Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

MedSeg: A Statistical Approach to Tokenization Assessment in Medical NLP

Min-Yung Yu[1], Jae-Chern Yoo[2]

[1] *Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea Email: 6unhuiw@gmail.com*

[2]*Professor, Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea Email: yoojc@skku.edu*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

With the rapid adoption of Large Language Models (LLMs) in healthcare, accurate tokenization of complex medical terms has become increasingly critical. Improper segmentation leads to high unidentified words and suboptimal performance, particularly in medical Natural Language Processing (NLP) tasks. While subword tokenization methods like WordPiece and Byte Pair Encoding (BPE) have been widely used to mitigate Out-of-Vocabulary issues, there remains a lack of specialized metrics for evaluating their effectiveness in the medical domain. In this study, we propose MedSeg, a novel statistical evaluation metric designed to assess tokenizer performance by analyzing the Token Split Rate and Out-of-Vocabulary distribution across word lengths. MedSeg introduces a domain-aware, regression-based scoring mechanism that compares each tokenizer's output to an estimated population distribution, quantified using normalized root mean square error (NRMSE). Experimental results using BioBERT and BioLlama on CTCAE data demonstrate that MedSeg effectively captures the trade-off between segmentation granularity and medical vocabulary preservation. The proposed metric provides a robust and interpretable framework for assessing tokenization strategies in domain-specific NLP applications.

Keywords: Medical NLP, Tokenization Evaluation, Subword Tokenizer, WordPiece, BPE, Out-of-Vocabulary, Token Split Rate, MedSeg

INTRODUCTION

With the growing advancement of Large Language Models (LLMs), their potential application in the healthcare industry has become a significant research focus(Shool et al., 2025; Ullah et al., 2024). Healthcare demands that LLMs achieve a higher standard of understanding, specifically tailored to accurately interpret medical terminologies and complex clinical contexts, far surpassing general language comprehension(Friedman et al., 1994; Meystre & Haug, 2006). Accuracy and reliability become especially critical, as incorrect interpretations can lead directly to compromised patient safety and even fatalities.

The tokenizer plays an important role in LLM's effectiveness, which segments text into manageable units based on its vocabulary(Khattak et al., 2019). Poor segmentation of medical terms by the tokenizer can significantly elevate the rate of Out-of-Vocabulary tokens, subsequently reducing the model's interpretative ability(Benamar et al., 2022). Tokenizers that divide words into excessively small segments unnecessarily increase input length, while overly broad segmentation can obscure important distinctions among specialized medical concepts(Dotan et al., 2024; Nayak et al., 2020). Hence, precision in tokenization processes is vital for optimizing the performance of LLMs in medical domains.

Incorporating specialized medical terminology into existing vocabularies significantly expands overall vocabulary size, presenting challenges in terms of computational complexity and resource management(Tao et al., 2024). Traditional word-level tokenization methods inadequately address these challenges, often replacing unknown medical terms with generic tokens, leading to information loss(Pfeiffer et al., 2021). Conversely, subword tokenization has emerged as a more viable solution, breaking words into smaller, more meaningful units and effectively reducing OOV issues without excessive computational burdens(He et al., 2014).

Tokenization strategies deeply influence an LLM's ability to handle domain-specific vocabulary and ultimately its overall accuracy and utility(Bolton et al., 2024; Goldman et al., 2024). Building on this, Wang et al. (2021)(Sachidananda et al., 2021) demonstrated that incorporating domain-specific subword units through adaptive tokenization—without additional pretraining—substantially improves model performance across biomedical, legal, and scientific domains. Their method highlights that tailoring tokenization to a target domain can yield performance gains comparable to domain-adaptive pretraining, emphasizing the critical role of domain-aware vocabulary design in efficient and accurate language model adaptation.

Conventional tokenization evaluation metrics, such as Average Token Length and Sparsity, do not adequately reflect the complexities inherent in medical texts, which frequently contain long and intricately structured terminology. To address this limitation, this research introduces Medical Segmentation Score, an innovative quantitative evaluation metric specifically designed for medical NLP applications. MedSeg systematically evaluates tokenizers by comparing their Token Segmentation Rates (TSR) and OOV distributions against statistical baselines derived from approximated population distributions, enabling a more nuanced and precise assessment of tokenizer performance in medical contexts.

