Multi-Level Summarization of General English and Scientific Texts: First Phase Report

Niloofar Karimi, Rimsha Kayastha, Yanzheng Liu
Northeastern University
Email: karimi.ni@northeastern.edu, kayastha.r@northeastern.edu, liu.yanz@northeastern.edu

***Abstract*—This project explores multi-level text summarization using transformer-based models to generate simplified summaries for readers of different proficiency levels. In this first phase, the BART model was fine-tuned on the OneStopEnglish (OSE) dataset, which contains 189 aligned articles across elementary, intermediate, and advanced levels. Comprehensive data cleaning, normalization, and readability analysis were performed to ensure linguistic consistency. Readability metrics and cosine similarity confirmed clear separation in text complexity while preserving semantic meaning. The fine-tuned model achieved strong readability control with coherent simplifications and a ROUGE-L score of 0.53. These findings validate the OSE dataset as a suitable foundation for multi-level summarization and establish a baseline for the next phase, which will extend the approach to biomedical text simplification.**

***Index Terms*—Text Summarization, BART, Transformer, Simplification, Readability**

I. Introduction and Motivation

A. Problem

Research dissemination focuses on peer-to-peer communication within academia rather than public outreach. Comprehension barriers remain severe with academic papers require grade 16–17 reading levels [2], while the average American adult reads at grade 7–8 [5]. This prevents researchers, students, patients, policymakers, and the general public from understanding scientific breakthroughs directly relevant to their lives, health decisions, and work.

B. Goals

The project goals are three-fold:

1) Develop a multi-level summarization system that adapts scientific and general texts for different audiences (students, laypersons, and experts) using fine-tuned transformer models such as BART.
2) Simplify domain-specific language through controlled abstraction and readability modeling, ensuring summaries remain both accurate and accessible.
3) Compare summarization performance between general (OneStopEnglish) and biomedical (M3) datasets using readability and semantic similarity metrics to evaluate model generalization.

C. Research Questions

1) Can fine-tuned BART models generate semantically accurate, multi-level summaries with statistically distinct readability levels across student, layperson, and expert audiences?
2) Does fine-tuning on general-domain datasets (OneStopEnglish) provide comparable multi-level summarization performance to biomedical-domain training (M3) when evaluated on scientific research papers?

II. Related Work

Transformer-based models have revolutionized text summarization and simplification. BART [3] and T5 [7] demonstrate strong abstract summarization performance through encoder-decoder architectures, while domain-specific models like BioGPT[4] handle specialized biomedical terminology. Multi-level summarization has advanced through datasets like BioLaySumm, MS², and OneStopEnglish, which provide parallel texts at different complexity levels enabling supervised learning of audience-adaptive generation. Evaluation combines content fidelity metrics (ROUGE, BERTScore) with readability measures (Flesch-Kincaid, SMOG), with studies showing plain language summaries achieve over 90% satisfaction across both expert and non-expert audiences. However, existing systems predominantly focus on single-level summarization or binary simplification (expert vs. lay), lacking fine-grained control across multiple distinct audience levels. Furthermore, limited research examines whether models trained on general-domain simplified texts can effectively generalize to specialized scientific and biomedical content, representing a critical gap in understanding domain transfer for accessibility applications.

III. Methodology

A. Modeling

- Base model: We use BART (Bidirectional and Auto-Regressive Transformers) as the foundation for abstractive multi-level summarization. Fine-tuning is performed on both general and medical texts to generate summaries for general and expert audiences. Summary detail is controlled through prompt design, decoding strategies, and fine-tuning objectives. Readability is evaluated using the

Flesch–Kincaid Grade Level (FKGL) metric to ensure that summaries are appropriately tailored to different audience literacy levels.

$$\text{FKGL} = 0.39 \times \frac{\text{total words}}{\text{total sentences}} + 11.8 \times \frac{\text{total syllables}}{\text{total words}} - 15.59. \tag{1}$$

- Level control via special tokens or prompts: We implement level control mechanisms using special tokens or prompt-based conditioning, enabling the model to adapt its summaries to different reader groups-general versus expert audiences. Prompts explicitly indicate the desired abstraction level and guide the decoder to adjust lexical and syntactic complexity accordingly.
- Two-phase: general-domain (OSE) and biomedical (M3): We conduct fine-tuning in two stages. The first stage trains the model on the OneStopEnglish corpora to capture broad summarization patterns and readability diversity. The second stage focuses on biomedical adaptation using the M3 dataset, which includes multi-level medical research summaries. This stage allows the model to specialize in domain-specific terminology and hierarchical information structuring.

