# Towards Knowledge-Guided Biomedical Lay Summarization using Large Language Models

**Shufan Ming** and **Yue Guo** and **Halil Kilicoglu***
School of Information Sciences
University of Illinois at Urbana-Champaign
{shufanm2, yueg, halil}@illinois.edu

## Abstract

The massive size, continual growth, and technical jargon in biomedical publications make it difficult for laypeople to stay informed about the latest scientific advances, motivating research on lay summarization of biomedical literature. Large language models (LLMs) are increasingly used for this task. Unlike typical automatic summarization, lay summarization requires incorporating background knowledge not found in a paper and explanations of technical jargon. This study explores the use of MeSH terms (Medical Subject Headings), which represent an article's main topics, to enhance background information generation in biomedical lay summarization. Furthermore, we introduced a multi-turn dialogue approach that more effectively leverages MeSH terms in the instruction-tuning of LLMs to enhance the quality of lay summaries. The best model improved the state-of-the-art on the eLife test set in terms of the ROUGE-1 score by nearly 2%, with competitive scores in other metrics. These results indicate that MeSH terms can guide LLMs to generate more relevant background information for laypeople. Additionally, evaluation on a held-out dataset, one that was not used during model pre-training, shows that this capability generalizes well to unseen data, further demonstrating the effectiveness of our approach.

## 1 Introduction

Biomedical publications contain valuable research findings on health topics that may interest a wide range of audiences, including laypeople. PubMed, the biomedical bibliographic database, contains more than 37 million articles as of January 2025, with an increase of almost one million articles in less than a year [1]. Despite the abundance of health-related scientific information available in these articles, it is challenging for laypeople to make sense of this information due to the enormous size and growth of the literature and the specialized jargon used in these publications (August et al., 2023). Summarizing lengthy literature into concise, jargon-free lay language that explains the article's background and motivation can help alleviate information overload for laypeople (Goldsack et al., 2022).

| **Abstract** |
| --- |
| Plasmodium sporozoites, the mosquito-transmitted forms of the malaria parasite, first infect the liver for an initial round of replication before the emergence of pathogenic blood stages. Sporozoites represent attractive targets for antimalarial preventive strategies, yet the mechanisms of parasite entry into hepatocytes remain poorly understood. Here we show that ... |
| **Lay Summary** |
| Malaria is an infectious disease that affects millions of people around the world and remains a major cause of death, especially in Africa. It is caused by Plasmodium parasites, which are transmitted by mosquitoes to mammals. Once in the mammal, the parasites infect liver cells, where they multiply. ... |

Table 1: Comparison of the first few sentences of the abstract and lay summary from an eLife article.

Table 1 demonstrates how lay summarization requires explaining jargon and providing background information to contextualize the study, which cannot always be fully derived from the source article alone. Text highlighted in blue from the abstract was simplified into two sentences highlighted in green in the lay summary. Text highlighted in yellow in the lay summary explains the term "Malaria" and background information missing from the abstract but necessary for laypeople.

To address this gap, previous work has explored the use of auxiliary inputs to incorporate relevant background knowledge from external resources (Guo et al., 2024; Goldsack et al., 2023) or to elicit hidden knowledge from LLMs through a two-stage inference process (Goldsack et al., 2025). For in-

*Corresponding author
[1] https://pubmed.ncbi.nlm.nih.gov/about/



*Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 285–297

stance, Guo et al. (2024) employed a separate retriever model to extract biomedical term definitions from Wikipedia, augmenting the input source articles. Similarly, Goldsack et al. (2023) constructed a graph-based knowledge representation, where biomedical concepts served as nodes and their relationships as edges, derived from the UMLS Semantic Network (McCray et al., 2001). This synthesized knowledge was then integrated with language models during fine-tuning. Both approaches demonstrated improvements in the relevance (i.e., alignment with gold-standard summaries) and readability of lay summaries.

In another line of research, keywords, length, readability, or other aspects of control have been used as non-parametric knowledge to modify prompts, rather than changing the parameters of the model, to generate desirable summaries (Fonseca and Cohen, 2024; He et al., 2022). Such modifications to the input prompt guide the model's conditional generation process during decoding, influencing the content, tone, or structure of the model output. However, the use of controllability in LLMs for the lay summarization task has achieved limited success compared to generic scientific summarization tasks, due to the highly abstractive nature of lay summaries and their particular emphasis on research background information (Jahan et al., 2024).

MeSH (Medical Subject Headings), developed at the National Library of Medicine, is a standardized terminology used to index medical and life science articles, offering relevant topical information and reflecting the broader context of the entire document. In this study, we hypothesize that using descriptive prompts consisting of a set of MeSH terms can guide the model's generation to provide tailored background information in lay summaries. To test this hypothesis, we designed a sequence of experiments using LLaMA-3[2] as the base model (Dubey et al., 2024) and instruction-response pairs constructed from the eLife dataset (Goldsack et al., 2022).

Specifically, we investigate the following research questions:

- What is the most effective approach for incorporating MeSH knowledge into the fine-tuning process to achieve high performance?
- How does the choice of MeSH terms (gold standard, predicted by another model, or a more focused subset of gold standard MeSH terms) affect the quality and relevance of lay summaries?
- Does the performance on articles published after LLaMA's knowledge cutoff date remain consistent when compared to the eLife test set, which contains articles published before the release date of the LLaMA model?

