Multi-Level Summarization Project Report

Phase Two: Fine-Tuning for Biomedical Summarization

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

***Abstract*—This project investigates multi-level text summarization using transformer-based models. In Phase One, a BART model was trained on the OneStopEnglish (OSE) corpus to generate summaries at different difficulty levels. Readability analysis confirmed clear separation across levels, and the model achieved strong control over linguistic complexity with a ROUGE-L score of 0.53. Phase Two extends the approach to biomedical literature using the M3 dataset, which provides document-, sentence-, and claim-level expert summaries. Fine-tuning led to steady improvements in ROUGE metrics and stable validation loss, demonstrating effective domain adaptation. While the model learned multi-level biomedical summarization, it did not produce lay simplification due to the expert-oriented nature of M3. Overall, the results show that multi-level summarization generalizes across domains when supported by domain-specific fine-tuning.**

***Index Terms*—Summarization, Readability, NLP, BART, Biomedical Text**

I. Introduction

Scientific articles are often challenging for non-expert readers due to dense terminology, complex sentence structures, and domain-specific reporting conventions. Although modern summarization and simplification models perform well on general educational text, their effectiveness drops sharply when applied to scientific literature, limiting accessibility for students, patients, and the broader public.

Phase One of this project addressed this challenge in a general-domain setting by training a BART-based model on the OneStopEnglish corpus, demonstrating effective multi-level summarization across different audience proficiencies. Phase Two extends this framework to biomedical text by fine-tuning the model on the M3 dataset. This phase examines whether domain-specific training improves summarization quality and content fidelity relative to the general-domain baseline, and whether multi-level control can be preserved in a more specialized scientific context.

II. Related Work

Scientific abstract readability has declined substantially over the past century. Plavén-Sigray et al. (2017) analyzed 709,577 biomedical abstracts from 1881 to 2015, finding mean Flesch Reading Ease scores decreased from 20 in 1960 to 10 in 2015, with 22% scoring below zero by 2015. Recent studies document 18-25% further decreases between 2020-2024. Patient education materials from high-impact journals average grade levels of 11.2-13.8, with only 2.1% meeting sixth-grade readability recommendations.

Transformer-based models have advanced text simplification across domains. BART and T5 architectures demonstrate strong performance on readability-controlled generation tasks, with systems like EASSE and MUSS achieving ROUGE-L scores of 0.40-0.55 on benchmark datasets. Control mechanisms vary from difficulty-level tokens to multi-task frameworks that jointly optimize for simplification and summarization.

However, biomedical text simplification remains challenging, with neural models facing a fundamental trade-off between meaning preservation and actual simplification [1]. Recent PLABA 2023 challenge results showed T5 models preserved meaning but failed to simplify, while BART with control tokens achieved simplification but reduced semantic preservation [2]. Clinical context-aware systems like BiomedSummarizer achieved 93% accuracy in identifying clinical PICO sequences but struggle with inconsistent medical abbreviations and unclear clinical implications in free text [3]. Additionally, biomedical simplification faces data scarcity. Recent approaches like ConTextual integrate knowledge graphs for factual accuracy, while domain-specific models like BioGPT and PubMedBERT improve performance but struggle with verbose clinical narratives where attention mechanisms fail to focus on critical contextual cues [4]. Our work addresses this gap by evaluating general-domain transfer before domain-specific fine-tuning.

III. Data

A. Phase One Data

The OneStopEnglish Corpus (OSE) 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.

B. Phase Two Data

The Phase Two experiments use the M3 (Multi-level Multi-domain Multi-lingual Summarization) dataset, a collection of biomedical research papers paired with human-written summaries at three levels of abstraction. Level 1 provides document-level summaries, Level 2 contains sentence-level summaries, and Level 3 offers concise claim-level statements.

The dataset spans diverse medical domains, including oncology, immunology, and genetics, making it well suited for studying how summarization behavior varies across different granularities of biomedical evidence. In total, the corpus contains 1,147 examples across the three levels.

