Tokenization Granularity and Medical Term Representations in Language Models

Vojtěch Lanz and Pavel Pecina
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{lanz,pecina}@ufal.mff.cuni.cz

Abstract

We investigate how tokenization granularity affects the representation of medical terminology in language models. Prior work links tokenization granularity to downstream performance under contextualized settings for specifically pretrained and fine-tuned models. We instead ask whether this relationship already emerges at the level of isolated term representations across existing pretrained models. We introduce an intrinsic definition retrieval task using UMLS term-definition pairs, with comparison to WordNet. We show that despite substantially heavier fragmentation of medical terminology, the models remain relatively robust in maintaining semantic alignment between medical terms and their definitions. At the same time, tokenization granularity still correlates with retrieval performance, indicating that effects previously observed in downstream biomedical tasks are already reflected at the level of isolated term representations. Encoder models benefit primarily from whole-token preservation, while for decoder LLMs, tokenization effects emerge mainly at deeper retrieval ranks.

1 Introduction

Tokenization is the first processing step for language models and directly shapes their input representations, thereby influencing overall performance. In practice, tokenizers are typically trained on general-domain corpora (Grattafiori et al., 2024; Jiang et al., 2023; Team et al., 2025; Abdin et al., 2024) and reused when adapting models to specialized domains (Christophe et al., 2024; Labrak et al., 2024b; Sellergren et al., 2025; Corbeil et al., 2025). This mismatch is particularly pronounced in the clinical and biomedical setting, where terminology often differs substantially from general language. As a result, medical terms are frequently segmented into subword units that carry little or no standalone semantic or morphological meaning (Cruz Díaz and Maña López, 2015; Jimenez Gutierrez et al., 2023).

Previous work suggests that contemporary language models are relatively robust to suboptimal tokenization, eventually relying on contextual cues to recover the meaning of poorly segmented clinical terms (Jimenez Gutierrez et al., 2023; Jeong et al., 2023). However, other studies report modest improvements from more suitable segmentation (Ashfaq et al., 2025; Xie, 2026; Jeong et al., 2023). At the same time, recent findings indicate that tokenization granularity can correlate with downstream task performance, particularly in encoder-based architectures trained for domain-specific biomedical applications (Labrak et al., 2024a).

In this work, we revisit this question from a different perspective. Rather than evaluating performance on standard downstream tasks that provide rich contextual signals, we focus on the ability of models to represent medical terms in isolation. We investigate whether tokenization granularity correlates with how effectively models encode medical terminology in their embedding spaces. The emphasis is on the consistency and stability of these representations in semantic alignment and embedding-space proximity, and whether models preserve a coherent representation of a medical concept despite differences in tokenization granularity. We conduct a comparison across publicly available models, including both encoder and decoder architectures, to assess their behavior under inefficient tokenization of medical terms.

To this end, we use UMLS term–definition pairs (Bodenreider, 2004) and frame the task as definition retrieval, where a model must associate a medical term with its correct definition. This setup allows us to assess whether the embedding of a term is relatively closer to its correct definition than to semantically related but distinct ones, avoiding reliance on absolute similarity scores, which could otherwise place many medical concepts close to-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 559–571

	UMLS	WordNet
term–definition pairs	289,531	87,379
words per term	4.19	1.37
words per definition	31.05	10.41

Table 1: Dataset statistics for UMLS and WordNet, reporting number of final term–definition pairs and average lengths of terms and definitions in words.

gether and obscure meaningful differences. We then analyze how performance on this task correlates with tokenizer granularity across models. To provide a general-domain reference point for interpretation, we include a parallel evaluation based on WordNet term–definition pairs (Miller, 1995). The complete source code is available at GitHub[1].

2 Setup

2.1 Data

For our experiments, we use term–definition pairs extracted from UMLS. For comparison, we also consider a general-domain term–definition dataset derived from WordNet. To ensure comparability across models, including those primarily designed for English, we restrict the data to English terms with English definitions.

To maintain a fair evaluation setup for the definition retrieval task, we retain only one term per semantic concept. Specifically, in UMLS we select a single term per concept unique identifier (CUI), and in WordNet we select a single term per synset. To avoid ambiguity arising from multiple surface forms with different senses, we remove duplicate term occurrences.

The final datasets consist of hundreds of thousands of pairs for UMLS and tens of thousands of pairs for WordNet. Table 1 reports the number of final pairs as well as the average number of words per term and per definition. We observe that both terms and definitions in UMLS are approximately three times longer than those in WordNet.

2.2 Tokenization Granularity

To measure how well tokenization preserves word integrity, or conversely how much it fragments words into subword units, we use two metrics: average number of tokens per word (*Avg tokens per word*) and the percentage of out-of-vocabulary words (*OOV words (%)*). OOV words (%) is computed as the proportion of words that are split into more than one token by the tokenizer.

