NLP4Health at MultiClinSum-2 2026: Multilingual Clinical Case Report Summarization with Qwen Prompting and QLoRA

BioASQ Lab at CLEF 2026

Moutushi Roy[1], Dipankar Das[1]

[1]Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

Abstract

This paper presents a system submitted to the MultiClinSum-2 shared task, organized as part of the BioASQ Lab at CLEF 2026. The task addresses multilingual clinical case report summarization, where long and information-rich clinical case reports must be transformed into concise summaries while retaining the most relevant clinical facts. The system was developed for four language tracks: English, Spanish, French, and Portuguese. All experiments were conducted under a resource-constrained setting using an NVIDIA RTX A4000 GPU with 16 GB memory. Three configurations were studied: a prompt-only Qwen2.5-7B system, a Qwen2.5-1.5B model adapted with QLoRA, and a Qwen2.5-7B model adapted with QLoRA. According to the official evaluation results, the prompt-only configuration achieved the best average performance among the submitted runs, obtaining a ROUGE-2 score of 0.1662, BERTScore of 0.8707, faithfulness score of 0.7372, and completeness score of 0.8745. These results indicate that, for long multilingual clinical case reports under limited computational resources, carefully designed prompting, deterministic decoding, and strict output validation can provide a reliable alternative to lightweight adaptation, especially when clinical completeness and factual preservation are important.

Keywords

Multilingual clinical summarization, clinical case report summarization, large language models, parameter-efficient fine-tuning, QLoRA, BioASQ, CLEF, MultiClinSum-2

1. Introduction

Clinical case reports are important sources of medical knowledge because they describe patient history, clinical presentation, diagnostic findings, treatment decisions, complications, and outcomes in narrative form. They are especially useful for rare diseases, unusual symptoms, unexpected complications, and treatment observations. However, such reports are often too long, dense, and difficult to read quickly. Automatic summarization can support faster access to central clinical information, but the task is challenging because a useful clinical summary must be concise without losing essential facts or introducing unsupported medical claims.

The difficulty increases in multilingual settings. Medical terminology, abbreviations, sentence structure, and reporting conventions vary between languages. A summarization system must preserve clinical meaning while maintaining the language of the source report. This is more demanding than general-domain summarization because small factual changes can alter diagnosis, treatment pathway, severity, or patient outcome. In clinical summarization, fluency alone is not sufficient; faithfulness and completeness are also necessary.

MultiClinSum was introduced as a multilingual clinical case report summarization task within the BioASQ-CLEF evaluation setting. The first edition focused on long clinical case reports in English, Spanish, French, and Portuguese and provided a shared benchmark for studying clinical summarization methods across languages [1, 2]. MultiClinSum-2 extends this direction by evaluating multilingual clinical case report summarization across 15 languages in the BioASQ Lab at CLEF 2026. The present submission is therefore connected to both the BioASQ 2026 lab overview and the task-specific MultiClinSum-2 overview [3, 4].

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

moutushiroy123@gmail.com (M. Roy); dipankardipnil2005@gmail.com (D. Das)

This submission addresses four language tracks: English, Spanish, French, and Portuguese. The system was developed in a restricted computational environment, where large-scale full fine-tuning was not feasible. Therefore, the study focused on instruction-based prompting and parameter-efficient adaptation using QLoRA. The aim was not only to generate valid test summaries, but also to understand which type of configuration was more reliable under limited GPU memory.

This submission is also related to previous work on multilingual clinical dialogue summarization and question answering using mT5 with LoRA-based parameter-efficient fine-tuning [5]. However, the present task differs in both the input type and the modelling requirement. Instead of dialogue-style clinical text, the input consists of long clinical case reports. The system must therefore identify and compress the patient presentation, diagnostic evidence, treatment, clinical progression, and outcome from a much longer document.

The main contributions of this paper are as follows:

1. A compact multilingual clinical case report summarization pipeline is described for English, Spanish, French, and Portuguese.
2. Three system configurations are compared under the same resource-constrained setting: prompt-only Qwen2.5-7B generation, Qwen2.5-1.5B QLoRA adaptation, and Qwen2.5-7B QLoRA adaptation.
3. Official results are analysed using lexical overlap, semantic similarity, and LLM-as-judge quality dimensions.
4. A qualitative English example is provided to show how the submitted runs differed in clinical completeness and output style.

