casas, D.d.L.; Hendricks, L.A.; Welbl, J.; Clark, A.; Hattingh, J.; et al. Improving language models by retrieving from trillions of tokens. *arXiv* **2022**, arXiv:2112.04426.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
23. Wahyudi; Akbar, R.; Suharsono, T.N.; Indrapriyatna, A.S. Essay Test Based E-Testing Using Cosine Similarity Vector Space Model. In Proceedings of the 2022 International Symposium on Information Technology and Digital Innovation (ISITDI), Padang, Indonesia, 27–28 July 2022; pp. 80–85.
24. Davis, B.P.; Mitchell, S.A.; Weston, J.; Dragon, C.; Luthra, M.; Kim, J.; Stoddard, H.; Ander, D. Situation, Background, Assessment, Recommendation (SBAR) Education for Health Care Students: Assessment of a Training Program. *MedEdPORTAL* **2023**, *19*, 11293. [CrossRef]
25. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
26. Kudo, T.; Yamamoto, K.; Matsumoto, Y. Applying Conditional Random Fields to Japanese Morphological Analysis. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, 25–26 July 2004; pp. 230–237.
27. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
28. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.
29. Park, J.; Shin, J.; Woo, S.; Lee, J.; Jang, M.; Lee, H.; Ham, D. KLUE: Korean Language Understanding Evaluation. *arXiv* **2021**, arXiv:2105.09680.
30. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. *arXiv* **2018**, arXiv:1803.08808.