2.3 Definition Retrieval Task

The retrieval task aims to find the most relevant definition for a given term based on cosine similarity among all available definitions. To evaluate performance, we report the *Success@k* curve ($k \in 1, 10, 100, 1000, 2000$), which measures the percentage of queries where the correct definition is within the top k retrieved results.

2.4 Models

We evaluate encoder-based models and decoder large language models, covering general-purpose, biomedical, and multilingual pretraining regimes.

Encoders include general-domain models such as BERT (cased and uncased) (Devlin et al., 2019) and ModernBERT (Warner et al., 2025), medical models such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and Clinical ModernBERT (Lee et al., 2025), as well as multilingual models including mBERT, Distil mBERT (Devlin et al., 2019), mmBERT (Marone et al., 2025), and XLM-R (base and large) (Conneau et al., 2019).

Decoders include general instruction-tuned LLMs such as LLaMA3 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Phi-3.5 Mini (Abdin et al., 2024), and Gemma3 (Team et al., 2025), as well as biomedical variants such as BioMistral (Labrak et al., 2024b), Med42 (Christophe et al., 2024), MediPhi (Corbeil et al., 2025), and MedGemma (Sellergren et al., 2025).

For all models, we extract representations from the final hidden layer and apply mean pooling over token embeddings to obtain a single vector. We use this representation for both encoders and decoders to ensure consistency.

3 Results and Analysis

3.1 Tokenization Granularity Across Domains

Building on our evaluation setup, we first analyze how tokenization granularity varies across domains and model families.

We analyze UMLS terms and their definitions separately and compare them with their general-domain counterpart based on WordNet in Table 2, reporting averages across all models (Section 2.4).

The UMLS terms are consistently segmented into more fine-grained subword units than their cor-

[1] https://github.com/lanzv/medical-tokenization-granularity

	UMLS		WordNet		Full Text Tokenization				
	Terms	*Defs*	*Terms*	*Defs*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
Avg tokens per word	2.53	1.79	2.31	1.32	1.41	1.57	1.36	1.80	2.38
OOV words (%)	53.04	31.33	61.13	18.09	24.20	26.83	17.37	33.83	48.74

Table 2: Tokenization statistics averaged across all encoder and decoder models for UMLS and WordNet (terms vs. definitions) and general corpora, including BookSum chapters, Wikipedia pages, PubMed articles, emrQA clinical reports, and Stroke Discharge Summaries (SDS) from UK hospitals.

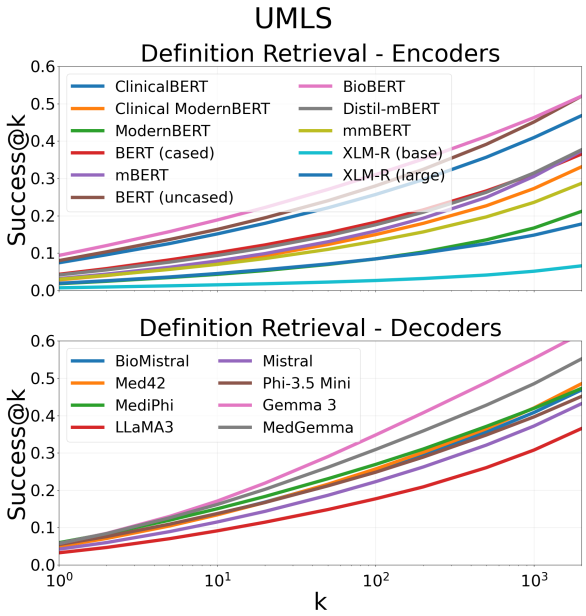

Figure 1: Definition retrieval performance on UMLS measured by Success@k. Encoder models are shown on the top and decoder models on the bottom.

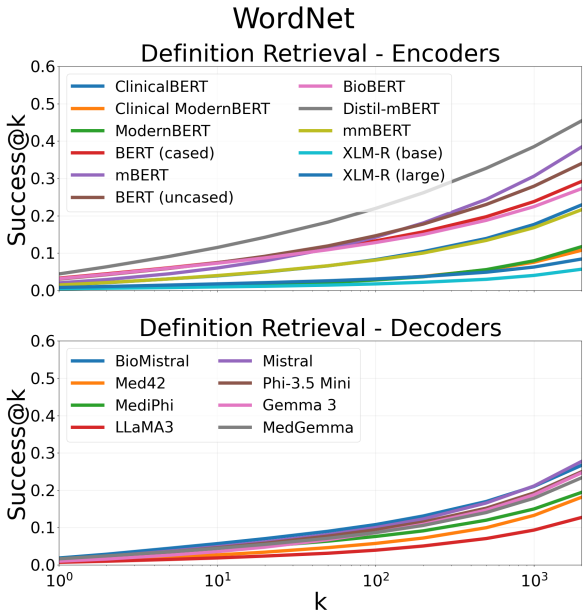

Figure 2: Definition retrieval performance on WordNet measured by Success@k. Encoder models are shown on the top and decoder models on the bottom.

