Agentic MCS: A Multilingual Clinical Summarization Framework

Notebook for the MultiClinSum Lab at CLEF 2025

Johanna Angulo[1,*], Víctor Yeste[1]

[1] School of Science, Engineering and Design, Universidad Europea de Valencia, Paseo de la Alameda, 7, 46010 Valencia, Spain

Abstract

This research presents Agentic MCS, a LangGraph-based framework for multilingual biomedical text summarization, evaluated across English, Spanish, French, and Portuguese clinical case report datasets from bioASQ MultiClinSum Challenge. The investigation implements a multi-agent workflow integrating extractive preprocessing, neural abstractive summarization generation, and quality enhancement through comparative experimental runs.

The framework employs diverse architectural approaches including hybrid BM25-neural pipelines, NER-integrated entity preservation systems, knowledge graph-guided summarization, and dense retrieval re-ranking mechanisms. Core models encompass domain-specific transformers (BioMistral-7B), multilingual fine-tuned models (mBART-large-50, mT5-base), general-purpose language models (Mistral 7B, Llama 3.1 8B), and proprietary LLMs as enhancement systems (GPT-4o).

The framework achieved comprehensive multilingual coverage with language-specific optimizations. Evaluation reveals consistently high BERTScore results (0.71-0.84) indicating strong semantic fidelity, while lower ROUGE scores (0.17-0.25) reflect high abstraction rather than lexical extraction. This work establishes a robust foundation for intelligent agentic systems in biomedical text processing.

Keywords

Multilingual summarization, Multi-agent systems, Medical text mining, Transformer models, Knowledge graphs, Entity preservation, Clinical NLP

1. Introduction

The biomedical domain faces rapid growth in data generation, characterized by massive volume, linguistic diversity, and modal heterogeneity. This exponential growth of scientific literature has been well-documented [1], creating significant challenges for information synthesis. Critical information is scattered across research articles, clinical reports, databases, and grey literature. This fragmentation creates substantial documentation burden for healthcare professionals [2, 3].

Recent advances in Generative AI and Large Language Models (LLMs) offer significant potential [4]. Technologies such as Retrieval-Augmented Generation (RAG) and Knowledge Graphs (KGs) are promising for improving reliability and accuracy in specialized domains like biomedicine [5, 6]. Recent surveys demonstrate the growing importance of RAG with LLMs for biomedical applications [6]. However, optimal integration of these components within coherent AI architectures remains an active research area.

2. Related Work

Biomedical summarization has evolved from traditional extractive methods to sophisticated neural approaches. Comprehensive surveys of document summarization techniques [7, 8] show this evolution from rule-based to neural approaches. Early work focused on rule-based systems and statistical methods,

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

[*] Corresponding author: Johanna Angulo (johanna.angulo@gmail.com)

✉ johanna.angulo@gmail.com (J. Angulo); victor.yeste@universidadeuropea.es (V. Yeste)
🌐 https://www.linkedin.com/in/johannaangulo/ (J. Angulo); https://victoryeste.com/ (V. Yeste)
iD 0009-0005-6965-0604 (J. Angulo); 0000-0002-3660-8347 (V. Yeste)

while recent advances leverage transformer architectures and domain-specific models. Multilingual biomedical summarization remains under-explored, with most systems designed for English. The MultiClinSum challenge addresses this gap by providing standardized evaluation across four languages. Large language models are increasingly being integrated into clinical decision support systems [9], though questions remain about their role in enhancing versus replacing human expertise.

3. MultiClinSum

MultiClinSum represents a shared task addressing the automatic summarization of clinical case reports across four major languages: English, Spanish, French, and Portuguese. This challenge responds to the critical need in healthcare and biomedical research domains, where the exponential growth of clinical documentation creates significant barriers for healthcare professionals, researchers, and patients attempting to extract essential medical knowledge from extensive clinical texts [10].

The challenge is structured into four independent sub-tracks (MultiClinSum-en, -es, -fr, -pt), providing participants the flexibility to focus on specific languages or develop comprehensive multilingual approaches [4]. Teams may submit up to five runs per language, with evaluation conducted using ROUGE-L-Sum and BERTScore metrics against human-generated reference summaries. This task addresses critical healthcare applications including clinical decision support, discharge summary generation, medical literature review, multilingual clinical communication, and patient-oriented summary creation, connecting natural language processing research with practical clinical applications across linguistic boundaries [10].

3.1. Gold and Large-Scale Dataset

The MultiClinSum challenge utilizes a multilingual corpus comprising both training and evaluation datasets across four major languages: English, Spanish, French, and Portuguese. The corpus is structured into two distinct training components and corresponding test sets, all made publicly available through the Zenodo repository to ensure reproducibility and accessibility for the research community [10].