IV. Data

A. Datasets

General English Text:

1) OneStopEnglish Corpus (OSE) [6] [8]: Contains 189 aligned article triplets across three reading levels: elementary, intermediate, and advanced, designed to study linguistic complexity and readability progression. Each article is rewritten by professional educators to maintain semantic equivalence while increasing syntactic and lexical sophistication. This dataset serves as a benchmark for evaluating models' ability to control and adapt output complexity across language proficiency levels. In the first phase of this project, the OSE dataset is used to train the BART model for general English text multi-level summarization.

Biomedical Text:

2)M3 (Multi-level Multi-domain Multi-lingual Summarization) Dataset: Comprises biomedical research papers with human-written summaries at three expertise levels: layperson, medical student, and expert. The dataset covers diverse medical domains such as oncology, immunology, and genetics, facilitating the study of how summarization style and terminology vary by reader expertise. The dataset will be fine-tuned and used to assess the model's capability to generate accurate, domain-specific summaries tailored to different professional audiences. In addition, it enables quantitative evaluation of readability shifts and lexical simplification across specialized biomedical sub-fields. The multi-lingual component further allows analysis of cross-lingual consistency in medical terminology and summarization quality.

B. Exploratory Data Analysis

OneStopEnglish:

- Data Quality Validation: To ensure that level-to-level text pairs are sufficiently distinct, pairwise textual similarity was computed using `difflib.SequenceMatcher`. A random sample of 50 pairs showed an average similarity of 41.8%, indicating that advanced, intermediate, and elementary versions differ substantially in structure and vocabulary. This confirms that the OneStopEnglish dataset is suitable for multi-level simplification training.
- Alignment check: confirms that each article correctly aligns across the three proficiency levels, ensuring that all versions of the same text discuss the same topic with different linguistic complexity.
- Average word count: increases as the text level rises, elementary texts contain about 534 words, intermediate texts 678, and advanced texts 825 words in a single article. This indicates the text complexity scales as expected (advanced texts are more detailed and consistently longer).
- Sentence length: increases by complexity level, from 21.6 words in elementary text to 28.8 words in advanced text. Also, the number of unique words increases as the linguistic complexity rises, indicating richer vocabulary and greater lexical diversity in high-level texts.
- Readability metrics: Flesch–Kincaid Grade Level (FKGL: 8.4–11.3) and Simple Measure of Gobbledygook (SMOG: 10.1–13.0) increase as the level increases, indicating that higher-level texts require a higher level of education and higher reading proficiency to comprehend.

TABLE I
Descriptive Statistics: OSE by Level (mean)

Metric	Elementary	Intermediate	Advanced
Word Count	534.6	677.9	824.8
Unique Words	279.2	351.3	433.9
Avg. Sentence Len	21.6	22.2	28.8
FKGL	8.40	10.08	11.26
SMOG Index	10.99	12.24	13.05

- **Cosine similarity:** was initially low (0.02–0.05) for a random article, indicating strong lexical divergence but topic consistency. This is expected because the goal for the model is to learn to simplify ideas, not just shorten sentences. To evaluate overall topic consistency, the average cosine similarities across all 189 aligned articles were computed. The scores were 0.80 (elementary–intermediate), 0.75 (intermediate–advanced), and 0.70 (elementary–advanced), confirming that while higher-level texts introduce more complex vocabulary and structure, all three levels effectively preserve semantic content.

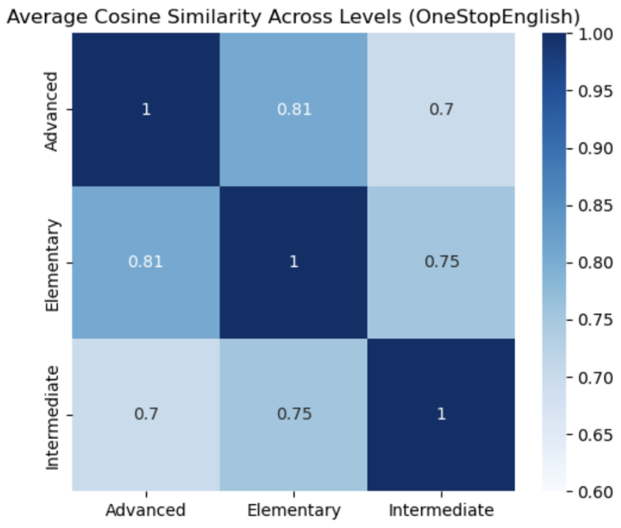

Fig. 1. Average cosine similarity across proficiency levels in the OneStopEnglish dataset. Higher similarity values indicate consistent semantic content across aligned articles.

