# Beyond Abstracts: A Biomedical MeSH Indexing Corpus Incorporating Summarized Methods Sections

**Sujoy Datta**[1], **Robert E. Mercer**[1], **Xindi Wang**[2]
[1]Department of Computer Science, University of Western Ontario, London, Ontario, Canada
[2]School of Artificial Intelligence, Shandong University, Jinan, China
sdatta46@uwo.ca, mercer@csd.uwo.ca, xindi.wang@sdu.edu.cn

**Abstract**

Automated Medical Subject Heading (MeSH) indexing systems rely predominantly on titles and abstracts, while human indexers at the National Library of Medicine examine full-text articles—particularly Methods sections—that often contain crucial experimental terminology absent from abstracts. This information asymmetry limits model performance and prevents detection of methodologically-grounded MeSH descriptors. We introduce a novel biomedical MeSH indexing corpus comprising over one million English biomedical articles, each annotated with title, abstract, journal metadata, publication year, expert-curated MeSH terms, and—uniquely—extractive summaries of Methods sections. Using LLaMA 3 with an iterative re-prompting strategy, we generated high-fidelity summaries. To avoid label leakage, evaluation labels are inferred using journal-specific MeSH frequency profiles rather than gold annotations. This publicly accessible dataset addresses a critical gap in full-text MeSH indexing research. Building upon this resource, we propose an extended multi-channel neural architecture that incorporates Methods-derived representations. Empirical results demonstrate consistent performance gains across both example-based and label-based evaluations, indicating better retrieval of infrequent terms. These findings highlight that procedural knowledge in the Methods section encodes critical semantic cues overlooked by title-abstract only models.

**Keywords:** Text Classification, MeSH Indexing, Extractive Summarization, Large Language Model

## 1. Introduction

PubMed[1] and MEDLINE[2] are among the most authoritative and extensively used repositories of biomedical literature, both curated by the United States National Library of Medicine (NLM). MEDLINE serves as the central bibliographic database, encompassing scholarly records from diverse domains within the biomedical and life sciences, including—but not limited to—clinical medicine, molecular biology, epidemiology, public health, and evolutionary sciences. According to recent statistics, MEDLINE contains over 31 million citations sourced from more than 5,200 peer-reviewed journals across the globe[2].

In 2022 alone, MEDLINE incorporated nearly one million new citations, averaging approximately 3,000 additions per day[3]. PubMed functions as the publicly accessible search interface for MEDLINE, enabling efficient retrieval of bibliographic metadata and, where available, access to full-text articles via linked repositories such as PubMed Central.

Central to MEDLINE's indexing framework is the use of **Me**dical **S**ubject **H**eadings (MeSH), a hierarchically structured and manually curated controlled vocabulary developed by the NLM. MeSH enables semantically consistent categorization of biomedical concepts across varying levels of abstraction. As of the 2019 release, the thesaurus comprised 28,939 descriptors, including 29 check tags—specialized terms used to denote fundamental study attributes such as organism type, demographic factors, or experimental conditions.

MeSH constitutes the cornerstone of NLM's indexing infrastructure. It is employed by both human indexers and, since April 2024, by the MTIX (Medical Text Indexer–NeXt Generation)[4] system, which now automates preliminary annotation for all incoming MEDLINE citations. While MTIX generates candidate labels based primarily on titles and abstracts, human indexers remain responsible for quality assurance, reviewing and finalizing the assigned descriptors (Fernandez-Llimos et al., 2024).

Historically, MeSH term assignment relied almost exclusively on manual review conducted by expert annotators, who examined full-text articles to determine concept relevance. However, this process is both labor-intensive and financially costly, with the average annotation estimated at $9.40 per article (Wang et al., 2022). Given the scale at which new literature enters MEDLINE—often exceeding 3,000 articles per day—fully manual indexing is no longer viable as a standalone strategy. Hence, the development of scalable, high-fidelity automated MeSH indexing methods has become a critical priority.

[1]pubmed.ncbi.nlm.nih.gov/about/
[2]www.nlm.nih.gov/medline/medline_overview.html
[3]www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
[4]www.nlm.nih.gov/pubs/techbull/ma24/ma24_mtix.html



*Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026*, pages 32–43
12 May 2026. 

Despite significant advances in neural architectures for MeSH indexing, most automated systems remain constrained by the limited scope of input data available during training. Existing public corpora overwhelmingly provide only titles and abstracts, whereas human indexers at the National Library of Medicine routinely examine the full text when assigning descriptors. This discrepancy creates an inherent information asymmetry: crucial contextual cues—often embedded in structurally informative sections such as Methods, Results, or Discussion—are inaccessible to current models. Prior work by Mork et al. (2017) underscores that the Methods section, in particular, contains experimentally grounded terminology that can substantially enhance indexing accuracy. Our own preliminary analysis supports this observation. For example, in the biomedical article with PMID 27456232[5], the assigned MeSH term "Cholecalciferol" appears exclusively within the methodology description and is absent from both the title and abstract. Such cases illustrate that models restricted to title–abstract inputs are structurally incapable of recovering many legitimate MeSH terms. Consequently, there is a clear need for a publicly accessible corpus that includes other sections paired with human-curated MeSH annotations to facilitate the development of next-generation context-aware indexing systems.

In this study, we introduce a newly curated open-access MeSH indexing dataset[6] in which each entry integrates the PubMed identifier (PMID), article title, abstract, journal metadata, publication year, annotated MeSH terms, and the corresponding summarized Methods section. Building upon this resource, we further propose an extended multi-channel architecture that incorporates feature representations derived from multiple document sections. We benchmark this enhanced framework and report its baseline performance as a foundation for future research in full-text-aware MeSH indexing.