Our contributions are:

- Our enhanced instruction-tuning approach, using two-turn conversations, produces more diverse background information that is aligned with the source document and accessible to non-expert readers.
- We incorporate structured knowledge (MeSH) into the supervised fine-tuning (SFT) model, serving as classifier-free guidance that is easier to obtain compared to previous approaches relying on auxiliary retrieval-augmented generation (RAG) models or graph structures.
- We constructed a recent dataset from the eLife corpus, using a cutoff date of June 30, 2024, to compare and assess the generalizability of different approaches.

## 2 Methods

In this section, we first describe the datasets we use. Next, we discuss our proposed main approach, multi-turn instruction tuning, followed by several ablation studies to verify the effectiveness of each model component and our hypothesis. Finally, we outline the experimental setup and evaluation metrics used to compare different settings.

### 2.1 Dataset and Data Collection

We trained and tested our model on the eLife dataset (Goldsack et al., 2022), which consists of 4,346 pairs of full-text articles and lay summaries for training, along with 241 pairs each for validation and testing. Compared to the PLOS dataset (Goldsack et al., 2022), another commonly used biomedical lay summarization dataset, eLife contains much longer lay summaries written by expert editors and exhibits a strong content bias toward research background (You et al., 2024). This characteristic makes the summaries easier for a lay audience to understand but presents a greater challenge for the LLM to generate (Fonseca and Cohen, 2024).

[2] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

The eLife corpus may have been included in the LLaMA model's training data, potentially giving the model an advantage by allowing it to memorize and reproduce information it has already encountered. To test the generalizability of our proposed method, we collected 71 articles published in the eLife corpus after June 30, 2024, through the open-source repository[3] on November 15, 2024, and used them as a held-out dataset for evaluation.

### 2.2 Multi-turn Instruction Tuning

Figure 1 illustrates the overall model architecture. The main method involves two back-and-forth conversational turns: the first turn focuses on MeSH prediction as an auxiliary task, while the second turn generates a lay summary conditioned on both the source document and the generated MeSH terms. During training, both MeSH terms and lay summaries are learned by minimizing the cross-entropy loss between the generated outputs and their respective gold standards for each article. Gold standard MeSH terms for each article were extracted by querying the PubMed database through the Entrez package[4].

The loss function is defined as follows:

$$\mathcal{L} = -\sum_{j=1}^{J}\sum_{t=1}^{T_j} \mathbb{1}_{[y_{t,j}\in y_{a,j}]} \log P\Big(y_{t,j} \mid y_{<t,j}, y_{\leq,j-1}, X, I; \theta\Big) \quad (1)$$

In this formulation, the loss function $\mathcal{L}$ uses cross-entropy to compare the model's generated responses at each turn with the gold-standard outputs, which consist of both MeSH terms and lay summaries. Here, $y_{t,j}$ denotes the token at time step $t$ during the $j$-th conversation turn, while $y_{a,j}$ represents the set of tokens specific to the model's output in turn $j$. The indicator function $\mathbb{1}_{[y_t\in y_a]}$ checks whether the token $y_{t,j}$ belongs to the set of target tokens $y_{a,j}$. If it does, the indicator returns 1, allowing the token to contribute to the loss computation; otherwise, it returns 0, excluding irrelevant tokens such as those from the user prompt.

The conditional probability term $\log P\Big(y_{t,j} \mid y_{<t,j}, y_{\leq,j-1}, X, I; \theta\Big)$ represents the likelihood of predicting token $y_{t,j}$, given all preceding tokens in the same turn $y_{<t,j}$, all tokens from previous turns $y_{\leq,j-1}$, the input article $X$, and any additional instructions $I$. This setup ensures that tokens from the current turn $j$ are conditioned on both intra-turn context and inter-turn history, enabling the model to incorporate contextual information from the entire conversation.

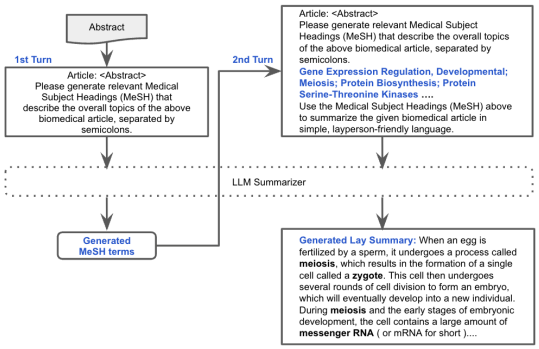


Figure 1: Workflow of the multi-turn instruction tuning at inference time. The generated MeSH terms serve as external guidance for the second forward pass. The upper portion of the figure illustrates the input prompts, while the lower portion displays the model's outputs.

We observed that some overly general MeSH terms, such as "Human" or "Animals," could mislead the model into generating irrelevant or overly broad background information. Our main approach involved applying a filtering strategy based on the hierarchical structure of the MeSH tree and its associated tree numbers[5] to retain only a subset of gold-standard MeSH terms. Specifically, if multiple terms shared the same tree prefix, we included the term with the longest identifier (indicating the highest specificity) and excluded others with shorter identifiers. For example, consider the full set of gold-standard MeSH terms separated by semi-colons: "Animals; Behavior, Animal; Cerebellum; Conditioning, Eyelid; Cues; Extinction, Psychological; Feedback; Learning; Male; Movement; Purkinje Cells; Rabbits; Time Factors". After applying the filtering strategy, more generic terms were removed in favor of more specific ones. For instance, "Animals" was eliminated because a more specific term, "Rabbits," from the same hierarchical branch (sharing the same prefix), was retained.