C. Data Cleaning

- **Initial Structural Cleaning:** After loading the OneStopEnglish CSV files, several column-level inconsistencies were identified. Specifically, the dataset contained duplicate columns such as `Intermediate` and `Intermediate ` (with a trailing space), as well as unnamed empty columns generated during merging. The valid data from both intermediate columns were merged into a single unified column, and redundant or unnamed columns were dropped. This step ensured a consistent and well-structured dataset before applying text-level cleaning and normalization.
- **Verifying completeness and structure:** to ensure that triplets for training are balanced, OneStopEnglish is checked so that each article contains all three readability levels (elementary, intermediate, and advanced) and there are no missing or duplicate entries.
- **Pre-cleaning Inconsistency Scan:** To quantify the data inconsistencies across all three levels (elementary, intermediate, and advanced), a diagnostic function (`scan_issues`) is applied, which detects problems including missing end punctuation, multiple spaces, strange non-ASCII characters, suspicious short tokens, invisible Unicode spaces, double punctuation marks, and unbalanced quotation marks or parentheses. These sources of noise are identified because they can distort tokenization and reduce summarization accuracy, especially while fine-tuning the Biomedical dataset in the second phase. The most frequent issues included over 1,400-2,000 strange characters and large missing end punctuations per level; therefore, normalization steps are applied to standardize encoding, punctuation, and text structure. The `find_strange_characters()` function detected non-ASCII symbols such as curly quotes and dashes. Normalizing these issues ensures consistent tokenization in general text and is especially critical for biomedical datasets that include domain-specific symbols such as μ, β, and ≤.
- **Text Normalization and Typo Correction:** The normalize_text function standardize punctuation, spacing, and encoding while protecting domain-specific tokens such as acronyms and chemical symbols. Also, Unicode errors are fixed, typographic marks (e.g., curly quotes, dashes) are replaced, common typos are corrected, and proper sentence endings is ensured.
- **Final Cleaning Summary:** After normalization, a final scan was performed to quantify improvements in text quality. After this process, strange characters and spacing issues were reduced, with moderate gains in punctuation and token regularity, ensuring consistent and well-structured text across all reading levels, ready for summarization and biomedical fine-tuning.

TABLE II
SUMMARY OF IMPROVEMENTS AFTER CLEANING

Type	Total Drop	Avg. (% Improvement)
Missing end punctuation	61	27.9%
Strange characters	3601	68.6%
Multiple spaces	649	100.0%
Suspicious short tokens	250	10.0%

D. Preprocessing

- **Align OSE to form pairs:** adv→inter, adv→elem, inter→elem: to ensure proper alignment, the OneStopEnglish is restructured so each article across difficulty levels, elementary, intermediate, and advanced, appears side by side in one row. Using the position index, 189 triplets are produced (since 189 articles are summarized in three levels), to compare readability and training models for learning level-to-level text simplification.
- **Data Split:** To prepare for model training and evaluation, the aligned level pairs were split into training (85%) and validation (15%) sets. A stratified split based on direction labels (adv→inter, adv→elem, inter→elem) ensured that each simplification path was proportionally represented in both sets, preventing bias toward any single difficulty transition.
- **Tokenizer and Length Coverage Check:** The BART tokenizer was used to measure token lengths for both source and target texts prior to training to ensure compatibility with the model's input constraints. This diagnostic step showed that approximately 53% of the source texts fit within the 1,024-token limit, with an average length of 1,017 tokens and a maximum of 1,849. Target summaries averaged 759 tokens, with a maximum of 1,464.
  Accordingly, both the maximum source and target sequence lengths were set to 1,024 tokens. Truncation was applied only during tokenization for training, affecting

a small subset of longer samples without noticeable information loss.

V. Training

A. Setup

We set up a pre-trained base BART model, as our Sequence-to-sequence model, to train with tokenized source texts as encoder inputs and tokenized target texts as decoder outputs. We used a seed of 42 for reproducibility during the training. The training also involved implementing a data collator to add padding to shorter tokenized texts and ensure consistency in the length of data. Furthermore, control tokens were added in the form of '<TO_TARGET>' to guide the model on which level each training data has been simplified to.

