Using Decoder-Based Distillation for Enhancing Multilingual Clinical Case Report Summarization*

Nicolay Rusnachenko[1,*,†], Xiaoxiao Liu[1,†], Jian Chang[1,†] and Jian Jun Zhang[1,†]

[1] *Centre for Applied Creative Technologies (CFACT+), Faculty of Media and Communications, Bournemouth, United Kingdom*

Abstract

Automatic summarization of clinical reports represent an important field of studies that contribute to shortening long textual narratives written in various languages. Effective report summarization poses numerous challenges, including density of medical terms mentions, semantic interdependency among mentioned entities. The most recent advances of instruction-tuned models illustrate promising capabilities of models at various scale across numerous fields of Natural Language Processing, including textual summarization. A hybrid teacher-student distillation process leverages the power of knowledge distillation by transferring knowledge from a large model (teacher) to a smaller model (student). To our best knowledge, numerous existing studies broadly exploit Seq2seq models. Despite their effectiveness for dialogues and summarization of short texts, such techniques have not become common for supporting multilingual and long input contexts. To bridge the gap in exploring distillation tuning, this paper proposes an adaptation of the teacher-student framework for decoder based systems. In this paper, we experiment with a teacher-student framework for summarising clinical case reports. We adopt the Qwen2.5 models family and evaluate our setup on the MultiClinSum^{small} dataset. We demonstrate that fine-tuning the 0.5B model with the knowledge transferred from the 72B model results in 2.4%-4% performance increment by Rouge metrics compared to the conventional fine-tuning process, highlighting our model's practical benefits in clinical information processing. Our framework is publicly available: https://github.com/nicolay-r/distil-tuning-llm

Keywords

Large Language Model, Hybrid Distillation, Clinical Report Summarization, Multilingual Summarization

1. Introduction

Text summarization is a task of shortening textual content while preserving crucial information. The approaches on automated shortening of the textual content are commonly divided into: *extractive* methods (keeping salient segments) and *abstractive* methods (essay generation). As a task within the clinical domain, textual summarization lies at the intersection of various information retrieval challenges, including but not limited by question-answering [1], entities extraction [2]. The texts to be summarised may vary in length, ranging from *short texts* (conversational dialogues [3]) to *long narratives* (clinical case reports [4]).

The advent of *transformer-based architectures* [5] with appearance of *self-attention* [5] caused a significant impact on automated text translation systems and as a result *Seq2seq* systems [6, 7, 8] and *decoder-based* solutions [9]. However, the benefit of attention comes at the cost of quadratic complexity with respect to the input sequence length. Such tradeoff raised a number of further works on *attention sparsification* techniques [8, 10]. However, the most recent tendency towards exploiting pretrained generalized systems [9, 11, 12, 13] shaped architectural concepts towards such factors as (i) *scalability*, and (ii) *alignment with next-token prediction training*; for which *decoder-based* systems are suited better. The generalized approach of adoption models for various problems results in so-called *instruction-tuned*

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

n.rusnachenko@bournemouth.ac.uk (N. Rusnachenko); xliu@bournemouth.ac.uk (X. Liu); jchang@bournemouth.ac.uk (J. Chang); jzhang@bournemouth.ac.uk (J. J. Zhang)

https://nicolay-r.github.io/ (N. Rusnachenko); https://jzhang.bournemouth.ac.uk/ (J. J. Zhang)

0000-0002-9750-5499 (N. Rusnachenko); 0000-0002-5906-264X (X. Liu); 0000-0003-4118-147X (J. Chang); 0000-0002-7069-5771 (J. J. Zhang)

models [9]. Despite the vast amount of benefits and adaptation for the downstream tasks, the trade-off of such systems is their scale. Such factor requires adoption of the specific fine-tuning techniques.

