Toward Patient Education with NLP: Readability and Informativeness in Clinical Summarization

Pinaki Mohanty, Alejandro Sandoval, Shaswat Shukla *

Due: May 10, 2025

1 Introduction

The transformative progress in Natural Language Processing (NLP) has unlocked unprecedented potential for improving healthcare communication, particularly through the use of large transformer-based models like T5 [RSR+20], BART [LLG+20], and Pegasus [ZZSL20]. Given the abundance of unstructured textual data in clinical settings—such as discharge summaries, physician notes, and patient instructions—there is an urgent need to translate this content into formats that are understandable and actionable for patients. This challenge has elevated the role of medical text summarization, especially for tasks related to patient education and continuity of care [oHS11].

Crucially, low health literacy is a well-documented public health concern. In the United States, nearly 9 out of 10 adults struggle to understand and use health information that is routinely available in clinical settings [KGJP06, Net24]. These difficulties are not evenly distributed: Black, Hispanic, and low-income populations, as well as individuals without high school diplomas, are disproportionately affected by low health literacy, contributing to delayed diagnoses, poorer treatment adherence, and higher rates of preventable hospitalizations [BSD+11, RMB+05]. Furthermore, socioeconomic disparities contribute to health inequalities. Individuals with lower educational attainment are more likely to engage in risky health behaviors and have limited access to healthcare resources, exacerbating the prevalence of chronic diseases in disadvantaged communities [Wik23].

While general-domain summarization has witnessed significant advances, adapting these methods to the medical domain introduces unique challenges. Clinical texts are replete with specialized jargon, structurally heterogeneous, and often omit explicit discourse markers that facilitate content condensation [JX24]. Moreover, conventional metrics like ROUGE, despite their widespread adoption, exhibit weak correlation with human comprehension and judgment in high-stakes domains like healthcare [TSI+23, LL08a]. As a result, there's increasing consensus that models for medical summarization should be evaluated not only on informativeness and conciseness, but also on readability, factual accuracy, and patient utility [KvD75, TR24].

In this project, we investigate how modern transformer-based models can be leveraged to generate patient-friendly summaries from complex clinical narratives. Specifically, we benchmark general-purpose models on datasets curated for patient education, incorporating readability and accessibility metrics to evaluate how well these summaries serve a lay audience. By anchoring the task in patient education, our work contributes toward the broader goal of equitable and trustworthy medical AI systems that empower patients through inclusive and actionable information delivery [PPTC17, ELK+12].

2 Related Works

2.1 Early Approaches and Evaluation Numerous studies have been carried out on summarization of medical documents with the aim of simplifying complex clinical information. Initially, extractive techniques like TF-IDF and TextRank were applied, which purely used identification without construction in obtaining key phrases from the text [MT04]. However, these models do not incorporate

*Names presented in alphabetical order of last name.

contextual understanding, which is why T5 [RSR+20], BART [LLG+20], and Pegasus [ZZSL20] developed abstractive summarization models that transform the original content into more fluent summaries.

2.2 Clinical Evaluation Evaluating these performances poses unique challenges. One method comes from ClinicBench, who developed a benchmark for radiology report summarization and hospitalization summarization among others [SCK+24]. Another non summarization but domain specific approach comes from the Healthcare Prompt-a-thon [SCK+24], which focuses on expert clinical evaluation and error analysis in emergency department notes to evaluate their quality. These approaches highlight the need for evaluation standards to improve the reliability, safety, and interpretability of summarization in healthcare. However, they are not focused on summarization but rather the use of LLMs in medical environments. This goes to show how little work has been done on the task of summarization for healthcare.

2.3 Evaluation Metrics ROUGE scores continue to be a popular choice for summarization evaluation due to their straightforward calculation processes and word overlap reflection [Lin04], but more work has drawn attention to their shortcomings in clinical and patient-centered studies [WOD+23]. These metrics ignore crucial criteria such as summarization readability, factual accuracy, and clinical relevance pertaining to the patients that profoundly influence the patient education process about their condition. Previous work found that in many cases users did not prefer or comprehend summarizations that scored highly on ROUGE metrics—this was especially true for meeting summarizations which included numerous disfluent, incoherent, and repetitive phrases. ROUGE is also inadequate in measuring sophisticated elements of human evaluation such as dialogue progression and contextual relevance to the speaker [LL08b]. Consequently, other metrics have been adopted to compute sum readability such as Flesch-Kincaid, expert judgement, and some other newer ones like the coverage content ratio [KFRC75]. Our project falls within this approach, allowing us to bridge objective measurement and patient-centered perception of summary quality.

