2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822640

# Enhancing Medical Dialogue Summarization: A MediExtract Distillation Framework

1st Xiaoxiao Liu
*National Centrefor Computer Animation*
*Bournemouth University*
Bournemouth, UK
xliu@bournemouth.ac.uk

2nd Mengqing Huang
*National Centrefor Computer Animation*
*Bournemouth University*
Bournemouth, UK
mhuang@bournemouth.ac.uk

3rd Nicolay Rusnachenko
*National Centrefor Computer Animation*
*Bournemouth University*
Bournemouth, UK
nrusnachenko@bournemouth.ac.uk

4th Julia Ive
*School of Electronic Engineering*
*and Computer Science*
*Queen Mary University of London*
London, UK
j.ive@qmul.ac.uk

5th Jian Chang
*National Centrefor Computer Animation*
*Bournemouth University*
Bournemouth, UK
jchang@bournemouth.ac.uk

6th Jian Jun Zhang
*National Centrefor Computer Animation*
*Bournemouth University*
Bournemouth, UK
jzhang@bournemouth.ac.uk

***Abstract*—Automatic summarization of medical dialogues, which converts colloquial doctor-patient conversations into concise notes, is increasingly important due to the growing complexity of healthcare data. However, the complexity of medical language and the lack of annotated datasets pose challenges for summarization models. In this paper, we propose a MediExtract Distillation Framework (MEDF), a novel hybrid teacher-student distillation process that leverages the power of Large Language Models (LLMs) in information capturing to enhance the performance of a smaller student model. Utilizing medical key information generated by GPT-3.5-Turbo, the model training involves two feedforward branches per iteration: one using ground truth as labels and another using generated structured medical key information as an auxiliary supervision. We validated our method on the MTS-Dialogue dataset, achieving a +2.1% improvement in BLEURT compared to previous methods, demonstrating its effectiveness in summarizing medical dialogues. Additionally, using UMLS-based BERTScore, we observed a +1.8% increase in MedBERTScore for medical term extraction, highlighting our model's practical benefits in clinical information processing. Our framework is publicly available at: https://github.com/Xiaoxiao-Liu/distill-d2n.git**

***Index Terms*—Hybrid Distillation, Medical Dialogue Summarization**

## I. Introduction

As a problem, *Medical Dialogue Summarization* aims at reporting the most essential information from extensive clinical texts in a form of generated concise summaries [1]. This task finds its application in various healthcare settings, including: documenting patient encounters, enhancing information retrieval for decision-making processes [2]. Clinical notes, created by clinicians after each patient visit are: time-consuming and costly to write, requires awareness of medical documentation. Therefore, the advances in automatic summarisation domain that bridging colloquial conversations and formal medical documentation are vital.

Traditionally, automatic summarization is divided into the *extractive* and *abstractive* paradigms [3]. The conventional methods for extractive summarization task primarily relied on rule-based algorithms [4], [5] and statistical techniques [6], [7]. Such approaches usually struggle with the complexity and variability of texts, as well as the handling of large datasets. The appearance of attention mechanisms [8] revolutionized this field by enabling models to dynamically focus on the most relevant parts of the input text. It results in the appearance of the vast amount of models that showcase enhanced performance across the various text summarization problems [9]–[11]. The emergence of instruction-tuned models referred to as *Large Language Models* (LLMs), allow significantly enhanced abstractive summarization, leading researchers to focus on improving summarization tasks in target-oriented domains [12].

In the medical summarization area, previous work has made considerable progress in enhancing the automated generation of medical notes by effectively capturing the critical elements of medical dialogue. Due to the scarcity of medical datasets, researchers have adapted LLMs pre-trained on extensive general-domain datasets [13], [14].These models utilize their deep learning capabilities to boost performance on smaller, specialized datasets by transferring knowledge from extensive general-domain corpora to nuanced medical contexts [15], [16]. This knowledge transfer is pivotal as it substantially augments the models' comprehension and summarization of medical information, despite the scarcity of domain-specific data [17]–[19].

Nevertheless, a significant limitation of these models is their inability to seamlessly incorporate domain-specific medical knowledge into their frameworks [20]. Consequently, these models frequently fall short on medical-specific evaluation metrics, including the precise capture of medical terminology, the understanding of contextually relevant nuances, and adherence to clinical documentation standards [21]. This shortfall underscores the crucial need for embedding explicit medical

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

knowledge within the architecture of LLMs. Such integration would enhance the models' proficiency in grasping intricate medical concepts, thereby boosting their practical utility in medical settings and ensuring a more accurate alignment with the requirements of medical summarization tasks.