The training data encompasses both gold-standard and large-scale variants, with the gold-standard datasets containing 592 document-summary pairs per language, providing high-quality reference materials for supervised learning approaches.

The large-scale training datasets expand the available data with 25,902 document-summary pairs per language, enabling the development and fine-tuning of data-intensive models while maintaining consistent cross-lingual coverage [10].

3.2. Test Dataset

The evaluation framework (test dataset) employs language-specific test datasets containing between 3,396 and 3,469 full-text clinical cases per language, with slight variations reflecting the natural distribution of available clinical literature across linguistic domains. All datasets maintain consistent organizational structure, with full-text documents and their corresponding summaries stored in separate directories as UTF-8 encoded plain text files [10].

4. Methodology

MCS is a multi-agent and multimodal system designed for summarizing complex biomedical texts from various languages [11]. It leverages LangGraph to create a stateful, resilient workflow, LlamaIndex for Retrieval-Augmented Generation (RAG) [12], and a suite of specialized agents and tools for different tasks using Large Language Model (LLM) [13].

This framework functions as an abstraction, defining a flexible multi-agent system wherein different architectural configurations for summarization can be implemented and compared [8]. The architectures detailed below represent specific instances developed and deployed for the challenge. It is important

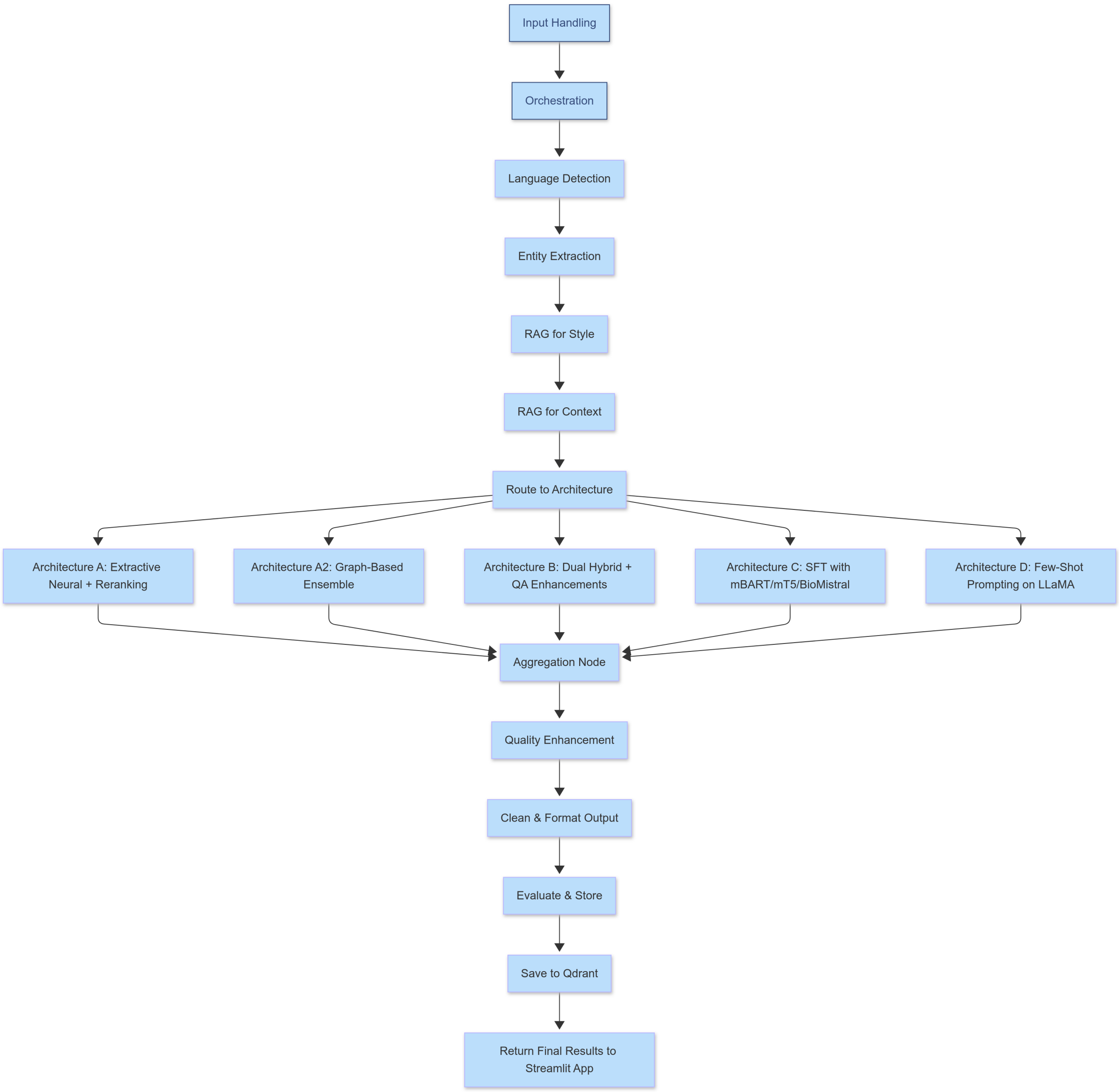

Input Handling
Orchestration
Language Detection
Entity Extraction
RAG for Style
RAG for Context
Route to Architecture
Architecture A: Extractive Neural + Reranking
Architecture A2: Graph-Based Ensemble
Architecture B: Dual Hybrid + QA Enhancements
Architecture C: SFT with mBART/mT5/BioMistral
Architecture D: Few-Shot Prompting on LLaMA
Aggregation Node
Quality Enhancement
Clean & Format Output
Evaluate & Store
Save to Qdrant
Return Final Results to Streamlit App

Figure 1: LangGraph architecture orchestrates the collaboration of distinct, interconnected components—mediated by AI agents and tools—to execute different architectures. (Source: elaborated by the authors).

to note that the runs submitted for the challenge may align with one or more of these conceptual architectures.

The scope of the MCS system is considerably broader than what was implemented for this specific challenge. Consequently, only the architectures directly relevant to our submitted runs are presented herein.

Agentic Decision Making: The Route to Architecture component uses similarity analysis of input documents against stored performance vectors to automatically select optimal architectures. The system calculates cosine similarity between document embeddings and historical performance patterns, routing to the architecture with highest expected performance (similarity threshold ≥ 0.85).

RAG Configuration: Retrieval uses k=5 most similar patterns, with BM25 + dense retrieval hybrid ranking. Deduplication removes patterns with $\geq$90% overlap before final selection.

4.1. Architecture A: Extractive with Neural Reranking

This architecture leverages a sequential pipeline of well-established Natural Language Processing (NLP) techniques. The process commences with an extractive phase, where an initial set of candidate summaries is generated using traditional methods such as BM25[14]. These candidates are subsequently processed by an LLM to produce more fluent, abstractive versions [15].

