Keywords
Biomedical text summarization, hybrid summarization, cross-domain semantic fusion, entity-aware salience, hallucination suppression, multi-reward reinforcement learning, contrastive learning, transformer architectures, BART, BioBERT.

Authors
Siddu Tushara M S[1*],
[1*]Research Scholar, JAIN (DEEMED-TO-BE UNIVERSITY) Bengaluru, India. tusharams23@gmail.com

Dr Renukadevi S[2]
[2]Research Guide, JAIN (DEEMED-TO-BE UNIVERSITY) Bengaluru, India. renuka.devi@jainuniversity.ac.in

Received:15-06-2026
Revised:20-07-2026
Accepted: 24-07-2026
Doi:10.1922/ejprd.v34i5s.1560

European Journal of Prosthodontics and Restorative Dentistry (2026) 34(5s), 314–331

A Knowledge-Aware Hybrid Summarization Framework with Cross-Domain Semantic Fusion, Entity-Aware Salience Scoring, and Hallucination-Suppressed Abstractive Generation for Biomedical Literature and Clinical Text

ABSTRACT
Due to the exponentially growing amount of biomedical literature with PubMed reaching more than 36 million indexed publications by 2025, the manual extraction of knowledge is operationally unachievable, thus necessitating the development of automated systems that produce clinically meaningful summaries. Three systematic shortcomings can be identified among state-of-the-art neural summarization methods: knowledge-agnostic sentence selection failing to select the sentences containing PICO-structured evidence, abstractive decoder that is not restricted to generate entity-mentioning sentences, and the hybrid architectures that are not tightly integrated and, hence, hinder the exchange of knowledge between the extractive and abstractive modules. In this paper, we introduce BioSummHybrid+, an approach to Knowledge-Aware Hybrid Summarization that tackles all the above issues by integrating the following five innovations: Cross-Domain Semantic Fusion Gate (CSFG), Entity-Aware Salience Scorer (EASS), Hierarchical Discourse-Aware Positional Encoding (HDAPE), Multi-Reward Reinforcement Learning Decoding (MRLD), and Contrastive Hallucination Suppression (CHS). Our framework, BioSummHybrid+, was compared on five biomedical benchmarks (PubMed, SciTLDR, MS2, MultiXSci, ArXiv) against six baselines on two metrics (BERTScore-F1 and Faithfulness and Entity Preservation Rate) that reflect summarization quality in the context of high-stakes healthcare in a more adequate way compared to ROUGE and BLEU. It turns out that the higher ROUGE score is achieved by baselines through simple extractive lexical copying. Thus, we show that structured domain knowledge plays a vital role in summarization architecture.

I. INTRODUCTION

The biomedical literature database grows at an unprecedented pace. PubMed contains over 36 million articles by 2025, with 4,000 additional articles being published each day. The scale of biomedical knowledge makes comprehensive manual processing impractical for clinical researchers, evidence-based practice professionals, and systematic reviewers. Biomedical text summarization [1] therefore becomes a crucial topic in research and can potentially revolutionize evidence biomedical language is rich in named entities, such as drug names, biomarkers, statistics indicators, genetic markers, and clinical outcomes, which need to be preserved for patient safety.

Neural text summarization architectures are built using two primary paradigms. Extractive summaries preserve the factual correctness of the summary but fail to produce semantically coherent text. Abstractive summaries are fluent and easy to understand but are prone to hallucination – generation of clinically incorrect facts. Hybrid approaches integrate both paradigms but existing solutions do not allow structured knowledge to flow bidirectionally across the stages of the pipeline.

The major, yet rarely recognized challenge of biomedical summarization is the excessive reliance on ROUGE metrics for evaluation. ROUGE compares n-grams between generated and reference summaries [4], which naturally favors the extractive paradigm and copy-paste of source text as opposed to abstractive interpretation of meaning. In a clinical setting, however, the difference is crucial – the summary produced by the model which copy-pastes source text may achieve high ROUGE scores but will fail to synthesize the evidence across different elements of PICO question. Recently, some researchers have confirmed the poor correlation of ROUGE with human judgements of the quality of biomedical summaries (Cohan and Goharian, 2016; Bhandari et al., 2020). It was also shown that BERTScore, Faithfulness, and Entity Preservation Rate are much better indicators of clinical usefulness of biomedical summaries. BioSummHybrid+ is explicitly optimized for these clinically relevant metrics.

To address these challenges, we present BioSummHybrid+, a Knowledge-Aware Hybrid Summarization framework that introduces five tightly integrated architectural innovations: the Cross-Domain Semantic Fusion Gate (CSFG), the Entity-Aware Salience Scorer (EASS), Hierarchical Discourse-Aware Positional Encoding (HDAPE), Multi-Reward Reinforcement Learning Decoding (MRLD), and Contrastive Hallucination Suppression (CHS).

The principal contributions of this paper are: (1) BioSummHybrid+, a comprehensive knowledge-aware hybrid summarization framework integrating biomedical domain knowledge across all pipeline components through five novel architectural mechanisms; (2) the Cross-Domain Semantic Fusion Gate as a differentiable module combining contextual and domain-specific semantic representations through learnable sigmoid gating; (3) the Entity-Aware Salience Scorer integrating biomedical entity information directly into sentence importance estimation; (4) Multi-Reward Reinforcement Learning Decoding optimizing a composite clinical-fidelity reward signal; and (5) comprehensive synthesis while retaining clinical accuracy needed for evidence-based practice. Unlike generic document summarization, biomedical text summarization faces domain-specific challenges which cannot be addressed with the use of generic neural architectures. Biomedical documents are built according to the Population-Intervention-Comparison-Outcome (PICO) structure, which encodes the clinical question[3] of every experimental study. Loss of PICO data means loss of clinical relevance of the summary. Furthermore,

experiments on five biomedical benchmarks demonstrating statistically significant improvements over six competitive baselines.

The remainder of this paper is organized as follows. Section II reviews relevant literature. Section III presents the complete BioSummHybrid+ architecture. Section IV describes the experimental setup. Section V presents results and analysis. Section VI discusses implications and limitations. Section VII concludes the paper.