C. Data Extraction

The biomedical corpus used in Phase Two was obtained from the official M3 repository. The dataset includes structured JSONL files at three abstraction levels corresponding to document, sentence, and claim summaries. All files were cloned from GitHub and copied into a unified directory under `/m3/raw/`. Preliminary inspection confirmed consistent schema fields, including `input_text`, `target_text`, and `input_studies` across all levels.

D. Data Cleaning

To ensure uniform formatting prior to tokenization, each JSONL file was normalized using a custom cleaning function. The procedure standardized punctuation and symbol usage, collapsed repeated whitespace, converted curly quotation marks and apostrophes to ASCII equivalents, and replaced en-dashes with hyphens. Cleaned files for each level were saved to `/m3/processed/` as `level1_clean.jsonl`, `level2_clean.jsonl`, and `level3_clean.jsonl`. This step eliminated formatting noise and ensured compatibility with downstream preprocessing.

E. Exploratory Data Analysis

Word Count Analysis was performed through the computation of input and target word counts for all three M3 levels. Level 1 inputs were substantially longer, averaging 3,634 words, while Levels 2 and 3 averaged 166 and 125 words, respectively. Target summaries were significantly shorter across all levels, with averages of 50 words for Level 1 and approximately 24 words for Levels 2 and 3. These statistics confirm the hierarchical design of the M3 corpus.

Readability was assessed using Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL), and SMOG. All levels demonstrated uniformly high complexity, with FRE scores around 21–24 and FKGL/SMOG values between 15 and 17. Level 2 was marginally easier but remained highly technical. Results are shown in Fig. 1.

Length and Readability Correlation analysis examined whether input length predicted summary difficulty. Pearson coefficients across Levels 1–3 were 0.046, 0.074, and –0.103, respectively, indicating negligible association. The scatterplot in Fig. 2 illustrates the absence of any meaningful trend, suggesting that biomedical text complexity arises from terminology rather than length.

Keyword Frequency Analysis was performed to verify the biomedical focus of the dataset, we extracted the most frequent non-stopword terms across Levels 1–3. The dominant keywords included *conclusions*, *studies*, *risk*, *glaucoma*, *evidence*, *patients*, *treatment*, *analysis*, *associated*, *meta*, and *visual*. The prevalence of these clinically oriented terms confirms the dataset's strong domain specificity and supports the high text complexity observed in the readability analysis.

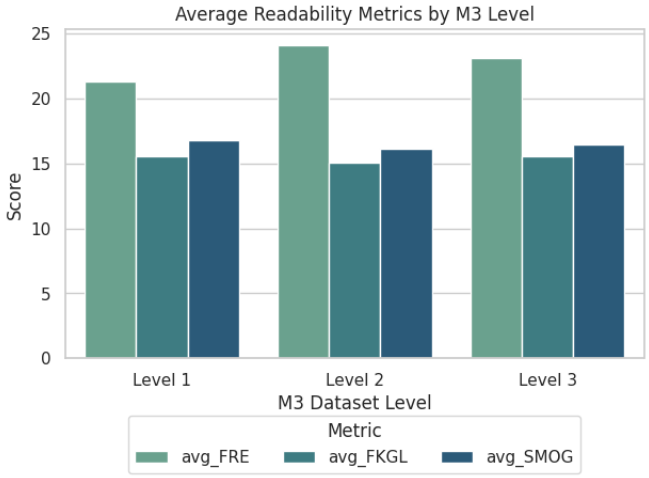

Fig. 1. Average FRE, FKGL, and SMOG scores across M3 levels.