## 2. Related Work

A variety of biomedical corpora have been developed to facilitate MeSH-related research. Early resources such as OHSUMED (Hersh et al., 1994) and GENIA (Kim et al., 2003) provided MEDLINE abstracts annotated with MeSH-derived concepts, primarily supporting document retrieval and entity recognition. Subsequent corpora such as CHEMDNER (Krallinger et al., 2015), BC5CDR (Li et al., 2015), and NLM-CHEM (Islamaj et al., 2021) focused more narrowly on chemical and disease annotation. While valuable, these datasets are restricted to abstracts or specific entity types and thus do not fully capture the complexity of large-scale MeSH indexing. In contrast, MeSHup (Wang et al., 2022) is the first corpus to provide full-text articles with expert-assigned MeSH terms, enabling exploitation of richer sections such as Methods and Results.

Automatic MeSH indexing has evolved from heuristic systems to increasingly sophisticated neural architectures. Early efforts such as MTI (Aronson et al., 2004) relied on UMLS-based concept mapping and k-NN retrieval, while subsequent learning-to-rank approaches like MeSHLabeler (Liu et al., 2015) and DeepMeSH (Peng et al., 2016) combined multiple evidence sources through weighted ranking schemes. With the rise of deep learning, attention-based encoders over titles and abstracts, for example AttentionMeSH (Jin et al., 2018) and MeSHProbNet (Xun et al., 2019), demonstrated that concise textual inputs already offer strong predictive signals. Full-text methods such as FullMeSH (Dai et al., 2020) and BERTMeSH (You et al., 2021) further incorporated additional sections beyond abstracts, though typically through late fusion or uniform context modeling. KenMeSH (Wang et al., 2022) advanced this line by combining masked attention with GCN-based label embeddings, yet still restricted its representation space to title and abstract. More recently, the National Library of Medicine replaced its long-standing rule-based MTIA system with MTIX, a neural indexing model trained on millions of MEDLINE citations. While MTIX marks an important acknowledgment of neural architectures for large-scale indexing, its deployment remains opaque and limited to title–abstract inputs. Furthermore, the model is not publicly released, hindering reproducibility and preventing the community from evaluating its performance beyond NLM's internal metrics.

Most existing MeSH indexing systems rely solely on titles and abstracts. Full-text models such as FullMeSH and BERTMeSH have shown that incorporating additional article sections can yield notable performance gains, reinforcing the value of richer document structure. However, their underlying corpora were not released, limiting reproducibility and fair comparison. This highlights the pressing need for publicly accessible, section-structured biomedical datasets to enable systematic evaluation of multi-section MeSH indexing approaches.

## 3. Methodology

In this section, we outline the overall system design and introduce the deep learning architecture

[5] pubmed.ncbi.nlm.nih.gov/27456232/

[6] https://doi.org/10.5281/zenodo.19221828

determined to be most effective for our task. We first describe the dataset construction process, emphasizing the practical challenges encountered and the strategies adopted to address them—a contribution that stands on its own merit. We then detail the model architecture, with a focus on the role and interaction of its key components.

## 3.1. Dataset Construction

We build our dataset using the MeSHup corpus (Wang et al., 2022), which is derived from the November 2021 release of BioC-PMC (Comeau et al., 2019). The corpus contains 1,342,667 full-text biomedical articles in English, each annotated with MeSH terms assigned by professional indexers, providing reliable supervision for downstream learning tasks. In addition to the manual MeSH annotations, each record includes rich metadata—such as publication venue, author list, and publication year—stored in a structured JSON format.

For our work, we extracted seven key fields: PMID, title, abstract, Methods section, journal name, publication year, and the associated MeSH terms (Figure 1). Analysis of the Methods field revealed 1,135,757 articles with valid content, with lengths ranging from 8 to 645,880 characters. To ensure consistency in document representation, we retained only entries with non-empty title, abstract, and Methods sections for further processing and model training.


```
{
  "pmid": "10023767",
  "title": "An antiviral mechanism of nitric oxide: inhibition of a
     viral protease.",
  "abstractText": "Although nitric oxide (NO) kills or inhibits the
     replication of a variety of intracellular pathogens, the
     antimicrobial mechanisms of NO are unknown. Here, we identify a
      viral protease as a target of NO. ...",
  "mesh": {
    "D019943": "Amino Acid Substitution",
    "D000998": "Antiviral Agents",
    "D001665": "Binding Sites",
    "D003545": "Cysteine",
    "D003546": "Cysteine Endopeptidases",
    ...
  },
  "journal": "Immunity",
  "year": "1999",
  "METHODS": "Experimental Procedures. Cells and Viruses.
     Coxsackievirus B3 (CVB3) ... Measurement of 3Cpro Activity. ...
      Determination of S-Nitrosylation of 3Cpro. ..."
}
```


Figure 1: Sample Record Extracted from the MeSHup Corpus.

### 3.1.1. Requirement of Text Summarization

Titles and abstracts in academic publications are typically constrained by journal policies, resulting in relatively uniform length and structure across articles. Abstracts are commonly limited to 150–250 words to ensure clarity and conciseness (Pottier et al., 2024), while titles are similarly regulated through character limits to aid discoverability[7]. In contrast, Methods sections are rarely subject to strict length constraints and therefore exhibit substantial variability depending on experimental complexity. A large-scale analysis[8] of over 61K articles reported lengths ranging from 7 to 18,517 words, with a median of 1,126 words, a trend also observed in our dataset.

Although such detail in the Method section is essential for transparency and reproducibility, their length makes them impractical for efficient retrieval and downstream MeSH-based classification. Summarization offers a pragmatic compromise by retaining semantically relevant content while discarding redundant material. Prior work shows that summaries maintain task-relevant information (Mandale-Jadhav, 2025), lower computational cost during processing (Luo et al., 2024), and, in MeSH indexing settings, can even outperform full-text inputs for concept detection (Jimeno-Yepes et al., 2013). These observations motivate our decision to adopt summarization as a central component of our pipeline rather than treating it as an auxiliary preprocessing step.