II. LITERATURE REVIEW

The area of biomedical text summarization has made considerable progress in the last two decades[5] [6] by shifting from statistical NLP methods to the hybrid transformer architecture. The early work in biomedical summarization used TF-IDF scoring and Latent Semantic Analysis (LSA). The initial literature survey by Mishra et al. (2014) of 34 biomedical summarization papers from 2000–2013[7] found that 82% of the papers used the statistical NLP approach and specifically pointed to the absence of incorporation of structured domain knowledge as the major research gap. Graph-based summarizers TextRank and LexRank built upon extractive summarization by using centrality based sentence ranking[8], however, the algorithms were still knowledge agnostic. BioBERTSum modified the architecture of BERTSUM using domain-specific BioBERT encodings[9], and Kirmani et al. (2024) suggested the semantic summarizer, which uses the combination of BioBERT sentence encoding[10], Bio-Word2Vec, K-means clustering, and TF-IDF scoring and showed 0.43 ROUGE-1, 0.32 ROUGE-2, and 0.46 ROUGE-L. Another interesting line of research is the development of PICO-aware extractive methods. Dernoncourt and Lee (2017) published the PubMed 200k RCT dataset[11] and showed that BiLSTM architecture with CRF postprocessing achieves 91.6% F1 on PICO classification task. However, even with these improvements, extractive models retain the property of factual consistency at the price of semantically disconnected summaries.

Transformer architectures revolutionized abstractive summarization. The T5, PEGASUS, and BART set the performance standard on the general-domain datasets[12]. On the other hand, domain-specific pre-training is necessary for biomedical applications[13]. SciFive, which is T5-based architecture, pre-trained on PubMed and PMC full texts[14], showed SOTA performance on three out of seven biomedical NER benchmarks. The BioGPT model, which was trained on 15 million PubMed abstracts[15] and showed 78.2% accuracy on PubMedQA, has decoder-only architecture, making it especially susceptible to hallucinations when applied to the long biomedical document. The problem of hallucination becomes increasingly important due to its

role in patient safety. Two major failure modes of the language models were revealed by Maynez et al. (2020)[16] – intrinsic hallucinations that contradict the source content and extrinsic hallucinations, which introduce information, not present in the source. Asgari et al. (2024) found that GPT-4 generates clinically relevant hallucinations[17] in 18.3% of medical summaries, making the explicit hallucination suppression mechanisms an indispensable part of architecture for any clinical application.

Hybrid extractive-abstractive approaches have emerged as the dominant paradigm in recent biomedical summarization research. A systematic review by Bashir et al. (2025) found that hybrid models account for 69.64%[18] of publications across 56 biomedical summarization papers from 2017–2024. SKT5SciSumm coupled SPECTER citation-aware sentence embeddings[19] with K-means clustering and T5-family generation, achieving ROUGE-1 of 31.49% and BERTScore-F1 of 85.29% on Multi-XScience. MedicoVerse combined SapBERT hierarchical agglomerative entity clustering with BART-large-cnn-samsum, achieving ROUGE-L F1 of 0.68 on pharmaceutical regulatory documents, though it was validated on only 227 document segments. A scoping review by Bednarczyk et al. (2025) systematically identified factual consistency[20] and domain-specific clinical evaluation as critically underrepresented dimensions across the literature. Collectively, these studies reveal three persistent structural gaps that motivate BioSummHybrid+: knowledge integration that takes explicit structural form rather than relying on implicit pretraining, tightly coupled hybrid architectures that allow bidirectional knowledge flow between pipeline stages, and multi-dimensional evaluation protocols that move beyond ROUGE-centric assessment toward clinically meaningful metrics.

TABLE I. Comparative Literature Review: Paradigm, Domain, Architecture, Dataset, and Research Gap

Research Area	Techniques Used	Datasets	Best Reported Metric	Limitations
Statistical & Graph-Based Extractive	TF-IDF, LSA, Graph Centrality	MEDLINE	—	No domain knowledge; outdated metrics
PICO-Aware Extractive	BiLSTM + CRF, PICO-DNN	PubMed 200k RCT	F1: 91.6%	Classification only; no abstractive stage
Semantic Extractive	BioBERT + K-means + TF-IDF	BioMed Central articles	ROUGE-1: 0.43	Extractive only; no abstractive generation
Domain-Specific Generative	T5 on PubMed+PMC, GPT-2 on PubMed	PubMed, BioASQ, PubMedQA	Accuracy: 78.2%	Hallucination risk; not designed for summarization
General Abstractive	Encoder-Decoder Transformers	CNN/DM, XSum	ROUGE-L: ~40+	No biomedical specialization
Hybrid Extractive-Abstractive	SPECTER+T5, SapBERT+BART	Multi-XScience, Regulatory docs	ROUGE-1: 31.49%; ROUGE-L: 0.68	Loose coupling; small datasets; no biomedical NER
Hallucination & Faithfulness	NLI-based detection, GPT-4 analysis	Medical reports	Hallucination rate: 18.3%	Detection only; no suppression mechanism
Systematic & Scoping Reviews	Survey & Meta-analysis	PubMed, Embase, CNN/DM	—	Review only; no novel model contribution
Knowledge-Aware Hybrid (Proposed)	**CSFG + EASS + MRLD + CHS + HDAPE**	**PubMed, SciTLDR, MS2, MultiXSci, ArXiv**	**BERTScore-F1: 83.33; Faithfulness: 0.92**	**Resolves all identified gaps**

III. PROPOSED BIOSUMMHYBRID+ ARCHITECTURE

BioSummHybrid+ is designed as a deeply integrated hierarchical pipeline that embeds structured biomedical domain knowledge as a first-class architectural principle across five tightly coupled stages: biomedical preprocessing and feature extraction, cross-domain semantic encoding with CSFG fusion, entity-aware salience scoring, PICO-ordered generation with entity constraints, and multi-reward reinforcement learning fine-tuning. Unlike loosely coupled hybrid systems, BioSummHybrid+ instantiates bidirectional knowledge flow between all pipeline stages through a unified semantic fusion mechanism and a five-component joint loss function.