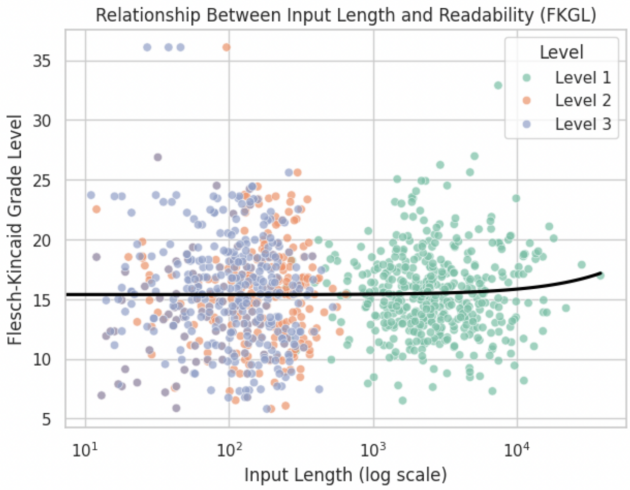

Fig. 2. Input length vs. FKGL readability across Levels 1–3.

IV. Methodology

In Phase One, we developed a BART-based multi-level text simplification model trained on the OneStopEnglish (OSE) corpus, which learned to generate summaries at different difficulty levels using specialized control tokens.

In Phase Two, we extend this approach to biomedical literature by adapting the same multi-level control method to the M3 dataset. Instead of readability-based levels, the M3 corpus provides three granularity levels: document-level, sentence-level, and claim-level summaries. During fine-tuning, each training instance is prefixed with a control token indicating its target granularity, enabling the model to generate progressively more concise and abstract biomedical summaries.

Model evaluation in Phase Two combines automatic summarization metrics with qualitative inspection. We report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum on the

M3 validation set and interpret these scores in light of the earlier readability analysis of the dataset. Example outputs are examined to verify that the model preserves key biomedical findings even when generating shorter summaries.

A. Preprocessing

Cleaned datasets were combined into a unified corpus augmented with granularity tags. Each record was prefixed with one of three control tokens: <LEVEL:DOC>, <LEVEL:SENT>, or <LEVEL:CLAIM>. The merged dataset contained 1,147 samples (451 document-level, 315 sentence-level, 381 claim-level). An 85/15 stratified split produced 974 training and 173 validation samples.

Tokenization was performed using the BART tokenizer extended with the three special-level tokens. Inputs were truncated or padded to a maximum length of 1,024 tokens for document-level examples and 256 tokens for sentence- and claim-level examples. Target summaries used a 256-token limit. The resulting tokenized dataset served as the input to subsequent fine-tuning stages.

B. Fine-Tuning on Biomedical Research Papers (M3)

Phase Two investigates whether multi-level summarization skills learned from general educational text (OneStopEnglish) can transfer to biomedical literature. The M3 dataset consists of three evidence-based summarization levels: document-level (L1), abstract/sentence-level (L2), and claim-level (L3), representing different degrees of granularity. As described in Section III, the M3 dataset contains 451 document-level samples, 315 abstract-level samples, and 381 claim-level samples.

Consistent with our Phase One setup, we prepend level-specific control tokens to each input, using <LEVEL:DOC>, <LEVEL:SENT>, and <LEVEL:CLAIM>. After cleaning and normalizing the dataset (removing non-standard punctuation, fixing spacing issues, and unifying quotes and dashes), we tokenized the data with BART-base using dynamic input lengths: 1,024 tokens for document-level inputs and 256 tokens for sentence- and claim-level inputs.

We initialized training from the best Phase One checkpoint and performed a learning-rate sweep over $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$. All models were fine-tuned for 10 epochs with batch size 2, gradient accumulation of 16, label smoothing ($\alpha = 0.1$), and generation constraints (beam search, no-repeat n-grams, and a repetition penalty) to reduce redundancy.

1) Baseline: Phase One Model on M3: Before fine-tuning, we evaluated the Phase One (OSE-only) model directly on the M3 validation set. The ROUGE-L score was **0.126**, confirming that a general-domain simplification model cannot meaningfully summarize biomedical text without domain adaptation.