A critical component of this architecture is the final reranking stage, which serves as a quality control mechanism [16]. Reranking is conceptualized as a two-stage process [17]. This methodology allows for the strategic application of powerful models on a pre-filtered set of promising candidates, thereby optimizing the trade-off between computational cost and output quality.

4.2. Architecture A2: Graph-Based Ensemble

This architecture introduces an approach by constructing a structured representation of the source document's content. It begins by extracting key entities and their semantic relations to build a document-specific knowledge graph (KG) using the `networkx` library [5, 18]. This graph is then condensed or serialized back into a textual format, which effectively serves as a dense, structured summary of the original document [19]. Knowledge graph-guided retrieval approaches [20] have shown effectiveness in maintaining factual consistency.

An LLM is then prompted to generate an abstractive summary, with the serialized knowledge graph provided as strong contextual guidance [21]. This ensures that the generated summary is not only fluent but also factually anchored to the core information captured in the KG. The final output is a hybrid summary that synergistically combines the KG-guided abstractive text with a traditional extractive summary produced by BM25, thus balancing narrative coherence with factual fidelity [19, 14].

4.3. Architecture B: Dual Hybrid + QA Enhancements

Building upon the principles of Architecture A2, this configuration incorporates a highly optimized post-processing system. The objective is to merge the strengths of fine-tuned models, such as mBART and mT5 [22], with the advanced reasoning and generation capabilities of GPT-4o [23]. This enhancement is informed by a detailed analysis of gold-standard summaries.

4.4. Architecture C: Supervised Fine-Tuning (SFT) of Multilingual Models

This architecture is centered on exploring the efficacy of domain adaptation via Supervised Fine-Tuning (SFT) [24]. This process adapts the model's general capabilities to the specific nuances of the summarization task at hand. The models subjected to SFT in this study were mBART, mT5 [25], and BioMistral [26].

4.5. Architecture D: Few-Shot Prompting on General-Purpose LLMs

This set of architectures leverages the powerful in-context learning capabilities of large-scale, non-fine-tuned LLMs. The core technique employed is Few-Shot Prompting [27], which involves conditioning the model by providing a small number of illustrative examples or example patterns (i.e., "shots") of the task directly within the prompt. This guides the model to generate output that conforms to the desired format and style, all without the need for updating the model's underlying parameters. To enhance the effectiveness of the prompts, they were enriched with several layers of information such as NER Extraction [28] and gold document patterns. This approach was systematically tested on two prominent models: `Llama 3.1 8B` and `Mistral 7B`. Additionally, for a subset of the Spanish-language runs, a post-processing step was implemented using GPT-4o. This step, which also falls under the paradigm of few-shot prompting, was designed to further refine and enhance the quality of the generated summaries.