We used a parameter-efficient fine-tuning technique (PEFT), low-rank adaptation (LoRA) (Hu et al., 2021), to fine-tune large language models

[3] https://github.com/elifesciences/elife-article-xml

[4] https://biopython.org/docs/1.75/api/Bio.Entrez.html

[5] https://hhs.github.io/meshrdf/tree-numbers

efficiently. LoRA achieves efficiency by inserting trainable matrices and updating only a small subset of weights while keeping the original model parameters frozen. In Equation 1, $\theta$ denotes the subset of parameters updated during fine-tuning via LoRA.

### 2.3 Multi-turn Instruction Modeling

We conducted another experiment, INSTRUCTION_MODELING, based on previous research findings (Shi et al., 2024). These findings suggest that training the model to generate both instructions and responses, mimicking how humans provide task descriptions and guidance, yields more robust and higher-performing results, especially when the number of training examples is limited. Unlike INSTRUCTION_TUNING, which focuses on training the model to follow instructions and generate high-quality contextual responses, INSTRUCTION_MODELING introduces a modified loss function that applies to both the user input and the assistant's response. The updated loss function is defined as follows:

$$\mathcal{L} = -\sum_{j=1}^{J}\sum_{t=1}^{T_j} \log P(y_{t,j} \mid y_{<t,j}, y_{\leq,j-1}, X, I; \theta) \tag{2}$$

The key distinction from Equation 1 is the absence of the indicator function. This omission allows the model to be trained on both the user's input and the assistant's responses. The goal of this approach is to evaluate whether it improves the model's ability to understand and distinguish the linguistic differences between a scientific article and its lay summary, as well as the translations between them, thereby enhancing lay summary generation.

### 2.4 Ablation Study on Adaptation Methods and Knowledge Integration

We conducted various ablation studies to understand the contribution of each component to the overall performance of the main model, including the impact of integrating MeSH terms as guidance, the role of different training objectives, and the effect of MeSH term selection on summary quality.

#### 2.4.1 In-context Learning

Another technique for adapting the pre-trained model to a domain-specific downstream task is in-context learning, which is a more lightweight alternative to PEFT. We tested three experimental setups: (1) an instruction-only setting without any external knowledge or guidance (Experiment 0-SHOT). (2) An approach in which the instruction was augmented with a pair consisting of an abstract and its corresponding lay summary selected from the training data (Experiment 1-SHOT). Specifically, for each source article, we retrieved the most similar abstract from the training set using SimCSE (Gao et al., 2021). The corresponding abstract and its associated lay summary from the training set are then provided as an exemplar to guide the generation. (3) An external knowledge-guided setting in which ground truth MeSH terms were explicitly integrated into the prompts (Experiment MESH_GUIDANCE). Unlike the main approach, which strictly requires the model to predict MeSH terms that closely match the gold standard, this method acts as a guiding framework, allowing the model greater flexibility to interpret and utilize MeSH terms based on its learned knowledge.

The prompt template is shown as below:

- 0-SHOT:
  *Article: <Abstract>*
  *Summarize the above biomedical article in simple, layperson-friendly language.*

- 1-SHOT:
  *Article: <Abstract>*
  *Summarize the above biomedical article in simple, layperson-friendly language. Use the example below to guide the tone, structure, and the inclusion of relevant background context in your summary.*
  *Example abstract:<Example Abstract>*
  *Example lay summary:<Example Lay Summary>.*

- MESH_GUIDANCE:
  *Article: <Abstract>*
  *Summarize the above biomedical article in simple, layperson-friendly language. Use the following Medical Subject Headings (MeSH) as guidance for providing relevant background context where appropriate: <List of MeSH terms>.*

#### 2.4.2 Single-turn Instruction Tuning

In the multi-turn experiment setting, MeSH term generation is trained as an auxiliary task. We also designed two single-turn experimental setups that

do not include training on MeSH terms: (1) SINGLE_TURN: instruction tuning using the same template as 0-SHOT for lay summary generation only, and (2) MESH_SINGLE_TURN: Instruction tuning that incorporates ground truth MeSH terms retrieved from the PubMed database into the user input prompt as non-parametric guidance, using the same template as MESH_GUIDANCE. Similarly, the training objective is shown below. The cross-entropy loss $\mathcal{L}$ is computed exclusively on the model's generated summaries.

$$\mathcal{L} = -\sum_{t=1}^{T} \mathbb{1}_{[y_t \in y_a]} \log P(y_t \mid y_{<t}, I, X; \theta) \quad (3)$$

#### 2.4.3 MeSH term selection

MeSH terms serve as a signal for identifying which topics are essential and relevant to the source article, guiding the model to incorporate these concepts as background knowledge in the lay summary. Our main approach, described in Section 2.2, uses a heuristic-based curation method to select a subset of ground truth MeSH terms as the gold standard during the fine-tuning process. We also used all the ground-truth MeSH terms, without applying our filtering strategy, to investigate how training with the complete set of ground-truth MeSH terms versus a subset affects performance. We refer to this experiment as INSTRUCTION_TUNING_FULL_LIST.