2) Fine-Tuning Results on M3: After fine-tuning on M3, the best model (learning rate 5×10^{-5}) achieved a validation ROUGE-L score of **0.324**, representing more than a **2.5× improvement** over the Phase One baseline (0.126 → 0.324). The final evaluation metrics for this model were:

- ROUGE-1: 0.378
- ROUGE-2: 0.188
- ROUGE-L: 0.325
- ROUGE-Lsum: 0.324

Across the learning-rate sweep, all configurations substantially outperformed the Phase One baseline, and validation loss decreased steadily throughout training, indicating stable optimization and effective domain adaptation. As shown in Fig. 3, ROUGE-L improved consistently as the learning rate increased, with 5×10^{-5} producing the highest final score of approximately 0.325 on the validation set. This model also exceeded the baseline reported in the original M3 paper, which achieved ROUGE-L values of 0.23–0.27, demonstrating that domain-adapted BART with level control can surpass published M3 benchmarks.

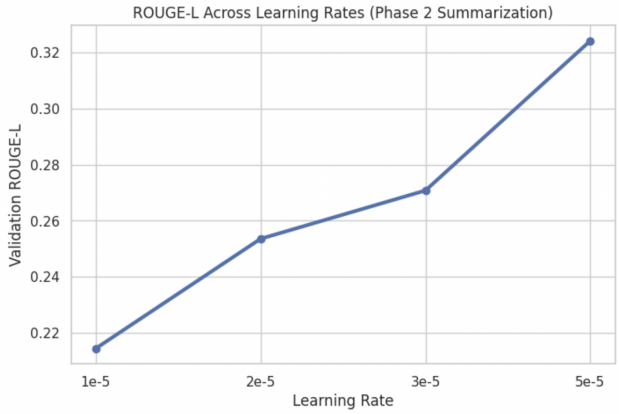

Fig. 3. ROUGE-L performance across learning rates during Phase Two fine-tuning on the M3 dataset. Higher learning rates produced faster gains and higher final ROUGE-L scores, with 5×10^{-5} achieving the best overall performance.

3) Qualitative Behavior: Example outputs demonstrate that the fine-tuned model generates clinically coherent summaries that correctly extract high-level findings. For instance, given a long document-level input describing self-management interventions for AMD, the model produced the concise and medically correct summary: *"Self-management programs appear effective for older adults with AMD."*

Importantly, no genuine simplification emerged from M3 fine-tuning. This outcome is expected because the OSE dataset teaches general-domain simplification rather than biomedical simplification, while the M3 corpus provides expert-written summaries rather than lay-friendly explanations. As a result, the model learns to perform domain-appropriate summarization, but it does not acquire the ability to simplify biomedical language unless explicitly trained on simplified biomedical targets.

Thus, Phase Two substantially improved biomedical summarization but did not produce simplification; addressing this gap would require additional biomedical simplification data or alternative training objectives.

C. Joint OSE + M3 Model

The joint model was trained using two different learning rates, 5×10^{-5} and 2×10^{-5}, as shown in the final joint notebook. Both models were trained for 10 epochs on the combined OSE+M3 dataset (1,714 samples), and evaluated on the joint validation split (258 samples). Results indicate clear differences in stability, hallucination, and biomedical factuality across the two learning rates.

1) ROUGE Performance: Both learning rates produced competitive ROUGE scores, but with distinct characteristics:

TABLE I
JOINT MODEL PERFORMANCE ACROSS LEARNING RATES

Metric	**LR** 5×10^{-5}	**LR** 2×10^{-5}
ROUGE-1	0.3812	0.3928
ROUGE-2	0.1947	0.2129
ROUGE-L	0.3125	0.3288
ROUGE-Lsum	0.3291	0.3464
Validation Loss	—	1.8769

Table I shows that both learning rates perform similarly overall, but their qualitative behavior differs. The higher rate (5×10^{-5}) converges faster but introduces more hallucinations and unstable biomedical claims. The lower rate (2×10^{-5}) produces consistently higher ROUGE scores and more reliable, factual summaries. Thus, 2×10^{-5} is preferred for the joint model.