responding definitions. A similar effect is observed in WordNet, indicating that lexical specificity also plays a role in general-domain resources. This consistent gap between terms and definitions suggests that isolated terminological expressions are generally more difficult to represent in subword-based tokenization schemes than more descriptive, sentence-level text, even when the latter is drawn from the same domain and contains many common lexical items. However, fragmentation is most pronounced in UMLS, suggesting higher terminological density in the medical domain.

OOV rates show a different pattern: WordNet terms have slightly higher OOV rates than UMLS. As shown in Table 1, UMLS terms are three times longer, which increases the likelihood of including vocabulary-friendly components such as numbers or stopwords, likely explaining this difference.

To assess how representative the term–definition results are in relation to larger and more practical text sources, we extend the analysis to document-level corpora. We evaluate tokenization on emrQA clinical reports (Pampari et al., 2018), BookSum chapters (Kryscinski et al., 2022), Wikipedia (Foundation), and PubMed articles (Cohan et al., 2018). To control for corpus size, we sample PubMed, BookSum, and Wikipedia documents to match the number of words in emrQA. Tokenization in BookSum and PubMed closely matches WordNet definitions, while Wikipedia shows slightly higher fragmentation. In contrast, emrQA clinical reports align more closely with UMLS definitions. We further analyze a specialized set of 600 stroke discharge summaries from UK hospitals (SDS), which shows fragmentation approaching UMLS-level values, highlighting the effect of clinical specificity on subword segmentation.

A detailed per-model breakdown is provided in Appendix A, along with qualitative tokenization examples in Appendix B.

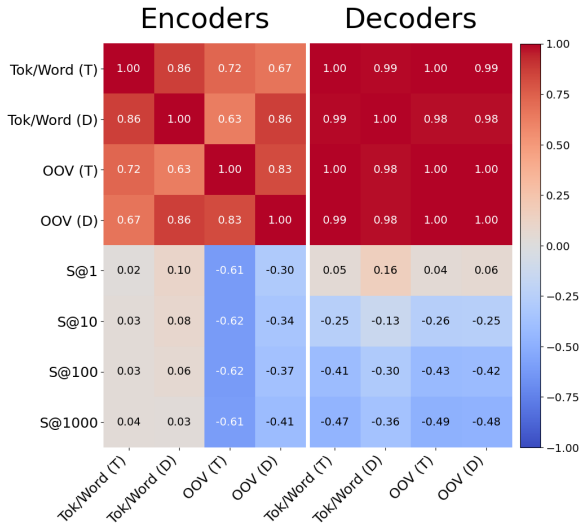

Figure 3: Pearson correlation matrices for UMLS models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (Success@k, denoted as S@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

3.2 Definition Retrieval Performance

We evaluate how well term representations align with their corresponding definitions using a definition retrieval task, which measures the robustness of the embedding space to semantic similarity of the same medical concept represented in different forms. Figures 1 and 2 report Success@k curves for encoder and decoder models on UMLS and WordNet term–definition pairs, allowing a direct comparison of models in terms of retrieval performance and the strength of their semantic alignment.

On UMLS, clinically pre-trained encoders consistently outperform general-domain and multilingual models. This advantage is less pronounced for decoder models, which nevertheless show relatively stable retrieval behavior across architectures.

In contrast, performance on WordNet decreases, despite the smaller candidate pool making the task easier. Drop is more evident for decoder models, which exhibit weaker discrimination in this setting.

These results suggest that models are well adapted to the medical domain, even though tokenization is more fragmented for UMLS compared to WordNet. This indicates a degree of robustness of models to tokenization granularity in semantic representation and alignment of concepts.

3.3 Correlation Between Tokenization and Definition Retrieval

To investigate whether tokenization granularity is related to how well models embed medical terms, we analyze UMLS pairs and measure the correlation between tokenization metrics and the ability of models to place term representations close to their correct definitions in the embedding space. We report Pearson correlations in Figure 3 for encoder and decoder models separately. This allows us to assess whether a consistent trend exists across models, i.e., whether models that split words into more subword units tend to exhibit weaker alignment between terms and their corresponding definitions.