3 Methods

Our project investigates the adaptation of transformer-based summarization models—T5-small (60M parameters), BART-base (140M parameters), and PEGASUS-XSUM (568M parameters)—to the domain of patient education, evaluating their performance across both real-world and semi-synthetic clinical datasets. We focus on three datasets: (1) **Hospitalization Summarization**, consisting of 382 de-identified discharge notes derived from MIMIC-IV [JBS+23]; (2) **Patient Education**, comprising 3,621 lay-oriented instruction-summary pairs extracted from MIMIC-III [JPS+16]; and (3) **Augmented Clinical Notes**, a semi-synthetic 30K-instance dataset from Hugging Face [Bon23].

To maintain data privacy and reduce token noise, input texts are preprocessed to remove placeholder tokens (e.g., `[name]`, `[hospital]`), which are common in clinical corpora but contribute little semantically. For T5 models, we prepend each input with the prompt `summarize:`, while BART and PEGASUS use the raw text. All models generate summaries via beam search (num_beams = 4), with maximum input length set to 512 tokens and output capped at 150 tokens. No fine-tuning is performed; this zero-shot setup allows us to assess how well pretrained summarizers generalize to low-resource medical summarization tasks [YC22].

We evaluate performance along two axes: (1) content fidelity and (2) readability. Content metrics include ROUGE-1/2/L [Lin04], BERTScore F1 [ZKW+20], and coverage ratio, a measure of similarity between generated summary and gold summary . Readability is assessed using Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL) [KFRC75], and Character-Per-Token (CPT), a custom proxy for lexical complexity [TR24]. To better capture abstraction and completeness, we perform dual evaluation: generated summaries are compared both against gold references and the original long-form inputs [NSK23].

All code is implemented in Python using Hugging Face's `transformers`, `evaluate`, `textstat`, and standard NLP libraries like `NumPy` and `pandas`. Tokenization is handled via `AutoTokenizer` for T5 and BART, while PEGASUS requires the `PegasusTokenizer` and an external `sentencepiece` installation due to its custom subword vocabulary.

4 Results

Results are averaged across the full dataset for Hospitalization and Patient Education, and Augmented Notes and reported with standard deviation in Tables 1, 2. Due to computational constraints, we report runtime only for the Hospitalization dataset: total inference times were 964.61 seconds for T5-small, 1107.78 seconds for PEGASUS, and 1643.77 seconds for BART.

Dataset	Model	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	Coverage ↑	BERTScore ↑
Hospital	T5-small	0.1840 ± 0.0752	0.0770 ± 0.0583	0.1390 ± 0.0608	**0.9760 ± 0.0292**	**0.8280 ± 0.0218**
	PEGASUS	0.1030 ± 0.0666	0.0420 ± 0.0387	0.0810 ± 0.0511	0.6840 ± 0.2485	0.8240 ± 0.0268
	BART	**0.2570 ± 0.0648**	**0.1020 ± 0.0492**	**0.1490 ± 0.0426**	0.9470 ± 0.0230	0.8070 ± 0.0140
Patient	T5-small	0.0330 ± 0.0188	0.0060 ± 0.0069	0.0210 ± 0.0111	0.8770 ± 0.0659	0.7780 ± 0.0152
	PEGASUS	0.0110 ± 0.0094	0.0020 ± 0.0029	0.0090 ± 0.0070	0.5320 ± 0.2685	0.7710 ± 0.0137
	BART	**0.0600 ± 0.0305**	**0.0110 ± 0.0108**	**0.0330 ± 0.0155**	**0.9870 ± 0.0215**	**0.7840 ± 0.0157**
Augmented	T5-small	0.1630 ± 0.0457	0.0960 ± 0.0386	0.1160 ± 0.0369	0.9720 ± 0.0261	**0.7960 ± 0.0128**
	PEGASUS	0.0520 ± 0.0208	0.0170 ± 0.0148	0.0390 ± 0.0158	0.6950 ± 0.1189	0.7830 ± 0.0120
	BART	**0.3050 ± 0.0465**	**0.1830 ± 0.0467**	**0.1970 ± 0.0465**	**0.9840 ± 0.0180**	**0.7960 ± 0.0123**

Table 1: Semantic fidelity metrics across three datasets. Bolded values represent best-performing models per metric per dataset.