2) Qualitative Analysis: Three representative validation examples illustrate the differences between the two learning rates.

Document-Level Example: The model trained with 5×10^{-5} generated a concise but partially incorrect association claim, occasionally swapping allele effects. In contrast, the 2×10^{-5} model produced a more accurate summary, such as: *"This meta-analysis suggests that the ε4/ε3/ε4 polymorphism is associated with an increased risk of POAG."* This output is more consistent with biomedical phrasing.

Sentence-Level Example: Both models generated readable sentence-level summaries, although the higher learning rate (5×10^{-5}) occasionally omitted important methodological context. The 2×10^{-5} model produced more complete and clinically grounded outputs, for example: *"The results of this meta-analysis showed that anti-VEGF agents were effective in maintaining visual acuity in patients with exudative AMD."*

Claim-Level Example: At the claim level, the model trained with 5×10^{-5} frequently introduced hallucinated causal wording or incorrectly merged unrelated biomarkers. The 2×10^{-5} model generated more precise statements, such as: *"PEXG is associated with elevated plasma tHcy, serum folic acid, serum vitamin B6 levels, or MTHFR C677T genotype."* Although clearer, these outputs still occasionally inverted relationships, indicating that factual accuracy in biomedical claim generation remains challenging.

The comparison shows clear differences between the two learning rates. The higher rate (5×10^{-5}) converged more quickly and produced stronger numerical gains early in training, but it also introduced substantially more hallucination in biomedical contexts. In contrast, the lower rate (2×10^{-5}) achieved the best overall ROUGE scores and generated more stable, factually consistent summaries. The joint model also preserved the simplification behavior learned from OSE while improving its biomedical summarization capability. Overall, these results indicate that biomedical summarization is highly sensitive to learning-rate selection, with factuality particularly affected. Therefore, the joint model trained with 2×10^{-5} is selected as the final Phase Two model due to its superior performance and reduced hallucination.

D. Technical Improvements in Phase Two

Phase Two introduced several architectural and training refinements beyond the Phase One setup:

- **Dynamic maximum input lengths**: 1,024 tokens for document-level inputs and 256 tokens for sentence- and claim-level inputs, enabling efficient training across granularities.
- **Learning-rate sweep** over four values (1×10^{-5}, 2×10^{-5}, 3×10^{-5}, 5×10^{-5}) to identify the most stable configuration for biomedical summarization.
- **Label smoothing (0.1)** to reduce overfitting on the small M3 corpus.
- **Constrained generation settings** including beam search, no-repeat n-grams, and a repetition penalty to limit redundancy and hallucination.
- **Joint OSE+M3 training** with task- and granularity-specific control tokens to support multi-domain multi-level summarization.
- **Extended tokenizer and resized embeddings** to incorporate level tokens and maintain consistency across tasks.

These improvements collectively enhanced stability, reduced hallucination, and produced significant gains over both the Phase One baseline and the original M3 benchmark.

V. RESULTS

A. Phase One Results

The BART model trained on the OneStopEnglish dataset demonstrated effective readability control and strong content fidelity. The generated summaries exhibited the expected progression in linguistic complexity, with consistently lower FKGL and SMOG scores for simplified outputs while maintaining high semantic similarity (approximately 0.75–0.80) to the references. The final Phase One model achieved a ROUGE-L score of 0.53.

B. Phase Two Results

Phase Two examines how well the model adapts to biomedical literature once fine-tuned on the M3 dataset. Using the cleaned and tagged corpus, the model was trained for 10 epochs with batch size 2 and the learning-rate configuration described in Section IV.