2. Task Description

MultiClinSum-2 focused on the automatic summarization of long clinical case reports across 15 languages. Given a complete case report, the system had to generate a shorter summary that preserved the main clinical content. The official MultiClinSum-2 overview paper describes the data, evaluation setting, participating systems, results, and analysis for multilingual clinical case report summarization at CLEF 2026 [4]. The task follows the previous MultiClinSum setting, where the organizers formalized multilingual clinical case report summarization as a benchmark and described the task design, data, evaluation resources, participating systems, and results [1]. The broader BioASQ overview papers describe the shared-task setting and the role of MultiClinSum within the BioASQ evaluation framework [2, 3].

In this submission, English, Spanish, French, and Portuguese were considered. Each evaluated language-run pair contained 1,000 test instances. Reference summaries were not available during test-time generation, so the system could not use reference-based filtering or checkpoint selection for the final test outputs. This made output validation important: generated files had to be present, non-empty, concise, and written in the correct language.

The official result file contained both traditional automatic metrics and quality-oriented evaluation dimensions. ROUGE-1, ROUGE-2, and ROUGE-Lsum measured lexical overlap with reference summaries [6]. BERTScore estimated semantic similarity using contextual embeddings [7]. In addition, LLM-as-judge dimensions were reported, including faithfulness, completeness, fluency, and consistency. These dimensions were useful for interpreting whether the generated summaries preserved clinically important facts and remained readable.

3. Methodology

The system followed a reproducible multilingual summarization pipeline. First, each clinical case report was loaded and lightly normalized. Second, a language-aware summarization instruction was

prepared. Third, summaries were generated using one of the submitted configurations. Finally, the generated outputs were cleaned and validated before submission. The same broad pipeline was followed for all runs, but the model size and adaptation strategy differed across the three configurations.

3.1. Run 1: Prompt-only Qwen2.5-7B Generation

Run 1 was a prompt-only configuration using `Qwen/Qwen2.5-7B-Instruct` loaded in 4-bit mode. No task-specific fine-tuning was applied. The final selected instruction format, referred to as Prompt V6 in the implementation, asked the model to generate a concise clinical case summary in the same language as the input report. The prompt explicitly required one complete paragraph, 3 to 5 complete sentences, and an approximate length of 70 to 110 words.

To reduce unsupported generation, the system message instructed the model to use only the provided case report and not to invent details. The prompt also required the summary to preserve patient age, sex, important findings, management, treatment, and outcome when available. For long clinical reports, the preprocessing stage applied the balanced excerpt strategy described in Section 4.2, so that both the initial presentation and later outcome information could be retained within the input budget.

3.2. Run 2: Qwen2.5-1.5B QLoRA Adaptation

Run 2 used `Qwen/Qwen2.5-1.5B-Instruct` with QLoRA-based parameter-efficient fine-tuning. This run was designed as a lightweight adapted model that could be trained reliably under the available GPU memory. Instead of updating all model parameters, the base model was loaded with quantization and only low-rank adapter parameters were trained. LoRA provides a lightweight adaptation method by training low-rank matrices [8], while QLoRA reduces memory use by combining quantization with adapter-based fine-tuning [9].

For Run 2, language-specific QLoRA adapters were fine-tuned for English, Spanish, French, and Portuguese using the same `Qwen/Qwen2.5-1.5B-Instruct` base model and training configuration. For each language, 2,000 training examples and 100 validation examples were selected from the official MultiClinSum-2 training corpus, giving 8,000 training and 400 validation examples in total. Each example used a supervised instruction-tuning format, where the input was a clinical case report and the target was the corresponding reference summary. Training used a maximum sequence length of 2048 tokens, LoRA rank 8, LoRA alpha 16, LoRA dropout 0.05, per-device batch size 1, gradient accumulation steps 8, and an effective batch size of 8. Each adapter was trained for 500 optimization steps with a learning rate of 2×10^{-4}. The optimizer was AdamW, and training was step-based rather than epoch-based because a fixed number of optimization steps was used for each language-specific adapter. With 2,000 training examples per language and an effective batch size of 8, 500 optimization steps corresponded to approximately two passes over the selected language-specific subset. This setting made fine-tuning feasible on a 16 GB NVIDIA RTX A4000 GPU.

During inference, greedy decoding was used with `do_sample=False`, `max_new_tokens=150`, and a repetition penalty of 1.1. Although temperature, `top_p`, and `top_k` were set to 0.7, 0.8, and 20, respectively, these sampling parameters did not affect generation because sampling was disabled. Therefore, the outputs were deterministic.