In addition, we designed an ablation study to evaluate the impact of MeSH terms on the model's performance in a single-turn setting. Instead of providing ground truth MeSH terms in the prompt, we used predicted MeSH terms generated by a BERT-based MeSH classifier (BERTMeSH (You et al., 2021)), which achieves a Micro-F1 score of 63%. This comparison aimed to assess how both the quality and inclusion of different sets of MeSH terms in the input affect the model's performance. We refer to this experiment as BERT_MESH_SINGLE_TURN.

### 2.5 Experimental Settings

We used the *LLaMA-3.2-3B-Instruct* as the base model for all experiments. Due to computational resource limitations and the high memory requirements for fine-tuning large language models, we set the maximum input length to 2,500 tokens. We integrated the Accelerate (Gugger et al., 2022) and DeepSpeed (Rasley et al., 2020) libraries for fine-tuning. In addition, we employed an early stopping strategy based on validation performance, restricting training to a maximum of 3 epochs. The checkpoint that achieved the best performance on the validation set was then selected for inference on the test set. During inference, we set the temperature to 0 to ensure consistency in our summarization experiments. We set *max_new_tokens* to 512 to allow sufficient space for complete summaries while preventing excessively long outputs that may introduce irrelevant information.

### 2.6 Evaluation

The experiments were assessed solely for lay summary generation, using two sets of commonly applied metrics in previous lay summarization work: relevance and readability. Specifically, we employed ROUGE scores (Lin, 2004), including ROUGE-1, ROUGE-2, and ROUGE-L, which measure n-gram overlaps, as well as BERTScore (Zhang et al., 2019), which evaluates semantic similarity in the embedding space, to assess relevance. For readability evaluation, we used the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), and the Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948).

We did not include factuality-related metrics because previous findings show that existing automatic evaluation metrics for faithfulness do not align well with human evaluation in the context of biomedical plain language summarization (Fang et al., 2024; Chen et al., 2024). For example, fact-checking or natural language inference (NLI)-based evaluations, such as SummaC (Laban et al., 2022) and AlignScore (Zha et al., 2023), are designed and trained at the sentence level. These methods are highly sensitive to benign modifications and perturbations, which limits their ability to evaluate abstractive summarization tasks that often require text rewriting and paraphrasing (Tang et al., 2022; Ramprasad and Wallace, 2024). Moreover, those evaluations focus on whether the content aligns with the source, whereas in our case, lay summarization requires incorporating new external knowledge not present in the source article.

We assessed the statistical significance of the differences between the generated summaries across several experimental settings using the Wilcoxon signed-rank test (Woolson, 2005) in a pairwise manner, following the methodology of previous studies (Van Veen et al., 2024).

| | Relevance | | | | Readability | | |
|---|---|---|---|---|---|---|---|
| | **R-1** | **R-2** | **R-L** | **BERTScore** | **FKGL ↓** | **CLI ↓** | **DCRS ↓** |
| SINGLE_TURN | 0.5003 | 0.1374 | 0.4718 | 0.8518 | 10.6904 | 10.8585 | 8.6497 |
| INSTRUCTION_TUNING_FULL_LIST | 0.5004 | 0.1395 | 0.4714 | 0.8516 | 10.5369 | 10.7346 | 8.5793* |
| INSTRUCTION_TUNING | 0.5021 | **0.1408*** | 0.4733 | **0.8524** | **10.4203**** | **10.6960**** | **8.5705**** |
| INSTRUCTION_MODELING | **0.5026** | 0.1399 | **0.4747** | 0.8520 | 10.5381* | 10.9456 | 8.6068 |

Table 2: Results for the multi-turn conversation and single turn approach on the eLife test set. ↓ indicates that lower scores are better for that metric. Asterisks indicate statistical significance relative to the baseline model without MeSH (SINGLE_TURN), as determined by the Wilcoxon signed-rank test (* $p < 0.05$, ** $p < 0.01$).

| | Relevance | | | | Readability | | |
|---|---|---|---|---|---|---|---|
| | **R-1** | **R-2** | **R-L** | **BERTScore** | **FKGL ↓** | **CLI ↓** | **DCRS ↓** |
| 0-SHOT | 0.3284 | 0.0781 | 0.3022 | 0.8399 | **9.2091** | **10.2627** | 8.5570 |
| 1-SHOT | 0.3949*** | 0.0851*** | 0.3675*** | 0.8409* | 9.4726* | 10.3205 | **8.3313***** |
| MESH_GUIDANCE | **0.4186***** | **0.0907***** | **0.3895***** | **0.8412**** | 10.1078*** | 11.0801*** | 8.6826* |

Table 3: In-Context Learning Experiments: Comparison of 0-SHOT, 1-SHOT, and MESH_GUIDANCE Results. Asterisks denote statistical significance relative to 0-SHOT: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

| | Relevance | | | | Readability | | |
|---|---|---|---|---|---|---|---|
| | **R-1** | **R-2** | **R-L** | **BERTScore** | **FKGL ↓** | **CLI ↓** | **DCRS ↓** |
| SINGLE_TURN | 0.5003 | 0.1374 | **0.4718** | 0.8518 | 10.6904 | 10.8585 | 8.6497 |
| MESH_SINGLE_TURN | **0.5007** | **0.1405** | 0.4714 | **0.8521** | **10.4605*** | **10.7297*** | **8.6129** |
| BERT_MESH_SINGLE_TURN | 0.4983 | 0.1388 | 0.4695 | 0.8517 | 10.5622 | 10.8195 | 8.6464 |