### 3.1.2. Selection of Summarization Technique

Text summarization methods are commonly divided into extractive and abstractive paradigms. Extractive methods identify salient sentences from the source and concatenate them without altering phrasing, while abstractive methods generate rephrased summaries through intermediate representations (Luo et al., 2024). In this work, we adopt the extractive paradigm, as abstractive summarization requires sophisticated linguistic modeling—ranging from syntax and semantics to discourse planning—which makes it considerably more resource-intensive to implement reliably (Moradi and Ghadiri, 2019).

Extractive summarization further offers practical advantages for classification tasks: it acts as an implicit feature selection mechanism by reducing input length and thereby lowering the dimensionality of the representation space. This leads to improved computational efficiency and more interpretable downstream models (Kolcz et al., 2001; Ghodratnama et al., 2020; Nallapati et al., 2017).

Summarization techniques can also be categorized as supervised or unsupervised. Supervised approaches rely on human-annotated reference summaries and achieve strong performance when such resources exist (Basyal and Sanghvi, 2023). Unsupervised methods, in contrast, operate without labeled summaries by ranking and

[7] https://pmc.ncbi.nlm.nih.gov/articles/PMC6942168

[8] https://quantifyinghealth.com/methods-section-length/

selecting sentences using statistical or semantic heuristics. This makes them particularly suitable when curated summaries are unavailable (Basyal and Sanghvi, 2023). Because our corpus does not contain human-generated summaries, and creating them at scale would be prohibitively expensive, we adopt an unsupervised strategy.

Recent advances in Large Language Models (LLMs) have further expanded the capabilities of unsupervised summarization. LLMs have been shown to produce summaries that are both preferred by human evaluators and more factually consistent than traditional systems, even when guided by minimal prompting (Goyal et al., 2022). Motivated by these findings, we employ an LLM-based summarization approach. Among available locals models, LLaMA 3 demonstrated the best trade-off between fluency, efficiency, and instruction adherence for our task (Dubey et al., 2024). In this study, we utilized the LLaMA 3 8B parameter model for summarization tasks. The model was deployed and executed locally using the Ollama[9] framework, enabling controlled and efficient inference without reliance on external APIs.

### 3.1.3. Summarization Process

Our summarization pipeline relies on MeSH terms as guidance signals for extracting relevant sentences from the Methods section of biomedical articles. The objective is to isolate sentences that explicitly reference, or are strongly associated with, the MeSH labels assigned to each record. To prototype this approach, we conducted preliminary experiments on 20 randomly selected samples to assess the behavior of the language model under different prompt formulations. The initial prompt design (Appendix A) instructed the model to return all matching sentences as a single paragraph; however, responses were frequently inconsistent in both structure and content.

In particular, the model often (i) returned unordered lists rather than continuous paragraphs, (ii) appended explanatory statements about the MeSH terms that were not part of the original text, (iii) duplicated sentences when multiple MeSH terms appeared within them, and (iv) paraphrased or partially reworded sentences instead of preserving them verbatim. As our aim was to retain the original textual form, these deviations were undesirable.

To mitigate these issues, we iteratively refined the prompt through a prescription-oriented strategy. The final version (Figure 2) explicitly enforced five constraints: (1) extract only sentences containing target MeSH terms or related expressions, (2) preserve sentence form exactly, (3) avoid duplication, (4) suppress commentary or labels in the output, and (5) return a single uninterrupted paragraph. This prompt was subsequently applied independently to each Methods section across the dataset.

*Extract all original sentences from the following text that contain any of the MeSH terms listed or their related words.*

*Text:*
*Experimental Procedures . Cells and Viruses .*
*Coxsackievirus B3 (CVB3) ... Measurement of 3Cpro Activity*
*... Determination of S- Nitrosylation of 3Cpro. ...*

*MeSH Terms:*
*Amino Acid Substitution, Antiviral Agents, Binding Sites, Cysteine,*
*Cysteine Endopeptidases*

*Task:*
*Your job is to extract all original sentences from the provided text that contain any of the MeSH terms listed or their related words. Ensure the following:*
*1. Sentences must be extracted in their original form without modification.*
*2. A sentence should only appear once, even if it contains multiple MeSH terms.*
*3. Matches should be case-insensitive but must respect the capitalization of the original text.*
*4. Do not include MeSH Terms from the list in output paragraph.*
*5. Return the extracted sentences as a paragraph.*

*Output Format:*
*- As a paragraph. No additional commentary or formatting.*

Figure 2: Final Prompt

Despite these refinements, two well-documented limitations of Large Language Models persisted. First, LLMs are prone to format drift, occasionally ignoring explicit output constraints (Tam et al., 2024). As illustrated in Figure 3, the model sometimes omitted header phrases or deviated from the expected textual shape. Second, the model often compressed sentences via ellipses or syntactic truncation, which is unsuitable for our objective of preserving complete, contextually intact statements.

| Unusual Output (No usual header) | Shrunken Output | Expected Output |
|---|---|---|
| The cDNA encoding CVB3 3Cpro was cloned between the NdeI and BamHI sites of the expression vector pSG04 so that a (His)6 amino terminal tag is fused to 3Cpro. | Here is the extracted paragraph with the original sentences containing Mesh terms:<br><br>The cDNA encoding CVB3 3Cpro was cloned ... terminal tag is fused to 3Cpro. | Here is the extracted paragraph with the original sentences containing Mesh terms:<br><br>The cDNA encoding CVB3 3Cpro was cloned between the NdeI and BamHI sites of the expression vector pSG04 so that a (His)6 amino terminal tag is fused to 3Cpro. |

Figure 3: Different Output Formats from LLM

To ensure output fidelity, we implemented an automated remediation mechanism: whenever the generated summary violated structural or completeness criteria, the model was re-prompted until a fully compliant version was obtained. Valid outputs were then parsed and stored for downstream integration. During the summarization process, Approximately 30% of instances required