BioASQ [14] represents one of the most recent competition challenges on biomedical semantic indexing and question answering. The MultiClinSum challenge [4] dedicated to advance automated long texts summarization systems in multilingual conditions. In this paper we propose a system that represent a decoder-based distillation framework for multilingual clinical case report summarization [4]. Our approach exploits distillation technique for transferring clinical key information derived from reports via large (teacher) model to its smaller scaled (student) model. The contribution of these studies are two fold:

- We propose distillation framework with *role-based dialogue modeling* notation [15, 9] for enhancing small-scaled models (student models) with clinical key information derived from reports via large-scaled model (teacher model); the designed system exploits *system*, *user*, and *assistant* roles which are commonly supported by instruction-tuned models [12, 11, 13, 16].
- We experiment with decoder-based distillation technique adaptation in clinical case report summarization task for Qwen-2.5 models family [12]; we demonstrate that extracting clinical key information from large-scaled (72B params) *teacher model* (ClinicalKeyInfo$^{\text{small}}$ dataset) and using this information in tuning of small-scaled (0.5B params) *student model* results in 2.4%-4% on MultiClinSum$^{\text{small}}$ [4] while at evaluation stage.

2. Related Works

Medical summarization research has evolved through two key approaches: addressing data scarcity and refining distillation techniques. Recent progress in Natural Language Processing (NLP), the field of building systems that understand and generate human language, has been driven by Large Language Models (LLMs)—neural networks trained on large-scale text corpora. These models have recently been applied in medical settings: for example, [17] introduced a hybrid distillation framework using LLMs to enhance medical term extraction. This builds on earlier pointer-generator models [18] and training-free methods like SummQA [19].

Knowledge Distillation has been effective for model compression [20, 21], with [22] proposing a foundational step-by-step method that uses LLMs-generated rationales to supervise small models. Liu et al.extended this approach to medical summarization with concept-level supervision [17]. However, both efforts focus on encoder-decoder architectures and leave the challenges of domain adaptation and decoder-only models underexplored.

Encoder-decoder models have been widely adopted for summarization tasks due to their strong sequence-to-sequence performance [23]. T5 [6] unified various NLP tasks into a text-to-text framework, while LongT5 [10] extended this architecture with sparse attention to better handle long sequences. mT5 [7] further scaled the T5 framework to cover over 100 languages, enabling multilingual summarization. These models have been applied to medical summarization benchmarks such as PubMed [2], MultiMedQA [24], and MTS-Dialogue [3]. Despite these advances, such models often face challenges in clinical summarization scenarios, where inputs are lengthy, domain-specific, and multilingual. Their fixed-length encoders and tokenizer limitations hinder generalization across diverse note types and clinical terminologies [23]. Besides T5-series, the other existing alternatives designed for long-input summarization, such as BigBird [25] and LED [8], require specialized pretraining and remain computationally intensive, making them less practical for real-world multilingual clinical applications.

In contrast, decoder-only models generate text sequentially, conditioning each token on the previously generated context. Recent work has explored applying these models to summarization tasks through prompting and fine-tuning, particularly in settings where large-scale instruction data is available [26, 27]. ChatGPT [9] and LLaMA-2-Chat [13] have been used as instruction-tuned models for abstractive summarization in zero-shot or few-shot settings, where the models are prompted with summarization tasks without further supervised training. These systems have shown competitive results on both general-domain and biomedical summarization benchmarks, including BioASQ [14] and MEDIQA [19].

In the context of knowledge distillation, decoder-only architectures have been leveraged as both teachers and students, allowing smaller generative models to learn the reasoning steps and instruction mapping behaviour demonstrated by larger models [28, 29, 30]. These studies show that decoder-only student models can benefit from rationale-augmented supervision and multi-task distillation [17], enabling effective transfer of reasoning ability and task generalization. Furthermore, such models offer practical advantages in handling longer input sequences and multilingual instruction formats, motivating recent extensions [16, 12], which support extended context lengths and multilingual tokenization.

3. Methodology

To enhance both faithfulness and medical relevance, we propose *two-stage distillation-based framework*, as illustrated in Figure 1. The framework takes as input: (1) training collection, (2) large-scale instruction-tuned teacher model, (3) small-scaled decoder-only student model. The result of the framework application is a fine-tuned student model capable of generating more accurate summaries.