EVALUATION SCORE

This study introduces the Medical Segmentation Score, a metric designed to assess tokenizer performance using two established indicators: Token Split Rate and Out-of-Vocabulary Rate. Due to the large number of words and the inherently statistical nature of subword segmentation, directly comparing tokenizers at the individual word level is neither feasible nor meaningful, especially when attempting to assign semantic value to subword units. To address this, MedSeg evaluates tokenizer performance by analyzing the distribution patterns of TSR and OOV across different word lengths. This word-length-aware statistical approach provides a more precise and informative evaluation of subword segmentation quality.

To evaluate tokenizer performance systematically, our proposed MedSeg follows a structured four-step process:

1. **Distribution Calculation**: Compute the TSR and OOV distributions separately for each word length.
2. **Ground Truth Estimation**: Derive a Ground Truth (GT) distribution by averaging the TSR and OOV distributions of the tokenizers under comparison.
3. **Non-linear Regression**: Apply Gaussian-based non-linear regression to estimate the GT distributions and calculate the deviation of each tokenizer's distribution from the GT using NRMSE.
4. **Metric Integration**: Integrate the resulting NRMSE values for TSR and OOV into a single normalized lexical segmentation score ranging from 0 to 1.

Token Split Rate(TSR) quantifies how often and to what extent words are broken down into subword tokens. For example, if a word is segmented by a WordPiece tokenizer into token1 + ##token2 + ##token3, its TSR would be 3. A high TSR reflects more aggressive token splitting, while a low TSR suggests that the tokenizer tends to preserve words as single units.

Instead of assessing each word individually, our approach organizes words by their character length and computes the average TSR for each group. That is, we calculate the mean number of tokens generated from words of length 5, 6, and so on. This results in a distribution that captures how tokenization behavior varies across different word lengths.

In general, the Out-of-Vocabulary (OOV) rate refers to how frequently a tokenizer encounters unknown words in a given dataset(Araabi et al., 2022). While subword-based tokenizers aim to eliminate traditional OOV issues by decomposing unfamiliar words into known subword units, we adopt a more comparative definition of OOV in this study. Specifically, a word is considered OOV if it is not found as a complete token in the tokenizer's vocabulary and must be split into multiple subword tokens.

Given the structural complexity of medical texts, we compute the OOV rate as a function of word length—measuring how often words of a specific length are absent from the tokenizer's vocabulary as complete tokens. This produces a word-length-specific OOV distribution that enables detailed, tokenizer-level comparisons.

Medical Segmentation Score (MedSeg)

Since it is inherently difficult to define a single "correct" tokenization, our study constructs a Ground Truth (GT) distribution by averaging the distributions of the two tokenizers being evaluated. Formally, we suppose the GT distribution as:

$$GT \approx \frac{D_{k_1} + D_{k_2} + \cdots + D_{k_n}}{n} \tag{1}$$

where $D[i]$ denotes the distribution (frequency or proportion) of words of length i tokenized by Tokenizers. We measure the similarity between each tokenizer's distribution and the GT distribution using the non-linear regression. MedSeg uses non-linear regression based on a Gaussian function to approximate the GT distributions of TSR and OOV rates. These regression equations serve as reference points for comparing different tokenizer distributions. The differences between the tokenizer's distributions and the GT distributions are quantitatively measured through an error function, specifically the Normalized Root Mean Squared Error ($NRMSE$). This process results in two distinct RMSE values for each tokenizer: one reflecting the TSR distribution ($NRMSE_{TSR}$) and another for the OOV distribution ($NRMSE_{OOV}$). The fitted Gaussian functions serve as baseline models, representing the optimal distributions of TSR and OOV rates. Lower RMSE values indicate that a tokenizer closely matches the optimal distribution, suggesting superior tokenization performance. The NRMSE is computed as:

$$NRMSE = \frac{1}{mean(GT)} \sqrt{\frac{\sum_{i=1}^{N} \left(GT_N - D_{k_N}\right)^2}{N}} \tag{2}$$