[9] https://ollama.com/

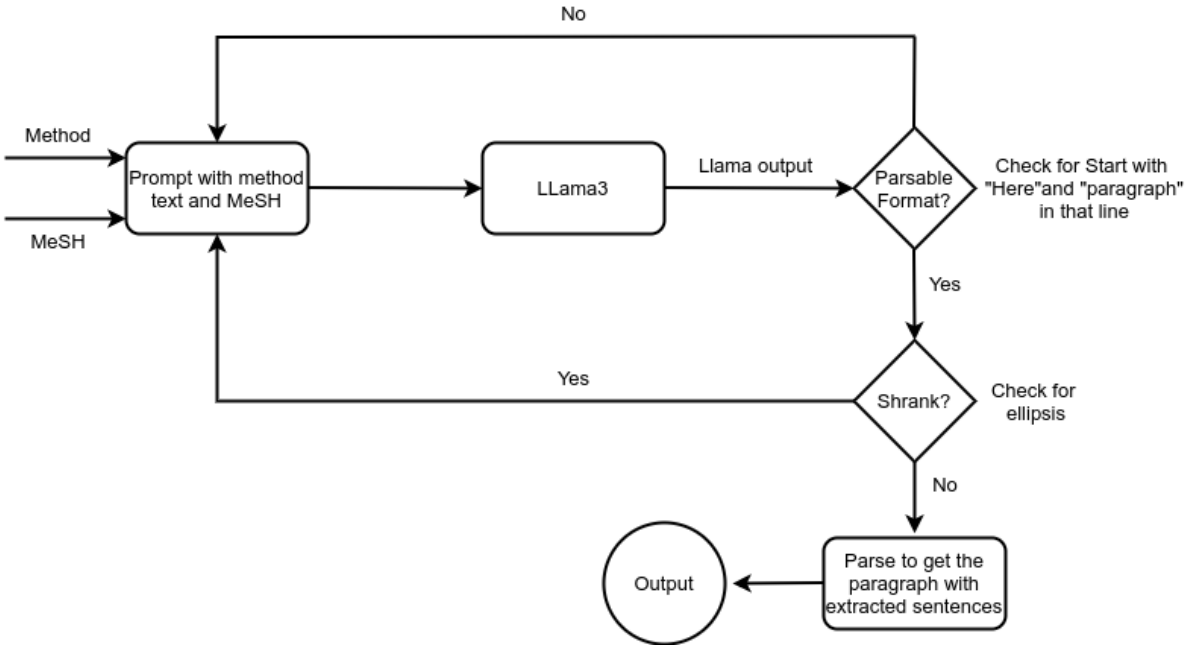


Figure 4: Summarization Pipeline

at least one re-prompt, with a small number of cases necessitating more than three iterations to obtain acceptable outputs. The end-to-end workflow—from prompt refinement to reprompting and persistence—is summarized in Figure 4.

For training instances, we used the human-annotated MeSH terms from MeSHup as guidance labels. However, using these same annotations during evaluation would constitute label leakage, since MeSH assignment is precisely the target of the predictive model. An early attempt to provide all $\sim$28,000 MeSH terms to the model alongside full Methods text proved infeasible: the excessive label set induced hallucinations and degraded the model's ability to maintain semantic coherence.

To balance coverage and reliability, we adopted an alternative strategy for evaluation: rather than supplying gold-standard annotations, we inferred journal-specific MeSH label sets based on metadata. Specifically, for each journal, we selected MeSH terms with a frequency threshold greater than 0.0345, resulting in approximately 100 dominant terms per journal. The threshold was determined empirically through preliminary analysis on 20 randomly sampled articles. We observed that substantially lower thresholds—which increased the number of candidate MeSH terms—led to prompt overload, causing the LLM to exhibit hallucination behavior, including reproducing large portions of the prompt or returning irrelevant lists of MeSH terms rather than extracting meaningful sentences. The selected threshold therefore reflects a practical trade-off between label coverage and generative stability.

These derived label sets were then used to guide summarization during validation and testing, enabling label-aware extraction without exposing ground truth. This approach provides a more realistic simulation of deployment in environments where annotations are sparse or unavailable. The corpus comprises a total of 28,070 unique MeSH terms. Following the dataset split strategy of Wang et al. (2022), we allocated 17,918 articles to the test set, 15,000 articles to the validation set, with the remaining instances reserved for training.

## 3.2. Model

We build upon KenMeSH (Wang et al., 2022), a state-of-the-art MeSH indexing framework, and extend it to incorporate methodological text. Below, we briefly summarize the base model, outline our modifications, and present the final architecture.

### 3.2.1. KenMeSH Overview

KenMeSH consists of three modules: (1) a dual-channel encoder for title and abstract using BioWordVec embeddings, BiLSTMs, and a dilated CNN (for abstract only); (2) a masked attention mechanism that uses journal metadata and $k$-nearest neighbour (KNN) label statistics to focus on label-relevant text spans; and (3) a two-layer Graph Convolutional Network (GCN) over the MeSH ontology to learn hierarchical label embeddings. Predictions are computed via similarity between document and label embeddings.

#### 3.2.2. Modifications

We introduce two extensions:

**(i) New input stream.** A third channel is added to encode the Methods section using the same BiLSTM+CNN structure as the abstract encoder.

**(ii) Joint attention.** The masked attention module is revised to attend to both abstract and method channels for improved label-specific focus.

#### 3.2.3. Enhanced Architecture

The final model (Figure 5) has four modules:

**Multi-channel Encoding** Title, abstract, and method inputs, each embedded as $\mathbf{E}$, encoded via:

$$\mathbf{H}_{\text{title}},\ \mathbf{H}_{\text{abstract}},\ \mathbf{H}_{\text{method}} = \text{BiLSTM}(\mathbf{E})$$

A dilated CNN is applied to $\mathbf{H}_{\text{abstract}}$ and $\mathbf{H}_{\text{method}}$ to obtain context-rich representations $\mathbf{D}_{\text{abstract}}$ and $\mathbf{D}_{\text{method}}$.