In this paper, we propose a MediExtract Distillation Framework (MEDF), which leverages the power of LLMs at scale as a knowledge base in enhancing the performance and accuracy of a smaller student model through a hybrid distillation process incorporating medically highlighted supervision. We design a domain-oriented *Teacher-Student Framework*, in which LLM at scale *teacher model* guides the student model through an auxiliary supervision setting for ensuring effective knowledge distillation. The MEDF methodology is featured by: a) *GPT-3.5-turbo as Teacher model* Leveraging the exceptional capabilities of GPT-3.5-turbo in medical content comprehension [22], [23], our framework employs this model to meticulously extract pivotal medical concepts. This approach, which does not require traditional training methods, significantly streamlines the learning process. b) *Auxiliary Supervision* The key medical concepts identified by the GPT-3.5-turbo are employed as auxiliary supervision in the fine-tuning of the student model. This strategy enriches the student model's medical knowledge and understanding, thereby enhancing its operational insight. c) *Medical-Specific Evaluation* Acknowledging the necessity of clinical validation to ensure real-world applicability, we incorporate both medical-specific evaluation metrics and human evaluation, in addition to generic summarization metrics, to comprehensively assess our results.

The contribution of this paper is as follows:

- We propose novel hybrid distillation method that leverages LLMs at scale as teacher model for enhancing the efficiency and accuracy of a smaller student model. This framework maintains high performance through a teacher-student setup with medically highlighted supervision.
- We experiment with Flan-T5$_{\text{large}}$ application as a teacher model in the proposed framework on the MTS-Dialogue; according to our experiments that involve UMLS-based BERTScore metrics, the application of the distilled model surpasses the previous approach by +2.1% improvement in human judgment and nuances, and +1.83% in medical terms extraction compared with the prior top submission.

## II. Related Work

Medical summarization research has significantly advanced automated note generation by addressing 1) resource limitations and 2) employing fine-tuning techniques. This is crucial for improving documentation efficiency and accuracy in healthcare. Knowledge Distillation (KD) has proven to be an effective method in various applications by transferring knowledge from a large teacher model to a smaller student model [24]–[28]. the method of KD enables the student model to perform comparably to the teacher model while reducing computational resources [29], making it ideal for medical summarization tasks.

### A. Medical Resources and Summarization Techniques

Exploiting the power of the Pointer Generator Model [30], Joshi et al. [15] utilized a pointer generator framework to generate clinical notes from medical dialogues in telemedicine. Their method is advantageous because it effectively captures essential information from dialogues, improving the accuracy of clinical note generation. However, the model has limitations in handling more complex dialogues and capturing nuanced medical details. Throughout this paper, we will refer to their findings as the state-of-the-art (SOTA) work. model [31] achieved the highest performance, this method faces challenges when applied to large medical datasets. In contrast, the work of SummQA [18] involved a training-free two-stage process where they selected similar dialogues from the training set based on input dialogues from the test set, using these selected dialogues as prompts for GPT-4. Authors propose to encode dialogues from the training set and the test set using MiniLM [32], then calculated cosine-similarity scores between encoded input dialogues. However, this work overlook the medical information embedding. The Cadence [19] team who use BART-large as the backbone, consider the importance of medical information grasping and expression, firstly augment the dataset using the MIMIC-IV-Note [33] dataset and the SAMSum dataset [14]. Although these methods have proven to perform well in the medical dialogue summarization task, challenges persist in accurately capturing and representing critical medical information in complex clinical dialogues. Chen et al. [34] explored the robustness of several medical dialogue systems by testing them on out-of-domain dataset, discovering that some hallucinations may exist. Therefore, they proposed a Subjective, Objective, and Assessment and Plan (SOAP) notes configuration to ensure the model generates information related to all essential categories.

methods may help to balance the hallucination problem. In addition, Abacha's work [20] indicates that when evaluating the language model through domain- and task-specific evaluation metrics, the language model may not be able to perform as good as on generic evaluation metrics. The authors suggest that language-model-based domain-specific metrics are effective solutions to hallucination. However, authors have not proposed one novel method which both performs well on a generic evaluation metrics by taking into account domain-specific metrics performance.