4.6. A Comprehensive Analysis of Medical Gold Summaries

A comprehensive analysis of medical gold summaries was conducted to extract optimal patterns for automated summarization systems. The analysis employed both traditional NLP metrics and advanced

Table 1
Fine-tuning configuration by language

Language	Fine-tuned Models	NER Model	Epochs	Batch	Grad Acc	LR
English	BioMistral-7B	Helios9/BioMed	3	1	16	2e-4
Spanish (1)	mBART-large-50, mT5-base	BSC-NLP4BIA/ bsc-bio-ehr-es	3	1	8	2e-5
Spanish (2)	BioMistral-7B	BSC-NLP4BIA/ bsc-bio-ehr-es	3	1	16	2e-4
French	mBART-large-50, Mixtral-8x7B*	TypicaAI/ HealthcareNER-Fr	3/2**	1	8/8	2e-5/2e-4
Portuguese	mBART-large-50, mT5-base	HUMADEX/ portuguese_medical_ner	3	1	8	2e-5

*Mixtral-8x7B faced device management issues during inference and was not successfully deployed.
**Mixtral used 2 epochs for lighter fine-tuning; values show mBART/Mixtral where different.

NER (Named Entity Recognition) [28] techniques to identify key characteristics that define high-quality medical summaries. For each language, the analysis was conducted independently, and the corresponding outputs were encoded into dense vector representations and stored in a vector database. These embeddings were later retrieved during inference using a Retrieval-Augmented Generation (RAG) [29, 6] architecture to provide contextual grounding for the language model [6].

Key Findings & Pattern Extraction

The quantitative analysis presented below focuses on Spanish gold-standard biomedical summaries as a representative example. Similar analyses were conducted for each language in our multilingual dataset, with language-specific patterns informing the respective summarization strategies.

- **Target Length:** 120.1 words (from gold dataset analysis)
- **Target Sentences:** 5 sentences
- **Average Sentence Length:** ~24 words per sentence

Words in Summaries. The distribution is centered around a mean of 120.1 words, indicating moderate summary length with a right-skewed tail due to a few longer samples.

Medical Terms. Most summaries contain fewer than 10 medical terms, with a mean of 4.0. These medical terms are obtained through regex rules designed to capture common medical terminology not included in the NER model. This suggests that while summaries are concise, they embed relevant domain-specific terminology.

NER Entities. The Named Entity Recognition (NER) distribution shows a broader spread, with a mean of 5.2 entities per summary. For NER processing, we employed different biomedical NER models per language, with the specific models detailed in the table above. This reflects consistent inclusion of structured biomedical concepts across different linguistic contexts.

4.7. Fine-Tuning Configuration

Table 1 presents the complete fine-tuning configuration for each language:

Hardware and Memory Optimization: All fine-tuning was conducted on NVIDIA GPUs with comprehensive memory optimization strategies. For large language models (BioMistral-7B, Mixtral-8x7B), we employed 4-bit quantization using BitsAndBytesConfig with NF4 quantization type and double quantization enabled. Gradient checkpointing was activated to trade computation time for memory efficiency.

Training Methodology: The fine-tuning process varied by model architecture. For sequence-to-sequence models (mBART, mT5), we used standard Seq2SeqTrainer with target-specific tokenization. For causal language models (BioMistral, Mixtral), we applied Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Low-Rank Adaptation) with rank=16, alpha=32, and dropout=0.1, targeting attention projections and feed-forward layers.

Language-Specific Configurations: mBART models were configured with appropriate language codes (en_XX, es_XX, fr_XX, pt_XX) for source and target languages. Mixtral employed instruction-tuned prompts with chat formatting. All models used a maximum sequence length of 1024 tokens for input and 256 for output summaries.

Training Hyperparameters:

- **Loss Function:** Cross-entropy loss for all models (sequence-to-sequence for mBART/mT5, language modeling for BioMistral/Mixtral)
- **Optimizer:** AdamW with weight decay 0.01
- **Warmup:** 25-50 steps depending on dataset size
- **Scheduler:** Linear decay after warmup
- **Precision:** FP16 enabled where supported, BF16 for Mixtral
- **Evaluation:** Loss-based with early stopping disabled
- **Saving:** Best model based on evaluation loss, limited to 1-2 checkpoints

Dataset Preparation: Training datasets consisted of 592 gold standard document-summary pairs per language, except for Spanish Run 2 which utilized an extended dataset of 12,100 pairs. For causal models, documents were formatted with instruction templates emphasizing biomedical summarization objectives and clinical relevance.

4.8. Language-Specific Implementation Details

We provide comprehensive implementation details for each run across all languages:

English Runs:

- **Run 1 (Architecture A):** Phase 1: BM25 Extractive preprocessing. Phase 2: BioMistral fine-tuning. Phase 3: Hybrid generation with semantic reranking. Strategy: 3-stage pipeline optimization.
- **Run 2 (Architecture B):** Phase 1: NER entity extraction. Phase 2: BioMistral generation. Phase 3: Coverage enhancement validation. Strategy: NER-integrated summarization.