3.3. Run 3: Qwen2.5-7B QLoRA Adaptation

Run 3 used `Qwen/Qwen2.5-7B-Instruct` with QLoRA-based adaptation. This configuration was intended to test whether a stronger 7B backbone could improve multilingual clinical summarization after parameter-efficient fine-tuning. Similar to Run 2, the system used LoRA rank 8, LoRA alpha 16, and LoRA dropout 0.05. The maximum sequence length was set to 2048 tokens, and the training was performed with a small batch size and gradient accumulation because of GPU memory limitations.

The optimizer, LoRA configuration, sequence length, and decoding setup were kept consistent with Run 2 wherever possible. AdamW was used as the optimizer, and training was performed with a learning rate of 2×10^{-4}, per-device batch size 1, and gradient accumulation steps of 8. Because of the

larger 7B backbone and the 16 GB GPU constraint, this run was more memory-sensitive than Run 2 and was treated as an additional adapted configuration.

Compared with Run 2, Run 3 used a larger base model and was therefore expected to have stronger multilingual generation capacity. However, it was also more sensitive to memory limits, input truncation, and decoding settings. In the evaluated result file, Run 3 results were available for English, Spanish, and French, while Portuguese results were not present. The Portuguese Run 3 output was completed after the official submission deadline and was therefore not included in the official evaluation. For this reason, Run 3 is analyzed only for the three evaluated languages and treated as an additional experimental adapted configuration rather than the primary submitted system.

4. Experimental Setup

4.1. Data and Languages

The submitted systems were evaluated on English, Spanish, French, and Portuguese. Each evaluated language-run pair contained 1,000 test examples. The input documents were full clinical case reports, and the expected output was a concise clinical summary. Since test references were not available during generation, the final outputs were not selected using reference-based metrics.

4.2. Preprocessing

Preprocessing was intentionally conservative. Repeated spaces and unnecessary line breaks were normalized, but the original clinical content was preserved. Aggressive cleaning was avoided because it could remove medically relevant information such as laboratory values, treatment details, or outcome statements.

For long reports, Run 1 used a balanced excerpt strategy rather than simple head-only truncation. The report was divided into three broad regions: the opening part, the middle part, and the ending or outcome-oriented part. Priority was given to the beginning and ending regions because the beginning usually contains patient background and presentation, while the ending often contains treatment response, complications, follow-up, and outcome. In practice, approximately 65% of the retained text was taken from the beginning and 35% from the ending when the report exceeded the input budget. A shorter middle segment was retained when space was available to preserve diagnostic and management details.

4.3. Prompt Format

Run 1 used a prompt-only instruction format without task-specific fine-tuning. The prompt was iterated during development to improve length control, language preservation, factual grounding, and clinical completeness. Listing 1 reports the final Prompt V6 template used for Run 1 generation.

During prompt development, six prompt variants were tested on selected development examples from English, Spanish, French, and Portuguese. Prompt V1 only asked the model to generate a concise clinical summary, but it produced unstable outputs such as overly long summaries, copied source fragments, bullet-style or labelled responses, incomplete endings, occasional language mixing, and missing treatment or outcome information. Prompt V2 added instructions to preserve the main presentation, diagnosis, treatment, and outcome. Prompt V3 introduced stronger conciseness constraints. Prompt V4 added explicit factual-grounding instructions to use only the provided case report and avoid unsupported details. Prompt V5 enforced a single-paragraph format and discouraged list-like responses. The final Prompt V6 combined these constraints with same-language generation, 3–5 complete sentences, an approximate length of 70–110 words, preservation of patient demographics and medication names, avoidance of excessive laboratory-value listing, and complete sentence endings. Prompt V6 was selected because it produced the most stable development outputs across the four languages, with fewer copied fragments, incomplete endings, bullet-style outputs, language-mixing issues, and unsupported

additions. No formal prompt ablation experiment was performed; selection was based on iterative qualitative inspection and submission-readiness checks.