Dataset	Model	FRES ↑	FKGL ↓	CPT ↓
Hospital	T5-small	**42.9440 ± 23.7590**	**11.4540 ± 4.4351**	**5.3780 ± 0.8238**
	PEGASUS	36.7800 ± 41.4667	12.3350 ± 8.6217	5.5890 ± 1.3490
	BART	26.2570 ± 16.1148	16.0470 ± 4.5818	5.6670 ± 0.4033
Patient	T5-small	57.8560 ± 16.9270	8.7760 ± 3.4134	4.9280 ± 0.6426
	PEGASUS	**65.1840 ± 32.3905**	**7.7280 ± 5.9976**	4.4680 ± 0.9599
	BART	54.4250 ± 19.5296	10.3660 ± 6.1471	**4.3280 ± 0.4003**
Augmented	T5-small	39.3730 ± 17.2199	11.8510 ± 2.9601	5.4550 ± 0.5351
	PEGASUS	**47.9800 ± 19.8876**	**11.0650 ± 3.2769**	**4.9230 ± 0.6609**
	BART	36.8040 ± 13.0327	12.5980 ± 2.4175	5.4430 ± 0.3929

Table 2: Readability metrics across three datasets. Bolded values represent best-performing models per metric per dataset.

The pictorial description of the results that we see in the two tables can be found in Figure 1.

5 Analysis

The results across all three datasets show us the clear superiority of BART over T5-small and PEGASUS in semantic fidelity. Of all the considered models, BART was noted to consistently score the highest on all fidelity metrics. These results indicate that BART is highly effective at capturing clinically relevant information and aligning closely with the content of the gold summaries. Its best performance occurred on the Augmented Notes dataset where it achieved ROUGE-1 equals to 0.305, showing its ability to create summaries which retain both lexical and semantic similarity to the reference. This could show that BARTs denoising pre-training enables more accurate content preservation, which makes it well suited for tasks which require fidelity to the original references such as patient education. On the other hand, T5 showed much worse results in regards to ROUGE socres. However, it performed almost as well and ocassionally better than BART on Coverage and BERTScores. PEGASUS performed considerably worse than both T5 and BART on every single fidelity metric, showing ROUGE scores which were fractions of what BART obtained.

While it had a very strong performance on semantic fidelity, BART underperformed both T5 and PEGASUS on nearly all readability metrics which might highlight a trade-off in clinical summarization. PEGASUS, on the other hand, consistently produced the most readable outputs. On the Patient Education and Augmented Notes datasets it overperformed both other models, on Hospitalization it still performed almost as well as T5. T5 was the best model for the Hospitalization dataset and performed well on the other two, better than BART but not as well as PEGASUS. Interestingly, this seems to suggest that dataset choice is more relevant than metric choice. No matter the metric, when using the Patient Education and Augmented Notes dataset PEGASUS performed best, the same is