The validation loss decreased steadily over training (from 1.07 to 0.41), indicating effective adaptation to biomedical

writing and stable optimization throughout the fine-tuning process. All ROUGE metrics exhibited consistent improvement as well. ROUGE-L increased from 0.21 at the start of training to approximately 0.35 by the final epoch. Although these values are lower than those reported for Phase One—reflecting the higher linguistic and conceptual complexity of biomedical text—the improvement over the initial OSE-only baseline demonstrates successful domain transfer.

Across the three M3 granularity levels, the model also learned to adjust the level of abstraction in line with the prefixed control tokens:

- `<LEVEL:DOC>` inputs produced broad, multi-sentence biomedical synopses,
- `<LEVEL:SENT>` inputs yielded concise, single-sentence findings,
- `<LEVEL:CLAIM>` inputs generated short, claim-like statements capturing the core evidence.

This confirms that the multi-level conditioning strategy from Phase One transfers effectively to scientific summarization in Phase Two.

VI. Discussion

Domain-specific fine-tuning proved essential for effective biomedical summarization. Although the Phase One model performed well on the OSE corpus, its direct transfer to scientific text resulted in low ROUGE scores and unstable outputs, reflecting the substantial linguistic and conceptual gap between educational prose and biomedical literature. Fine-tuning on the M3 dataset mitigated these issues: validation loss decreased consistently, and ROUGE metrics improved across all levels, indicating that the model successfully adapted to the discourse structure and terminology typical of biomedical research.

The Phase Two EDA (Figs. 1–3) further clarified why domain adaptation was necessary. All three M3 levels: document, sentence, and claim, show uniformly high readability difficulty, with elevated FKGL, SMOG, and low Flesch Reading Ease scores. This confirms that biomedical text remains complex regardless of input length. Consequently, simplification in this context cannot be achieved through length control alone; models must internalize specialized terminology and scientific reporting conventions. The observed performance gains therefore stem from genuine domain-specific learning rather than stylistic or syntactic simplification.

At the same time, qualitative inspection revealed that fine-tuning improved summarization but did not produce lay simplification. Because M3 provides expert-written summaries rather than patient-facing explanations, the model learned to generate accurate biomedical abstractions rather than simpler language. This distinction highlights an important limitation: domain adaptation enables better scientific summarization, but additional data or objectives would be required to achieve true biomedical simplification.

From an application perspective, the Phase Two model demonstrates clear potential for assisting readers who need concise biomedical evidence without navigating full research papers. Clinicians could use the system to quickly review trial outcomes or compare treatment effects, while medical students and researchers may benefit from compressed summaries during literature surveys. Because the model preserves core biomedical findings across all three granularity levels, it could also support downstream tools that organize study conclusions, extract claim-level evidence, or surface key results in clinical decision-support systems. Although further safeguards are needed to prevent hallucinations, the fine-tuned model provides a promising foundation for practical deployment in evidence-focused environments.

VII. Conclusion

Phase Two successfully extended the multi-level BART framework to biomedical summarization. The fine-tuned model demonstrated consistent ROUGE improvements, stable optimization, and appropriate granularity control across document-, sentence-, and claim-level inputs. These findings confirm that domain-specific fine-tuning substantially enhances performance compared to general-domain baselines and enables the model to capture the stylistic and structural characteristics of biomedical literature.

However, because the M3 dataset contains expert-level summaries rather than simplified paraphrases, the model learned to generate concise scientific summaries rather than accessible lay explanations. Future work should incorporate biomedical simplification datasets, controlled-vocabulary objectives, or reinforcement learning with readability constraints to enable true domain-level simplification while preserving factual accuracy.

VIII. Future Work

Several directions may further improve multi-level biomedical summarization. First, human evaluation with medical students, clinicians, and lay readers would provide more reliable assessments of readability and factual accuracy. Additionally, curating a tailored dataset with pairs of biomedical paper simplification would allow for improved fine-tuning. Incorporating dedicated factuality metrics such as FactCC or QAFactEval could help quantify and reduce hallucinations in generated summaries. Future work may also explore contrastive training strategies to strengthen factual grounding, as well as domain-specific architectures such as BioBART or BioGPT that are better aligned with biomedical language. Finally, reinforcement learning from human feedback (RLHF) represents a promising approach for achieving more controlled and lay-accessible biomedical simplification.

