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

BioASQ 2025
Task MultiClinSum

Nicolay Rusnachenko[1] Xiaoxiao Liu[1] Jian Chang[1] Jian Jun Zhang[1]

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

Key Takeaway

1. Distillation framework with *role-based dialogue modeling* notation for **Student Models** with **Clinical Key Info** derived from reports via **Teacher Model**
   - Exploit of **System**, **User**, and **Assistant** roles which are commonly supported by instruction-tuned models
2. We experiment with technique adaptation in clinical case report summarization task for **Qwen-2.5** models family
   - extracting clinical key information from *teacher model* (72B params) and using this information in tuning of small-scaled student model (0.5B params) **results in 2.4%-4% on MultiClinSum^{small}** while at evaluation stage.

YouTube WATCH ONLINE

Related Works

[1] Xiaoxiao Liu and et. al. Enhancing medical dialogue summarization: A mediextract distillation framework. pages 6466–6473, 2024.

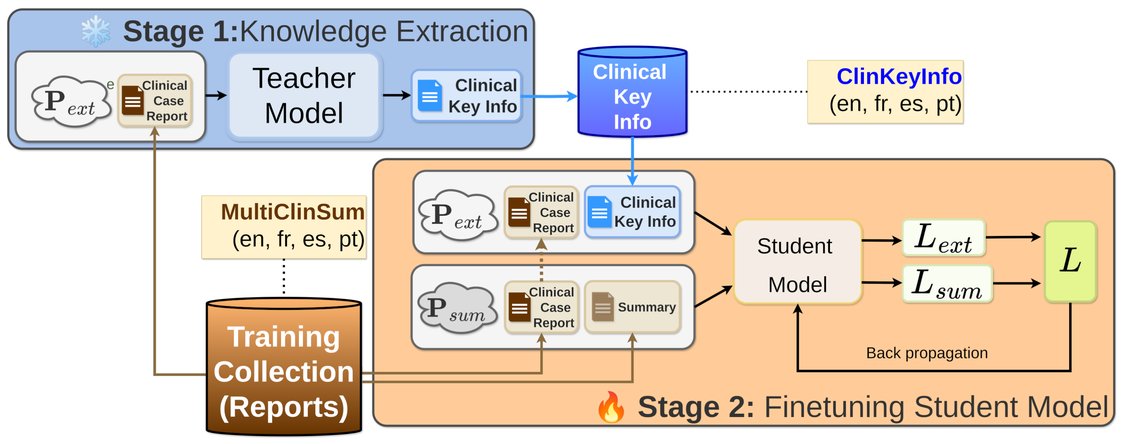
[2] M Rodríguez-Ortega and et. al. Overview of multiclinsum task at bioasq 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results.

Implementation

GitHub

nicolay-r/distil-tuning-llm

Two-stage distillation for role-based input systems



Definitions:

- – Clinical Case Report from Reports Collection
- – Clinical Case Report Summary (from Reports Collection)
- – Clinical Key information from Clinical Key Information collection
  - extracted from on Stage 1
- $\mathbf{P}_{ext}$ – ext. prompt: «*Extract the key information from clinical text:* »
- $\mathbf{P}_{sum}$ – summarization prompt: «*Summarize clinical text:* »
- (**S**, **U**, **A**) – role-based dialogue modelling notation.

Stage 1

Given: [$\mathbf{P}_{ext}$, Dataset of]

For each :

- Infer result from `Teacher Model`:

(**S:** $\mathbf{P}_{ext}$, **U:** , **A:** $\emptyset$)

Result: dataset of

Stage 2

Given:

- $\mathbf{P}_{ext}$ – extraction prompt;
- $\mathbf{P}_{sum}$ – summarization prompt;
- Dataset of (, ,)

For each (, ,):

1. Extraction Supervision (L_{ext}):

(**S:** $\mathbf{P}_{ext}$, **U:** , **A:**)

$$L_{ext} = \sum_{t=i_{\text{start}}}^{T_e} l(\mathbf{y}_e, \hat{\mathbf{y}}_e, t, \mathbf{x}_e)$$