Spanish Runs:

- **Run 1 (Architecture A2):** Phase 1: Model fine-tuning (mBART, mT5). Phase 2: Knowledge graph ensemble. Phase 3: NER preservation validation. Strategy: Dual-model hybrid approach.
- **Run 2 (Architecture A+D):** Phase 1: BM25 extractive preprocessing. Phase 2: Llama 3.1 8B generation. Phase 3: GPT-4o quality enhancement. Strategy: 3-stage quality assurance.
- **Run 3 (Architecture A+D):** Phase 1: BM25 extractive preprocessing. Phase 2: Mistral 7B generation. Phase 3: Dense retrieval reranking. Strategy: Dense retrieval optimization.
- **Run 4 (Architecture B+D):** Phase 1: NER-guided KG summarization. Phase 2: Hybrid generation. Phase 3: GPT-4o quality assurance. Strategy: Dual-model with proprietary enhancement.
- **Run 5 (Architecture C):** Phase 1: BioMistral fine-tuning on large-scale data. Phase 2: Hybrid summarization. Phase 3: Semantic reranking. Strategy: Parameter-efficient adaptation.

Portuguese Runs:

- **Run 1 (Architecture C+A):** Phase 1: Fine-tuning (mBART, mT5). Phase 2: Initial hybrid approach (failed). Phase 3: Revised BM25-neural hybrid. Strategy: Fine-tuning with fallback mechanism.
- **Run 2 (Architecture A):** Phase 1: Re-ranking with multiple summarization modalities. Phase 2: Gold pattern extraction. Phase 3: Quality control pipeline. Strategy: Pattern-based reranking.

Table 2
Complete evaluation results for all runs

Language	Run	BERTScore F1	ROUGE F1
English	1	0.842	0.176
	2	0.840	0.167
Spanish	1	0.702	0.218
	2	0.742	0.247
	3	0.727	0.227
	4	0.728	0.225
	5	0.710	0.191
Portuguese	1	0.689	0.197
	2	0.710	0.187
French	1	0.712	0.197

French Run:

- **Run 1 (Architecture C):** Phase 1: Fine-tuning (mBART, Mixtral). Phase 2: NER-guided generation. Phase 3: Quality normalization with BM25 fallback. Strategy: Quality threshold optimization.

Architecture Legend: A: Extractive with Neural Reranking; B: Knowledge Graph-Based Ensemble; C: Supervised Fine-Tuning; D: Few-Shot Prompting.

5. Results

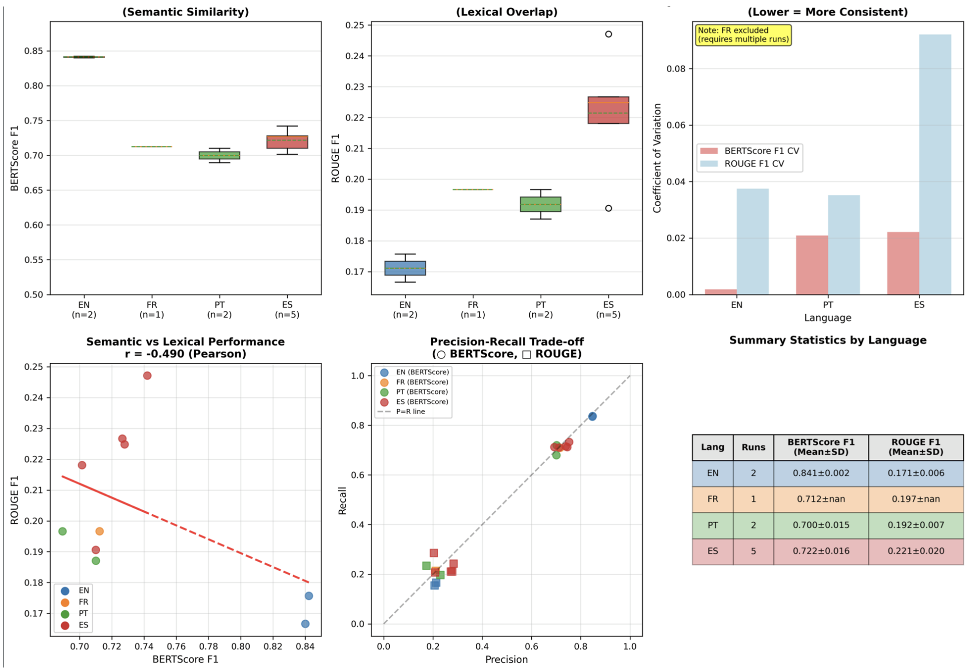
Table 2 presents the complete evaluation results across all runs and languages.