**Label Encoding via GCN** Each MeSH descriptor is embedded via averaged token embeddings and refined through a 2-layer GCN over the ontology:

$$\mathbf{H}_{\text{label}} = [\mathbf{v} : \text{GCN}(\mathbf{v})]$$

**Dynamic Masked Attention** Using journal-conditioned $P(L|J_j)$ statistics and KNN-based aggregation of neighbour labels a document-specific label mask is built. Masked attention computes:

$$\alpha_{\text{ch}} = \text{Softmax}(\mathbf{H}_{\text{ch}} \cdot \mathbf{H}_{\text{label}})$$

$$\mathbf{c}_{\text{ch}} = \alpha_{\text{ch}}^{\top}\mathbf{H}_{\text{ch}}, \quad \text{ch} \in \{\text{title}, \text{abstract}, \text{method}\}$$

$$\mathbf{D} = \mathbf{c}_{\text{title}} + \mathbf{c}_{\text{abstract}} + \mathbf{c}_{\text{method}}$$

**Classification** Document–label similarity is computed via:

$$\hat{y}_i = \sigma(\mathbf{D} \odot \mathbf{H}_{\text{label}})$$

with binary cross-entropy as training objective.

## 4. Experiment

MeSH indexing is commonly formulated as a multi-label text classification problem. Let $X = x_1, x_2, \ldots, x_N$ denote a collection of biomedical documents and $Y = y_1, y_2, \ldots, y_L$ denote the vocabulary of MeSH labels. The objective is to learn a function $f : X \to [0,1]^L$ that assigns to each document $x_i$ a vector of label likelihoods, where each dimension corresponds to the probability of assigning label $y_j$. The training corpus is defined as $\mathcal{D} = (x_i, Y_i)_{i=1}^{N}$, where $Y_i \subseteq Y$ denotes the ground-truth set of MeSH terms for $x_i$. When representing each document as a token sequence embedded in a matrix of size $n \times e$ (with $n$ tokens and $e$-dimensional embeddings), the goal is to learn a mapping $f$ that generalizes to unseen instances such that $f(x_k)$ accurately predicts the true label set $Y_k$ for any new document $x_k$ (Wang et al., 2022).

### 4.1. Implementation Details

Models were implemented in PyTorch (Paszke et al., 2019) and trained on a 48-core CPU with an NVIDIA RTX A6000 GPU (48 GB). We first trained the original KenMeSH configuration (title + abstract only), then trained our extended model by adding a third channel for methodological summaries, which consistently improved performance.

Text inputs were lowercased and normalized by removing punctuation, non-alphanumeric tokens, stop words, and single-character tokens. Word representations were initialized with 200-dimensional BioWordVec (Zhang et al., 2019) embeddings, followed by 0.2 dropout and early stopping (Yao et al., 2007). Each encoder used 200 hidden units and a 3-layer dilated CNN with dilation rates of $[1, 2, 3]$.

For label masking, we retrieved the top 1000 nearest neighbors via FAISS (Johnson et al., 2019). Models were optimized using Adam (Kingma and Ba, 2015) with a learning rate of $0.0003$ (exponentially decayed by 0.9 per epoch), gradient clipping at norm 5, and batch size 32. End-to-end training required approximately 3 days. All experiments were conducted over three independent runs. Reported results correspond to the mean performance and standard deviation, reflecting inter-run variability and model stability.

In the subsequent stage, we conducted randomized hyperparameter search to identify the configuration yielding the best overall performance. The optimization process focused on parameters known to strongly affect convergence behavior and generalization in convolution-based architectures. Table 1 summarizes the candidate search space along with the final selected values corresponding to the optimal configuration. Hyperparameter optimization was conducted using a held-out validation set comprising 15,000 instances. Model configurations were selected by maximizing the micro-averaged F1 score on the validation data, as micro-F1 is a widely adopted and robust evaluation metric in multi-label classification settings (Pal et al., 2020).

### 4.2. Model Evaluation Techniques

We evaluate MeSH indexing performance using standard bipartition-based metrics (Tsoumakas et al., 2010), considering both instance-level and label-level prediction quality. Example-based evaluation measures correctness per document using precision, recall, and F-score computed over the predicted and ground-truth label sets. Given $y_i$