This model was set up with batch sizes of 32, based on our computational limits, with AdamW as an optimizer for the model. The best model is chosen based on the ROUGE-L score. We used DataCollatorForSeq2Seq as our data collator for dynamic padding, but set it up to be ignored in the loss calculation.

B. Model Fine-tuning

We experimented with learning rate of $1e-5$, $2e-5$, and $3e-5$, and found the optimal rate to be $3e-5$. We ran it for 15 epochs and reached a plateau in validation loss between epoch 9-10, with diminishing returns after epoch 10.

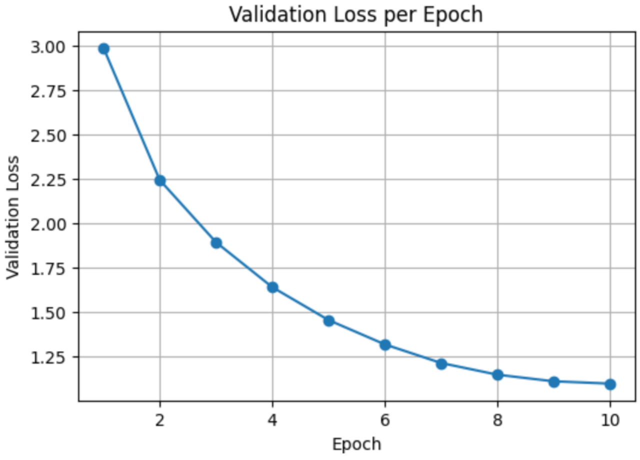

Fig. 2. Validation Loss Over Epochs

The ROUGE scores increased rapidly over epochs 1 to 3, with ROUGE-L improving from 0.507 to 0.528 and ROUGE-2 from 0.460 to 0.475. After epoch 3, the scores plateaued, with the final model achieving ROUGE-1 = 0.668, ROUGE-2 = 0.473, and ROUGE-L = 0.533. The checkpoint with the highest validation ROUGE-L score (epoch 10) was selected for all subsequent evaluation and analysis.

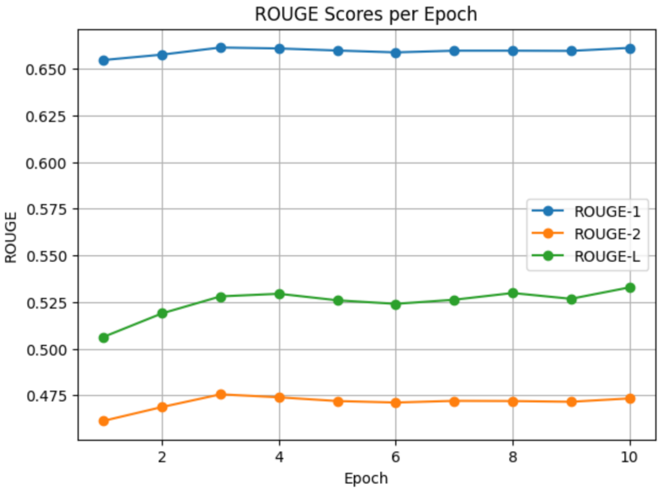

Fig. 3. ROUGE-1, ROUGE-2, ROUGE-L Scores Over Epochs

Initially, we also faced a challenge of generating near-identical copies of the source text. To overcome this, we improved upon the model by adding an anti-copy LogitsProcessor (ngram size of 3) to the generation parameters. Additionally, we also added a repetition penalty of 1.3 to reduce loops in text generation.

VI. Evaluation

A. Readability Metrics

Automatic metrics: ROUGE measures n-gram overlap between generated and reference summaries. This method verifies that each level maintains appropriate content fidelity to the source while varying in detail density, ensuring that critical information is not lost when simplifying from expert to general level. BERTScore uses contextual embeddings to measure semantic similarity beyond surface-level word matching. It captures cases where the same concepts are expressed using different vocabulary-for example, technical jargon at the expert level versus plain language at the general level. We use several readability metrics to ensure the output is accurate, including Flesch Reading Ease, Flesch-Kincaid Grade Level, and SMOG (Simple Measure of Gobbledygook). Flesch Reading Ease scores range from 0 to 100 based on sentence length and syllables per word. We adopt this method because it directly measures whether the "general audience" summary is easier to read than the "expert" summary. Flesch-Kincaid Grade Level converts readability into a U.S. school grade level, targeting grade 8 for the public and grade 16+ for clinicians.