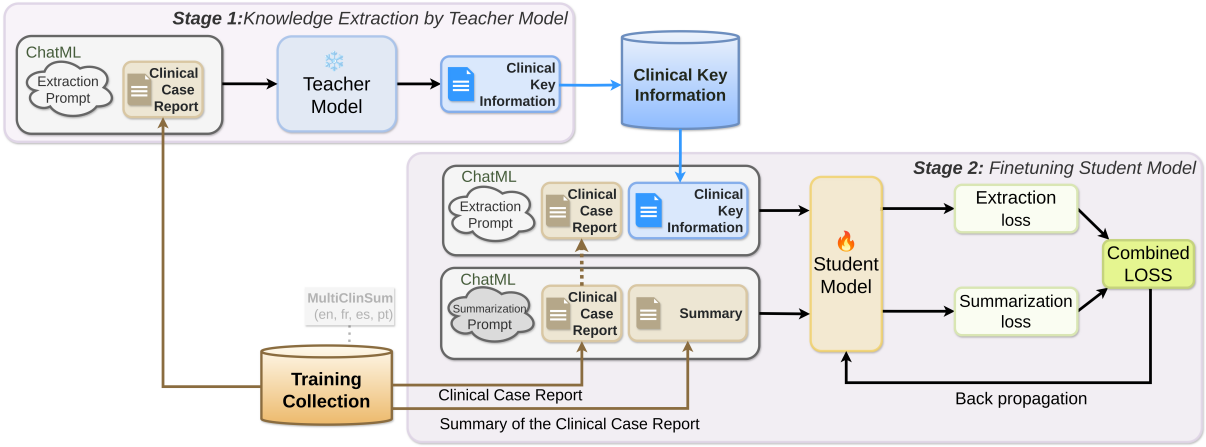

Figure 1: Overview of the two-stage distillation framework for decoder-based systems with *role-based dialogue* formatting input structuring [15, 9]; the process of the framework application represent a sequential completion of stages: (stage 1) using teacher model to extract clinical key information from Training Collection, and (stage 2) fine-tuning process with dual supervision, based on (i) clinical key information and (ii) clinical case summaries

Stage 1 refers to a process of teacher model application for clinical key information extraction from training collection. Stage 2 refers to student model fine-tuning process with dual supervision: (1) supervision from the reference summary to the original clinical case report, and (2) supervision from the extracted clinical key information to the one obtained from the teacher model. The design of the proposed system relies on models that support role-based dialogue modelling notation, and ChatML format[1] in particular. Specifically, the input is structured as a sequence of $system$, $user$, and $assistant$ roles.

3.1. Stage 1: Knowledge Extraction by the Teacher Model

Given (i) extraction prompt and (ii) *Training Collection* we adopt *Teacher Model* (Figure 1) in a zero-shot setting to infer textual responses (treated as *Clinical Key Information*). The format of the input data for Teacher Model of each clinical case report from Training Collection is as follows:

(**system:** extraction prompt, **user:** clinical case report, **assistant:** $\emptyset$)

Where «$\emptyset$» refers to the absence of the $assistant$ role. The results of this stage represent a *Clinical Key Information* collection (see Figure 1), which we use for the Stage 2.

[1] https://platform.openai.com/docs/guides/text

3.2. Stage 2: Fine-tuning Student Model

Given (i) reports from *Training Collection* and (ii) *Clinical Key Information* collection (obtained from Stage 1, Section 3.1), we adopt *Student Model* in the distillation tuning process. Further in this section we declare methodology for evaluating the alignment of the student model towards the expected output, followed by construction of the combined loss function.

Our methodology of student model fine-tuning assumes a hard alignment of the output towards raw textual content with no additional span annotations. We use *strict position-wise cross-entropy* (hard alignment) in the fine-tuning process. If the two sequences differ in length, comparison continues position-wise until either sequence ends, with any remaining positions scored against the end-of-sequence symbol. Let the input formatted sequence as $\mathbf{x}$, ground truth answer as $\mathbf{y} = (y_1, \ldots, y_T)$, inferred text from student model as $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{\hat{T}})$ (T and $\hat{T}$ denotes the total number of tokens for $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively). For the hidden state of the student model (θ), and step t (index of generated token), we define *strict position-wise loss calculation* ($loss$) as follows:

$$loss(\mathbf{y}, \hat{\mathbf{y}}, t, \mathbf{x}) = -\log P_\theta\big(\hat{y}_t = y_t \mid \hat{y}_{<t}, \mathbf{x}\big)$$

We use the formula above in separate calculations of *extraction loss* and *summarization loss* (see Figure 1).