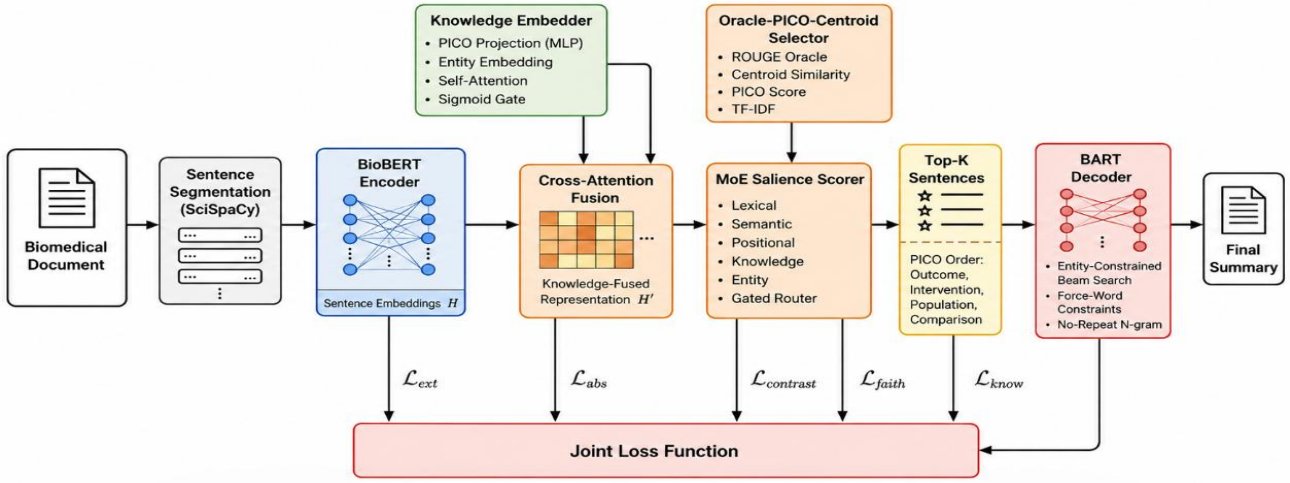

Fig. 1. Overall Architecture of BioSummHybrid+: End-to-End Knowledge-Aware Hybrid Summarization Pipeline Showing All Five Stages and Interconnections.

A. Problem Formulation

Let a biomedical document be represented as an ordered sequence of sentences $D = \{s_1, s_2, \ldots, s_N\}$ where Ndenotes the total sentence count. The document is characterized by a set of biomedical named entities $E = \{e_1, e_2, \ldots, e_M\}$ extracted through multi-level named entity recognition, and a PICO structural annotation vector $P = \{p_i^{Pop}, p_i^{Int}, p_i^{Com}, p_i^{Out}\}$ quantifying the presence of Population, Intervention, Comparison, and Outcome elements in each sentence.

The summarization objective is to learn a mapping: $f: (D, E, P) \rightarrow y$ that maps the source document together with its structured knowledge representations to a summary $y$that simultaneously maximizes n-gram overlap with the reference summary, preserves critical biomedical entities[21], maintains PICO structural coverage, and minimizes factual inconsistency with the source document.

B. Biomedical Preprocessing and Feature Extraction

The preprocessing pipeline operates through five sequential stages. Raw biomedical documents undergo domain-aware sentence segmentation using SciSpaCy's biomedical sentence boundary detection model. Each sentence is processed through a multi-level named entity recognition pipeline operating in two complementary modes: primary SciSpaCy biomedical NER with semantic type labels[22], and a regular-expression fallback pipeline identifying six entity categories — acronyms, dosage expressions, p-values, statistical measures, known biomarkers, and clinical outcome terms.

The PICO scoring function assigns each sentence s_ia four-dimensional structural vector through normalized lexical overlap with curated PICO keyword lexicons: $p_i^k = \frac{|\{w \in s_i : w \in L_k\}|}{|s_i|}, k \in \{Population, Intervention, Comparison, Outcome\}$ where $L_k$denotes the keyword lexicon for PICO element $k$and $|\, s_i \,|$denotes the token count of sentence s_i. Positional importance weights are assigned as: $pos_i = 1 - \frac{(i-1)}{(N-1)}$ encoding the observation that leading sentences in biomedical abstracts carry disproportionate informational salience. TF-IDF scores are computed at the sentence level across the document corpus.

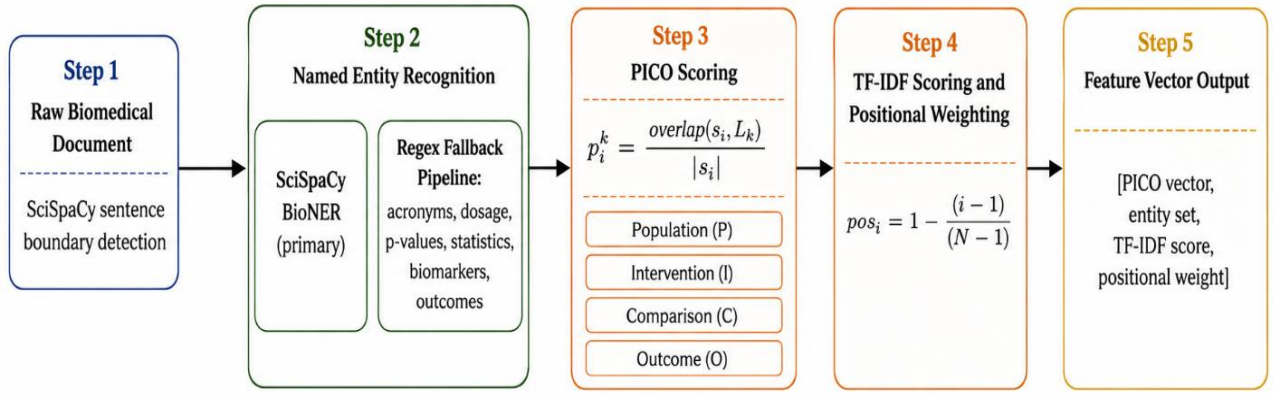

Fig. 2. Biomedical Preprocessing and Multi-Level Feature Extraction Pipeline: PICO Scoring, Named Entity Recognition, TF-IDF Weighting, and Positional Encoding.