Listing 1: Prompt V6 used for Run 1 prompt-only multilingual clinical summarization.

```
System message:
You are a careful biomedical summarization assistant.
You write faithful clinical case summaries.
Use only the provided case report.
Do not invent details.
Do not translate medication names incorrectly.

User message:
Write one concise clinical case summary in {language_name}.

Strict output rules:
- Write only in {language_name}.
- Write one complete paragraph.
- Use 3 to 5 complete sentences.
- Aim for 70 to 110 words.
- Start with the patient or case background, not with a copied treatment fragment.
- Include the main presentation, important findings, management/treatment, and outcome if
    available.
- Preserve exact age, species, sex, numbers, and medication names when clearly stated.
- If a medication name is unclear, use a general phrase such as "supportive treatment"
    instead of inventing.
- Do not list every laboratory value.
- Do not repeat the same treatment sentence.
- Do not start with labels, bullets, numbering, or lowercase fragments.
- Finish with a complete sentence.
- Output only the final summary paragraph.

Clinical case report:
{source_text}
```

4.4. Training and Inference Settings

The adapted runs used QLoRA-based supervised instruction tuning. Run 2 used a smaller Qwen2.5-1.5B-Instruct model to create a lightweight adapted baseline, while Run 3 used Qwen2.5-7B-Instruct to test whether a larger backbone improved the final summaries. Both adapted runs used a maximum sequence length of 2048 tokens and LoRA adapters with rank 8, alpha 16, and dropout 0.05. The training configuration was selected pragmatically because the experiments were conducted on a single NVIDIA RTX A4000 GPU with 16 GB of memory.

During Run 1 inference, deterministic decoding was used with `do_sample=False`, a maximum input length of 6144 tokens, a maximum generation length of 190 new tokens, and a repetition penalty of 1.05. The generated text was post-processed by removing summary labels, bullet markers, repeated spaces, and incomplete trailing fragments. A lightweight quality checker flagged empty outputs, very short or long summaries, incomplete sentence endings, bullet-like starts, lowercase fragment starts, and generic refusal outputs.

No external medical knowledge base was used to add information to the generated summaries. This decision was taken to reduce the risk of introducing unsupported medical claims.

Table 1
Main experimental settings used for the submitted systems.

Component	Setting
Task	Multilingual clinical case report summarization
Languages	English, Spanish, French, Portuguese
Hardware	NVIDIA RTX A4000, 16 GB GPU memory
Run 1	Qwen2.5-7B-Instruct, prompt-only, 4-bit
Run 1 decoding	Deterministic, `do_sample=False`
Run 1 input limit	6144 tokens
Run 1 generation limit	190 new tokens
Run 1 repetition penalty	1.05
Run 2	Qwen2.5-1.5B-Instruct with QLoRA
Run 3	Qwen2.5-7B-Instruct with QLoRA
QLoRA sequence length	2048 tokens
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.05
Training batch strategy	Batch size 1 with gradient accumulation
QLoRA optimizer	AdamW
QLoRA learning rate	2×10^{-4}
Run 2 training steps	500 per language-specific adapter
Run 2 approximate epochs	Approximately 2 over the selected subset
QLoRA inference decoding	Greedy, `do_sample=False`
QLoRA generation limit	150 new tokens
QLoRA repetition penalty	1.1
Validation	Non-empty output and length checks

Table 2
Average performance of the submitted runs.

Run	R-1	R-2	R-L	BERT	Faith.	Compl.	Flu.	Cons.
Run 1	0.3944	0.1662	0.2641	0.8707	0.7372	0.8745	0.7006	0.6524
Run 2	0.3624	0.1477	0.2399	0.8633	0.6822	0.6716	0.6954	0.4361
Run 3	0.3353	0.1048	0.2234	0.8467	0.6546	0.6308	0.6997	0.5562

5. Results

5.1. Overall Run-wise Performance

Table 2 shows the average performance of the three submitted configurations. Run 1 achieved the strongest average scores across most metrics. It obtained the highest ROUGE-1, ROUGE-2, ROUGE-Lsum, BERTScore, faithfulness, completeness, fluency, and consistency among the submitted runs.

The strongest result of Run 1 suggests that direct prompting was more reliable than the adapted variants in this particular setting. The largest difference was observed in completeness. Run 1 achieved an average completeness score of 0.8745, while Run 2 and Run 3 achieved 0.6716 and 0.6308, respectively. This suggests that the adapted systems often produced shorter or less complete summaries.

5.2. Language-wise Performance

Table 3 presents the official language-wise results. The same general pattern was observed across languages. Run 1 remained the strongest configuration for English, Spanish, French, and Portuguese in terms of BERTScore and completeness.

Table 3
Language-wise official results for the submitted runs.