For the extraction supervision, the input is structured as:

(**system:** extraction prompt, **user:** clinical case report, **assistant:** clinical key information)

Given the extraction-formatted input sequence ($\mathbf{x}_e$), output from student model $\hat{\mathbf{y}}_\mathbf{e}$, clinical key information $\mathbf{y}$, we compute extraction loss ($\mathcal{L}_{\text{ext}}$) as sum from step i_{start} that marks the first token of the assistant segment $\mathbf{x}_e$ to the final token position T_e:

$$\mathcal{L}_{\text{ext}} = \sum_{t=i_{\text{start}}}^{T_e} loss(\mathbf{y}_e, \hat{\mathbf{y}}_e, t, \mathbf{x}_e)$$

For the summarization supervision, the input is constructed as:

(**system:** summarization prompt, **user:** clinical case report, **assistant:** summarized case report)

Given the summarization-formatted input sequence ($\mathbf{x}_s$), output from student model $\hat{\mathbf{y}}_\mathbf{s}$, summarized case report $\mathbf{y}$, we compute summarization loss ($\mathcal{L}_{\text{sum}}$) as sum from step i_{start} that marks the first token of the assistant segment $\mathbf{x}_s$ to the final token position T_s:

$$\mathcal{L}_{\text{ext}} = \sum_{t=i_{\text{start}}}^{T_s} loss(\mathbf{y}_s, \hat{\mathbf{y}}_s, t, \mathbf{x}_s)$$

Finally, we calculate the combined loss ($\mathcal{L}$) as superposition of the losses with the decay coefficient (γ):

$$\mathcal{L} = \gamma\, \mathcal{L}_{\text{sum}} + (1 - \gamma)\, \mathcal{L}_{\text{ext}}$$

4. Experimental Setup

Data preparation. The complete set of the available annotated data represent texts with summaries written in four different languages: English, Portuguese, Spanish, and French. For each language, the originally provided reports with their summaries portioned into two groups [4]: *small* (≈ 500 texts per language), and *large* (≈ 25000 texts per language). In this work **we utilize only small groups** in our experiments, majorly due to both (i) limitation of computational resources and (ii) time required for experiments organization. In further, we refer to a small part of the collection as MultiClinSum$^{\text{small}}$. Table 1 illustrates the statistics of clinical case reports and summaries for MultiClinSum$^{\text{small}}$.

Table 1
Statistics of the publicly available MultiClinSum^{small} dataset

Language	# Reports	Clinical Case Report Characters Stat			Summary Characters Stat		
		Mean	Min	Max	Mean	Min	Max
English	592	3785.4	719	34071	725.3	90	3883
Spanish	592	4056.1	825	17602	792.6	125	4161
French	592	4783.2	827	37138	832.1	121	4542
Portuguese	592	4096.0	793	37351	809.5	116	28227

Knowledge Extraction by the Teacher Model (Stage 1). As for the teacher model we adopt Qwen-2.5-72B-instruct[2] for extracting clinical key information from clinical case reports. Given clinical case report (R) from Training Dataset (MultiClinSum^{small}), we use the following extraction prompt[3]: *"Extract the key information from clinical text: R"*. We denote the result dataset composed with the selected teacher model as ClinicalKeyInfo^{small}. Table 2 illustrates the statistic of the ClinicalKeyInfo^{small}, separately for reports written in each language.

Table 2
Statistics of the composed ClinicalKeyInfo^{small} by applying the first stage of the proposed methodology towards MultiClinSum^{small}, separately for reports written in each language

Language	# Reports	Characters Stat		
		Mean	Min	Max
English	592	2971.4	1088	6597
Spanish	592	2929.5	392	6040
French	592	2873.3	879	5472
Portuguese	592	2871.9	911	7961

Data-split. All the reports were divided into three *train/valid/test* subsets with the following proportion of 80%, 1%[4], and 19% respectively. For the *test*, the 456 of original reports were chosen (19% of MultiClinSum^{small}).