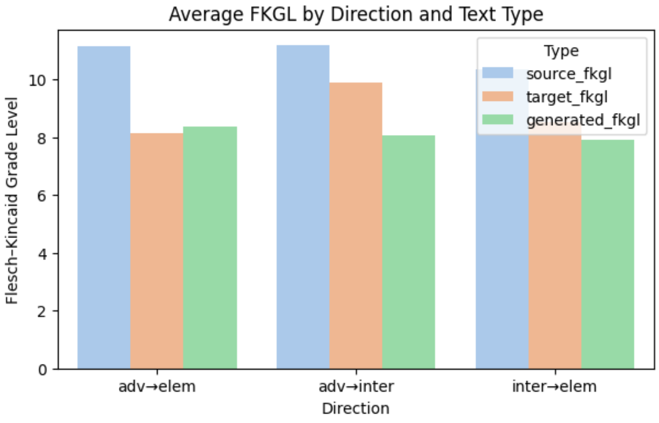

Fig. 4. Average FKGL by Direction and Text Type

In all three directions, the model's generated summaries (green bars) exhibit noticeably lower FKGL scores than the corresponding source texts (blue bars), indicating successful simplification. The generated FKGL values are also closely aligned with the human-written target summaries (orange bars), suggesting that the model produces readability levels consistent with the intended audience. Among the directions, adv→inter shows the largest readability gap between the target and generated summaries. The model sometimes oversimplifies intermediate-level content. In contrast, adv→elem and inter→elem display smaller differences, showing that the model maintains more stable readability control when simplifying to elementary-level outputs. Overall, the model effectively reduces linguistic complexity relative to the source.

B. Accuracy and Fidelity Metrics

Statistical testing:

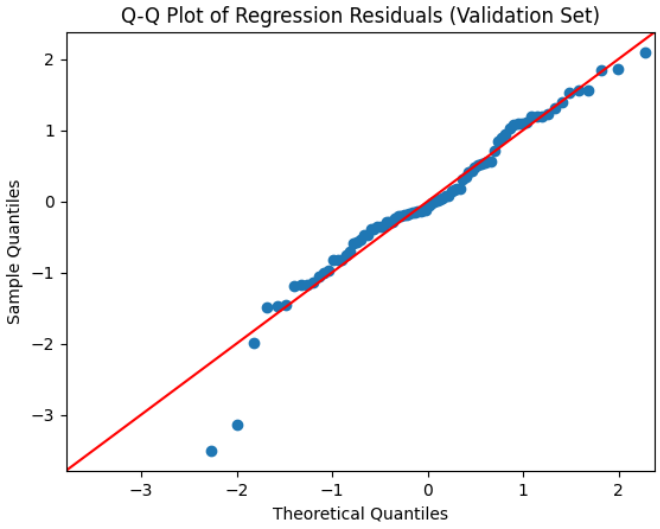

Fig. 5. Regression Residuals

The regression residuals plot (also known as a Q–Q plot) visualizes how well the residuals - the differences between the predicted and actual FKGL values - follow a normal distribution. In the plot, each point represents an observation's standardized residual compared against the theoretical quantiles of a normal distribution. If the residuals are approximately normal, the points should align closely along the 45° diagonal reference line. We use ANOVA and regression to assess readability differences and quality relationships. ANOVA tests whether there are statistically significant differences in readability among groups with different FKGL levels by comparing their F-values. Regression analysis examines the relationships between variables in the models, allowing us to determine whether the current level separation is appropriate or if there is a better way to define the multi-level structure. In the overall metrics, the BertScore is 0.381, which shows that the model captures semantic equivalence when wording diverges. ROUGE-1 $\approx$ 0.6 and ROUGE-2 $\approx$ 0.3 are strong for abstractive summarization tasks. It shows that the model rephrases effectively without losing information.

TABLE III
PER-DIRECTION AUTOMATIC EVALUATION METRICS

Direction	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore F1**
adv→elem	0.598	0.270	0.326	0.889
adv→inter	0.593	0.318	0.359	0.900
inter→elem	0.620	0.314	0.361	0.897

The ROUGE and BERTScore evaluations demonstrate that the model achieves meaningful compression and simplification while maintaining a high degree of semantic integrity across multiple readability levels.