From the results, we observe that for encoder models, the average number of tokens per word (Tok/Word) shows little to no correlation with retrieval performance. In contrast, the proportion of words represented as a single token (OOV) in both definitions and, in particular, terms exhibits a strong correlation with performance across all values of k in Success@k. For decoder models, performance begins to correlate more strongly with tokenization metrics only at higher values of k (i.e., $k = 100$ and $k = 1000$), where both Tok/Word and OOV rate become relevant.

An additional observation is that, for decoder models, tokenization metrics are almost perfectly correlated between terms and definitions, whereas encoders show greater variation.

Overall, the results indicate that tokenization granularity is reflected in how models represent medical terms in their vector space. For encoder models, retrieval performance is largely insensitive to how finely words are split once tokenized; the key factor is whether a word is preserved as a single token. For decoder models, tokenization has little effect on the proximity of term and definition representations, but its influence becomes more apparent when considering a broader neighborhood of nearby representations.

4 Discussion

Although we use definition retrieval as an intrinsic evaluation, it provides a direct and transparent way to assess how a model encodes medical concepts in its embedding space without additional training. It allows us to measure whether term embeddings are consistently and meaningfully aligned with term definitions and how effectively and consistently the model internalizes and represents these concepts.

Although tokenization granularity correlates with retrieval performance, this does not directly indicate its impact on downstream tasks. While previous work has studied such effects in downstream biomedical settings (Labrak et al., 2024a), the general significance and generality of this relationship remain unclear and might vary between tasks.

In addition, we find that English-domain-specific encoders outperform general-domain multilingual models for medical terminology, despite its strong Latin and Greek origins (Jimenez Gutierrez et al., 2023). This contrasts with the findings in span-based clinical question answering (Lanz and Pecina, 2025), where multilingual models tend to perform better. This discrepancy suggests that improvements in multilingual models in downstream question-answering tasks are likely driven by their stronger ability to model contextual interactions and relational structure, rather than by the superior quality of isolated term representations in the embedding space of encoder models. In contrast, for decoder models, no clear advantage of domain-specific pretraining is observed in terms of consistent semantic representation quality.

5 Conclusion

We studied how tokenization granularity relates to the ability of language models to represent medical terms without contextual support, using a definition retrieval task built on UMLS and a WordNet comparison. Our results show that medical terms are more fragmented than general-domain ones, yet models remain largely robust. However, tokenization still plays a role: for encoders, performance depends mainly on whether terms are preserved as single tokens, while for decoders, its effect appears at larger retrieval depths. Overall, tokenization does not fully determine performance, but it influences how medical terms are structured in the embedding space.

Limitations

Our analysis focuses on intrinsic representation properties and does not evaluate downstream task performance; thus, it does not directly capture effects previously observed in task-based studies of tokenization granularity. While definition retrieval provides a controlled proxy for assessing term–definition alignment in embedding space, it remains an indirect measure of representation quality. Finally, our experiments are limited to UMLS-derived term–definition pairs, which restricts coverage to curated medical terminology and may not reflect broader medical and clinical language.

Acknowledgments

This research was partially supported by the SVV project number 260 821 and the Charles University GAUK grant No. 284125. It has also received support and funding from the European Union's Horizon Europe research and innovation programme project *RES-Q plus* (Grant Agreement No. 101057603). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Marah Abdin and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Farzeen Ashfaq, NZ Jhanjhi, Navid Ali Khan, Danish Javed, Mehedi Masud, and Mohammad Shorfuzzaman. 2025. Enhancing ecg report generation with domain-specific tokenization for improved medical nlp accuracy. *IEEE Access*, 13:85493–85506.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *Preprint*, arXiv:2408.06142.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria. Association for Computational Linguistics.

Noa P. Cruz Díaz and Manuel Maña López. 2015. An analysis of biomedical tokenization: Problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 40–49, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Geunyeong Jeong, Juoh Sun, Seokwon Jeong, Hyunjin Shin, and Harksoo Kim. 2023. Improving automatic KCD coding: Introducing the KoDAK and an optimized tokenization method for Korean clinical documents. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 96–101, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang and 1 others. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bernal Jimenez Gutierrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to sub-optimal tokenization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 350–362, Toronto, Canada. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Mickael Rouvier, and Richard Dufour. 2024a. How important is tokenization in French medical masked language models? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8223–8234, Torino, Italia. ELRA and ICCL.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024b. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, {Chan Ho} So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. Publisher Copyright: © 2020 Oxford University Press. All rights reserved.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *Preprint*, arXiv:2504.03964.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Andrew Sellergren and 1 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Gemma Team and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and

Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Wenran Xie. 2026. Optimizing biomedical text processing: A comparative analysis of tokenization methods and context-aware representation learning. In *Advanced Data Mining and Applications*, pages 148–160, Singapore. Springer Nature Singapore.