Fine-tuning. In these studies we consider fine-tuning a single model instance for all languages / subtasks. We use GoogleColab service and publish Jupyter-Notebook at the project repository. We rent a single instance with NVidia A100 40GB VRAM with 80GB RAM. To accomplish this goal, we consider model instruction Qwen-2.5-0.5B which both (1) covers a set of languages utilized in MultiClinSum^{small} and (2) support long context input. Towards the setup of the model parameters for the fine-tuning process. We majorly refer to the initial list of parameters proposed for Qwen-2.5-VL[5]. In particular, we use `bf16` mode precision for model weight representation in memory. We set γ (see Section 3.2) to 0.8 according to the earlier organized studies [17]. We limit[6] the amount of input tokens to 3078, among which 2566 were assigned altogether for *system* and *user* roles and remaining 512 for the *assistant* role. For the output, we set a limit of 768 tokens in accordance to the *mean-length* statistic for summaries, mentioned in Table 1. The statistics of the MultiClinSum^{small} dataset *train* and *valid* subsets, cropped by max threshold presented in Table 3. Towards the formatting of the input data, for the given clinical report (R) we use the following text summarization prompt[7]: "*Summarize clinical text: R*". We pad input data by maxim length calculated across all the formatted texts at the dataset tokenization stage.

Models Evaluation. We assess the performance of the fine-tuning models every 250 optimization steps by using texts from a *valid* set (see Table 1). As for the metrics, we assess *rouge-1*, *rouge-2*, *rouge-L*,

[2] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[3] In this work we limit our analysis on relevant types of information in depth.
[4] Such a small amount majorly due to computational issues caused by OOM Trainer Exception using Huggingface library
[5] https://github.com/QwenLM/Qwen2.5-VL
[6] Limitation of the utilized computational resources
[7] The prompt for the explanation is similar to the one utilized at the Stage 1 (see Section 3)

Table 3
Statistics of the of the MultiClinSum^{small}-based processed resources, separately for each subset utilized in fine-tuning process (i.e. *train* and *valid* subsets)

Subset	Type	Source	# Reports	Characters Stat	
				Mean	Range
train	**Clinical Case Report**	MultiClinSum^{small}	1892	2435.6	719-2560
	Summary	MultiClinSum^{small}	1892	490.6	90-512
	Clinical Key Information	ClinicalKeyInfo^{small}	1892	511.9	392-512
valid	**Clinical Case Report**	MultiClinSum^{small}	20	2560.0	2560-2560
	Summary	MultiClinSum^{small}	20	510.2	486-512
	Clinical Key Information	ClinicalKeyInfo^{small}	20	512.2	512-512

rouge-Lsum. We adopt the policy of keeping the best performing instance of the model throughout the whole fine-tuning process.

Inference: We use T4 GPU with 16GB VRAM hosted by GoogleColab to infer the results of the following versions of the Qwen2.5-0.5B-Instruct[8]. Since the Qwen2.5 models input context window size significantly exceed the assigned amount of tokens for input, we increase this threshold for up to 16384 tokens. We follow a similar policy of the restrictions towards the amount of output tokens. We set the max amount of output generated tokens to 1024 which surpasses the mean amount of characters in average per summary (see Table 1). **We use localized summarization prompts** to align output with the source language utilized in original report. The following templates for summarization of non-English clinical case report (R) were used: *"Resumir texto clínico: R"* (Portuguese), *"Resumir el texto clínico: R"* (Spanish), *"Résumer le texte clinique: R"* (French).

5. Result Analysis and Discussion

Following the fine-tuning procedure organization described in Section 4, we prepare models:

- Qwen2.5-0.5B: https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct
- Qwen2.5-0.5B$_{standard}$: https://huggingface.co/nicolay-r/qwen25-05b-multiclinsum-standard
- Qwen2.5-0.5B$_{distil}$: https://huggingface.co/nicolay-r/qwen25-05b-multiclinsum-distil