Subsequently, these two NRMSE values are combined into a single comprehensive metric, MedSeg, defined as:

$$\mathrm{MedSeg} = 1 - [\lambda(\mathrm{NRMSE}_{TSR}) + (1 - \lambda)\mathrm{NRMSE}_{OOV}] \tag{3}$$

Here, λ is a hyperparameter that balances the relative importance of TSR and OOV; in this study, λ is set at 0.8, emphasizing TSR slightly more than OOV. By integrating both TSR and OOV aspects, MedSeg penalizes tokenizers that significantly deviate in either dimension. As a result, a MedSeg value near 1 suggests that a tokenizer closely matches the optimal tokenization pattern, whereas a score approaching 0 highlights significant deficiencies in tokenizer performance. This metric, therefore, provides a detailed and nuanced evaluation tailored specifically to address the unique tokenization challenges within medical NLP.

MATERIALS AND METHODS

Subword tokenization aims to strike a balance between minimizing vocabulary size and ensuring comprehensive text coverage by segmenting text into units smaller than full words but larger than individual characters. This study centers on two widely used subword tokenization algorithms: WordPiece and Byte Pair Encoding (BPE). WordPiece(Schuster & Nakajima, 2012), initially created for machine translation and later widely adopted through BERT(Devlin et al., 2019), incrementally builds its vocabulary by selecting subword units that optimize the language model's likelihood. In contrast, BPE(Sennrich et al., 2016) follows a greedy approach by repeatedly merging the most frequent adjacent character pairs, and was originally developed for neural machine translation tasks. Despite their different mechanisms, both methods ultimately produce a fixed subword vocabulary that balances representing rare words effectively while preserving longer, semantically rich tokens.

For evaluation, we employed the Common Terminology Criteria for Adverse Events (CTCAE) dataset (*National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE)*, n.d.) , curated by the National Institutes of Health (NIH). The CTCAE is a standardized resource used widely in clinical oncology trials to classify adverse events by type and severity.

From version 5.0 of this dataset, we extracted a list of adverse event terms, such as disease conditions and symptom names (e.g., sore throat, anal fistula, Stevens-Johnson syndrome). These terms often consist of multiple words or complex compounds, posing a challenge for tokenization systems. Because such terms are typically domain-specific and relatively rare in general corpora, they test the tokenizer's ability to preserve medical meaning through effective segmentation.

To better understand how tokenizers process specialized terminology, we chose to work with term-level data instead of full sentences. Prior to tokenization, we applied minimal preprocessing: all non-textual metadata was removed, and the text was converted to lowercase to maintain consistency, given that the tokenizers used in this study are uncased models.

We implemented both tokenizers using pre-trained BioBERT (WordPiece, 110M parameters, 30K Vocabulary size) (DMIS Lab., n.d.) and BioLlama(BPE, 137M parameters, 1.2M Vocabulary size)(iRASC., n.d.) models. We analyzed tokenization outputs in terms of segmentation patterns and semantic preservation.

RESULTS

In this section, we present the results of our experiments. Within each, we compare the performance of WordPiece-based vs BPE-based models. We also include analyses of tokenization-specific metrics to interpret the results.

Table 1 presents sample tokenization outputs for selected medical terms. The WordPiece tokenizer preserves contextual meaning by indicating connections between subwords using the prefix '##' when splitting words, whereas the Byte-Pair Encoding (BPE) tokenizer retains structural text information by marking word boundaries with the 'Ġ' symbol to indicate whitespace.

Overall, WordPiece demonstrated superior performance in preserving medically meaningful segments. For example, in the phrase "cardiac arrest," WordPiece successfully maintained two distinct medical terms ("cardiac" and "arrest"), whereas BPE fragmented "cardiac" into "card" and "iac," potentially obscuring its medical significance. However, for terms like "myocarditis," both WordPiece and BPE produced semantically ambiguous subword splits, making interpretation more difficult.