Lang.	Run	R-1	R-2	R-L	BERT	Faith.	Compl.	Flu.	Cons.
En	Run 1	0.3819	0.1550	0.2653	0.8689	0.7648	0.9393	0.7014	0.6980
En	Run 2	0.3552	0.1440	0.2403	0.8635	0.7007	0.7669	0.6984	0.6635
En	Run 3	0.3426	0.1159	0.2388	0.8559	0.6994	0.7419	0.6998	0.6958
Es	Run 1	0.4060	0.1760	0.2726	0.8727	0.7219	0.8392	0.7004	0.6296
Es	Run 2	0.3825	0.1637	0.2571	0.8655	0.6776	0.6338	0.6982	0.3699
Es	Run 3	0.3347	0.0977	0.2213	0.8450	0.6322	0.5752	0.6996	0.5293
Fr	Run 1	0.4054	0.1777	0.2611	0.8709	0.7237	0.8283	0.6998	0.6463
Fr	Run 2	0.3562	0.1424	0.2243	0.8598	0.6746	0.6344	0.6884	0.3386
Fr	Run 3	0.3285	0.1008	0.2102	0.8390	0.6322	0.5753	0.6996	0.4435
Pt	Run 1	0.3842	0.1559	0.2574	0.8702	0.7384	0.8913	0.7008	0.6355
Pt	Run 2	0.3558	0.1408	0.2378	0.8641	0.6757	0.6514	0.6966	0.3723

Table 4
Example English output generated by the three submitted runs for the same case report.

System	Generated summary
Run 1	A 34-year-old man presented with pain and swelling in his right leg after being buried in a landslide for 5 hours. Initial examination revealed tenderness, tension, and visible bruising, with no signs of fracture. He was diagnosed with rhabdomyolysis-induced acute kidney injury (AKI) stage II and underwent anterolateral fasciotomy for compartment syndrome. Despite aggressive fluid and pharmacological management, his condition worsened, leading to AKI stage III and metabolic acidosis. Hemodialysis was performed but the patient remained hemodynamically unstable and ultimately died 5 days later due to AKI complications and septic shock.
Run 2	We present a case of a 34-year-old man who presented with acute kidney injury (AKI) due to rhabdomyolysis induced by a fall into snowslide. His initial laboratory test showed elevated serum creatinine and blood urea nitrogen levels. Renal replacement therapy (RRT) was initiated immediately after the onset of AKI. However, the patient's condition worsened rapidly despite RRT. Eventually, the patient died from multiple organ failure. In this case, we discuss the pathophysiology of rhabdomyolysis-induced AKI, its management, and prognosis. Our case is reported according to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline.
Run 3	We reported a case of a 34 years old male who suffered from rhabdymyolysis due to trauma caused by landslide. The patient initially presented with severe pain and swelling of the right thigh. The patient underwent antero-lateral fasciotomies for compartment syndrome and was subsequently treated with aggressive fluid therapy, hemodynamic support, and hemodialysis. Despite these treatments, the patient ultimately succumbed to multiple organ failure.

5.3. Qualitative Example

Table 4 shows one English test example and the summaries generated by the three submitted runs. The source case describes a 34-year-old man who developed traumatic rhabdomyolysis, compartment syndrome, acute kidney injury, metabolic acidosis, and hemodynamic instability after being buried in a landslide. The patient deteriorated despite fasciotomy, fluid management, vasopressor support, and hemodialysis, and died after five days of hospitalization.

In this example, Run 1 produced the most complete summary. It preserved the traumatic cause, affected limb, rhabdomyolysis-induced AKI, compartment syndrome, fasciotomy, progression to AKI stage III, metabolic acidosis, hemodialysis, hemodynamic instability, and final fatal outcome. Run

2 captured the central diagnosis and death, but it used broader article-style phrasing and added a guideline-related statement that was less central to the clinical summary. Run 3 was concise and clinically focused, but it omitted several important details such as AKI staging, metabolic acidosis, septic shock, and the clinical deterioration sequence. This example supports the quantitative trend in which Run 1 achieved the strongest completeness and semantic similarity scores.

6. Discussion

The results show that the prompt-only system was the most effective configuration for this submission. This may appear unexpected because fine-tuning is often assumed to improve task-specific performance. However, clinical summarization is sensitive to missing or distorted information. If adaptation is performed with limited data, short training, strong truncation, or suboptimal decoding, the model may learn to generate summaries that are fluent but incomplete.

A useful observation from the submitted runs is that a larger or adapted model did not automatically lead to better clinical summarization. Run 2 used a smaller 1.5B model with QLoRA adaptation, which made training feasible but limited the model capacity. Run 3 used a stronger 7B backbone with QLoRA, but the improvement expected from model size was not reflected in the final average scores. This suggests that, for long clinical case reports, model size and adaptation are not sufficient by themselves. The way the input is truncated, the stability of the prompt, the decoding strategy, and the ability to preserve outcome information are also important.