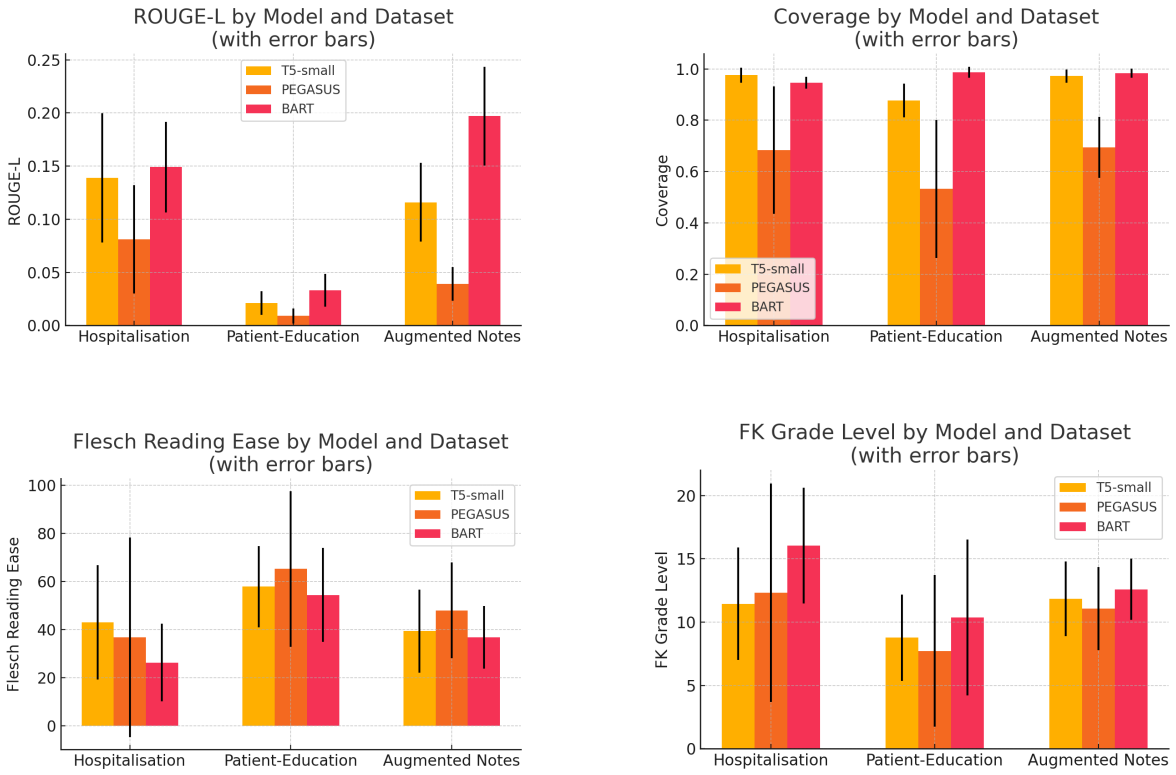

Figure 1: Key semantic fidelity and readability metrics across three datasets for all three models.

true for T5 and Hospitalization. It would seem that the concise structure of the Hospitalization dataset worked in favor of T5 while the more diverse datasets highlighted PEGASUS' strength in simplification.

It is also interesting to see the differences in both semantic fidelity and readability scores based on the dataset used rather than just the model. We can see that on average models scored best on the Hospitalization dataset for semantic fidelity, very closely followed by the Augmented Notes dataset. Patient Education, on the other hand, gave us much worse results. This seems to suggest that models were better at summarizing more structured, formal medical records. For readability scores we can see that the opposite happens, Patient Education comes out on top in every readability metric with Augmented Notes and Hospitalization Summarization falling behind. This is probably due to the datasets intention to communicate ideas to a patient with little medical knowledge which results in simpler language. Overall, there seems to be a inverse relationship between readability and semantic fidelity as can be seen in Figure 2. In order to more easily compare scores along different metrics we normalized the scores. We did this by calculating $\frac{x-x_{min}}{x_{max}-x_{min}}$ for scores that ascended as the input improved and $\frac{x_{max}-x}{x_{max}-x_{min}}$ for scores that seeked lower values. When performing a regression on these normalized scores for readability and fidelity we can see a strong R^2 score of 0.685, suggesting this inverse relationship holds.

PEGASUS exhibits the highest variance across our evaluations—a pattern traceable to its pre-training objective and data. Specifically, PEGASUS-XSUM was trained on the XSUM dataset, which consists of single-sentence news summaries designed to capture the main idea in a headline-like format. This encourages concise, abstractive outputs but also introduces variability when applied to longer, unstructured clinical narratives. In our zero-shot setting, where no domain-specific fine-tuning is applied, PEGASUS often produces readable but brief summaries that omit clinical nuance and fluctuate in coverage and content selection. While this behavior makes PEGASUS well-suited for plug-and-play use in low-resource environments, its full potential likely requires fine-tuning on medical corpora to ensure semantic fidelity (see Figure 3, Figure 4).