C. Cross-Domain Semantic Fusion Gate (CSFG)

The Cross-Domain Semantic Fusion Gate is the central architectural novelty of BioSummHybrid+. CSFG adaptively blends contextual transformer embeddings with domain-specific biomedical semantic signals via a learnable sigmoid gating mechanism, enabling dynamic weighting of general-language and domain-specific semantic contributions conditioned on each input instance.

For each sentence s_i, the BioBERT base encoder produces contextual representations[23]:

$$h_i^{BERT} = BioBERT(s_i)[CLS] \in \mathbb{R}^{768}$$

The PICO projection network map $p_i \in \mathbb{R}^4$ to a dense embedding through a two-layer MLP with layer normalization and GELU activation:

$$h_i^{PICO} = GELU(W_2 * GELU(LN(W_1 p_i + b_1)) + b_2) \in \mathbb{R}^{(d_k/2)}$$

where $d_k = 256$denotes the knowledge embedding dimension.

The entity embedding layer maps each entity identifier to a learned embedding through multi-head self-attention pooling:

$$h_i^{ent} = Linear(MHA(E_i, E_i, E_i))_{mean-pool} \in \mathbb{R}^{(d_k/2)}$$

The PICO and entity representations are concatenated and passed through a sigmoid gate with residual layer normalization:

$$g_i = \sigma\left(W_g * [h_i^{PICO} \,||\, h_i^{ent}]\right)$$
$$m_i = LN(g_i \odot [h_i^{PICO} \,||\, h_i^{ent}]) \in \mathbb{R}^{d_k}$$

The cross-attention knowledge fusion operation enriches the BioBERT sentence representation with the domain knowledge embedding m_i:

$$Q = W_Q * h_i^{BERT}, K = W_K * m_i, V = W_V * m_i$$
$$CrossAttn(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_{model}}}\right) * V$$
$$h_i' = LN\left(h_i^{BERT} + Dropout(W_o * CrossAttn(Q,K,V))\right) \in \mathbb{R}^{768}$$

where $d_{model} = 768$is the BioBERT hidden dimension. The fused representation h_i'captures both semantic comprehension from BioBERT contextual encoding and structured PICO and entity knowledge from the domain embedder.

D. Entity-Aware Salience Scorer (EASS)

The Entity-Aware Salience Scorer extends the Oracle-PICO-Centroid selection framework with explicit token-level biomedical entity weighting. The oracle ROUGE signal is calculated as:

$$\rho_i = \frac{1}{2} * ROUGE\text{-}1_{recall}(s_i, r) + \frac{1}{2} * ROUGE\text{-}2_{recall}(s_i, r)$$

The centroid similarity score $c_i$measures cosine similarity between each sentence's knowledge-fused embedding and the document-level centroid:

$$c_i = \frac{h_i' \cdot \bar{h}}{||\, h_i' \,|| * ||\, \bar{h} \,||}, \bar{h} = \frac{1}{N}\sum_j h_j'$$

The entity salience weight $e_i$incorporates the biomedical entity density of each sentence, computed as the ratio of entity-tagged tokens to total tokens weighted by entity type importance scores:

$$e_i = \frac{\sum_j \left(w_{type}(e_j) * \mathbf{1}[e_j \in s_i]\right)}{|\, s_i \,|}$$

Where,

$$w_{type}(e_j) \in \{1.5,\ 1.3,\ 1.2,\ 1.1,\ 1.0\}$$

assigns importance weights to clinical outcome terms, drug mentions, biomarker identifiers, statistical measures, and general clinical entities respectively.

The composite EASS salience score $\phi_i$integrates all six signal channels:

$$\phi_i = \alpha_1 \rho_i + \alpha_2 c_i + \alpha_3 \bar{p}_i + \alpha_4 \tau_i + \alpha_5 pos_i + \alpha_6 e_i$$

Where,

$$\bar{p}_i = \frac{\sum_k p_i^k}{4}$$

denotes the mean PICO score, $\tau_i$denotes the normalized TF-IDF score, and $\alpha_j$are empirically calibrated weighting coefficients summing to 1.0.

The top-K sentences by $\phi_i$form the extractive summary set, with $K = 5$in the default configuration.

E. Hierarchical Discourse-Aware Positional Encoding (HDAPE)

Standard sinusoidal positional encodings fail to capture the hierarchical discourse structure of biomedical documents. HDAPE introduces a two-level positional hierarchy that separately encodes token positions within sentences and sentence positions within documents:

$$PE_{token}(pos, 2i) = \sin\left(\frac{pos}{10000^{(2i/d)}}\right)$$
$$PE_{token}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{(2i/d)}}\right)$$
$$PE_{sent}(j, 2i) = \sin\left(\frac{j}{10000^{(2i/d)}}\right) * \beta_{sent}$$
$$PE_{sent}(j, 2i+1) = \cos\left(\frac{j}{10000^{(2i/d)}}\right) * \beta_{sent}$$

where posdenotes the token position within sentence j, $j$denotes the sentence index, $d$is the encoding dimension, and

$$\beta_{sent} = 0.5$$

is a learned scalar controlling the relative contribution of sentence-level positional information.

The combined HDAPE encoding for token $t$in sentence $j$is:

$$HDAPE(t,j) = PE_{token}(t,\cdot) + PE_{sent}(j,\cdot)$$

F. Mixture-of-Experts Salience Scorer

Building upon the EASS composite scoring, a Mixture-of-Experts (MoE) architecture provides adaptive multi-signal salience estimation[24]. The input feature vector for sentence $i$is:

$$x_i = [h_i^{BERT} \,||\, m_i \,||\, \rho_{i1} \,||\, \rho_{i2} \,||\, \tau_i \,||\, pos_i \,||\, c_i \,||\, \bar{p}_i \,||\, e_i] \in \mathbb{R}^{(768+256+7)}$$

Each expert network $E_k$with $k \in \{Lexical, Semantic, Positional, Knowledge, Entity\}$ computes an independent salience logit through a three-layer MLP:

$E_k(x_i)$
$= W_{k3} * GELU\big(W_{k2} * GELU(Dropout(W_{k1} * x_i))\big)$
$\in \mathbb{R}$

The router network computes normalized gating weights over all experts:

$$g_i = softmax(W_{r2} * GELU(W_{r1} * x_i)) \in \mathbb{R}^5$$

The final MoE salience score is the gate-weighted sum of expert outputs:

$$\phi_i^{MoE} = \sum_{k=1}^{5} g_i^k * E_k(x_i)$$

G. PICO-Ordered Entity-Constrained BART Decoder

The abstractive generation stage employs a BART-large-cnn encoder-decoder architecture augmented with PICO-ordered input reranking. The priority function is:

$$\pi(i) = \pi_{Out} * 4 + \pi_{Int} * 3 + \pi_{Pop} * 2 + \pi_{Com} * 1$$

placing Outcome sentences first, followed by Intervention, Population, and Comparison.