English Results. Two experimental runs demonstrated consistent performance with minimal variation. BERTScore results showed good semantic similarity (F1: 0.842, 0.840) with Run 1 slightly outperforming Run 2 (Precision: 0.848, Recall: 0.837). ROUGE metrics indicated more constrained lexical overlap (F1: 0.176, 0.167), with Run 1 achieving superior performance (Precision: 0.213, Recall: 0.165). Lower ROUGE scores compared to Spanish suggest greater lexical diversity in English biomedical references.

Spanish Results. Five runs showed consistent performance with moderate variance. BERTScore F1 ranged 0.701–0.742, with `es_run_2` achieving highest performance (F1: 0.742, Precision: 0.752, Recall: 0.732). All runs maintained balanced precision-recall trade-offs with recall above 0.710. ROUGE F1 scores ranged 0.190–0.247, with Run 2 demonstrating superior lexical overlap (F1: 0.247, Precision: 0.283, Recall: 0.242). Lower ROUGE compared to BERTScore indicates effective semantic capture despite challenging exact lexical matching.

Portuguese Results. Run 2 achieved higher BERTScore F1 (0.71 vs 0.689) while Run 1 demonstrated superior ROUGE F1 (0.196 vs 0.187). Both runs showed identical BERTScore precision (0.70), with Run 2 exhibiting higher recall (0.71 vs 0.67) and Run 1 displaying greater ROUGE precision (0.23 vs 0.173).

French Results. Single run `fr_run_1` achieved solid semantic performance (BERTScore F1: 0.712, Precision: 0.716, Recall: 0.709), positioning between Spanish (0.701–0.742) and English (0.84+) results. ROUGE metrics yielded F1: 0.196 (Precision: 0.210, Recall: 0.213), with recall exceeding precision, indicating comprehensive summary generation with some redundancy.



Lang	Runs	BERTScore F1 (Mean±SD)	ROUGE F1 (Mean±SD)
EN	2	0.841±0.002	0.171±0.006
FR	1	0.712±nan	0.197±nan
PT	2	0.700±0.015	0.192±0.007
ES	5	0.722±0.016	0.221±0.020

Figure 2: Performance comparison across all runs showing BERTScore F1 vs ROUGE F1

6. Analysis and Discussion

6.1. Cross-Lingual Performance Patterns

The results reveal distinct performance hierarchies across languages as shown in Figure 2.

- **Semantic Similarity (BERTScore):** English (0.842) > Spanish (0.742) > French (0.712) > Portuguese (0.710)
- **Lexical Overlap (ROUGE):** Spanish (0.247) > French/Portuguese (0.197) > English (0.176)

6.2. Architecture Effectiveness

Three-stage pipelines (Spanish Runs 2-3) achieved best performance by combining extractive preprocessing, neural generation, and quality enhancement. The integration of GPT-4o for post-processing (Spanish Run 2) yielded the best overall results across both metrics. The GPT-4o quality assurance operates as an intelligent agent within our LangGraph multi-agent architecture framework, functioning as a specialized tool that performs automated quality validation, content refinement, and error correction on generated summaries. While the system architecture supports both GPT-4o and GPT-4.1 models, we utilized GPT-4o for all quality assurance operations in this implementation.

Knowledge Graph approaches (Spanish Runs 1, 4) demonstrated competitive performance, with GPT-4o enhancement providing improvements (+3.2% ROUGE, +3.7% BERTScore over baseline KG approaches). The multi-agent framework enables seamless integration where the knowledge graph extraction agent passes structured medical entities to the GPT-4o quality assurance agent for validation and summary enhancement.

Fine-tuning strategies showed mixed results. BioMistral fine-tuning (Spanish Run 5) underperformed despite extensive parameter-efficient training, likely due to the English-centric pre-training conflicting with Spanish medical terminology requirements.

6.3. Semantic vs. Lexical Trade-offs

The analysis reveals a trade-off between semantic coherence and lexical overlap (Pearson r = -0.490, $p < 0.05$). English achieves maximum semantic performance but lowest lexical overlap, while Spanish demonstrates the most balanced profile across both metrics.

This pattern suggests that different languages require distinct optimization strategies: English benefits from semantic preservation approaches, while Spanish allows for better balance between abstractive generation and lexical alignment.

6.4. Error Analysis

The challenges of building multilingual language models for medicine [30] are reflected in our cross-lingual performance variations. Key limitations identified include:

- Language-specific model adaptation challenges, particularly for domain-specific models trained primarily in English
- Entity preservation vs. fluency trade-offs when implementing NER-guided generation
- Computational resource constraints affecting architecture selection, especially for large multilingual models

7. Conclusions

This work presents Agentic MCS, a multilingual framework for biomedical text summarization evaluated across four languages. Key findings include:

English achieves high semantic similarity (BERTScore: 0.842) through optimized neural architectures, demonstrating the effectiveness of domain-specific fine-tuning with BioMistral and hybrid reranking approaches.