### B. Knowledge Distillation

Traditionally, KD involves training a student model using both ground truth labels and pseudo labels generated by a larger teacher model [24]. But it still falls short in replicating the predictive accuracy and the nuanced decision-making capabilities of the larger model [35], particularly in complex tasks [27] or datasets with high variability [36]. This challenge later be tackled through another typical approach that involves transferring logits and feature maps instead of soft labels from

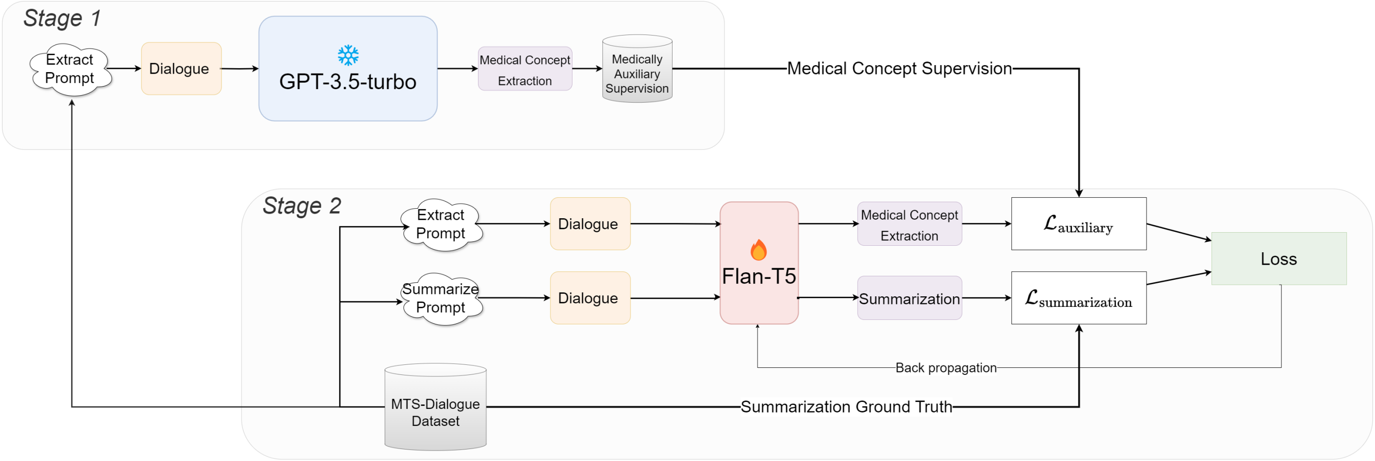
Stage 1
Extract Prompt
Dialogue
GPT-3.5-turbo
Medical Concept Extraction
Medically Auxiliary Supervision
Medical Concept Supervision
Stage 2
Extract Prompt
Summarize Prompt
Dialogue
Dialogue
Flan-T5
Medical Concept Extraction
Summarization
$\mathcal{L}_{\text{auxiliary}}$
$\mathcal{L}_{\text{summarization}}$
Loss
Back propagation
MTS-Dialogue Dataset
Summarization Ground Truth

Fig. 1. A two-stage MediExtact distillation framework for fine-tuning student model; *Stage 1*, the input dialogue data from the MTS-Dialogue dataset is fetched along with an extraction prompt and passed to the LLM at scale (GPT-3.5-turbo); *Stage 2*: a Flan-T5 model is fine-tuned as the student model.

the teacher model [27], [37], [38]. For example, Liu et al. [39] introduced the Instance Relationship Graph method, which distills knowledge by extracting instance features, instance relationships, and feature space transformations from the teacher model. Although these methods improved the performance of smaller models and reduce the need for extensive training resources [40], they initially required fine-tuning the large model on high-performance computing devices [27], [41].

Recently, the capabilities of these models have been significantly expanded due to the development of plug-in frameworks that leverage the extensive abilities of LLMs. One success paradigm is the Chain-of-thought (CoT) prompting method [42], which refers to the strategy of prompting language models to generate intermediate reasoning steps before arriving at a final answer. Inspired by the insight of CoT, Hsieh et, al [43] have proposed an efficient framework using LLMs to generate rationales as the CoT for knowledge distillation to the student model. This method saves the GPU cost by directly collecting task rationales from LLMs inferencing. In addition, the method is proved to be efficient when reducing the training dataset, even with no dataset. Hsieh's work did not discuss the domain-adaptation application of their methods. Hence, following Hsieh's work, Zhou et al [44] explore the CoT knowledge distillation methods in low-budget scenarios. Previous studies have not explored the application of CoT Distillation in domains or tasks that require the highly accurate extraction and generation of domain-specific terms. In this research, we focus on improving the extraction and processing of precise medical terms. We employ our MEDF method within a medical dialogue summarization task. Utilizing a small medical dialogue dataset [16], we discuss whether our training-free hybrid distillation methodis both efficient and accurate.

## III. MediExtract Distillation Framework

Inspired by [43], our MediExtract Distillation Framework (MEDF) engages in a two-stage knowledge distillation process, for medical dialogue summarization. We exploit *teacher-student model framework* specifically adapted for capturing nuances of medical dialogue. Fig. 1 illustrates the proposed method. In the following subsections, we first describe how medical knowledge is extracted using LLM. Next, we discuss the process of knowledge transfer to the student model.