Table 4: Results for the single-turn conversation approach augmented with different sets of MeSH terms (ground truth vs. MeSH classifier) on the eLife test set. Asterisks indicate statistical significance relative to the baseline model without MeSH (* $p < 0.05$)

## 3 Results and Discussion

### 3.1 MeSH Prediction as an Auxiliary Task vs. No MeSH

The main results on the test set are presented in Table 2. Both INSTRUCTION_TUNING and INSTRUCTION_MODELING incorporate MeSH term prediction as an auxiliary task within a multi-turn instruction tuning framework. Compared to the baseline approach (SINGLE_TURN), INSTRUCTION_TUNING achieved statistically significant improvements in ROUGE-2 ($p < 0.05$) and all readability metrics ($p < 0.01$). In contrast, INSTRUCTION_MODELING, which trains the model to generate both user inputs (scientific articles) and assistant responses (lay summaries), achieved the highest ROUGE-1 ($p = 0.23$) and ROUGE-L ($p = 0.06$) scores but did not show significant improvements over the baseline approach. Notably, the ROUGE-1 score represents state-of-the-art performance, improving by nearly 2% compared to prior work (which reported results of approximately 0.48–0.49) (Jahan et al., 2024).

In prior assessments on benchmark datasets, although the INSTRUCTION_MODELING approach has proven effective in language understanding tasks, as evidenced by high BLEU scores in benchmarks such as OpenBookQA (Mihaylov et al., 2018) and MMLU (Hendrycks et al., 2020), significance tests indicate that INSTRUCTION_MODELING offers no improvements over INSTRUCTION_TUNING in ROUGE-1 and ROUGE-L, and it even performs significantly worse on readability metrics in CLI ($p < 0.001$).

We also compare training with the full list of ground truth MeSH terms (Experiment INSTRUCTION_TUNING_FULL_LIST) versus a selectively chosen subset as the gold standard (INSTRUCTION_TUNING) in a multi-turn setting. Improvements were observed across all metrics but were not statistically significant.

### 3.2 Ablation Results

**In-context Learning** The results without instruction tuning are shown in Table 3. We compared experiments using a basic prompt only (0-

SHOT), incorporating ground truth MeSH terms (MESH_GUIDANCE), or using an exemplar pair of a scientific abstract and lay summary as guidance (1-SHOT). When using MeSH as guidance, all the relevance metrics showed statistically significant improvement over the basic prompt ($p < 0.001$ for ROUGE scores; $p < 0.01$ for BERTScore). However, the readability scores significantly decreased when the user prompt became more complex due to the augmentation with MeSH terms ($p < 0.001$ for FKGL and CLI; $p < 0.05$ for DCRS).Using the most similar example from the training data, which serves as the standard approach in a few-shot learning setting, yielded significant improvements across all ROUGE scores compared to the 0-SHOT setting ($p < 0.001$), also achieving the best DCRS score. Notably, when comparing 1-SHOT and MESH_GUIDANCE, all the ROUGE scores were significantly improved ($p < 0.001$), as well as the BERTScore ($p < 0.01$), but all readability scores decreased ($p < 0.001$ for FKGL and CLI; $p < 0.05$ for DCRS).

When incorporating prompts with MeSH terms, even without any fine-tuning, the model achieves higher lexical overlap and improved semantic alignment with the gold-standard lay summary, suggesting that it can effectively distill useful topical information from these terms.

**The Effect of MeSH Term Selection in Single-Turn Instruction Tuning.** As shown in Table 4, the SINGLE_TURN experiment, which uses only the abstract as input for instruction tuning, demonstrated less competitive performance than the MESH_SINGLE_TURN experiment, which incorporates ground truth MeSH terms in the prompt and improves results on all metrics except ROUGE-L. When using predicted MeSH terms from a BERT-based classifier (You et al., 2021), BERT_MESH_SINGLE_TURN, the improvement was less pronounced, with only a non-significant increase observed in ROUGE-2 ($p = 0.3$), FKGL ($p = 0.2$), CLI ($p = 0.8$), and DCRS ($p = 0.7$).

We selected the BERT-based MeSH classifier for its strong performance and ease of implementation, providing a reliable baseline for comparison. While using a more recent model could have yielded slightly better results, it is unlikely to reach the performance achieved with ground truth terms. Although the improvements with machine-generated MeSH terms were not statistically significant, they suggest potential for applying our method to articles without ground truth MeSH terms. With further refinement of MeSH prediction models and more sophisticated term selection strategies, this approach could be extended to biomedical literature beyond PubMed.