Spanish delivers the most balanced performance across both metrics (BERTScore: 0.742, ROUGE: 0.247), with three-stage pipelines and GPT-4o enhancement proving most effective. The diversity of architectural approaches (5 runs) provides valuable insights into optimal strategy selection.

Portuguese and **French** achieve good results (BERTScore: 0.71+, ROUGE: 0.197) through adapted architectures, though with distinct optimization patterns reflecting language-specific characteristics. Portuguese benefits from pattern mining approaches, while French shows effectiveness of quality normalization strategies.

Architecture effectiveness varies significantly by language, with ensemble approaches and quality enhancement mechanisms outperforming single-model systems across all languages. The systematic trade-off between semantic and lexical optimization provides insights for future multilingual biomedical summarization systems.

The consistent semantic performance (BERTScore 0.71-0.84) across languages demonstrates the framework's robust abstractive capabilities, while varying lexical overlap reflects language-specific adaptation requirements.

8. Future Work

The current analysis reveals several promising directions for advancing multilingual biomedical summarization capabilities. Building upon our demonstrated semantic fidelity across languages, future research will pursue three complementary directions.

Comprehensive Evaluation Framework: Comprehensive baseline comparisons and ablation studies are planned for the extended journal version to quantify the individual contributions of each architectural component and validate the necessity of multi-agent complexity.

Enhanced Quality Assurance Framework: We plan to develop enhanced RAG-based quality assurance mechanisms that leverage our multi-agent architecture to maintain consistency across languages while addressing the identified limitations in entity preservation and fluency trade-offs.

Multimodal Integration: The evolution toward a multimodal clinical summarization framework represents our primary long-term objective. This development will focus on efficiently processing multimodal biomedical inputs prevalent in clinical and research settings, aligning with advances in multimodal biomedical foundation models [31] that can process both text and image data effectively.

Advanced Cross-lingual Adaptation: Future work will develop systematic approaches to model fine-tuning with incrementally larger datasets combined with quality-supervised training methodologies. This includes expanded cross-lingual transfer learning strategies to better address language-specific medical terminology requirements and optimize the semantic versus lexical trade-offs identified in our analysis.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Bornmann, R. Haunschild, R. Mutz, Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases, Humanities and Social Sciences Communications 8 (2021) 1–15.

[2] A. Holmgren, J. Adler-Milstein, N. Apathy, Electronic health record documentation burden crowds out health information exchange use by primary care physicians: Article examines electrnoic health record documentation burden, Health Affairs 43 (2024) 1538–1545. doi:10.1377/hlthaff.2024.00398.

[3] E. Asgari, J. Kaur, G. Nuredini, J. Balloch, A. Taylor, N. Sebire, R. Robinson, C. Peters, S. Sridharan, D. Pimenta, Impact of electronic health record use on cognitive load and burnout amongst clinicians: A narrative review (preprint), JMIR Medical Informatics 12 (2023). doi:10.2196/55499.

[4] D. Veen, C. Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Blüthgen, A. Pareek, M. Polacin, E. Reis, A. Seehofnerova, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. Langlotz, J. Hom, S. Gatidis, A. Chaudhari, Clinical text summarization: Adapting large language models can outperform human experts, Research square (2023). doi:10.21203/rs.3.rs-3483777/v1.

[5] X. Ge, Y. C. Wang, B. Wang, C.-C. J. Kuo, et al., Knowledge graph embedding: An overview, APSIPA Transactions on Signal and Information Processing 13 (2024).

[6] M. Arslan, H. Ghanem, S. Munawar, C. Cruz, A survey on RAG with LLMs, Procedia Computer Science 246 (2024) 3781–3790.

[7] H. Y. Koh, J. Ju, M. Liu, S. Pan, An empirical survey on long document summarization: Datasets, models and metrics, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3545176. doi:10.1145/3545176.

[8] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, D. R. I. M. Setiadi, Review of automatic text summarization techniques & methods, Journal of King Saud University-Computer and Information Sciences 34 (2022) 1029–1046.

[9] J. Li, Z. Zhou, H. Lyu, Z. Wang, Large language models-powered clinical decision support: Enhancing or replacing human expertise?, Intelligent Medicine 05 (2025) 1–4. doi:10.1016/j.imed.2025.01.001.

[10] 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.

[11] D. B. Acharya, K. Kuppan, B. Divya, Agentic AI: Autonomous intelligence for complex goals—a comprehensive survey, IEEE Access 13 (2025) 18912–18936. doi:10.1109/ACCESS.2025.3532853.

[12] S. Liu, A. B. McCoy, A. Wright, Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines, Journal of the American Medical Informatics Association (2025) ocaf008.