### A. Medical Knowledge Extraction by the Teacher Model

In stage 1, inspired by [43], we employ a teacher training-free approach in the distillation process, leveraging the outstanding performance of LLMs in medical content comprehension [22], [23] and text extraction [45]. Specifically, we use GPT-3.5-turbo to analyze doctor-patient conversation scripts. Carefully designed prompts guide the LLM to extract key medical concepts such as family history, medical history, and medications. the design of prompts can be seen in IV-B. The extracted information is then structured in a format suitable for knowledge transfer. This involves converting the free-text outputs from the LLM into structured data format and then integrating them with the raw input data for the medically auxiliary supervision in training pipeline III-B. By following these steps, our approach leverages the strengths of the LLM in comprehending and extracting medical information without the need for a pre-trained teacher model. This method facilitates efficient and accurate knowledge transfer, enhancing the performance of the student model in medical contexts.

### B. Enhanced Supervision in Student Model Training

In stage 2, the focus is on fine-tuning the student model through a dual-supervision strategy. During each training iteration, the student model learns two key tasks: 1) extracting key medical information, using the medical concepts extracted by the teacher model as supervised labels; and 2) summarizing the dialogue into clinical notes, using reference summaries as labels. This dual focus is facilitated by employing distinct prompts tailored to each task within instruction-tuning settings, detailed further in IV-B.

Additionally, an auxiliary supervision method is introduced via a specialized loss function. This function emphasizes medical accuracy, ensuring that the student model not only generates accurate summaries but also reproduces the critical medical information contained in the dialogue. This enhanced supervision is critical for improving the model's understanding of complex medical concepts, which in turn leads to better generalization and faster convergence during the training phase. The loss function is a weighted sum of the label loss and the auxiliary supervision :

$$\mathcal{L} = \gamma\mathcal{L}_{\text{summarization}} + (1-\gamma)\mathcal{L}_{\text{extra-supervision}} \tag{1}$$

where $\gamma$ is a *weight parameter* tuned during the optimization process serving as the weighting factor for the label loss, while $(1-\gamma)$ is the corresponding factor for the extra-supervision loss. The best performance is observed when $\gamma$ is set to 0.8 (as shown in Fig. 2). Both of the $\mathcal{L}_{\text{label}}$ and $\mathcal{L}_{\text{rationale}}$ are Cross Entropy loss as shown in (2):

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

where $N$ is the number of samples, $y_i$ is the true label for the $i$-th sample, and $\hat{y}_i$ is the predicted probability for the $i$-th sample.

This approach ensures that the model can efficiently assimilate critical information, significantly improving performance with reduced computational resources.

## IV. Experimential Setup

### A. Dataset

We use the MTS-Dialog dataset [46], a comprehensive collection of doctor-patient conversations accompanied by corresponding notes, curated for the MEDIQA-Chat 2023 challenge. This dataset includes 1,201 entries in the training set, 100 in the validation set, and 200 in the test set, providing a robust foundation for empirical analysis.

The MEDIQA-Chat 2023 challenge introduces several different tasks, including dialogue summarization, which requires participants to generate clinical notes from doctor-patient dialogues. In this study, we focus exclusively on the dialogue summarization aspect of TaskA, aiming to explore and evaluate methodologies tailored specifically to this challenge.

### B. Models

In the development of our hybrid distillation method as outlined in Section III, we adopt a teacher-student framework for the training process. Within this framework, the teacher model is employed to extract structured medical information from input dialogues. This extraction task is crucial for generating the supervision needed by the student model. To design effective extraction prompts, we draw inspiration from the MEDIQA-23 Challenge [47]. Based on our observations and consultations with medical experts, we have included 'age' and 'sex' in the prompts, as these details are crucial in real-world clinical note scenarios. The prompts utilized by both the teacher and student models are crucial for the success of our distillation method, ensuring accurate and relevant medical information extraction. These prompts are detailed in:

**Extract Prompt**: Extract the key information from the dialogue, Include all medically relevant information, such as age, sex, medications or drugs, smoking, alcohol, family history, diagnosis, past medical (and surgical) history, immunizations, lab results, and known allergies. DIALOGUE: {dialogue}

**Summarize Prompt**: Summarize the following patient-doctor dialogue in a clinical note style. DIALOGUE: {dialogue}

For the student model, we select the Google Flan-T5$_{\text{large}}$[1], which undergoes instruction-tuning using prompts for medical key information extraction and dialogue summarization. The structured data generated by the teacher model serves as extra supervision for the student model.