**The Effectiveness of Incorporating MeSH in Multi-Turn Conversations vs. Single-Turn Approach.** Comparing the MESH_SINGLE_TURN approach in Table 4 with the multi-turn instruction tuning experiments in Table 2, INSTRUCTION_TUNING, which was fine-tuned on a selectively chosen subset of MeSH terms, demonstrated improvements across all metrics. This is likely due to two key factors: (1) iterative interactions, where the second-turn summary generation builds upon the previously predicted MeSH terms, allowing the model to engage in a step-by-step reasoning process that mirrors the chain-of-thought strategy, and (2) improved calibration of MeSH term selection during fine-tuning, which ensures that a more focused subset of gold standard MeSH terms are incorporated into the generation process.

Overall, we observed performance gains by incorporating MeSH terms in both in-context learning and PEFT settings, including single-turn and multi-turn approaches. Our results suggest that MeSH terms can serve as an effective proxy for guiding the LLM in generating coherent, relevant, and readable lay summaries with essential background explanations. Moreover, the fact that the most significant improvement is more pronounced in a simpler, training-free setting (see Table 3) motivates the development of a more sophisticated method for selecting gold-standard MeSH terms as an auxiliary task during multi-turn instruction tuning, which could further improve the quality of lay summary generation.

### 3.3 Performance on the held-out evaluation set

LLMs are often pretrained on vast datasets. If the test set overlaps with pretraining data, the model might perform well due to memorization rather than generalization. To fairly and accurately evaluate the effectiveness of our approach, we further investigate whether the fine-tuned summarizer can achieve comparable results when applied to a held-out dataset consisting of articles published after the release date of *LLaMA-3.2-3B-Instruct*.

As shown in Table 5, the best performance was achieved in Experiment INSTRUCTION_TUNING, the multi-turn approach with instruction tuning, yielding results that closely align with the

| | Relevance | | | | Readability | | |
|---|---|---|---|---|---|---|---|
| | **R-1** | **R-2** | **R-L** | **BERTScore** | **FKGL ↓** | **CLI ↓** | **DCRS ↓** |
| INSTRUCTION_TUNING | **0.4954** | **0.1346** | **0.4580** | **0.8550** | **10.4887** | **10.8483** | **8.6276** |
| INSTRUCTION_MODELING | 0.4863 | 0.1273 | 0.4508 | 0.8537 | 10.8408 | 11.1756 | 8.7201 |
| SINGLE_TURN | 0.4843 | 0.1237 | 0.4466 | 0.8521 | 10.8084 | 11.0800 | 8.7638 |
| MESH_SINGLE_TURN | 0.4876 | 0.1270 | 0.4501 | 0.8526 | 10.6718 | 10.8514 | 8.7322 |

Table 5: Held out evaluation results for relevance and readability. ↓ denotes the scores that need to be minimized for those metrics.

test data. This suggests that the model is not simply memorizing the training data from pre-training stage. However, a pronounced decrease in all ROUGE scores and readability metrics was observed in Experiments SINGLE_TURN and MESH_SINGLE_TURN on the held-out dataset compared to the test set. These findings indicate that multi-turn conversation instruction-tuning, with MeSH generation as an auxiliary task, ensures better generalizability to unseen data than other approaches.

## 4 Qualitative Analysis

Tables 6 and 7 in the Appendix compare the generated summaries across different experimental settings. In this example, Experiment INSTRUCTION_TUNING achieved the best relevance score, followed by Experiment INSTRUCTION_MODELING and Experiment MESH_SINGLE_TURN. Notably, in the abstract, the first sentence begins with the study design of the approach, whereas the gold standard lay summary includes additional sentences introducing the importance of the topic, the symptoms of the disease, and the current research gap, which are highlighted in different colors. Both multi-turn conversation approaches closely follow the same information flow and context as the gold standard. They also state the method precisely as conveyed in the abstract's first sentence. The SINGLE_TURN approach contained more technical jargon, which is harder for laypeople to understand, and lacked sufficient background information.

## 5 Related Work

Current research in biomedical plain language summarization focuses on two main subtasks: text simplification and explanation and background generation. Text simplification involves linguistic transformations, such as rewording and replacing biomedical terminology with less technical terms, to make content more accessible (Attal et al., 2023; Devaraj et al., 2021). On the other hand, explanation and background information generation leverage external knowledge to enhance the informativeness of summaries (Guo et al., 2024).

Two main model architectures are commonly used for plain language summarization: encoder-decoder models (e.g., T5 (Raffel et al., 2020), BART (Lewis, 2019), Longformer (Beltagy et al., 2020)) and generative models such as the GPT family (Radford et al., 2019) and LLaMA (Touvron et al., 2023). Generative LLMs have demonstrated strong zero-shot and few-shot summarization capabilities, producing coherent and relevant text from demonstrations alone, without the need for fine-tuning or parameter updates (Zhao et al., 2024).

While LLMs are inherently capable of following natural language instructions, instruction-tuned models, such as Flan-T5 (Chung et al., 2024), demonstrate improved generalization to unseen tasks. This fine-tuning allows LLMs to better understand and respond to user requests, enhancing both zero-shot and few-shot learning capabilities. PEFT techniques have been developed to address the challenges posed by the growing number of trainable parameters in LLMs (Xu et al., 2023).

## 6 Conclusion

In this study, we aimed to improve the biomedical lay summarization of scientific publications by augmenting article text with MeSH terms. We introduced a novel method for integrating this knowledge into a generative LLM, providing guidance for background information generation through a multi-turn conversation. Our results demonstrated that MeSH terms offer a broader perspective on the content of a biomedical article, helping the model generate more focused and relevant background information specific to the article's topic.