[13] S. Thapa, S. Adhikari, ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls, Annals of Biomedical Engineering 51 (2023) 2647–2651. URL: https://doi.org/10.1007/s10439-023-03284-0. doi:10.1007/s10439-023-03284-0.

[14] A. Trotman, A. Puurula, B. Burgess, Improvements to BM25 and language models examined, in: Proceedings of the 19th Australasian Document Computing Symposium, ADCS 14, Association for Computing Machinery, New York, NY, USA, 2014, p. 58–65. URL: https://doi.org/10.1145/2682862.2682863. doi:10.1145/2682862.2682863.

[15] W. Lan, Z. Tang, M. Liu, Q. Chen, W. Peng, Y. P. Chen, Y. Pan, The large language models on biomedical data analysis: A survey, IEEE Journal of Biomedical and Health Informatics (2025).

[16] M. Rybinski, J. Xu, S. Karimi, Clinical trial search: Using biomedical language understanding models for re-ranking, Journal of Biomedical Informatics 109 (2020) 103530. URL: https://www.sciencedirect.com/science/article/pii/S1532046420301581. doi:https://doi.org/10.1016/j.jbi.2020.103530.

[17] E. Yoo, G. Kim, S. Kang, Summary-sentence level hierarchical supervision for re-ranking model of two-stage abstractive summarization framework, Mathematics 12 (2024) 521.

[18] M. Kaur, H. Kaur, Implementation of enhanced graph layout algorithm for visualizing social network data using NetworkX Library, International Journal of Advanced Research in Computer Science 8 (2017).

[19] X. Wang, L. Chen, T. Ban, M. Usman, Y. Guan, S. Liu, T. Wu, H. Chen, Knowledge graph quality control: A survey, Fundamental Research 1 (2021) 607–626.

[20] X. Zhu, Y. Xie, Y. Liu, Y. Li, W. Hu, Knowledge Graph-Guided Retrieval Augmented Generation, 2025. doi:10.48550/arXiv.2502.06864. arXiv:2502.06864.

[21] P.-E. Genest, G. Lapalme, Framework for abstractive summarization using text-to-text generation, in: Proceedings of the workshop on monolingual text-to-text generation, 2011, pp. 64–73.

[22] P. Wilman, T. Atara, D. Suhartono, Abstractive english document summarization using BART model with chunk method, Procedia Computer Science 245 (2024) 1010–1019. URL: https://www.sciencedirect.com/science/article/pii/S1877050924031375. doi:https://doi.org/10.1016/j.procs.2024.10.329, 9th International Conference on Computer Science and Computational Intelligence 2024 (ICCSCI 2024).

[23] J. Gallifant, A. Fiske, Y. A. Levites Strekalova, J. S. Osorio-Valencia, R. Parke, R. Mwavu, N. Martinez, J. W. Gichoya, M. Ghassemi, D. Demner-Fushman, et al., Peer review of GPT-4 technical report and systems card, PLOS digital health 3 (2024) e0000417.

[24] L. Wang, S. Chen, L. Jiang, et al., Parameter-efficient fine-tuning in large language models: a survey of methodologies, Artificial Intelligence Review 58 (2025). URL: https://doi.org/10.1007/s10462-025-11236-4. doi:10.1007/s10462-025-11236-4.

[25] K. Pająk, D. Pająk, Multilingual fine-tuning for grammatical error correction, Expert Systems with Applications 200 (2022) 116948. URL: https://www.sciencedirect.com/science/article/pii/S0957417422003773. doi:10.1016/j.eswa.2022.116948.

[26] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, R. Dufour, BioMistral: A collection of open-source pretrained large language models for medical domains, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 5848–5864. URL: https://aclanthology.org/2024.findings-acl.348/. doi:10.18653/v1/2024.findings-acl.348.

[27] T. Schick, H. Schütze, True few-shot learning with Prompts—A real-world perspective, Transactions of the Association for Computational Linguistics 10 (2022) 716–731. URL: https://aclanthology.org/2022.tacl-1.41/. doi:10.1162/tacl_a_00485.

[28] N. Perera, M. Dehmer, F. Emmert-Streib, Named entity recognition and relation detection for biomedical information extraction, Frontiers in cell and developmental biology 8 (2020) 673.

[29] Y. Yang, et al., Pseudo-Knowledge Graph: Meta-Path Guided Retrieval and In-Graph Text for RAG-Equipped LLM, https://arxiv.org/html/2503.00309v1, 2025. doi:10.48550/arXiv.2503.00309. arXiv:2503.00309.

[30] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, W. Xie, Towards building multilingual language model for medicine, Nature Communications 15 (2024) 8384. doi:10.1038/s41467-024-52417-z.

[31] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, H. Poon, BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. doi:10.48550/arXiv.2303.00915. arXiv:2303.00915.