The training of the student model is performed on an NVIDIA A100 GPU with 80G of memory, and involves a double feed-forward process in each training iteration, with a batch size set to 2. We fine-tuned the MeDistill model [2] for 20 epochs at a learning rate of $1 \times 10^{-5}$, focusing particularly on the efficient extraction and processing of medical terminology.

### C. Evaluation with Medical-Specific Metrics

Inspired by [20], we utilize several advanced clinical evaluation metrics, including precision, recall, and F1 score, as part of the MedBERTScore. The MedBERTScore calculates the similarity between tokens in a candidate summary and a reference summary. Specifically, the score is calculated by comparing each token in the candidate summary against each token in the reference summary, focusing particularly on clinically significant words as defined by the Unified Medical Language System (UMLS) [48]. The UMLS corpus is accessed through the MedCat tool [3].

For a pair of a candidate summary $\hat{x}$ and reference summary $x$, the MedBERTScore [20] is defined as:

$$\text{MedBERTScore } = \frac{1}{|\hat{x}|} \sum_{\hat{x_i} \in \hat{x}} w_x \max_{x_j} x_i^{\top} \hat{x}_j \tag{3}$$

where, for both metrics, $w_x = 1$ for all the non-medical words and $w_x = 2$ for all the medical words. By adopting these medical-specific evaluation metrics, we can achieve a more nuanced assessment of the relevance and accuracy of clinical details. As generic evaluation, we also compute ROUGE scores [49], including: ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Sum. Furthermore, we incorporate advanced metrics such as BLEURT [50] and BERTScore [51] for enhancing text quality assessment.

[1] https://huggingface.co/google/flan-t5-large

[2] https://huggingface.co/Xiaolihai/flan-t5-large-MeDistill

[3] https://github.com/CogStack/MedCAT

Given that automatic evaluation methods sometimes fall short in capturing the nuances required in medical contexts, human evaluation was crucial to assess if the generated summaries were accurate and practically useful for real-world medical practice. For the human evaluation, we engaged four medical experts, who are currently or have been a doctor, nurse or healthcare administrator, to review clinical notes from doctor-patient conversations. All experts assess the clinical notes by selecting the most suitable note, with the option to choose multiple notes if equally comprehensive or ”none of the above” if none met the criteria. Each evaluation included 10 random samples with one dialogue and three summaries (reference, our model, existing model [17]) in random order. Detailed descriptions of the specifically designed evaluation questions can be accessed through:

**Comprehensive Assessment**: Of these three notes, which one most accurately matches the factual information from the clinical note and covers the important information from the dialogue most completely?

**Practicality and Thoroughness**: Of these three notes, which one best summarizes the essential diagnostic details, includes relevant contextual information, and is most practical for real-world clinical settings while minimizing irrelevant content?

## V. Results

### A. Automatic Summarization Evaluation

The evaluation metrics and rules adopted on the generated notes follow the MEDIQA-Challenge. For the comparison, we include the application of: OpenAI models (GPT-3.5-turbo, GPT-4), well-performed model of the MEDIQA-Challenge (WangLab-run2) [17]. In Table I, Zero-shot learning mode indicates model performance without prior training on task-specific data (GPT-3.5-turbo, GPT-4, Flan-T5$_{\text{large}}$), whereas fine-tuning mode shows performance after model adjustment with task-specific training (MeDistill). According to the obtained results, MeDistill demonstrates notable improvements over GPT-3.5-turbo and GPT-4 and competitive performance. In particular, MeDistill achieves the highest ROUGE-1 score and ROUGE-L score, indicating better unigram overlap and longest common subsequence match with the reference summaries. Furthermore, MeDistill showcases the +2.1% improvement in BLEURT over the prior officially best submission results, suggesting a better alignment with human judgment. Overall, the results highlight the effectiveness of our proposed framework, which consistently delivers strong performance across the summarization evaluation metrics.

### B. Automatic Medical-Specific Evaluation

The proposed MeDistill significantly outperforms other models by MedBERTScore (Table I). With the highest Precision at 74.62, which is 2.62% higher than the second rank model, and the highest F1-Score at 72.96, which is 1.83% higher than the existing model (Table I). This suggests the MeDistill model as significantly more accurate in identifying medical relevant instances identification. Both GPT-3.5-turbo and GPT-4 showcasing a lag behind both in all categories, with GPT-4 having the lowest scores, particularly in Precision at 66.34%, which is significantly lower by 8.28% compared to our MeDistill.