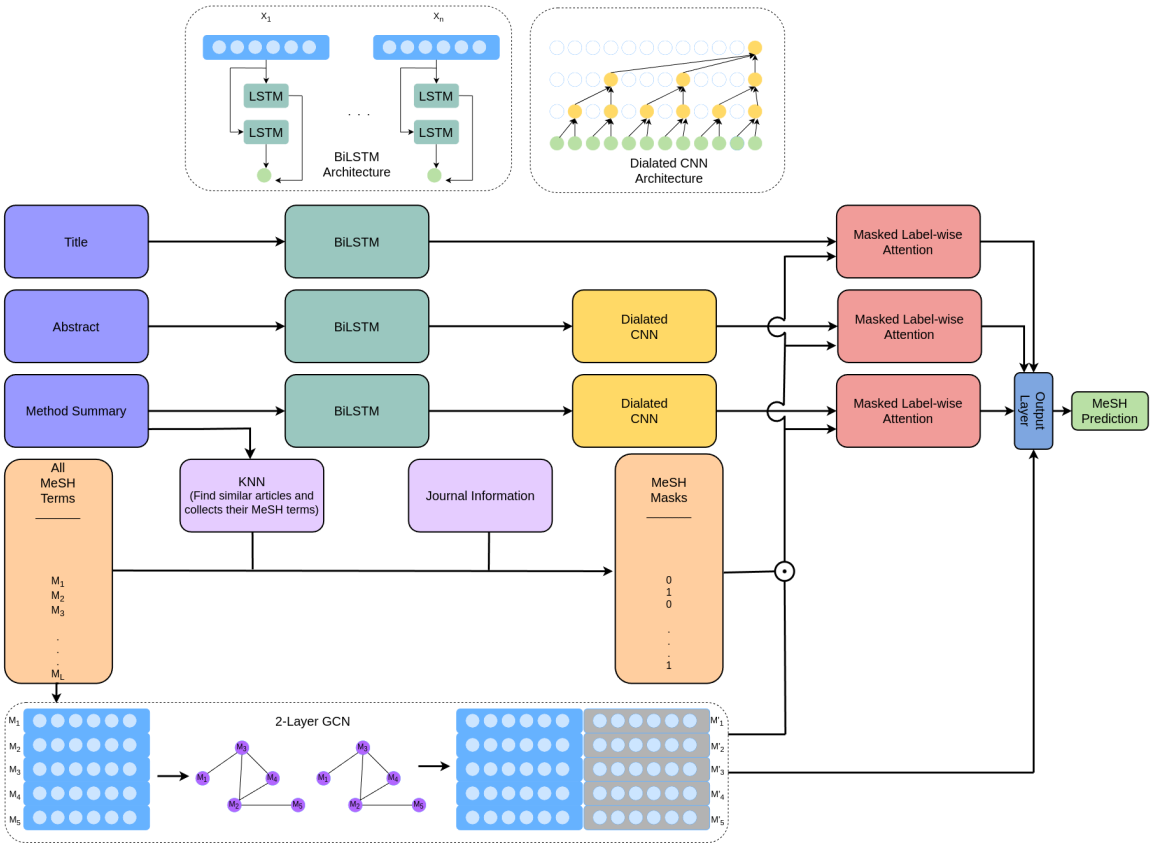


Figure 5: Architecture of Our Model

| Hyperparameters | Values |
|---|---|
| Learning Rate | $0.001, 0.0001,$ **0.0003**, $0.0005$ |
| Decay Rate | $0.8$, **0.9** |
| Dropout Probability | **0.2**, $0.5$ |
| Batch Size | $8, 16,$ **32** |
| Number of dilated CNN layers | **3**, $4, 5, 6$ |

Table 1: Hyperparameter settings used for optimization. Selected parameter values are in bold.

and $\hat{y}_i$ as the gold and predicted labels for instance $i$, the metrics are defined as follows:

$$\text{EBP} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|\hat{y}_i|} \quad (1)$$

$$\text{EBR} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i|} \quad (2)$$

$$\text{EBF} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \cdot |y_i \cap \hat{y}_i|}{|y_i| + |\hat{y}_i|} \quad (3)$$

To evaluate performance at the label level, we consider two common aggregation strategies: macro-averaging and micro-averaging. Macro-averaging computes metrics independently for each label and then averages them uniformly, thereby treating all labels equally regardless of their frequency. In contrast, micro-averaging aggregates true positives, false positives, and false negatives across all labels before computing precision and recall, which implicitly assigns greater weight to high-frequency labels (Yang, 1999). Let $\text{TP}_j$, $\text{FP}_j$, and $\text{FN}_j$ denote the true positives, false positives, and false negatives, respectively, for label $l_j$, and let $L$ denote the total number of labels. The metrics are then defined as:

$$\text{Macro-P} = \frac{1}{L}\sum_{j=1}^{L}\frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j} \quad (4)$$

$$\text{Micro-P} = \frac{\sum_{j=1}^{L}\text{TP}j}{\sum j = 1^{L}(\text{TP}_j + \text{FP}_j)} \quad (5)$$

$$\text{Macro-R} = \frac{1}{L}\sum_{j=1}^{L}\frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j} \quad (6)$$

$$\text{Micro-R} = \frac{\sum_{j=1}^{L}\text{TP}j}{\sum j = 1^{L}(\text{TP}_j + \text{FN}_j)} \quad (7)$$

$$\text{Macro-F1} = 2 \cdot \frac{\text{Macro-P} \cdot \text{Macro-R}}{\text{Macro-P} + \text{Macro-R}} \quad (8)$$

$$\text{Micro-F1} = 2 \cdot \frac{\text{Micro-P} \cdot \text{Micro-R}}{\text{Micro-P} + \text{Micro-R}} \quad (9)$$

Since threshold selection critically affects bipartition outcomes, we assign an individual decision threshold $\tau_i$ per label and predict label $i$ as positive when $\hat{p}_i \geq \tau_i$. Thresholds are optimized using the micro-F maximization strategy of Pal et al. (2020):

$$\tau_i = \arg\max_T \text{MiF}(T) \quad (10)$$

where $T$ represents the set of candidate threshold values for label $i$.

### 4.3. Experimental Results and Analysis

Table 2 compares the performance of two model configurations: one operating solely on title and abstract inputs, and another augmented with methodology-derived representations. Across nearly all bipartition-based evaluation metrics, the inclusion of methodology information yields consistent improvements. Example-based F1 (EBF) increases from 0.53 to 0.56, primarily driven by an increase in recall (0.43 to 0.50), while precision decreases slightly (0.68 to 0.65), suggesting that additional contextual evidence enables broader coverage of relevant labels at the cost of marginal over-generation. A similar pattern is observed in micro-averaged metrics, where micro-F1 improves from 0.54 to 0.56 due to a gain in micro-recall, indicating enhanced retrieval of positive labels across frequent categories.

More notably, gains are amplified in macro-level metrics—Macro-F1 rises from 0.18 to 0.22—indicating improved detection of less frequent labels. This is further corroborated by the increase in the number of distinct MeSH terms detected during inference, from 7,740 to 9,596, suggesting that methodology-aware representations facilitate coverage of a broader portion of the long-tail label space.