Run 1 benefited from the general instruction-following ability of the base model and from a highly constrained prompt. Because the model parameters were not changed, the system retained its broad multilingual generation capacity. The controlled prompt helped guide the model toward a concise but clinically complete answer. In contrast, QLoRA adaptation was more sensitive to training data balance, learning dynamics, sequence length, and checkpoint selection.

The qualitative example also supports this interpretation. Run 1 preserved the sequence of clinical events more completely, including compartment syndrome, progression of acute kidney injury, metabolic acidosis, hemodialysis, and final fatal outcome. Run 2 and Run 3 generated fluent summaries, but they omitted some details or shifted toward article-style phrasing. For a clinical summarization task, this type of omission can reduce completeness even when the summary remains readable.

The lower performance of the adapted configurations does not mean that QLoRA is unsuitable for this task. Rather, it suggests that lightweight adaptation must be tuned carefully for long clinical documents. Future improvements may require better multilingual sampling, longer-context modeling, validation-based checkpoint selection, more careful handling of long reports, and factuality-aware decoding. Completeness should be monitored during development because it was the metric where the adapted systems showed the largest drop.

Several practical constraints should also be considered. The experiments were limited to a single NVIDIA RTX A4000 GPU with 16 GB memory. This restricted the model size, batch size, sequence length, and range of hyperparameters that could be tested. Only four language tracks were included, although the full shared task covered a broader multilingual setting. In addition, Run 3 was evaluated for three languages rather than four because the Portuguese Run 3 output was completed after the official submission deadline and was not included in the official evaluated result file. Finally, LLM-as-judge scores are useful for analysis, but they should not be treated as a replacement for expert clinical review. The analysis is based on the submitted system outputs and the official result file; therefore, it focuses on system behavior under limited computational resources rather than claiming general superiority of one modeling approach over another.

7. Conclusion

This paper presented a multilingual clinical case report summarization system submitted to the MultiClinSum-2 task at the BioASQ Lab in CLEF 2026. Three configurations were investigated under a

resource-constrained setting: prompt-only Qwen2.5-7B generation, Qwen2.5-1.5B QLoRA adaptation, and Qwen2.5-7B QLoRA adaptation. The official results showed that the prompt-only configuration achieved the strongest average performance across lexical, semantic, and quality-oriented metrics.

The findings suggest that, when computational resources are limited, careful prompting and strict output validation can remain competitive for multilingual clinical summarization. The adapted configurations generated fluent outputs but were less reliable in preserving completeness and consistency. Future work will focus on improved multilingual training, stronger long-context handling, validation-based checkpoint selection, and factuality-aware evaluation for clinical summarization.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools for language polishing, grammar checking, LaTeX formatting assistance, and presentation of manuscript text. The tools were not used to create the experimental dataset, generate the submitted system outputs, fabricate evaluation results, or produce unsupported scientific claims. All experimental design decisions, system configurations, official scores, result analysis, and final manuscript content were reviewed, verified, and approved by the authors.

Acknowledgements

The authors thank the BioASQ and MultiClinSum-2 organizers for providing the shared task, datasets, evaluation framework, and official results. The authors also thank the reviewers for their constructive comments, which helped improve the clarity of the camera-ready version.

References

[1] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, C. Escolano, A. Mash, M. Melero, L. Pratesi, L. Vigil-Giménez, L. Fernandez-Lopez, 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: Working Notes of CLEF 2025, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4038/paper_2.pdf.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodríguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. 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, arXiv preprint arXiv:2508.20554 (2025). URL: https://arxiv.org/abs/2508.20554.

[3] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The Fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2026.

[4] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, A. Mash, M. Krallinger, Overview of MultiClinSum-2 at CLEF 2026: Data, evaluation, and results for multilingual clinical case report summarization across 15 languages, in: CLEF, 2026.

[5] M. Roy, D. Das, NLP4Health: Multilingual clinical dialogue summarization and QA with mT5 and LoRA, in: Proceedings of the NLP-AI4Health Workshop, Association for Computational Linguistics, Mumbai, India, 2025, pp. 93–97. URL: https://aclanthology.org/2025.nlpai4health-main.10/.

[6] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[7] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: Proceedings of the 8th International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: Proceedings of the 10th International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[9] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems, 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.