Consequently, these results underscore the inherent difficulty of consistently producing semantically meaningful subword splits in medical texts. Although WordPiece demonstrates relatively greater semantic coherence overall, despite a relatively compact vocabulary size, challenges remain in ensuring fully interpretable segmentation, emphasizing the need for further refinement in tokenizer strategies tailored specifically for medical terminology.

text	BPE (BioLlama)	WordPiece (BioBERT)
cardiac arrest	[card, iac, Ġarrest]	[cardiac, arrest]
myocarditis	[my, ocard, itis]	[my, ##oc, ##ard, ##itis]

Table 1. Experimental Results of WordPiece and BPE on Two Text Samples

Results of Proposed Segmentation Score

In this section, we present detailed results derived from our proposed Medical Segmentation Score. We systematically analyze each component of MedSeg—TSR and OOV Rate—and interpret the overall MedSeg scores by comparing the performance of BioBERT and BioLlama tokenizers.

1) TSR

Figure 1 illustrates the normalized average TSR for each word length produced by two tokenizers—BioLlama (blue) and BioBERT (orange)—along with a fitted Gaussian curve (dashed red line) representing the average TSR trend. The green and red lines show the differences between each tokenizer's TSR and the fitted curve, which are used to compute NRMSE values.

As described in the Evaluation Score section, TSR reflects how often a word is broken into subword units. For words shorter than 15 characters, both tokenizers produce comparable segmentation patterns. However, with longer medical terms (15 characters or more), BioLlama applies more aggressive splitting than BioBERT, suggesting that it breaks complex terms into finer subword units. This detailed segmentation can help reveal morphological or semantic components in medical vocabulary, which may aid interpretability. Still, excessive splitting can result in fragmented tokens, potentially reducing the coherence of the representation. Compared to BioBERT, BioLlama shows a smoother, more gradual increase in TSR as word length grows, which may help maintain semantic consistency across longer terms.

2) OOV

Figure 1 presents a similar analysis for OOV rates by word length. The fitted Gaussian function models the OOV distribution, and the deviations from the curve are again used to calculate NRMSE for each tokenizer. The visualization illustrates how closely each tokenizer aligns with the modeled ground truth across varying word lengths. The BioLlama tokenizer generally exhibits higher OOV rates, particularly for words between 5 and 10 characters long. This pattern suggests that BioBERT includes a richer set of medical-specific subwords in its vocabulary, especially for medium- to long-length terms. Conversely, BioLlama's elevated OOV rate for shorter words indicates that its general-domain vocabulary lacks sufficient specialized medical tokens. Despite this, BioLlama demonstrates strong coverage of medical terminology overall, even with a more compact vocabulary size.

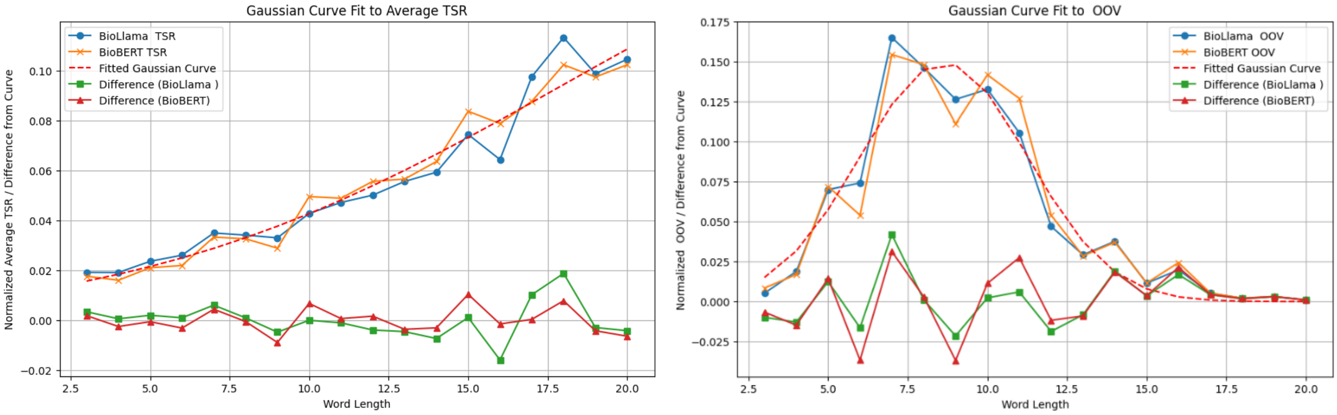

Figure 1. Gaussian curve fitting of TSR (left) and OOV (right) distributions by word length for BioLlama and BioBERT. The plots show each tokenizer's normalized values, the fitted Gaussian curve, and the differences used to compute NRMSE.