A Tokenization Granularity

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
ClinicalBERT	2.52	1.82	2.44	1.32	1.36	1.58	1.34	1.63	2.18
Clinical ModernBERT	2.34	1.60	2.39	1.26	1.42	1.52	1.22	1.70	2.34
ModernBERT	2.34	1.60	2.39	1.26	1.42	1.52	1.22	1.70	2.34
BERT (cased)	2.80	1.86	2.36	1.27	1.34	1.47	1.34	1.84	2.39
mBERT	2.66	1.87	2.35	1.37	1.43	1.47	1.37	1.82	2.35
BERT (uncased)	2.33	1.73	2.17	1.23	1.31	1.41	1.27	1.55	2.07
BioBERT	2.52	1.82	2.44	1.32	1.36	1.58	1.34	1.63	2.18
Distil-mBERT	2.66	1.87	2.35	1.37	1.43	1.47	1.37	1.82	2.35
mmBERT	2.06	1.60	1.78	1.17	1.35	1.51	1.30	1.77	2.38
XLM-R (base)	2.58	1.88	2.35	1.46	1.44	1.58	1.46	1.75	2.15
XLM-R (large)	2.58	1.88	2.35	1.46	1.44	1.58	1.46	1.75	2.15

Table 3: Average tokens per word across encoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum chapters, Wikipedia pages, PubMed articles, emrQA clinical notes, and Stroke Discharge Summaries (SDS).

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
ClinicalBERT	44.13	29.83	58.65	16.47	22.68	27.52	15.97	30.40	45.11
Clinical ModernBERT	50.33	25.03	66.50	15.12	23.13	24.45	10.55	29.21	47.71
ModernBERT	50.33	25.03	66.50	15.12	23.13	24.45	10.55	29.21	47.71
BERT (cased)	56.25	30.63	55.07	14.11	22.14	23.43	15.97	34.15	50.52
mBERT	58.23	34.87	63.10	21.28	27.02	26.22	19.84	36.73	52.35
BERT (uncased)	40.18	27.96	49.73	12.35	21.26	21.62	13.12	27.18	41.66
BioBERT	44.13	29.83	58.65	16.47	22.68	27.52	15.97	30.40	45.11
Distil-mBERT	58.23	34.87	63.10	21.28	27.02	26.22	19.84	36.73	52.35
mmBERT	37.18	23.64	41.00	9.48	19.92	21.34	10.04	25.54	40.60
XLM-R (base)	63.41	40.81	67.65	29.49	28.72	34.35	33.41	44.43	54.59
XLM-R (large)	63.41	40.81	67.65	29.49	28.72	34.35	33.41	44.43	54.59

Table 4: OOV words (%) across encoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum, Wikipedia, PubMed, emrQA, and Stroke Discharge Summaries (SDS).

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
BioMistral	2.87	1.95	2.42	1.39	1.49	1.75	1.49	2.06	2.77
Med42	2.35	1.65	2.31	1.25	1.36	1.49	1.28	1.67	2.09
MediPhi	2.94	2.01	2.50	1.45	1.53	1.77	1.54	2.13	2.85
LLaMA3	2.35	1.65	2.31	1.25	1.36	1.49	1.28	1.67	2.09
Mistral	2.87	1.95	2.42	1.39	1.49	1.75	1.49	2.06	2.77
Phi-3.5 Mini	2.94	2.01	2.50	1.45	1.53	1.77	1.54	2.13	2.85
Gemma 3	2.14	1.60	2.07	1.19	1.37	1.52	1.29	1.76	2.42
MedGemma	2.14	1.60	2.07	1.19	1.37	1.52	1.29	1.76	2.42

Table 5: Average tokens per word across decoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum, Wikipedia, PubMed, emrQA clinical notes, and Stroke Discharge Summaries (SDS).

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
BioMistral	63.30	36.01	64.01	21.20	25.90	30.64	20.30	38.54	54.91
Med42	48.98	27.69	63.95	14.57	22.29	24.07	12.12	29.80	42.14
MediPhi	67.29	39.03	68.78	25.28	27.52	32.70	23.59	41.94	58.03
LLaMA3	48.98	27.70	63.95	14.56	22.29	24.07	12.12	29.80	42.14
Mistral	63.30	36.01	64.01	21.20	25.90	30.64	20.30	38.54	54.91
Phi-3.5 Mini	67.29	39.03	68.78	25.28	27.52	32.70	23.59	41.94	58.03
Gemma 3	41.41	23.22	55.20	10.49	20.98	21.73	9.71	26.88	41.78
MedGemma	41.41	23.22	55.20	10.49	20.98	21.73	9.71	26.88	41.78

Table 6: OOV words (%) across decoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum, Wikipedia, PubMed, emrQA, and Stroke Discharge Summaries (SDS).