Despite achieving high semantic fidelity, many of the generated summaries—particularly those produced by BART—consistently scored above the recommended readability thresholds, confirming

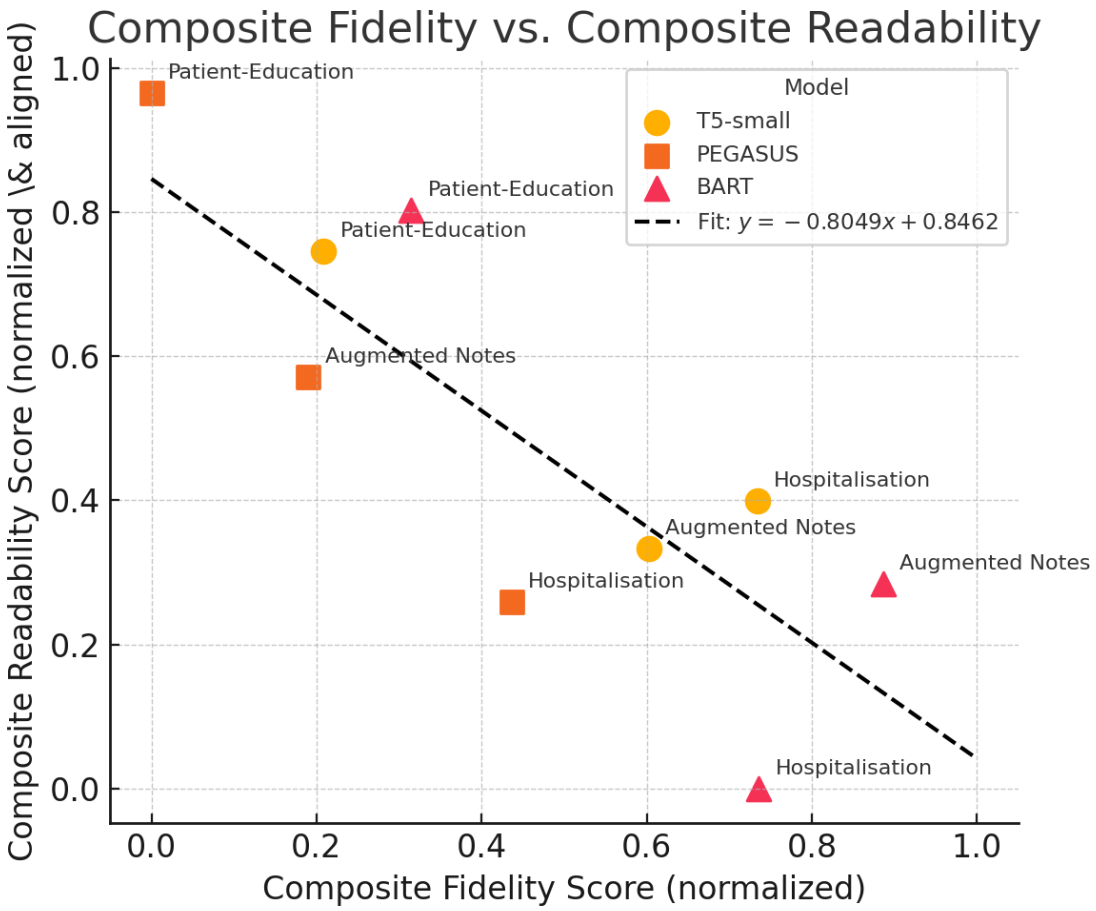

Figure 2: Composite fidelity vs. readability scores for all models and datasets.

our concern that such outputs may be inaccessible to the intended patient populations. National health agencies such as the CDC and NIH recommend that patient-facing materials be written at a 6th to 8th grade reading level to accommodate the average American adult, who typically reads at a 7th to 8th grade level [KGJP06, Net24]. However, across all datasets, our summaries frequently exceeded these thresholds. For example, BART's summaries on the hospitalization dataset reached a Flesch-Kincaid Grade Level of 16.0 and a Character-Per-Token average exceeding 5.6—indicative of lexically dense, complex language unsuitable for low-literacy audiences. This is evident from summary presented in Figure 4. Words like stenosis, progressive, radiating may not be perceived patient friendly. This is further problematic given the disproportionately low health literacy found among older adults, Black populations, rural communities in Appalachia, and Native American groups [BSD+11, WLR+13]. These groups often face not only limited educational attainment but also higher rates of chronic illness, making accessible aftercare information essential.