3) Comparison Using the Proposed Score

	$NRMSE_{TSR}$	$NRMSE_{OOV}$	**MedSeg**
BPE (BioLlama)	0.36179	0.44640	0.57051
WordPiece (BioBERT)	0.37818	0.42398	0.58517

Table 2. summarizes the results of the MSE and the final MedSeg calculated based on the TSR and OOV distributions for both tokenizers.

This section presents the evaluation results for the BioLlama and BioBERT tokenizers, based on the normalized distributions of TSR and OOV, as well as the corresponding NRMSE values and final MedSeg scores in **Table 2**. BioBERT achieves a higher MedSeg score (0.58517) than BioLlama (0.57051), indicating superior overall segmentation behavior. Since MedSeg values closer to 1 represent stronger alignment with the Ground Truth (GT) distribution, a higher MedSeg score reflects better tokenization performance.

MedSeg is computed by integrating NRMSE values from both TSR and OOV. Because higher Mean Squared Error (MSE) values indicate greater deviation from the GT distribution, lower NRMSE values suggest better alignment and thus better tokenizer quality. In this regard, the MedSeg metric effectively captures these differences, and the trends are clearly observable in both **Figure 1** and the numerical results—demonstrating strong agreement between quantitative and visual evaluations.

In **Figure 1**, BioLlama exhibits TSR behavior that more closely follows the fitted Gaussian curve, particularly for longer words (≥15 characters), where its segmentation is smoother and more consistent than BioBERT's. This visual pattern is consistent with BioLlama's lower $NRMSE_{TSR}$ (0.36179), indicating more stable token splitting.

In contrast, **Figure 1** shows that BioBERT's OOV distribution better fits the Gaussian curve, especially in the 5–10 character word length range. This suggests that BioBERT includes a richer set of medical subword tokens, resulting in a lower OOV rate. The corresponding $NRMSE_{OOV}$ (0.42398) confirms this, as it is lower than that of BioLlama (0.44640), indicating stronger vocabulary coverage.

These results illustrate a clear trade-off between segmentation stability (TSR) and vocabulary preservation (OOV). BioLlama performs better in terms of consistent segmentation, while BioBERT excels in preserving domain-specific vocabulary. MedSeg effectively combines these two aspects, and the alignment between numerical scores and visual patterns supports its validity as a robust and interpretable metric for evaluating tokenizers in medical NLP.

CONCLUSION AND FUTURE WORK

In this study, we introduced the Med, a novel evaluation metric designed to assess tokenizer performance in the context of medical NLP. Unlike traditional metrics that focus solely on token counts or OOV rates, MedSeg provides a more fine-grained assessment by incorporating TSR and OOV distributions across word lengths. These distributions are compared to a Gaussian-approximated GT to quantify how closely a tokenizer aligns with optimal segmentation behavior.

Our results demonstrate that BioLlama shows stronger consistency in segmentation (lower $NRMSE_{TSR}$), while BioBERT performs better in vocabulary preservation (lower $NRMSE_{OOV}$). The trade-offs between these two dimensions are clearly captured by MedSeg, and the metric's values closely align with the visual and statistical patterns observed in the experimental data—highlighting MedSeg as a reliable and interpretable metric.

MedSeg offers a comprehensive and intuitive way to evaluate tokenizers by balancing segmentation granularity and vocabulary coverage, both of which are critical in domain-specific NLP applications. Future work may refine the definition of the GT distribution or extend the application of MedSeg to a broader range of domains and languages, further enhancing its generalizability and practical value.