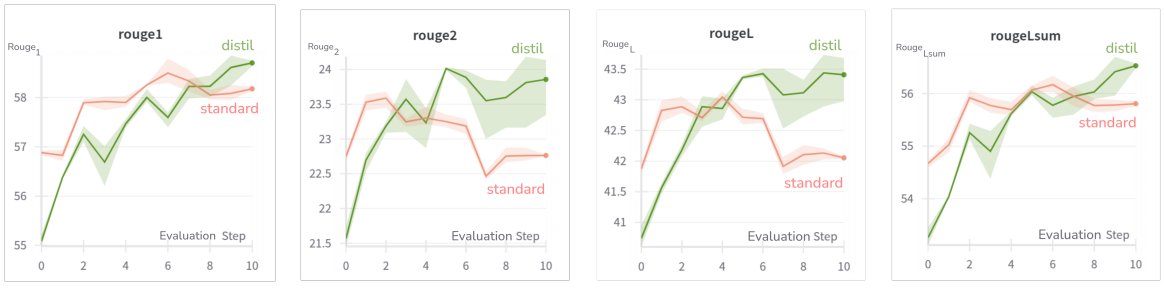

Figure 2: The variation for the results on *valid* of MultiClinSum^{small} set between Qwen2.5-0.5B$_{distil}$ (green) and Qwen2.5-0.5B$_{standard}$ (red) fine-tuning procedures for the 10 evaluation steps (each measurement performed after 250 optimization steps) for 3 individual runs on MultiClinSum^{small}

To fit in the 40GB VRAM limitation, we set BatchSize=2 for the Qwen2.5-0.5B$_{standard}$ version[9] and BatchSize=1 for the Qwen2.5-0.5B$_{distil}$. Figure 2 illustrates the analysis of variation of the results obtained on MultiClinSum^{small} (valid) for 3 individual fine-tuning runs and 10 evaluation steps. According

[8] https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct
[9] We noticed that attempting to fine-tune Qwen2.5-0.5B$_{standard}$ with smaller BatchSize of 1 results in worse performing model among all the rouge metrics (except *Rouge-1*).

to the related analysis, using Qwen2.5-0.5B$_{\text{distil}}$ results in 2.4%-4% of the improved performance in comparison with the conventional fine-tuning (Qwen2.5-0.5B$_{\text{standard}}$).

To obtain the results we follow the inference setup mentioned in Section 4. We infer our models on non-official *test* subset (TEST$_{\text{non-official}}$) and on officially provided test sets (TEST$_{\text{official}}$). Due to limited amount of time, **for TEST$_{\text{official}}$ we reduced the total number of generated token by half** (up to 512) from the initially defined limit (see *inference* details, Section 4) to gain inference performance. We use `bulk-chain`[10] library to perform inference with the custom implementation of the Qwen2.5 provider based on the *HuggingFace pipelines* API[11]. We manually implement evaluation for TEST$_{\text{non-official}}$ which differs from one utilized by competition organizers for TEST$_{\text{official}}$. In particular, to evaluate results on TEST$_{\text{non-official}}$ we adopt Rouge metrics (see Evaluation, Section 4) and BertScore based on DistilBERT$_{\text{base-uncased}}$ [31].

Non-official Evaluation Results. Table 4 provides the results on TEST$_{\text{non-official}}$ for baseline, Qwen2.5-0.5B$_{\text{standard}}$, Qwen2.5-0.5B$_{\text{distil}}$ models. According to the obtained results, we first noticed the gap in the results on English subtask in comparison with all the other languages. Towards the results of the particular models, using fine-tuning techniques outperform the Qwen2.5-0.5B approach by ≈1% (BertScore$_{\text{F1}}$) and ≈8%-20% in average for Rouge results. Towards the individual subtasks, using Qwen2.5-0.5B$_{\text{distil}}$ for summarization clinical reports in English results in ≈1-2% (Rouge) over Qwen2.5-0.5B$_{\text{standard}}$. In the case of all the other non-English subtask, both fine-tuned version of the student model (Qwen2.5-0.5B$_{\text{standard}}$ and Qwen2.5-0.5B$_{\text{distil}}$) illustrate relatively similar performance. As for the content comparison, we noticed that Qwen2.5-0.5B$_{\text{distil}}$ model tend contains a high proportion of words that are semantically similar to words in the reference sentence (high BertScore precision) unlike other models and in exchange of the lowered recall. We believe that such an effect is due to enhanced alignment of the composed report summarization to the ground truth texts in MultiClinSum$^{\text{small}}$ while at during Stage 2 (Section 3.2).