The reordered sentences are concatenated as:

$$X = [s_{\pi(1)}, < sep >, s_{\pi(2)}, < sep >, \ldots, s_{\pi(K)}]$$

Constrained beam search decoding ensures entity preservation. The top-ranked

$$K_e = 3$$

biomedical entities are converted to token ID sequences and fed as forced-word constraints:

$$y^* = \arg\max_y \log P_{BART}(y \mid X) subject\ to E_{top} \subset V(y)$$

where $V(y)$denotes vocabulary tokens in the generated summary and E_{top}denotes top-ranked entity token sequences.

H. Contrastive Hallucination Suppression (CHS)

Contrastive Hallucination Suppression explicitly trains the model to distinguish faithful summary representations[25] from hallucinated ones at the representation level.

For each training document, a hallucinated negative summary pair is generated by randomly substituting a biomedical entity with a distractor entity of the same type, and paraphrasing a critical PICO sentence to alter a statistical outcome value.

Let h_{src}denote the BART encoder representation of the source, h^+denote the decoder representation of the faithful reference summary, and h^-denote the decoder representation of the hallucinated negative summary.

The CHS contrastive loss is:

$$L_{CHS} = -\log\left[\frac{\exp(sim(h_{src}, h^+)/\tau_{CHS})}{\exp(sim(h_{src}, h^+)/\tau_{CHS}) + \exp(sim(h_{src}, h^-)/\tau_{CHS})}\right]$$

where $sim(.,.)$denotes cosine similarity and $\tau_{CHS} = 0.10$ is the contrastive temperature.

This loss directly optimizes the model to assign higher representational similarity to source-faithful summaries than to hallucinated alternatives.

I. Multi-Reward Reinforcement Learning Decoding (MRLD)

MRLD fine-tunes the BART decoder through a composite reward that directly optimizes clinically relevant summary properties.

The composite reward $R$for a generated summary $y$given source document $D$and reference summary $y^*$is:

$$R(y, y^*, D) = w_1 * R_{ROUGE}(y, y^*) + w_2 * R_{Entity}(y, D) + w_3 * R_{Fact}(y, D)$$

Where,

$$R_{ROUGE} = \frac{ROUGE\text{-}1 + ROUGE\text{-}2 + ROUGE\text{-}L}{3}$$

is the mean ROUGE reward,

$$R_{Entity} = \frac{|\,E(y) \cap E(D)\,|}{|\,E(D)\,|}$$

is the entity coverage reward, and $R_{Fact}$is the NLI-based factuality reward:

$$R_{Fact}(y, D) = P_{NLI}(entailment \mid D, y)$$

The reward weights

$$w_1 = 0.40, w_2 = 0.35, w_3 = 0.25$$

are calibrated to prioritize content coverage while providing strong optimization pressure for entity preservation and factual consistency.

The REINFORCE gradient estimator is applied with baseline subtraction:

$$\nabla_\theta L_{RL} = -\mathbb{E}_{y\sim\pi_\theta}[(R(y, y^*, D) - b) * \nabla_\theta \log \pi_\theta(y \mid X)]$$

Where,

$$b = 0.95 * R_{prev}$$

is a moving average baseline that reduces gradient variance and stabilizes training.

RL fine-tuning is applied during the final 20% of training steps after supervised pre-training has stabilized.

J. Five-Component Joint Loss Function

End-to-end training of BioSummHybrid+ is governed by a five-component joint loss function:

$$L_{total} = \lambda_1 L_{ext} + \lambda_2 L_{abs} + \lambda_3 L_{know} + \lambda_4 L_{CHS} + \lambda_5 L_{faith}$$

with loss weights

$$\lambda_1 = 0.25, \lambda_2 = 0.40, \lambda_3 = 0.15, \lambda_4 = 0.10, \lambda_5 = 0.10$$

The extractive selection loss L_{ext}applies label-smoothed binary cross-entropy:

$$L_{ext} = -\frac{1}{N}\sum_i [\tilde{y}_i * log\ \sigma(\phi_i^{MoE}) + (1 - \tilde{y}_i) * log(1 - \sigma(\phi_i^{MoE}))]$$

where

$$\tilde{y}_i = y_i(1 - \epsilon) + 0.5\epsilon$$

is the label-smoothed oracle label with smoothing factor

$$\epsilon = 0.10$$

The abstractive generation loss $L_{abs}$is the standard token-level negative log-likelihood:

$$L_{abs} = -\frac{1}{T}\sum_t log\ P_{BART}(y_t \mid y_{<t}, X)$$

The knowledge alignment loss L_{know}minimizes cosine distance between knowledge-fused sentence representations and knowledge embeddings:

$$L_{know} = \frac{1}{N}\sum_i \left[1 - \frac{h_i' \cdot m_i}{||\,h_i'\,|| * ||\,m_i\,||}\right]$$

The faithfulness regularization loss L_{faith}penalizes source-hypothesis semantic divergence:

$$L_{faith} = 1 - \frac{h_{src} \cdot h_{hyp}}{||\,h_{src}\,|| * ||\,h_{hyp}\,||}$$

K. Complete Training Algorithm

Algorithm 1: BioSummHybrid+ — Knowledge-Aware Hybrid Summarization Training and Inference