However, these results must be interpreted in the context of the highly skewed label distribution characteristic of MeSH indexing. High-frequency labels (e.g., “Humans”, “Male”, “Female”, and “Animals”) dominate the training signal, enabling the model to achieve high confidence on these ubiquitous categories, which disproportionately inflates precision- and recall-based metrics. Conversely, rare and domain-specific MeSH terms remain underrepresented, limiting the model’s ability to generalize beyond dominant patterns. As such, improvements in macro-averaged scores and unique label coverage provide a more meaningful indication of progress than micro-level metrics alone.

| **Bipartition Evaluation** | **Title and Abstract** | **Title, Abstract and Methodology** |
|---|---|---|
| EBF | 0.53 ($\pm$0.014) | **0.56** ($\pm$0.026) |
| EBP | **0.68** ($\pm$0.007) | 0.65 ($\pm$0.018) |
| EBR | 0.43 ($\pm$0.021) | **0.5** ($\pm$0.025) |
| Micro-F1 | 0.54 ($\pm$0.008) | **0.56** ($\pm$0.006) |
| Micro-P | **0.69**($\pm$0.009) | 0.65 ($\pm$0.027) |
| Micro-R | 0.44 ($\pm$0.007) | **0.49** ($\pm$0.023) |
| Macro-F1 | 0.18 ($\pm$0.008) | **0.22** ($\pm$0.016) |
| Macro-P | 0.22 ($\pm$0.011) | **0.25** ($\pm$0.021 ) |
| Macro-R | 0.12 ($\pm$0.018) | **0.2** ($\pm$0.038) |
| Unique MeSH terms detected | 7740 ($\pm$44) | **9596** ($\pm$102) |

Table 2: Evaluation Metrics Comparison Across Two Models

Most importantly, Methodological knowledge enables the model to correctly identify MeSH terms that appear exclusively—or more prominently—in sections beyond the title and abstract. For instance, in the biomedical article with PMID 27529679[10], the MeSH term “Agmatine” is clearly mentioned in both the abstract and the methodology section. However, the baseline KenMeSH configuration operating solely on title and abstract failed to assign this label. In contrast, the methodology-augmented variant successfully detected Agmatine, demonstrating that explicit methodological grounding reinforces recognition even when the term is present in multiple sections but under-emphasized in the abstract.

More critically, methodological content facilitates recovery of MeSH terms that are entirely absent from the title-abstract space. In the article with PMID 27472359[11], the assigned MeSH term “Nebraska” appears only within the Methods section. The title and abstract offer no lexical cues that could be leveraged by conventional title-abstract–based models. Yet, our multi-channel system correctly predicted this label by attending to geographic and procedural mentions embedded in methodological descriptions.

These case studies highlight inclusion of methodological evidence not only broadens label coverage but also enables inference of contextually grounded MeSH descriptors that are otherwise inaccessible.

[10] pubmed.ncbi.nlm.nih.gov/27529679/
[11] pubmed.ncbi.nlm.nih.gov/27472359/

## 5. Conclusion and Future Work

In this study, we presented a novel MeSH indexing dataset that, for the first time at scale, integrates summarized methodological content in addition to the widely used title and abstract fields. Summaries of the Methods sections were generated locally using the LLaMA3 model via the open-source Ollama framework, enabling full control over generation behavior without dependence on external APIs. To ensure consistency and reliability, we employed a structured prompt design coupled with an iterative re-prompting strategy to enforce format fidelity and mitigate common failure cases such as truncation or hallucination. To prevent label leakage during evaluation, MeSH terms for the test set were derived using journal-specific frequency thresholds rather than ground-truth annotations.

Leveraging this enriched dataset, we developed an extended multi-channel neural architecture designed to incorporate section-specific representations. Empirical results demonstrate that the inclusion of methodological summaries yields consistent performance gains across both example-based and label-based metrics, with particularly notable improvements in macro-level scores—highlighting enhanced robustness on infrequent labels. These observations substantiate our central claim that critical domain-specific cues frequently reside in procedural descriptions that are not captured by title-and-abstract-only models, underscoring the necessity of full-text-aware MeSH indexing frameworks.

While the methodology section offers focused technical insights that are directly relevant to MeSH term assignment, expanding the summarization process to include other key sections of biomedical papers—such as the introduction, results, and discussion—may further enrich the textual representation space. These sections frequently contain broader contextual framing, experimental outcomes, and interpretative commentary that can support the inference of higher-level or cross-disciplinary MeSH descriptors. Future research may therefore investigate hierarchical or section-aware summarization strategies that dynamically weight contributions from different sections based on their relevance to specific categories of MeSH terms.

## 6. Acknowledgements

We would like to thank all reviewers for their comments. This research is partially funded by The Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to R. E. Mercer.

## 7. Bibliographical References

Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. 2004. The nlm indexing initiative's medical text indexer. In *MEDINFO 2004*, pages 268–272. IOS Press.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: a comparative study of MPT-7B-Instruct, Falcon-7B-Instruct, and OpenAI Chat-GPT models. *arXiv preprint arXiv:2310.10449*.

Suyang Dai, Ronghui You, Zhiyong Lu, Xiaodi Huang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2020. Fullmesh: improving large-scale mesh indexing with full text. *Bioinformatics*, 36(5):1533–1541.

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*.

Fernando Fernandez-Llimos, Luciana G Negrão, Christine Bond, and Derek Stewart. 2024. Influence of automated indexing in medical subject headings (MeSH) selection for pharmacy practice journals. *Research in Social and Administrative Pharmacy*, 20(9):911–917.

Samira Ghodratnama, Amin Beheshti, Mehrdad Zakershahrak, and Fariborz Sobhanmanesh. 2020. Extractive document summarization based on dynamic feature space mapping. *IEEE Access*, 8:139084–139095.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, et al. 2021. Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. *Scientific data*, 8(1):91.

Antonio J Jimeno-Yepes, Laura Plaza, James G Mork, Alan R Aronson, and Alberto Díaz. 2013. MeSH indexing based on automatically generated summaries. *BMC Bioinformatics*, 14:1–12.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. Attentionmesh: Simple,

effective and interpretable automatic mesh indexer. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 47–56.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*.

Aleksander Kolcz, Vidya Prabakarmurthi, and Jugal Kalita. 2001. Summarization as feature selection for text categorization. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 365–370.

Jiao Li, Yueping Sun, R Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2015. Annotating chemicals, diseases, and their interactions in biomedical literature. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 173–182. The Fifth BioCreative Organizing Committee.