Table 4
Evaluation results on TEST$_{\text{non-official}}$ for various version of the experimenting models, separately for each language / subtask of MultiClinSum$^{\text{small}}$; for each subtask and each evaluation metric, the **best results are bolded**

Language / Subtask	Model	BERTScore			ROUGE			
		Precision	Recall	F1	R-1	R-2	R-L	R-Lsum
English	Qwen2.5-0.5B	78.59	**82.91**	80.62	32.49	11.88	20.97	22.06
	Qwen2.5-0.5B$_{\text{standard}}$	80.94	82.13	81.47	37.49	15.26	25.48	25.70
	Qwen2.5-0.5B$_{\text{distil}}$	**81.80**	81.67	**81.69**	**38.30**	**15.57**	**25.71**	**26.11**
Spanish	Qwen2.5-0.5B	80.26	**84.64**	82.35	33.66	13.33	20.69	23.75
	Qwen2.5-0.5B$_{\text{standard}}$	84.07	83.62	**83.80**	**40.50**	**17.14**	**26.74**	**27.03**
	Qwen2.5-0.5B$_{\text{distil}}$	**84.10**	83.48	83.76	40.26	16.72	26.38	26.77
French	Qwen2.5-0.5B	81.16	**84.36**	82.69	34.45	13.80	20.23	23.22
	Qwen2.5-0.5B$_{\text{standard}}$	84.05	83.80	**83.88**	**39.67**	**17.00**	**24.94**	**25.24**
	Qwen2.5-0.5B$_{\text{distil}}$	**84.34**	83.10	83.68	38.95	16.25	24.39	24.99
Portuguese	Qwen2.5-0.5B	80.65	**83.53**	82.02	30.81	11.44	19.53	21.64
	Qwen2.5-0.5B$_{\text{standard}}$	83.21	83.29	83.19	37.51	**15.02**	**24.30**	**24.51**
	Qwen2.5-0.5B$_{\text{distil}}$	**83.42**	83.10	**83.22**	**37.66**	14.92	24.19	24.40

Official Submission Results. In the case of official evaluations, organisers utilise BertScore and *rouge*-score that involve evaluation of precision, recall, and F1-measure. Due to limited amount of time during the test stage, it was decided to submit the results for Qwen2.5-0.5B$_{\text{distil}}$ model. The results evaluation on TEST$_{\text{official}}$ for the summaries obtained by Qwen2.5-0.5B$_{\text{distil}}$ model illustrated in Table 5. Similar to the results on TEST$_{\text{non-official}}$, we noticed significant gap in the results performance for the reports written in English comparing with the results obtained for reports written in the other languages.

Limitations Discussion and Further Works According to the obtained results, we believe that application of the proposed methodology could be enhanced in following directions: (1) enlarging of

[10] https://github.com/nicolay-r/bulk-chain
[11] https://huggingface.co/docs/transformers/main_classes/pipelines

Table 5
Evaluation results on $\text{TEST}_{\text{official}}$ for the Qwen2.5-0.5B_{distil} model.

Language /	BERTScore			ROUGE		
Subtask	Precision	Recall	F1	Precision	Recall	F1
English	85.54	85.70	85.59	27.53	27.53	25.87
Spanish	72.42	73.47	72.88	26.06	29.03	25.87
French	72.48	73.96	73.15	24.15	28.90	24.66
Portuguese	72.39	73.20	72.73	24.95	27.05	24.40

the training data, (2) highlighting relevant features for clinical key information, (3) overcoming hard alignment in student model fine-tuning (Section 3.2). In particular, we see no technical limitations in adaptation of larger dataset. According to the MultiClinSum statistic mentioned in *data preparation* of Section 4, we believe that switching from $\text{MultiClinSum}^{\text{small}}$ to $\text{MultiClinSum}^{\text{large}}$ result in 5-times longer process (excluding the time required for evaluation steps). However, we believe that addressing limitations for other directions (2) and (3) is crucial for employing $\text{MultiClinSum}^{\text{large}}$. With the existing extraction prompt, observations regarding the most *relevant features* mention in extracted *clinical key information* are considered out of scope. The extraction of such *relevant features* from outputs of student model could also address on *hard-alignment limitation* in model fine-tuning process (Section 3.2).