### C. Human Evaluation

Although, the human evaluation has subjectivity in assessing the reasons, it can demonstrate whether the summary generation results align with the actual work scenarios of healthcare professionals from their perspective. Together with the automatic evaluation, the assessment can more comprehensively reflect the effectiveness and applicability of the summaries in real-world medical settings. The results of the human evaluation can be seen in Table II. From the table, we can tell that our model MeDistill stands out as the top performer in both ”Comprehensive Assessment” and ”Practicality and Thoroughness,” with 57.5% and 47.5% of respondents rating it highest, respectively. Another existing model, while receiving the lowest score in comprehensive evaluation at 20%, fares better in practicality and thoroughness with 30% of the vote, suggesting its strengths lie more in application than in breadth. The Reference model shows moderate results, attaining 15% and 20% in the respective categories, indicating an average performance. The ”None of above” option, despite being frequently chosen in comprehensive assessment (23 times), only reflects a minority opinion overall (7.5%), and it is rarely considered best for practicality and thoroughness (2.5%), highlighting a significant preference for the named models over the null option.

## VI. Ablation Study/Discussion

To evaluate the efficacy of our proposed distillation framework, we conducted two ablation studies based on these two questions:

*a) What will be the difference if we only use the teacher model backbone and the student model backbone?*

*b) What are the differences for different weight parameter settings?*

### A. Distillation VS Standard Finetuning on Teacher and Student models

The goal of this ablation study is to quantify the performance improvements provided by the distillation process by comparing the results of the teacher model, student model, and the combined distillation framework. This will help in understanding the effectiveness of knowledge transfer from the teacher to the student model. In the experiments, we utilize GPT-3.5-turbo accessed from OpenAI API to inference the results. As for the student model, we fine-tune the backbone model of the Flan-T5$_{\text{large}}$ for 20 epochs. Other hyperparameters are all the same as our proposed distillation experiments. Both the student and teacher models are performed on the test dataset. Table III showcases the relative performance

TABLE I
PERFORMANCE COMPARISON OF MODELS ON GENERIC AND MEDICAL-SPECIFIC EVALUATION METRICS

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSum | BERTScore | BLEURT | MedBERTScore | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Precision | Recall | F1 score |
| Zero-shot-learning mode | | | | | | | | | |
| Flan-T5$_{\text{large}}$ | 28.53 | 12.61 | 24.13 | 24.13 | 67.00 | 50.95 | 74.52 | 70.34 | 71.89 |
| GPT-3.5-turbo | 30.32 | 12.09 | 24.20 | 24.20 | 65.97 | 50.32 | 65.50 | 68.37 | 66.40 |
| GPT-4 | 30.71 | 12.83 | 23.65 | 23.65 | 64.84 | 52.92 | 66.34 | 70.79 | 68.06 |
| Fine-tuning mode | | | | | | | | | |
| WangLab-run2 | 44.66 | **22.82** | 38.37 | 38.37 | **73.07** | 55.93 | 72.00 | 71.14 | 71.13 |
| MeDistill (ours) | **44.80** | 22.10 | **39.00** | **39.00** | 72.80 | **57.10** | **74.62** | **72.24** | **72.96** |

TABLE II
EVALUATION WITH MEDICAL-SPECIFIC METRICS

| | Comprehensive Assessment | | Practicality and Thoroughness | |
|---|---|---|---|---|
| Model | Freq | Percent | Freq | Percent |
| WangLab-run2 | 8 | 20.0 | 12 | 30.0 |
| Reference | 6 | 15.0 | 8 | 20.0 |
| MeDistill | 23 | **57.5** | 19 | **47.5** |
| None of above | 23 | 7.5 | 1 | 2.5 |

TABLE III
PERFORMANCE COMPARISON OF MODELS WITH AND WITHOUT DISTILLATION. GPT-3.5-TURBO REPRESENTS THE TEACHER MODEL, WHILE FLAN-T5$_{\text{LARGE}}$ AND MEDISTILL REPRESENT STUDENT MODELS.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSum | BERTScore | BLEURT |
|---|---|---|---|---|---|---|
| Zero-shot-learning mode | | | | | | |
| GPT-3.5-turbo | 30.32 | 12.09 | 24.20 | 24.20 | 65.97 | 50.32 |
| Fine-tuning mode | | | | | | |
| Flan-T5$_{\text{large}}$ | 41.30 | 20.50 | 35.50 | 35.50 | 71.50 | 54.00 |
| MeDistill (ours) | **44.80** | **22.10** | **39.00** | **39.00** | **72.80** | **57.10** |

enhancements observed in models when distillation techniques are applied.

The teacher model, GPT-3.5-turbo, is set as the zero-shot learning mode.Whereas the Flan-T5$_{\text{large}}$ and our MeDistill are set as fine-tuned mode. When distillation is employed, our MeDistill, exhibit significant improvements in their performance metrics, which is reflected in its superior performance in the BLEURT metric. These results underscore the effectiveness of model distillation in boosting the linguistic capabilities of student models, enhancing their performance in a range of evaluation metrics without direct reference to specific scores.