VII. Preliminary Results

A. Key Findings

The fine-tuned BART model demonstrated effective readability control and strong content fidelity on the OneStopEnglish dataset. The generated summaries exhibited the expected progression in linguistic complexity, showing consistently lower FKGL and SMOG scores for simplified outputs while preserving high semantic similarity (approximately 0.75–0.80) to the reference texts. The model achieved a ROUGE-L score of 0.53, indicating coherent abstraction and stable level adaptation. Overall, these findings validate the dataset's quality and establish a reliable baseline for the upcoming biomedical fine-tuning phase.

VIII. Discussion and Next Steps

Our model successfully performs text simplification across multiple difficulty levels, achieving ROUGE-L scores of 0.533 and compression ratios of 0.68, which are competitive with existing simplification systems. The model demonstrates the ability to reduce text complexity by approximately 2-3 grade levels (from FKGL 11.4 to 8.4) while maintaining semantic content.

The next steps would be comparing the model's performance on biomedical scientific research papers to ensure its consistency in maintaining domain-specific context and text simplification. This would require additional fine-tuning on clean scientific datasets. We aim for this model to be applicable for the general public to gain an understanding of scientific breakthroughs. This work shows the potential of automated multi-level summarization models to bridge the gap between expert research and public understanding. It will promote more accessible communication of scientific knowledge.

IX. Timeline

- Week 1: Conduct exploratory data analysis on OneStopEnglish Corpus
- Week 2-3: Preprocess and clean text (tokenization, labeling, readability scoring), and fine-tune BART for 3 readability levels: elementary, intermediate, advanced
- Week 3: Add readability control with ROUGE, BERTScore, and readability indices (FKGL, SMOG, Dale–Chall), and statistical testing (ANOVA, regression)
- Week 4–6: Fine-tune BART on the M3 dataset (expert, student, layperson summaries), and compare general vs. biomedical performance on readability and accuracy with statistical testing
- Week 7–8: Finalize model evaluation, visualizations, and deliverables

Appendix

Model generation samples

- **Source text 1:** <TO_ELEM> Noise from ships may disturb animals such as killer whales and dolphins much more than people previously thought. New research shows that underwater noise could disturb the animals' communication and ability to and prey. The low rumble of passing ships has, for a long time, been connected...
  Generated text 1: Scientists have found that underwater noise may disturb animals like killer whales and dolphins very much more than people had thought. New research has found noise at medium frequencies, including 20,000Hz, where killer whales hear best. The noise from ships may disturb the animals' communications...
- **Source text 2:** <TO_INTER> A Canadian man who sprang to fame after offering a free round-the-world trip to a woman with the same name as his ex-girlfriend has returned from the jaunt with his chosen namesake, although, to the dismay of those following the story, love did not blossom between the pair. Jordan Axani,...
  Generated text 2: Jordan Axani has returned from a trip with Elizabeth Quinn Gallagher. The Canadian man who made headlines in 2014 for offering a free round the world trip to a woman named Elizabeth Gallagher has returned to Canada with his chosen namesake. Jordan Axani said he did not want to have a relationship wi...
- **Source text 3:** <TO_ELEM> ”I got a Dyson vacuum cleaner but I don't even know if I want it. I just picked it up, ” Louise Haggerty, a 56-year-old hairdresser and waitress, said of her 1am trip to the Black Friday sales. ”It was mental in there. It was crazy. It was absolutely disgusting, disgusting.” Haggerty had v...
  Generated text 3: Louise Haggerty, 56, a hairdresser, said she was scared to go to the black Friday sales at Sainsbury's in Harringay. She didn't even know if she wanted it. She just picked it up. ”It was mentally in there,” she said. ”People were behaving like animals, and it was absolutely disgusting, ” she added. Hag...

References

[1] Scott A. Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 2022.

[2] K. Kasabwala, N. Agarwal, D. R. Hansberry, S. Baredes, and J. A. Eloy. Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae. *Frontiers in Human Neuroscience*, 12:308, 2018.

[3] 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. *CoRR*, abs/1910.13461, 2019.

[4] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), September 2022.

[5] National Center for Education Statistics. Program for international assessment of adult competencies (piaac). Survey, U.S. Department of Education, 2017.

[6] Vallabhi Nishkala. Onestopenglish corpus. https://github.com/nishkalavallabhi/OneStopEnglishCorpus, 2018. Accessed: [Oct. 03, 2025.].

[7] 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. *CoRR*, abs/1910.10683, 2019.

[8] Sowmya Vajjala and Ivana Lučić. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 297–304. Association for Computational Linguistics, 2018.