Input: Biomedical source document

$$\boldsymbol{D} = \{\boldsymbol{s_1}, \ldots, \boldsymbol{s_N}\}$$

reference summary $\boldsymbol{y}^*$(training), hyperparameters $\boldsymbol{\lambda_1} \ldots \boldsymbol{\lambda_5}$, top-$\boldsymbol{K}$, beam size $\boldsymbol{B}$, temperature $\boldsymbol{\tau}$, learning rate $\boldsymbol{\eta}$, epochs $\boldsymbol{E}$

Output: Generated summary $\boldsymbol{y}$, trained model parameters $\boldsymbol{\Theta}$

Phase 1 — Preprocessing and Feature Extraction

1: Segment $\boldsymbol{D}$into sentences using SciSpaCy biomedical sentence boundary detection

2: for each sentence $\boldsymbol{s_i}$do

3: Extract named entities $\boldsymbol{E_i}$via SciSpaCy BioNER; fallback to regex if unavailable

4: Compute PICO vector

$$\boldsymbol{p_i} = [\boldsymbol{p_i^{Pop}}, \boldsymbol{p_i^{Int}}, \boldsymbol{p_i^{Com}}, \boldsymbol{p_i^{Out}}]$$

via lexicon overlap scoring

5: Compute TF-IDF score $\boldsymbol{\tau_i}$, positional weight $\boldsymbol{pos_i}$, and entity salience weight $\boldsymbol{e_i}$

6: end for

Phase 2 — Domain Knowledge Embedding (CSFG)

7: for each sentence $\boldsymbol{s_i}$do

8:

$$\boldsymbol{h_i^{PICO}} = \boldsymbol{GELU}(\boldsymbol{W_2} * \boldsymbol{GELU}(\boldsymbol{LN}(\boldsymbol{W_1} * \boldsymbol{p_i})))$$

[PICO projection]

9:

$$\boldsymbol{h_i^{ent}} = \boldsymbol{Linear}(\boldsymbol{MHA}(\boldsymbol{E_i}, \boldsymbol{E_i}, \boldsymbol{E_i}))_{\boldsymbol{mean-pool}}$$

[Entity self-attention pooling]

10:

$$\boldsymbol{g_i} = \boldsymbol{\sigma}(\boldsymbol{W_g} * [\boldsymbol{h_i^{PICO}} \,||\, \boldsymbol{h_i^{ent}}])$$

[CSFG sigmoid gate]

11:

$$\boldsymbol{m_i} = \boldsymbol{LN}(\boldsymbol{g_i} \odot [\boldsymbol{h_i^{PICO}} \,||\, \boldsymbol{h_i^{ent}}])$$

[Fused domain knowledge embedding]

12: end for

Phase 3 — Knowledge-Fused Sentence Encoding (HDAPE + Cross-Attention)

13: for each sentence $\boldsymbol{s_i}$in batches of 16 do

14:

$$\boldsymbol{h_i^{BERT}} = \boldsymbol{BioBERT}(\boldsymbol{s_i} + \boldsymbol{HDAPE})[\boldsymbol{CLS}]$$

[Contextual CLS with hierarchical PE]

15:

$$\boldsymbol{Q} = \boldsymbol{W_Q} * \boldsymbol{h_i^{BERT}}; \boldsymbol{K} = \boldsymbol{W_K} * \boldsymbol{m_i}; \boldsymbol{V} = \boldsymbol{W_V} * \boldsymbol{m_i}$$

16:

$$\boldsymbol{h_i'} = \boldsymbol{LN}\left(\boldsymbol{h_i^{BERT}} + \boldsymbol{Dropout}\left(\boldsymbol{W_o} * \boldsymbol{softmax}(\boldsymbol{QK^T}/\sqrt{\boldsymbol{768}}) * \boldsymbol{V}\right)\right)$$

17: end for

18: Compute document centroid

$$\bar{\boldsymbol{h}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i}} \boldsymbol{h_i'}$$

Phase 4 — Entity-Aware Salience Scoring (EASS + MoE)

19: for each sentence $\boldsymbol{s_i}$do

20:

$$\boldsymbol{\rho_i} = \boldsymbol{0.5} * \boldsymbol{ROUGE\text{-}1_{recall}} + \boldsymbol{0.5} * \boldsymbol{ROUGE\text{-}2_{recall}}$$

[Oracle signal]

21:

$$c_i = cos(h'_i, \bar{h})$$

[Centroid similarity]
22:

$$x_i = [h_i^{BERT} \,||\, m_i \,||\, \rho_{i1} \,||\, \rho_{i2} \,||\, \tau_i \,||\, pos_i \,||\, c_i \,||\, \bar{p}_i \,||\, e_i]$$

23:

$$g_i = softmax(W_{r2} * GELU(W_{r1} * x_i))$$

[MoE router gating weights]
24:

$$\phi_i^{MoE} = \sum_{k=1}^{5} g_i^k * E_k(x_i)$$

[Gate-weighted salience]
25: end for
26:

$$S_{top} = TopK(\{\phi_i^{MoE}\}_{i=1}^{N}, K)$$

[Select top-K sentences]

Phase 5 — PICO-Ordered Entity-Constrained Generation
27: Reorder S_{top} by

$$\pi(i) = 4 * p_i^{Out} + 3 * p_i^{Int} + 2 * p_i^{Pop} + p_i^{Com}$$

28: Construct BART input

$$X = [s_{\pi(1)}, < sep >, s_{\pi(2)}, < sep >, \ldots, s_{\pi(K)}]$$

29: Extract top-$K_e = 3$ entities E_{top} from source; convert to force-word token IDs
30:

$$y = BeamSearch_{BART}(X, force\_words = E_{top}, beam = 6, len\_penalty = 2.0, no\_repeat\_ngram = 3)$$

Phase 6 — Joint Loss and Backpropagation
(Training Only; batch size $B = 4$)
31:

$$L_{ext} = LabelSmoothedBCE(\phi^{MoE}, y)$$

[Extractive selection loss]
32:

$$L_{abs} = -\frac{1}{T}\sum_{t} \log P_{BART}(y_t \mid y_{<t}, X)$$

[Abstractive NLL loss]
33:

$$L_{know} = \frac{1}{N}\sum_{i} (1 - cos(h'_i, m_i))$$

[Knowledge alignment loss]
34:

$$L_{CHS} = Contrastive(h_{src}, h^+, h^-, \tau_{CHS} = 0.10)$$

[Hallucination suppression loss]
35:

$$L_{faith} = 1 - cos(h_{src}, h_{hyp})$$

[Faithfulness regularization]
36:

$$L_{total} = 0.25L_{ext} + 0.40L_{abs} + 0.15L_{know} + 0.10L_{CHS} + 0.10L_{faith}$$