Despite focusing on a comparative evaluation, our work revealed critical challenges that underscore the complexity of clinical summarization. First, compute constraints limited our ability to experiment extensively across larger or more diverse corpora. Model inference, even in zero-shot settings, was time-intensive (e.g., over 1600 seconds for BART on the hospitalization dataset), and fine-tuning was not feasible in our setup. Second, access to high-quality clinical datasets remains a significant bottleneck. Many real-world datasets are locked behind strict access controls (e.g., PhysioNet credentialing), making it difficult to conduct reproducible or scalable research in this domain [NKA+21, Stu23]. Most importantly, we emphasize that clinical summarization should not be viewed solely as a technical task. The intersection of health literacy, sociocultural factors, and patient comprehension is essential. Populations with limited educational attainment may struggle to understand outputs generated by black-box systems. Our study reinforces that AI outputs must be contextually aware, and that interdisciplinary collaboration—with experts in sociology, public health, health communication, and ethics—is essential

Example Summarization Output

Input (truncated): A 31-year-old male with no significant past medical history who presented from an outlier hospital to our emergency department with an acute onset crushing substernal chest pressure and dyspnea with p...

Gold Summary (truncated): { "visit motivation": "Acute onset crushing substernal chest pressure and dyspnea with profuse sweating",
"admission": [{ "reason": "Acute onset crus... } }

Generated Summary : We report the case of a patient who suffered cardiac arrest while undergoing aortic valve replacement surgery.

Figure 3: Example comparison between input note, structured gold summary, and generated output for pegasus-xsum on augmented dataset

Example Summarization Output (Cycle Note)

Input (truncated): Input: age: 55. gender: M. Name: ...Service: ORTHOPAEDICS. Allergies: Ibuprofen / Penicillins / Sulfa ...Patient admitted with rapidly progressive back and radiating leg pain ...

Gold Summary (truncated): The patient is a 55-year-old male who was admitted to the hospital with a rapidly progressive syndrome of back and radiating leg pain, progressive lower extremity weakness ...

Generated Summary: The patient presented with a rapidly progressive syndrome of back and radiating leg pain with progressive lower extremity weakness. The progressive nature of his neurological deficit, the severity of his stenosis, and the progressive loss of the ability to walk. The patient was transferred to the PACU in a stable condition.

Figure 4: Example comparison between input note, structured gold summary, and generated output for t5-small on hospitalization dataset.

to design interventions that are equitable, not just efficient [KCZT10, SPG+03, SBC+21].

To enhance readability without sacrificing fidelity, future work could explore *controlled generation*—guiding models to prefer simpler tokens during decoding [KNWS16, DML+20]. *Lexical simplification* is another path, where complex phrases are replaced with clearer ones (e.g., substituting "microcytic anemia" with "too few healthy blood cells") while maintaining semantic alignment [Sha14, DMWL21]. A practical alternative is *ontology-guided paraphrasing*, which maps medical terms to patient-friendly equivalents using resources like UMLS or SNOMED [KCZT10, ZZG20]. Such strategies are essential to ensure medical AI tools are both accurate and accessible, particularly for low-literacy populations.

References

[Bon23] Antoine Bonnet. Augmented clinical notes dataset. https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes, 2023. Accessed: 2025-04-28.

[BSD+11] Nancy D. Berkman, Stacey L. Sheridan, Katrina E. Donahue, David J. Halpern, and Karen Crotty. Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine*, 155(2):97–107, 2011.

[DML+20] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinki, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR)*, 2020.

[DMWL21] Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online, June 2021. Association for Computational Linguistics.

[ELK+12] Jean Anderson Eloy, Sharon Li, Kavita Kasabwala, Nitin Agarwal, David R Hansberry, Steven Baredes, and Michael Setzen. Readability assessment of patient education materials on major otolaryngology association websites. *Otolaryngology–Head and Neck Surgery*, 147(5):848–854, 2012.

[JBS+23] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–11, 2023.

[JPS+16] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[JX24] Chao Jiang and Wei Xu. Medreadme: A systematic study for fine-grained sentence readability in medical domain, 2024.

[KCZT10] S Kandula, D Curtis, and Q Zeng-Treitler. A semantic and syntactic text simplification tool for health content. In *Proceedings of the AMIA Annual Symposium*, pages 454–463. American Medical Informatics Association, 2010.

[KFRC75] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[KGJP06] Mark Kutner, Elizabeth Greenberg, Ying Jin, and Christine Paulsen. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. Technical Report NCES 2006-483, National Center for Education Statistics, Washington, DC, 2006.

[KNWS16] Yuta Kikuchi, Graham Neubig, Tetsuji Watanabe, and Eiichiro Sumita. Controlling output length in neural encoder-decoders. In *Proceedings of EMNLP*, pages 1328–1338, 2016.