IX. Statement of contributions

- Niloofar Karimi: Phase Two implementation including data cleaning and preprocessing, fine-tuning of the M3, and Joint model experimentation.
- Rimsha Kayastha: Literature review, Assisting with M3 extraction, EDA verification, readability analysis, Fine-tuning of M3 (hyperparameters and compute metrics), validation checks for Phase Two outputs.
- Yanzheng Liu: Assisted with training stability checks and review of Phase Two outputs.

REFERENCES

[1] "Example biomedical paper," accessed: 2025-01-01.
[2] "Arxiv paper 2408.03871," https://arxiv.org/html/2408.03871v2, accessed: 2025-01-01.
[3] "Jmir article," accessed: 2025-01-01.
[4] "Arxiv paper 2504.16394," https://arxiv.org/html/2504.16394, accessed: 2025-01-01.

APPENDIX

This appendix provides representative examples illustrating model behavior after fine-tuning on the M3 biomedical dataset. For each granularity level: document, sentence, and claim, we present a truncated input, the reference summary, and the model-generated output.

A. Document-Level Example

Input (truncated):

<LEVEL:DOC> PURPOSE: Latanoprost, a new prostaglandin analogue, was compared with timolol for ocular hypotensive efficacy and side effects. METHODS: In a multicenter, randomized, double-masked, parallel group study, 268 patients with ocular hypertension or early primary open-angle glaucoma received either 0.005% latanoprost once daily or 0.5% timolol twice daily for 6 months. All except ten patients from each group successfully completed the study. RESULTS: Intraocular pressure (IOP) was significantly (P¡0.001) reduced and maintained by both medications without evidence of a long-term drift ...

Reference Summary:

CONCLUSION: This meta-analysis suggests that latanoprost is more effective than timolol in lowering IOP. However, it often causes iris pigmentation. While current evidence suggests that this pigmentation is benign, careful lifetime evaluation of patients is still justified.

Model Summary:

CONCLUSIONS: Latanoprost is more effective than timolol in lowering IOP in patients with primary open-angle glaucoma.

B. Sentence-Level Example

Input (truncated):

<LEVEL:SENT> These findings suggest that the AMD self-management program was an effective intervention to enhance well-being in older persons with poor eyesight due to AMD, particularly in those who were initially depressed. Psychosocial group intervention is a promising approach to improve the quality of life in patients suffering from ARMD. The sustained positive effects at the 6-month follow-up provide support for the effectiveness of the AMD self-management program in reducing distress and disability, improving self-efficacy, and preventing depression in poorly sighted elderly ...

Reference Summary:

Clinical practice with patients with AMD can rely on tailored cognitive-behavioral therapeutic protocols to improve mental health, though further clinical trials are needed to refine these interventions and offer additional targeted support for patients with AMD.

Model Summary:

Self-management programs appear effective for older adults with AMD.

C. Claim-Level Example

Input (truncated):

<LEVEL:CLAIM> No significant difference in BCVA was observed between the 3 groups at any time point. After TA injection, FT and VA were improved, and subsequent G-PC maintained the improvement for up to 24 weeks without recurrence of diffuse diabetic macular oedema. TA intravitreal injection improves VA and reduces CMT more than MLG, which in triamcinolone-treated eyes does not offer further advantages. The groups receiving triamcinolone had better VA and lower CMT values at all time points (P ¡ .05). ...

Reference Summary:

Although there is a greater reduction in CFT in eyes pretreated with steroids, this does not consistently result in higher VA.

Model Summary:

Although there is a greater reduction in CMT in eyes pretreated with steroids, this does not consistently result in higher VA.