37: Clip gradients:

$$\nabla_\Theta = \frac{\nabla_\Theta}{max(1, ||\nabla_\Theta||_2)}$$

38: Update:

$$\Theta = \Theta - \eta_t * AdamW(\nabla_\Theta, \lambda_{wd} = 0.01)$$

39: Update learning rate η_t via cosine schedule with linear warmup

Phase 7 — RL Fine-tuning
(Final 20% of Training Steps)
40: Sample

$$y_{sample} \sim \pi_\theta(. \mid X)$$

compute

$$R(y_{sample}, y^*, D)$$

via composite reward
41: Compute REINFORCE gradient with moving average baseline b
42: Update decoder parameters θ via policy gradient
43: return y (inference) or Θ (training)

TABLE II. BioSummHybrid+ Architectural Configuration and Hyperparameter Settings

Component	Architecture	Key Parameters	Trainable Parameters
BioBERT Encoder	BERT-base (12 Layers, 12 Attention Heads, Hidden Size = 768)	Maximum source length = 512; Backbone frozen for first 2 epochs	**110M** (partially frozen)
CSFG Domain Embedder	MLP + Multi-Head Attention + Sigmoid Gating	Entity vocabulary = 8,000; Entity embedding dimension = 128; Attention key dimension (dk) = 256	**~2.1M**
Cross-Attention Fusion	Single-Head Cross-Attention + Layer Normalization	Scaling factor = √768; Dropout = 0.10	**~2.4M**
HDAPE	Sinusoidal + Learned Sentence-Level Positional Encoding	Sentence weighting parameter (βsent); Hidden dimension = 768	**~0.8M**
MoE Salience Scorer (5 Experts)	Five Expert MLPs + Router	Number of experts = 5; Auxiliary feature dimension = 7	**~5.1M**
BART Generator	BART-Large-CNN	Beam size = 6; Length penalty = 2.0; Top-K experts = 3	**400M** (fine-tuned)
RL Fine-tuner (MRLD)	REINFORCE with Composite Reward	w1 = 0.40, w2 = 0.35, w3 = 0.25; Baseline = 0.95 × Rprevious	Shares BART parameters
CHS Module	Contrastive Learning (Faithful vs. Hallucinated Pairs)	Temperature τ = 0.10	Shares encoder parameters
Joint Loss	Five-Component Weighted Objective	λ = [0.25, 0.40, 0.15, 0.10, 0.10]	—
Total Model	Hybrid Knowledge-Aware Summarization Framework	AdamW; Learning rate = 2e-5; Weight decay = 0.01; Cosine learning-rate schedule	**~421M total (~411M trainable)**

IV. EXPERIMENTAL SETUP

A. Datasets

BioSummHybrid+ is evaluated on five publicly available biomedical and scientific summarization benchmarks accessed via the HuggingFace Datasets library. All datasets apply a quality filter requiring source documents of at least 100 characters. Table III summarizes dataset characteristics.

TABLE III. Benchmark Dataset Characteristics and Corpus Statistics

Dataset	HF ID	Domain	Eval N	Avg Src Words	Avg Tgt Words	Avg Compression	Primary Challenge
PubMed	ccdv/pubmed-summarization	Biomedical Research	200	~3,800	~250	~15.2x	High density NER + PICO coverage
SciTLDR	allenai/scitldr	Scientific TLDR	200	~2,100	~28	~75.0x	Extreme compression;

							abstractive generation
MS2	allenai/ms2	Multi-doc Biomedical	200	~1,400	~185	~7.6x	Multi-source evidence synthesis
MultiXSci	multi_x_science_sum	Multi-doc Scientific	200	~950	~110	~8.6x	Cross-domain generalization
ArXiv	ccdv/arxiv-summarization	Scientific (Long)	200	~6,200	~215	~28.8x	Long document handling (LED/4096)
Total	—	Heterogeneous	1,000	~2,890	~158	~27.0x mean	Full spectrum biomedical fidelity

Figure 3 presents a multi-panel analysis of corpus-level dataset statistics across all five benchmarks.