### B. Optimal distillation weight ($\gamma$) parameter searching

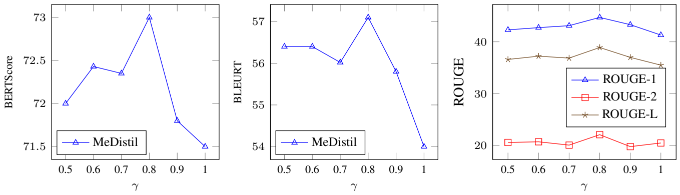

Fig. 2. The result performance of MeDistill depending on the distillation weight paramter ($\gamma$)

In this ablation study, we critically examine the effects of varying the alpha coefficient in our hybrid loss function, defined in 1. The objective is to identify the optimal $\gamma$ value that maximizes model performance, balancing between label-focused accuracy and the benefits derived from the extra-supervision. We conducted experiments with various alpha settings, systematically analyzing how changes in this coefficient influence overall model effectiveness. Results on Fig 2 indicate that the model achieves the best performance at $\gamma = 0.8$, suggesting higher weighting towards the label-focused loss component significantly enhances model accuracy.

We conducted a series of experiments with various alpha settings to systematically analyze how variations in this coefficient affect overall model performance. As shown in Fig 2, the model exhibits optimal performance at $\gamma = 0.8$. This finding suggests that assigning a higher weight to the label-focused loss component substantially improves model accuracy

## VII. CONCLUTION

In conclusion, our study introduces a pioneering two-stage distillation process tailored for the medical domain, which significantly enhances the capability of smaller models to accurately summarize complex medical dialogues. By utilizing a robust teacher model, GPT-3.5-turbo, to initially extract key medical concepts, and then distilling this knowledge into a student model, our approach not only reduces computational demands but also maintains high standards of accuracy. The integration of UMLS-based evaluation further ensures that the summaries generated are both medically precise and relevant. The success of our model, as demonstrated by surpassing SOTA models +1.83% in terms of UMLS precision and +2.1% improvement in general summarization metrics, underscores the effectiveness of our methodology. Ultimately, this research contributes to the field of medical informatics by improving the efficiency and accuracy of information processing in healthcare settings, offering substantial benefits for medical professionals and patients alike.

## Acknowledgment

This research was funded by Bournemouth University Intel PA Project grant number EPSRC Grant EP/L016540/1.

## References

[1] W. Yim, A. Ben Abacha, N. Snider, G. Adams, and M. Yetisgen, “Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations,” in *CLEF 2023 Working Notes*, ser. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, September 18-21 2023.

[2] J. C. Wyatt and J. L. Liu, “Basic concepts in medical informatics,” *Journal of Epidemiology & Community Health*, vol. 56, no. 11, pp. 808–812, 2002.

[3] A. P. Wibawa, F. Kurniawan *et al.*, “A survey of text summarization: Techniques, evaluation and challenges,” *Natural Language Processing Journal*, vol. 7, p. 100070, 2024.

[4] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, and S. Shah, “Combining semantics and social knowledge for news article summarization,” in *Data Mining and Analysis in the Engineering Field*. IGI Global, 2014, pp. 209–230.

[5] K. Janaki Raman and K. Meenakshi, “Automatic text summarization of article (news) using lexical chains and wordnet—a review,” *Artificial Intelligence Techniques for Advanced Computing Applications: Proceedings of ICACT 2020*, pp. 271–282, 2021.

[6] S. Song, H. Huang, and T. Ruan, “Abstractive text summarization using lstm-cnn based deep learning,” *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 857–875, 2019.

[7] K. S. Thakkar, R. V. Dharaskar, and M. Chandak, “Graph-based algorithms for text summarization,” in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*. IEEE, 2010, pp. 516–519.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[9] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in *International conference on machine learning*. PMLR, 2020, pp. 11 328–11 339.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[11] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, “PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5245–5263. [Online]. Available: https://aclanthology.org/2022.acl-long.360

[12] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, “Can large language models reason about medical questions?” *Patterns*, vol. 5, no. 3, 2024.

[13] S. Narayan, S. B. Cohen, and M. Lapata, “Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization,” *ArXiv*, vol. abs/1808.08745, 2018.