B Qualitative Tokenization Examples

Model	Infertility	Spectroscopy	Neurochemistry	Excipients	Bacteroides
bert-base-cased	In-fer-tility	S-pect-ros-copy	N-eur-och-em-ist-ry	Ex-ci-pie-nts	Ba-cter-oid-es
bert-base-multilingual-cased	Inf-erti-lity	Sp-ect-ros-co-py	Neu-roch-emi-str-y	Ex-ci-pien-ts	Ba-cter-oides
bert-base-uncased	in-fer-tility	spectroscopy	ne-uro-chemist-ry	ex-ci-pie-nts	ba-cter-oides
Bio_ClinicalBERT	in-fer-tility	s-pect-ros-copy	ne-uro-chemistry	ex-ci-pie-nts	b-act-ero-ides
biobert-base-cased-v1.2	in-fer-tility	s-pect-ros-copy	ne-uro-chemistry	ex-ci-pie-nts	b-act-ero-ides
Clinical_ModernBERT	Inf-ertility	Spect-rosc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-acter-oides
ModernBERT-base	Inf-ertility	Spect-rosc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-acter-oides
distilbert-base-multilingual-cased	Inf-erti-lity	Sp-ect-ros-co-py	Neu-roch-emi-str-y	Ex-ci-pien-ts	Ba-cter-oides
mmBERT-base	Infer-tility	Spectroscopy	Neuro-chemistry	Ex-cip-ients	Bacter-oides
xlm-roberta-base	Inf-ert-ility	Spec-tros-copy	Neuro-che-mist-ry	Exc-ipi-ents	Bac-tero-ides
xlm-roberta-large	Inf-ert-ility	Spec-tros-copy	Neuro-che-mist-ry	Exc-ipi-ents	Bac-tero-ides
BioMistral-7B	In-fer-t-ility	Spect-ro-sc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-acter-oid-es
gemma-3-4b-pt	In-fert-ility	Spect-roscopy	Neuro-chemistry	Ex-cip-ients	B-acter-oides
Llama3-Med42-8B	Inf-ertility	S-pect-ro-scopy	Ne-uro-chemistry	Exc-ipients	B-acter-oid-es
medgemma-4b-it	In-fert-ility	Spect-roscopy	Neuro-chemistry	Ex-cip-ients	B-acter-oides
MediPhi-Instruct	In-fer-til-ity	Spect-ro-sc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-act-ero-ides
Meta-Llama-3-8B	Inf-ertility	S-pect-ro-scopy	Ne-uro-chemistry	Exc-ipients	B-acter-oid-es
Mistral-7B-Instruct-v0.1	In-fer-t-ility	Spect-ro-sc-opy	Ne-uro-chemistry	Ex-cip-ients	B-acter-oid-es
Phi-3.5-mini-instruct	In-fer-til-ity	Spect-ro-sc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-act-ero-ides

Table 7: Qualitative examples of tokenization across encoder and decoder models. Medical terms are frequently fragmented into subword units that do not necessarily correspond to medically meaningful morphemes.