Ke Liu, Shengwen Peng, Junqiu Wu, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2015. Meshlabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347.

Mengqi Luo, Bowen Xue, and Ben Niu. 2024. A comprehensive survey for automatic text summarization: Techniques, approaches and perspectives. *Neurocomputing*, 603:128280.

Ashwini Mandale-Jadhav. 2025. Text summarization using natural language processing. *Journal of Electrical Systems*, 20:3410–3417.

Milad Moradi and Nasser Ghadiri. 2019. Text summarization in the biomedical domain. *arXiv preprint arXiv:1908.02285*.

James Mork, Alan Aronson, and Dina Demner-Fushman. 2017. 12 years on–is the nlm medical text indexer still useful and relevant? *Journal of biomedical semantics*, 8(1):8.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ankit Pal, Muru Selvakumar, and Malaikannan Sankarasubbu. 2020. Multi-label text classification using attention-based graph neural network. *arXiv preprint arXiv:2003.11644*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2016. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics*, 32(12):i70–i79.

Patrice Pottier, Malgorzata Lagisz, Samantha Burke, Szymon M Drobniak, Philip A Downing, Erin L Macartney, April Robin Martinig, Ayumi Mizuno, Kyle Morrison, Pietro Pollo, et al. 2024. Title, abstract and keywords: a practical guide to maximize the visibility and impact of academic papers. *Proceedings Biological Sciences*, 291(2027):20241222.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? A study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer.

Xindi Wang, Robert E. Mercer, and Frank Rudzicz. 2022. KenMeSH: Knowledge-enhanced end-to-end biomedical text labelling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2941–2951.

Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2019. Meshprobenet: a self-attentive probe net for mesh indexing. *Bioinformatics*, 35(19):3794–3802.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1:69–90.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive approximation*, 26(2):289–315.

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.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):52.

## 8. Language Resource References

Comeau, Donald C and Wei, Chih-Hsuan and Islamaj Doğan, Rezarta and Lu, Zhiyong. 2019. *PMC text mining subset in BioC: about three million full-text articles and growing*. Oxford University Press. PID https://arxiv.org/abs/1804.05957.

Hersh, William and Buckley, Chris and Leone, TJ and Hickam, David. 1994. *OHSUMED: An interactive retrieval evaluation and new large test collection for research*. Springer. PID https://doi.org/10.1007/978-1-4471-2099-5_20.

Kim, J-D and Ohta, Tomoko and Tateisi, Yuka and Tsujii, Jun'ichi. 2003. *GENIA corpus—a semantically annotated corpus for bio-textmining*. Oxford University Press, ISLRN 905-770-498-692-0.

Krallinger, Martin and Rabal, Obdulia and Leitner, Florian and Vazquez, Miguel and Salgado, David and Lu, Zhiyong and Leaman, Robert and Lu, Yanan and Ji, Donghong and Lowe, Daniel M and others. 2015. *The CHEMDNER corpus of chemicals and drugs and its annotation principles*. Springer. PID https://doi.org/10.1186/1758-2946-7-S1-S2.

Wang, Xindi and Mercer, Robert E. and Rudzicz, Frank. 2022. *MeSHup: A Corpus for Full Text Biomedical Document Indexing*. European Language Resources Association. PID https://aclanthology.org/2022.lrec-1.586/.

## A. Appendix: Initial Prompt

| | |
|---|---|
| **Methodology Text** | Experimental Procedures . Cells and Viruses . Coxsackievirus B3 (CVB3) ... Measurement of 3Cpro Activity ... Determination of S- Nitrosylation of 3Cpro. ... |
| **MeSH terms** | Amino Acid Substitution,<br>Antiviral Agents,<br>Binding Sites,<br>Cysteine,<br>Cysteine Endopeptidases |
| **Inital Prompt** | *Experimental Procedures . Cells and Viruses . Coxsackievirus B3 (CVB3) ... Measurement of 3Cpro Activity ... Determination of S- Nitrosylation of 3Cpro. ...*<br><br>*This is the text and this is the annotated list of mesh terms: Amino Acid Substitution, Antiviral Agents, Binding Sites, Cysteine, Cysteine Endopeptidases*<br><br>*Extract all original sentences from the text that contains mesh terms from the given list. Retain the sentences in their original form without altering or ensuring coherence. Provide the extracted sentences as a single paragraph.* |

Figure 6: Initial prompt for summarization

## B. Appendix: Final Corpus Characteristics

| Statistic | Value |
|---|---|
| Total Articles | 1,135,725 |
| Training Set Size | 1,102,807 |
| Validation Set Size | 15,000 |
| Test Set Size | 17,918 |
| Unique MeSH Terms | 28,070 |
| Summarization Model | LLaMA 3 (8B) |

Table 3: Final corpus statistics.

A summary of the final corpus statistics is provided in Table 3. To further ensure the validity and correctness of the summarization pipeline, we conducted a manual inspection of 50 randomly selected summaries. Given that the adopted approach is strictly extractive, the generated summaries preserve the original sentence structure and wording from the source text. Our verification confirmed that sentences containing MeSH-relevant content were retained verbatim, without semantic alteration or distortion. This qualitative assessment provides additional evidence of the accuracy, faithfulness, and reliability of the summarization process employed in constructing the dataset.

## C. Appendix: Limitation

A direct empirical comparison with full-text-based models such as FullMeSH and BERTMeSH would provide additional insight into the relative effectiveness of our approach. However, such a comparison was not feasible due to several practical constraints. First, the implementations of these

models are not publicly available, limiting reproducibility and preventing standardized benchmarking. Second, the datasets employed in those studies differ substantially in both structural composition and scale, making direct performance comparison methodologically inconsistent. Finally, the hardware configurations reported in those works are not fully accessible, and replicating their experimental setup would require computational resources beyond those available to us. These factors collectively preclude a controlled, one-to-one comparison and represent an inherent limitation of the present study.