2. Summarization Supervision (L_{sum}):

(**S:** $\mathbf{P}_{sum}$, **U:** , **A:**)

$$L_{sum} = \sum_{t=i_{\text{start}}}^{T_s} l(\mathbf{y}_s, \hat{\mathbf{y}}_s, t, \mathbf{x}_s)$$

3. Combined loss (L) with $\gamma = 0.8$:

$$L = \gamma\, L_{sum} + (1-\gamma)\, L_{ext}$$

Result: Fine-tuned `Student Model`

Loss Definition

Given:

- $\mathbf{x}$ – Input Sequence
- $\mathbf{y} = (y_1, \ldots, y_T)$ – GT Summary
- $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{\hat{T}})$ – Inferred Text
- T and $\hat{T}$ – total number of tokens for $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively

Strict position-wise loss calculation:

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

Dataset I

- **MultiClinSum^{small} – 592 reports** of each language from original MultiClinSum:

Lng	#	# Chars in			# Chars in		
		Avg	Min	Max	Avg	Min	Max
EN	592	3785.4	719	34071	725.3	90	3883
ES	592	4056.1	825	17602	792.6	125	4161
FR	592	4783.2	827	37138	832.1	121	4542
PT	592	4096.0	793	37351	809.5	116	28227

Dataset II

- **ClinKeyInfo^{small} – augm. of reports of MultiClinSum^{small}** with obtained from Stage 1

Lng	#	# Chars in		
		Avg	Min	Max
EN	592	2971.4	1088	6597
ES	592	2929.5	392	6040
FR	592	2873.3	879	5472
PT	592	2871.9	911	7961

- We use `Qwen-2.5-72B-instruct`.

Experimential Setup

Our setup based on content from MultiClinSum^{small} (,) and ClinKeyInfo^{small} ( ) for subsets:

Subset		#	# Chars Avg	# Chars Range
train		1892	2435.6	719-2560
		1892	490.6	90-512
		1892	511.9	392-512
valid		20	2560.0	2560-2560
		20	510.2	486-512
		20	512.2	512-512

We use `Qwen2.5-0.5B` as a student model for Stage 2 to prepare:

Hugging Face

`Qwen2.5-0.5B`$_{distil}$

Non-official Results

Models

- • – baseline `Qwen2.5-0.5B`
- † – finetuned `Qwen2.5-0.5B`$_{standard}$
- ‡ – distil-tuned `Qwen2.5-0.5B`$_{distil}$

Test Set

The 456 of original reports were used (19% of MultiClinSum^{small})

L		BERTScore			ROUGE		
		P	R	F1	R-1	R-2	R-L
EN	•	78.59	**82.91**	80.62	32.49	11.88	20.97
	†	80.94	82.13	81.47	37.49	15.26	25.48
	‡	**81.80**	81.67	**81.69**	**38.30**	**15.57**	**25.71**
ES	•	80.26	**84.64**	82.35	33.66	13.33	20.69
	†	84.07	83.62	**83.80**	**40.50**	**17.14**	**26.74**
	‡	**84.10**	83.48	83.76	40.26	16.72	26.38
FR	•	81.16	**84.36**	82.69	34.45	13.80	20.23
	†	84.05	83.80	**83.88**	**39.67**	**17.00**	**24.94**
	‡	**84.34**	83.10	83.68	38.95	16.25	24.39
PT	•	80.65	**83.53**	82.02	30.81	11.44	19.53
	†	83.21	83.29	83.19	37.51	**15.02**	**24.30**
	‡	**83.42**	83.10	**83.22**	**37.66**	14.92	24.19

Official Results

Model: `Qwen2.5-0.5B`$_{distil}$

Lng	BERTScore			ROUGE		
	P	R	F1	P	R	F1
EN	85.54	85.70	85.59	27.53	27.53	25.87
ES	72.42	73.47	72.88	26.06	29.03	25.87
FR	72.48	73.96	73.15	24.15	28.90	24.66
PT	72.39	73.20	72.73	24.95	27.05	24.40