6. Conclusion

In this paper we propose a system for automated clinical case report summarization within the scope of the MultiClinSum challenge. Our approach exploits distillation framework for fine-tuning small-scaled (student) decoder-based models by relying on clinical key information derived reports via large-scaled model (teacher model). Unlike previously existing work on distillation technique adaptation for Seq2seq architectures, our system is dedicated for decoder-based models that support role-based dialogue modelling notation. We assess our approach on MultiClinSum reports written in English, Portuguese, French and Spanish. According to the related analysis, the use of the proposed distillation framework for Qwen-2.5 model series results in a 2.4%-4% better performing model (Qwen2.5-0.5B_{distil}) on validation data compared to the one fine-tuned with the conventional approach (Qwen2.5-0.5B_{standard}). From our final evaluation on test data, we conclude that the Qwen2.5-0.5B_{distil} model surpasses Qwen2.5-0.5B_{standard} by $\approx$1-2% in the summarization clinical reports written in English.

7. Declaration on Generative AI

AI tools were used for: rephrasing sentence and paragraphs to enhance reading quality (abstract and introduction), grammar correction and spell check (all sections).

References

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

[2] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[3] A. Ben Abacha, W.-w. Yim, Y. Fan, 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, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2291–2302. URL: https://aclanthology.org/2023.eacl-main.168.

[4] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of multiclinsum task at bioasq 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results., in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.

[7] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41/. doi:10.18653/v1/2021.naacl-main.41.

[8] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[10] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, Y. Yang, LongT5: Efficient text-to-text transformer for long sequences, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 724–736. URL: https://aclanthology.org/2022.findings-naacl.55/. doi:10.18653/v1/2022.findings-naacl.55.

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[12] A. Yang, B. Yang, B. Zhang, B. H. et. al., Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

[13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[14] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. Maria Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: C.-d.-A. Jorge, G. Julio, P. Laura, G. S. d. H. Alba, M. Josiane, P. Florina, R. Paolo, S. Damiano, F. Guglielmo, F. Nicola (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[15] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, Dialogpt: Large-scale generative pre-training for conversational response generation, in: ACL, system demonstration, 2020.

[16] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[17] X. Liu, M. Huang, N. Rusnachenko, J. Ive, J. Chang, J. J. Zhang, Enhancing medical dialogue summarization: A mediextract distillation framework, in: 2024 IEEE International Conference on

Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 6466–6473.

[18] A. Joshi, N. Katariya, X. Amatriain, A. Kannan, Dr. summarize: Global summarization of medical dialogue by exploiting local structures., in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3755–3763. URL: https://aclanthology.org/2020.findings-emnlp.335/. doi:10.18653/v1/2020.findings-emnlp.335.

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

[20] A. Alkhulaifi, F. Alsahli, I. Ahmad, Knowledge distillation in deep learning and its applications, PeerJ Computer Science 7 (2020). URL: https://api.semanticscholar.org/CorpusID:220632998.

[21] L. Wang, 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 44 (2020) 3048–3068. URL: https://api.semanticscholar.org/CorpusID:215745611.

[22] C.-Y. Hsieh, C.-L. Li, C.-k. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, T. Pfister, Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 8003–8017. URL: https://aclanthology.org/2023.findings-acl.507/. doi:10.18653/v1/2023.findings-acl.507.

[23] N. Rusnachenko, N. D. Nguyen, et al., Pre-training longt5 for vietnamese mass-media multi-document summarization, Journal of Mathematical Sciences 285 (2024) 88–99.

[24] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al., Toward expert-level medical question answering with large language models, Nature Medicine (2025) 1–8.

[25] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.

[26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.

[27] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13484–13508. URL: https://aclanthology.org/2023.acl-long.754/. doi:10.18653/v1/2023.acl-long.754.

[28] N. Calderon, S. Mukherjee, R. Reichart, A. Kantor, A systematic study of knowledge distillation for natural language generation with pseudo-target training, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14632–14659. URL: https://aclanthology.org/2023.acl-long.818/. doi:10.18653/v1/2023.acl-long.818.

[29] K. Shridhar, A. Stolfo, M. Sachan, Distilling reasoning capabilities into smaller language models, Findings of the Association for Computational Linguistics: ACL 2023 (2023) 7059–7073.

[30] J. Ko, T. Chen, S. Kim, T. Ding, L. Liang, I. Zharkov, S.-Y. Yun, Distillm-2: A contrastive approach boosts the distillation of llms, arXiv preprint arXiv:2503.07067 (2025).

[31] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).