C Definition Retrieval - Mean Reciprocal Rank

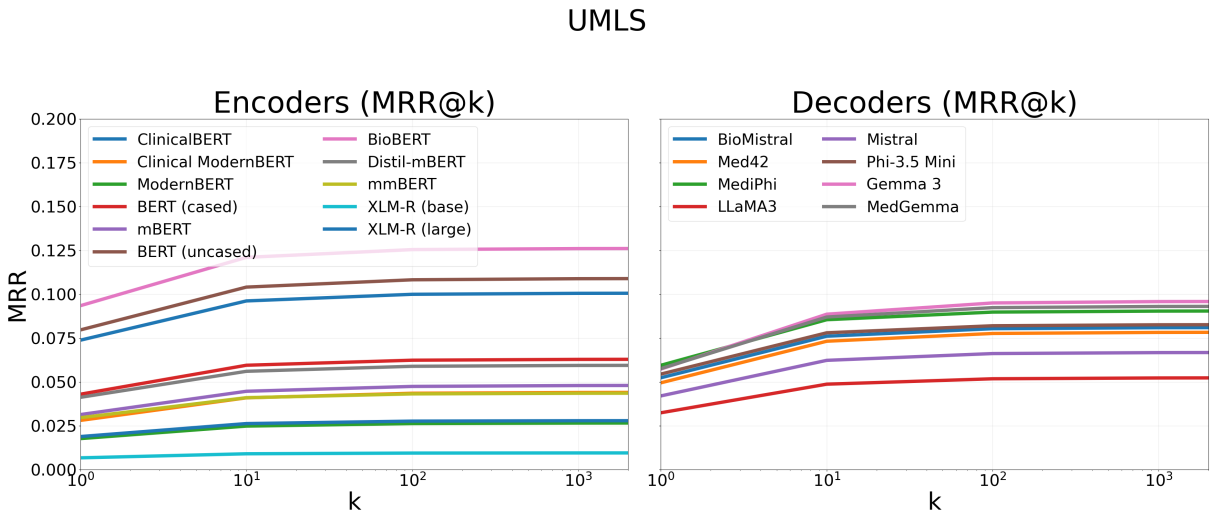

Figure 4: Definition retrieval performance on UMLS measured by MRR@k. Encoder models are shown on the left and decoder models on the right.

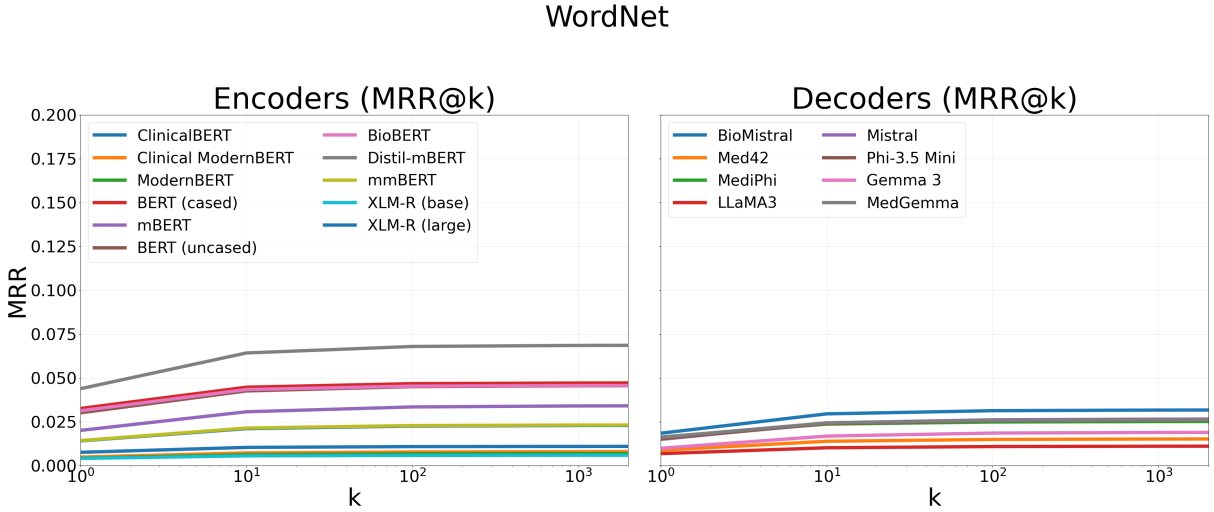

Figure 5: Definition retrieval performance on WordNet measured by MRR@k. Encoder models are shown on the left and decoder models on the right.

D Correlation Between Tokenization Granularity and Definition Retrieval

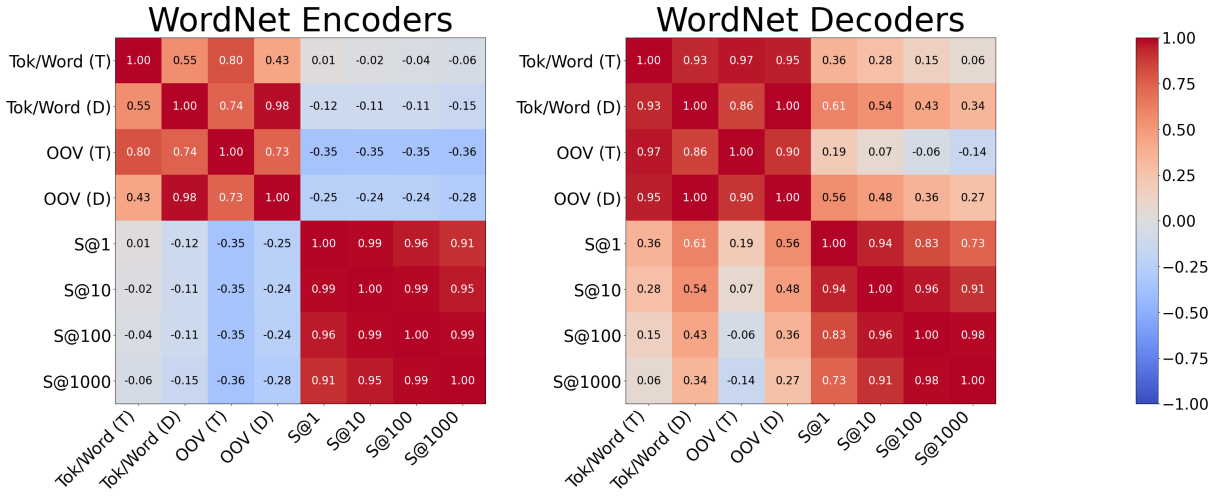

Figure 6: Pearson correlation matrices for WordNet models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (Success@k, denoted as S@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