## 7 Limitations

First, due to computational costs and memory limitations, we used only the abstract as input and tested our experimental design on a single dataset. Second, we evaluated performance based on relevance and readability metrics, as there is a lack of satisfactory evaluation for faithfulness that aligns well with human preferences, as revealed in previous studies (Fang et al., 2024). Although incorporating MeSH generation as an auxiliary task led to some improvements, its performance was not statistically significant different from the SINGLE_TURN approach. However, ablation studies indicate that MeSH selection plays a crucial role in guiding lay summary generation. In future work, we aim to further enhance its effectiveness by integrating it into the learning process with automatic feedback. Moving forward, we plan to conduct human evaluations to better assess how well model-generated summaries align with human judgments. Additionally, we will explore both closed- and open-source LLMs to evaluate the generalizability of our approach across different models.

## Acknowledgement

This work is partially supported by the National Library of Medicine (NLM) of the National Institutes of Health under the award number R01LM014292.

## References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*, 30(5):1–38.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Xiuying Chen, Tairan Wang, Qingqing Zhu, Taicheng Guo, Shen Gao, Zhiyong Lu, Xin Gao, and Xiangliang Zhang. 2024. Rethinking scientific summarization evaluation: Grounding explainable metrics on facet-aware benchmark. *arXiv preprint arXiv:2402.14359*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Ashwin Devaraj, Iain Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911.

Marcio Fonseca and Shay Cohen. 2024. Can large language model summarizers adapt to diverse scientific communication goals? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8599–8618, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2025. Leveraging large language models for zero-shot lay summarisation in biomedicine and beyond. *arXiv preprint arXiv:2501.05224*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. Enhancing biomedical lay summarisation with external knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8032.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. Ctrlsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 171:108189.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel: Inst Sim Trng*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. In *MEDINFO 2001*, pages 216–220. IOS Press.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Sanjana Ramprasad and Byron C Wallace. 2024. Do automatic factuality metrics measure factuality? a critical evaluation. *arXiv preprint arXiv:2411.16638*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Zhengyan Shi, Adam X Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction tuning with loss over instructions. *arXiv preprint arXiv:2405.14394*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. 2021. BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, 37(5):684–692.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLaySumm: An extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143, Bangkok, Thailand. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingmeng Zhao, Tongnian Wang, and Anthony Rios. 2024. Improving expert radiology report summarization by prompting large language models with a layperson summary. *arXiv preprint arXiv:2406.14500*.

| |
|---|
| **Gold Standard:** Although Parkinson's disease is the second most common neurodegenerative disorder, its cause is not known and there is no cure. The symptoms of Parkinson's disease, which include tremor and slowing of voluntary movements, get progressively worse over time. The numbers of neurons in certain brain regions also decrease, causing those parts of the brain to shrink; this is known as "atrophy". However, no conclusive signs of atrophy have been found in the brains of people in the early stages of the disease. One theory suggests that Parkinson's disease is caused by a toxic protein that is able to spread from neuron to neuron. Recent advances in brain imaging have made it possible to map networks in the living human brain—the so-called brain connectome. These networks could form the 'highways' through which a disease-causing agent might spread. The Parkinson's Progression Markers Initiative (PPMI) is a large study that collects data from hundreds of people in an effort to identify the causes of Parkinson's disease. Zeighami et al. have now analyzed MRI scans that were collected as part of this initiative, which show the structure of the brains of 230 people in the early stages of Parkinson's disease. Comparing these scans to those from age-matched healthy individuals allowed Zeighami et al. to identify the set of brain regions that show atrophy in the early stages of Parkinson's disease. These regions correspond to a normal brain network, and the relative extent of atrophy in each brain region supports the theory that the disease spreads through the connectome. The patients who were enrolled in this study will continue to be evaluated on a yearly basis. Zeighami et al. plan to continue mapping how the disease progresses throughout the brain and to relate this to the development of new symptoms of Parkinson's disease. |
| **Abstract:** We mapped the distribution of atrophy in Parkinson's disease (PD) using magnetic resonance imaging (MRI) and clinical data from 232 PD patients and 117 controls from the Parkinson's Progression Markers Initiative. Deformation-based morphometry and independent component analysis identified PD-specific atrophy in the midbrain, basal ganglia, basal forebrain, medial temporal lobe, and discrete cortical regions. The degree of atrophy reflected clinical measures of disease severity. The spatial pattern of atrophy demonstrated overlap with intrinsic networks present in healthy brain, as derived from functional MRI. Moreover, the degree of atrophy in each brain region reflected its functional and anatomical proximity to a presumed disease epicenter in the substantia nigra, compatible with a trans-neuronal spread of the disease. These results support a network-spread mechanism in PD. Finally, the atrophy pattern in PD was also seen in healthy aging, where it also correlated with the loss of striatal dopaminergic innervation. |
| **INSTRUCTION_TUNING** Parkinson's disease is a neurodegenerative disorder that affects millions of people around the world. It is characterized by the progressive loss of dopamine-producing neurons in the brain, which leads to symptoms such as tremors, rigidity, and difficulty with movement. However, it is not clear how the disease starts or how it spreads through the brain. One theory is that Parkinson's disease begins in a region of the brain called the substantianigra, which is responsible for producing dopamine. From there, the disease may spread to other brain regions via a network of nerve fibers. To better understand how Parkinson's spreads, researchers studied the brains of 232 patients with the disease and compared them to 117 healthy individuals. The brains of the patients were imaged using a technique called MRI, which allows researchers to measure the volume of different brain regions. This revealed that patients with Parkinson's had more atrophy than healthy individuals in several brain regions, including those involved in movement, emotion, and memory. Furthermore, the extent of this atrophy was related to the severity of the symptoms experienced by the patients. The brain regions affected by Parkinson's were also found to be connected to each other via nerve fibers, which supports the theory that the disease spreads through a network. The study also found that the pattern of brain atrophy seen in Parkinson's patients was similar to that seen in older adults without the disease, suggesting that age-related changes may also contribute to the development of Parkinson's. Overall, the findings of this study provide new insights into the mechanisms underlying Parkinson's, and may help researchers to develop new treatments for the disease in the future. |