[14] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, “Samsum corpus: A human-annotated dialogue dataset for abstractive summarization,” in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 70–79. [Online]. Available: https://www.aclweb.org/anthology/D19-5409

[15] A. Joshi, N. Katariya, X. Amatriain, and A. Kannan, “Dr. summarize: Global summarization of medical dialogue by exploiting local structures.” *ArXiv*, vol. abs/2009.08666, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221802479

[16] A. Ben Abacha, W.-w. Yim, Y. Fan, and T. Lin, “An empirical study of clinical note generation from doctor-patient encounters,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2291–2302. [Online]. Available: https://aclanthology.org/2023.eacl-main.168

[17] J. Giorgi, A. Toma, R. Xie, S. Chen, K. An, G. Zheng, and B. Wang, “Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models,” in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 323–334.

[18] Y. Mathur, S. Rangreji, R. Kapoor, M. Palavalli, A. Bertsch, and M. R. Gormley, “Summqa at mediqa-chat 2023: In-context learning with gpt-4 for medical summarization,” in *Clinical Natural Language Processing Workshop*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259309155

[19] A. Sharma, D. Feldman, and A. Jain, “Team cadence at mediqa-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models,” in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 228–235.

[20] A. Ben Abacha, W.-w. Yim, G. Michalopoulos, and T. Lin, “An investigation of evaluation methods in automatic medical note generation,” in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 2575–2588. [Online]. Available: https://aclanthology.org/2023.findings-acl.161

[21] F. Moramarco, A. P. Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov, “Human evaluation and correlation with automatic metrics in consultation note generation,” in *Annual Meeting of the Association for Computational Linguistics*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247922269

[22] D. Brin, V. Sorin, A. Vaid, A. Soroush, B. S. Glicksberg, A. W. Charney, G. Nadkarni, and E. Klang, “Comparing chatgpt and gpt-4 performance in usmle soft skill assessments,” *Scientific Reports*, vol. 13, no. 1, p. 16492, 2023.

[23] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash *et al.*, “How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment,” *JMIR Medical Education*, vol. 9, no. 1, p. e45312, 2023.

[24] A. Alkhulaifi, F. Alsahli, and I. Ahmad, “Knowledge distillation in deep learning and its applications,” *PeerJ Computer Science*, vol. 7, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220632998

[25] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, “Big self-supervised models are strong semi-supervised learners,” *ArXiv*, vol. abs/2006.10029, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219721239

[26] R. Smith, J. A. Fries, B. Hancock, and S. H. Bach, “Language models in the loop: Incorporating prompting into weak supervision,” *ArXiv*, vol. abs/2205.02318, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248524894

[27] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. J. Lin, “Distilling task-specific knowledge from bert into simple neural networks,” *ArXiv*, vol. abs/1903.12136, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:85543565

[28] S. Srinivas and F. Fleuret, “Knowledge transfer with jacobian matching,” in *International Conference on Machine Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3603145

[29] L. Wang and K.-J. Yoon, “Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3048–3068, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:215745611

[30] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. [Online]. Available: https://aclanthology.org/P17-1099

[31] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, “Scaling instruction-finetuned

language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[32] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020.

[33] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[34] Y.-W. Chen and J. Hirschberg, "Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain soap notes," *arXiv preprint arXiv:2406.02826*, 2024.

[35] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *arXiv preprint arXiv:1802.05668*, 2018.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[37] D. Walawalkar, Z. Shen, and M. Savvides, "Online ensemble model compression using knowledge distillation," *ArXiv*, vol. abs/2011.07449, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226841849

[38] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 925–10 934.

[39] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. feng Duan, "Knowledge distillation via instance relationship graph," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7089–7097, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198185886

[40] Y. Cheng, D. Wang, P. Zhou, and Z. Tao, "A survey of model compression and acceleration for deep neural networks," *ArXiv*, vol. abs/1710.09282, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:22163846

[41] X. Zhu, S. Gong *et al.*, "Knowledge distillation by on-the-fly native ensemble," *Advances in neural information processing systems*, vol. 31, 2018.

[42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[43] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister, "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes," *arXiv preprint arXiv:2305.02301*, 2023.

[44] Y. Zhou and W. Ai, "Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270371055

[45] J. Huang, D. M. Yang, R. Rong, K. Nezafati, C. Treager, Z. Chi, S. Wang, X. Cheng, Y. Guo, L. J. Klesse *et al.*, "A critical assessment of using chatgpt for extracting structured data from clinical notes," *npj Digital Medicine*, vol. 7, no. 1, p. 106, 2024.

[46] A. B. Abacha, W.-w. Yim, Y. Fan, and T. Lin, "An empirical study of clinical note generation from doctor-patient encounters," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2291–2302.

[47] A. B. Abacha, W.-w. Yim, G. Adams, N. Snider, and M. Yetisgen-Yildiz, "Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 503–513.

[48] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[49] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[50] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[51] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *ArXiv*, vol. abs/1904.09675, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:127986044