REFRENCES

[1] Araabi, A., Monz, C., & Niculae, V. (2022). How Effective is Byte Pair Encoding for Out-Of-Vocabulary Words in Neural Machine Translation? In K. Duh & F. Guzmán (Eds.), *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (pp. 117–130). Association for Machine Translation in the Americas. https://aclanthology.org/2022.amta-research.9/

[2] Benamar, A., Grouin, C., Bothua, M., & Vilnat, A. (2022). Evaluating Tokenizers Impact on OOVs Representation with Transformers Models. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4193–4204). European Language Resources Association. https://aclanthology.org/2022.lrec-1.445/

[3] Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., & Manning, C. D. (2024). *BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text* (No. arXiv:2403.18421). arXiv. https://doi.org/10.48550/arXiv.2403.18421

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

[5] DMIS Lab. (n.d.). *BioBERT (biobert-base-cased-v1.2)* [Dataset]. Hugging Face. Retrieved March 31, 2025, from https://huggingface.co/dmis-lab/biobert-base-cased-v1.2

[6] Dotan, E., Jaschek, G., Pupko, T., & Belinkov, Y. (2024). Effect of tokenization on transformers for biological sequences. *Bioinformatics*, *40*(4), btae196. https://doi.org/10.1093/bioinformatics/btae196

[7] Friedman, C., Alderson, P. O., Austin, J. H. M., Cimino, J. J., & Johnson, S. B. (1994). A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, *1*(2), 161–174. https://doi.org/10.1136/jamia.1994.95236146

[8] Goldman, O., Caciularu, A., Eyal, M., Cao, K., Szpektor, I., & Tsarfaty, R. (2024). *Unpacking Tokenization: Evaluating Text Compression and its Correlation with Model Performance* (No. arXiv:2403.06265). arXiv. https://doi.org/10.48550/arXiv.2403.06265

[9] He, Y., Hutchinson, B., Baumann, P., Ostendorf, M., Fosler-Lussier, E., & Pierrehumbert, J. (2014). Subword-based modeling for handling OOV words inkeyword spotting. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7864–7868. https://doi.org/10.1109/ICASSP.2014.6855131

[10] iRASC. (n.d.). *BioLlama-Ko-8B* [Dataset]. Hugging Face. Retrieved March 31, 2025, from https://huggingface.co/iRASC/BioLlama-Ko-8B

[11] Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, *100*, 100057. https://doi.org/10.1016/j.yjbinx.2019.100057

[12] Meystre, S., & Haug, P. J. (2006). Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, *39*(6), 589–599. https://doi.org/10.1016/j.jbi.2005.11.004

[13] *National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE)*. (n.d.). [Dataset]. Retrieved March 31, 2025, from https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.html

[14] Nayak, A., Timmapathini, H., Ponnalagu, K., & Gopalan Venkoparao, V. (2020). Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words. In A. Rogers, J. Sedoc, & A. Rumshisky (Eds.), *Proceedings of the First Workshop on Insights from Negative Results in NLP* (pp. 1–5). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.insights-1.1

[15] Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2021). *UNKs Everywhere: Adapting Multilingual Language Models to New Scripts* (No. arXiv:2012.15562). arXiv. https://doi.org/10.48550/arXiv.2012.15562

[16] Sachidananda, V., Kessler, J. S., & Lai, Y. (2021). *Efficient Domain Adaptation of Language Models via Adaptive Tokenization* (No. arXiv:2109.07460). arXiv. https://doi.org/10.48550/arXiv.2109.07460

[17] Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079

[18] Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural Machine Translation of Rare Words with Subword Units* (No. arXiv:1508.07909). arXiv. https://doi.org/10.48550/arXiv.1508.07909

[19] Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., & Tara, M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, *25*(1), 117. https://doi.org/10.1186/s12911-025-02954-4

[20] Tao, C., Liu, Q., Dou, L., Muennighoff, N., Wan, Z., Luo, P., Lin, M., & Wong, N. (2024). *Scaling Laws with Vocabulary: Larger Models Deserve Larger Vocabularies* (No. arXiv:2407.13623). arXiv. https://doi.org/10.48550/arXiv.2407.13623

[21] Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, *19*(1), 43. https://doi.org/10.1186/s13000-024-01464-7