[KvD75] Walter Kintsch and Teun A. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394, 1975.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[LL08a] Chin-Yew Lin and Yang Liu. A survey of automatic text summarization. *Handbook of Natural Language Processing*, 2008.

[LL08b] Feifan Liu and Yang Liu. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, pages 201–204, Columbus, Ohio, USA, 2008. Association for Computational Linguistics.

[LLG+20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, 2020.

[MT04] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of EMNLP*, 2004.

[Net24] Network of the National Library of Medicine. Introduction to health literacy, 2024. Accessed: 2025-04-28.

[NKA+21] Varun Nair, Namit Katariya, Xavier Amatriain, Ilya Valmianski, and Anitha Kannan. Adding more data does not always help: A study in medical conversation summarization with pegasus. In *Proceedings of the Machine Learning for Health (ML4H) Workshop at NeurIPS 2021*, 2021.

[NSK23] Varun Nair, Elliot Schumacher, and Anitha Kannan. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models, 2023.

[oHS11] U.S. Department of Health and Human Services. Health literacy online: A guide to writing and designing easy-to-use health web sites. https://health.gov/healthliteracyonline/, 2011.

[PPTC17] Timothy E Paterick, Nachiket Patel, A Jamil Tajik, and Krishnaswamy Chandrasekaran. Improving health outcomes through patient education and partnerships with patients. *Proceedings (Baylor University Medical Center)*, 30(1):112–113, 2017.

[RMB+05] Russell L. Rothman, Richele Malone, Bonita Bryant, Christopher Wolfe, Patricia Padgett, Darren A. DeWalt, Morris Weinberger, and Michael Pignone. The spoken knowledge in low literacy in diabetes scale: A diabetes knowledge scale for vulnerable patients. *The Diabetes Educator*, 31(2):215–224, Mar–Apr 2005.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[SBC+21] Dean Schillinger, Renu Balyan, Scott A Crossley, Danielle S McNamara, Jenny Y Liu, and Andrew J Karter. Precision communication: physicians' linguistic adaptation to patients' health literacy. *Science Advances*, 7(51):eabj4760, 2021.

[SCK+24] Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, and Edward Choi. Evaluation framework of large language models in medical documentation: Development and usability study. *Journal of Medical Internet Research*, 2024.

[Sha14] Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70, 2014.

[SPG+03] Dean Schillinger, John Piette, Kevin Grumbach, Frances Wang, Carl Wilson, Carina Daher, Karen Leong-Grotz, Carlos Castro, and Andrew B Bindman. Closing the loop: physician communication with diabetic patients who have low health literacy. *Archives of internal medicine*, 163(1):83–90, 2003.

[Stu23] Stanford CS224N Students. Clinical text summarization with llm-based evaluation, 2023. Available at https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1246/final-reports/256989380.pdf.

[TR24] Sean Trott and Pamela D. Rivière. Measuring and modifying the readability of english texts with gpt-4, 2024.

[TSI+23] Lianmin Tang, Zhen Sun, Bryan Idnay, John G. Nestor, Arman Soroush, Paul A. Elias, Zhihao Xu, Yujia Ding, Greg Durrett, Jean-François Rousseau, Chunhua Weng, and Yifan Peng. Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6(1):158, 2023.

[Wik23] Wikipedia contributors. Inequality in disease, 2023.

[WLR+13] Cyprian Wejnert, Binh Le, Charles E. Rose, Alexandra M. Oster, Amanda J. Smith, Jianmin Zhu, Gabriela Paz-Bailey, and NHBS Study Group. Hiv infection and awareness among men who have sex with men — 20 cities, united states, 2008 and 2011. *PLoS ONE*, 8(10):e76878, Oct 2013.

[WOD+23] Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey E Kuehl, Erin Bransom, and Byron C Wallace. Automated metrics for medical multi-document summarization disagree with human evaluations. *arXiv preprint arXiv:2305.13693*, 2023.

[YC22] Shweta Yadav and Cornelia Caragea. Towards summarizing healthcare questions in low-resource setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2892–2905, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics.

[ZKW+20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[ZZG20] Yuhao Zhang, Emily Zhang, and James Glass. Optimizing medical terminology simplification with bert embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1431–1442, 2020.

[ZZSL20] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 11328–11339, 2020.