Table 6: Comparison of experiments. The text highlighted in the same color conveys the same information.

| |
|---|
| **INSTRUCTION_MODELING** Parkinson's disease is a neurodegenerative disorder that affects millions of people around the world. It is characterized by the progressive loss of dopamine-producing neurons in the brain, which leads to symptoms such as tremors, stiffness and difficulty with movement. The disease is caused by the accumulation of abnormal protein clumps, known as Lewy bodies, in the cells of the brain. However, it is not clear how these clumps lead to the death of brain cells, or how they cause the symptoms of Parkinson's disease. One theory is that the clumps disrupt the connections between different brain regions, leading to a loss of communication between them. This theory is supported by the fact that some brain regions are more affected than others in Parkinson's patients. For example, the brain region known as the 'basal ganglia', which is involved in movement control, is one of the first regions to be affected by the disease, while other regions such as the hippocampus and the amygdala, which are involved in memory and emotion, are later affected. To better understand how the disease affects the brain and to develop new treatments, researchers need to map the extent of brain damage in patients with Parkinson's. One way to do this is to use a technique called MRI, which uses magnetic fields and radio waves to create detailed images of the inside of the body. Using this technique, a team of researchers has now mapped the amount of damage to the brain in over 350 patients and healthy individuals. The results show that the extent and location of the damage vary greatly from person to person, and that the brain regions that are most affected are those that are farthest from the area of brain where the disease is thought to start. This suggests that the disease spreads through the brain via a network of interconnected regions, rather than through a single pathway. The next step will be to investigate whether the same pattern of damage is seen in people who are at risk of developing Parkinson's, but have not yet been diagnosed. |
| **SINGLE_TURN** The human brain is made up of billions of neurons that communicate with each other through electrical and chemical signals. In Parkinson's disease, a protein called alpha-synuclein clumps together and forms abnormal structures called Lewy bodies, which are toxic to neurons. This leads to the death of many neurons in areas of the brain that control movement, leading to symptoms such as tremors, stiffness and slowness of movement. However, it is not clear how the disease affects the brain, or which areas are most affected. To address this question, researchers from the University of California, Los Angeles ( UCLA ) and the National Institute of Neurological Disorders and Stroke ( NINDS ) used a technique called diffusion tensor imaging ( DTI ) to map the structure of the brains of 232 people with Parkinson's, as well as 117 people who did not have the condition. The DTI scans showed that the brains affected by Parkinson's were smaller than those of people without the disease, and that the extent of this shrinkage was related to the severity of the symptoms. The researchers also found that the brain areas that were most affected by the disease were those that are most connected to each other, suggesting that the disease may spread from one area to another. Furthermore, the patterns of brain shrinkage seen in Parkinson's patients were similar to those seen in people who are aging normally. This suggests that the changes in brain structure that occur with aging may also contribute to the development of Parkinson's. The next step will be to use these findings to develop new treatments for Parkinson's that can target the specific brain regions that are affected. |
| **MESH_SINGLE_TURN** The human brain is made up of billions of neurons that communicate with each other through electrical and chemical signals. Parkinson's disease is a neurodegenerative disorder that affects millions of people worldwide. It is characterized by the progressive loss of dopamine-producing neurons in the brain, which leads to symptoms such as tremors, rigidity, and difficulty with movement. However, it is not clear how these neurons are affected by the disease, or how their loss leads to the symptoms that people with Parkinson's experience. To address this question, researchers from the University of California, Los Angeles ( UCLA ) used a technique called MRI to scan the brains of 232 people who had Parkinson's, as well as 117 people who did not have the condition. The scans were then analyzed to identify areas of the brain that were affected by Parkinson's. The results showed that the disease causes atrophy – or shrinkage – in several brain regions, including the basal ganglii, the basal fore-brain, the medial temporal lobes, and specific regions of the cortex. The extent of the shrinkage in each of these brain regions was related to the severity of the symptoms experienced by the patients. Furthermore, the regions that were most severely affected were those that are closest to the brain region known to be damaged in Parkinson's: the substantian nigra. These findings suggest that Parkinson's may be caused by a network of interconnected brain regions that are damaged as the disease progresses. The next step will be to investigate whether the same pattern of brain atrophy is seen in people who do not have Parkinson's but are aging. |

Table 7: Continued from Table 6