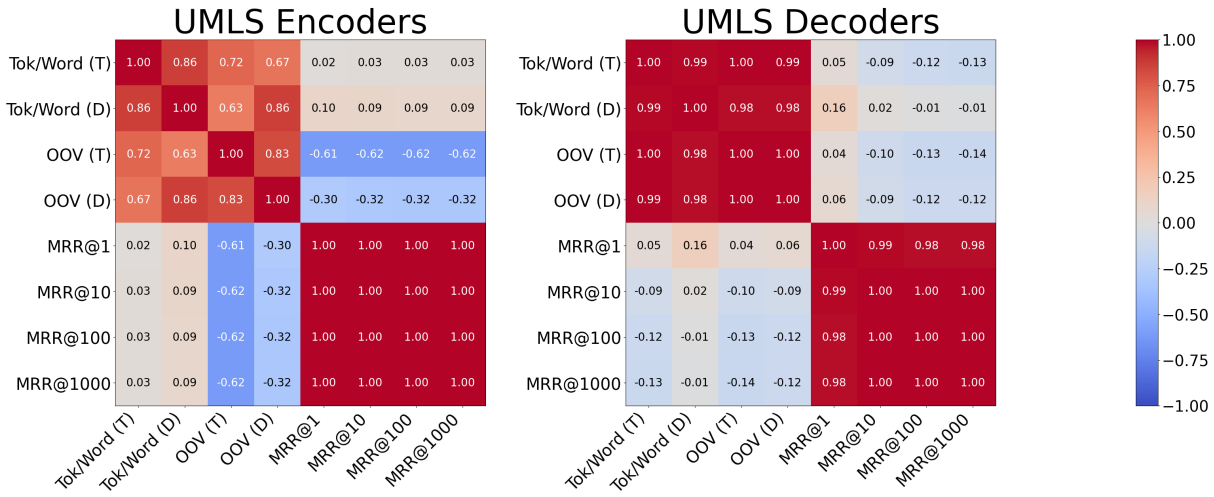

Figure 7: Pearson correlation matrices for UMLS models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (MRR@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

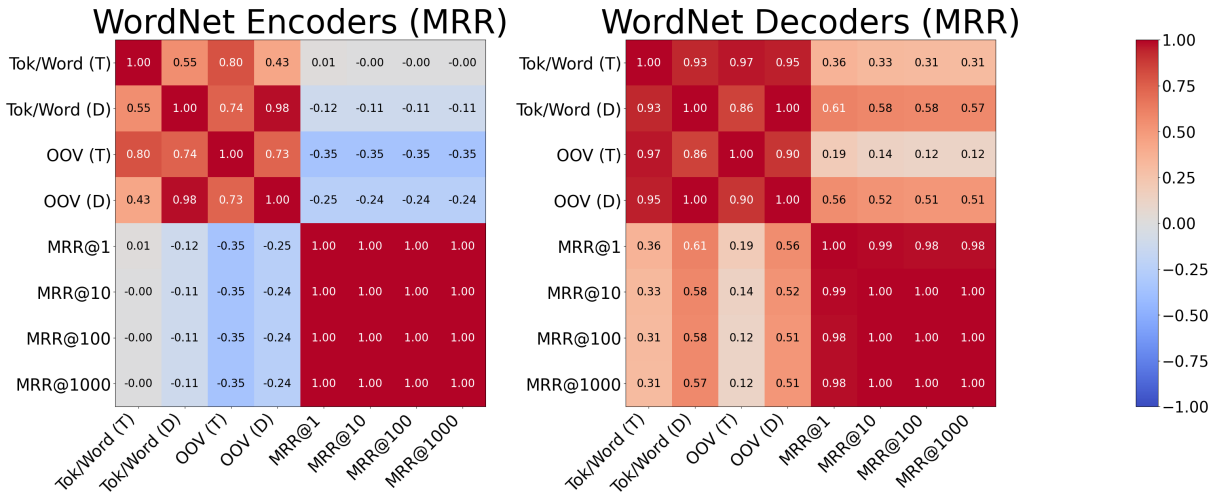

Figure 8: Pearson correlation matrices for WordNet models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (MRR@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).