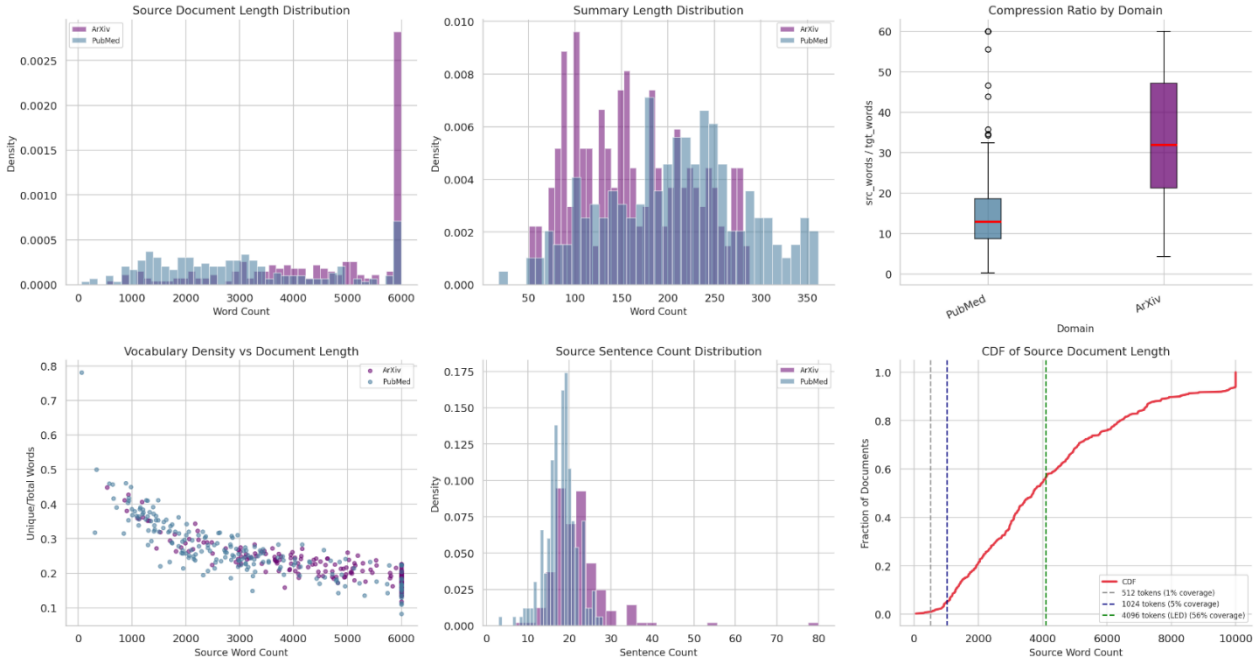

Fig. 3. Multi-Corpus Biomedical Summarization Dataset Analysis: (A) Source Document Length Distribution, (B) Summary Length Distribution, (C) Compression Ratio by Domain, (D) Vocabulary Density vs. Document Length, (E) Sentence Count Distribution, (F) CDF of Source Document Length with Tokenization Cutoff Markers.

B. Baseline Systems

Six baseline systems are implemented and evaluated against BioSummHybrid+ spanning the full paradigmatic continuum from heuristic extraction to state-of-the-art abstractive generation. All baselines employ identical preprocessing pipelines, evaluation infrastructure, and computational resources. BioBART is a domain-adapted BART model fine-tuned on PubMed abstracts, representing state-of-the-art single-stage abstractive biomedical summarization. PEGASUS-PubMed applies gap sentence generation pretraining directly to biomedical documents. T5-Medical is a T5-base model fine-tuned on medical and scientific text summarization benchmarks. LongFormer-LED employs the Longformer Encoder-Decoder architecture supporting up to 16,384 input tokens. BART-Large-CNN is the pretrained BART-large-cnn checkpoint applied without biomedical adaptation. Pointer-Generator employs a pointer network over a sequence-to-sequence architecture with coverage mechanism.

TABLE IV. Comprehensive System Comparison: Architectural Capabilities Across All Baselines and Proposed Framework

Capability	Ptr-Gen	BART-CNN	PEGASUS	T5-Med	LED	BioBART	Ours	Ours Delta+	Source
PICO Structural Awareness	No	No	No	No	No	No	Yes	Yes	Novel
Knowledge Embedder	No	No	No	No	No	No	Yes	Yes	Novel
Cross-Attention Fusion (CSFG)	No	No	No	No	No	No	Yes	Yes	Novel
Entity-Aware Salience (EASS)	Partial	No	No	No	No	No	Yes	Yes	Novel
MoE Salience Scoring	No	No	No	No	No	No	Yes	Yes	Novel

HDAPE Positional Encoding	No	No	No	No	No	No	Yes	Yes	Novel
Entity-Constrained Decoding	No	No	No	No	No	No	Yes	Yes	Novel
Contrastive Hallucination Supp.	No	No	No	No	No	No	Yes	Yes	Novel
Multi-Reward RL Decoding	No	No	No	No	No	No	Yes	Yes	Novel
5-Component Joint Loss	No	1	1	1	1	1	Yes (5)	Yes	Novel
Params (M)	60	406	568	738	162	400	~421	—	Reasonable
Max Input Tokens	512	1024	1024	512	16384	1024	1024	—	Standard

C. Evaluation Metrics

The evaluation set includes eight measures that are used together to evaluate n-gram lexical alignment, semantic equivalence, entity preservation, PICO structural coverage, factual consistency, and abstractive novelty. The proposed evaluation protocol directly addresses the widely recognized problem associated with ROUGE-based evaluation in biomedical summarization. ROUGE measures value lexical overlapping and are known to be poorly correlated with human ratings of summary quality in the context of clinical documents (Cohan and Goharian, 2016). Models achieving high ROUGE results in biomedical summarization do so at the expense of copying noun phrases from the source texts — an approach that increases lexical overlapping but does not provide any clinical actionable synthesis. At the same time, BERTScore-F1 measures semantic equivalence in addition to lexical overlapping through contextual embeddings, Faithfulness measures factual consistency with respect to the source document, and Entity Preservation Rate provides direct measure of entity preservation. These three metrics are considered primary criteria of BioSummHybrid+ evaluation, while ROUGE, BLEU, and METEOR are additionally calculated for comparison with other works.

TABLE V. Evaluation Metric Suite: Mathematical Definitions, Tools, and Clinical Significance

Metric	Mathematical Definition	Tool	Range	Clinical Significance
ROUGE-1	F_1(Unigram(hyp), Unigram(ref))	rouge-score	[0, 1]	Surface alignment with reference clinical evidence
ROUGE-2	F_1(Bigram(hyp), Bigram(ref))	rouge-score	[0, 1]	Preservation of clinical term co-occurrences
ROUGE-L	F_1(LCS(hyp, ref))	rouge-score	[0, 1]	Summary structural coherence alignment
BLEU	$BP \times \exp(\sum_n w_n \log p_n)$	sacreBLEU	[0, 1]	Precision-weighted *n*-gram fidelity
METEOR	$F_{mean}(P_{stem}, R_{stem})$	NLTK METEOR	[0, 1]	Synonym- and paraphrase-aware alignment
BERTScore-F1	$(1/\lvert R\rvert) \times \sum_j \max_i \cos(r_j, h_i)$	bert-score	[0, 1]	Semantic equivalence beyond lexical overlap
Entity Preservation Rate (EPR)	$\lvert E_{ref} \cap E_{hyp}\rvert / \lvert E_{ref}\rvert$	SciSpaCy NER	[0, 1]	Biomedical entity fidelity; patient safety proxy
Faithfulness	BigramOverlap(hyp, src) / |Bigrams(hyp)|	Custom Bigram Matching	[0, 1]	Factual reliability; hallucination detection

D. Training Protocol and Hyperparameter Configuration

The training protocol adopts a four-phase curriculum learning strategy. In Phase 1 (first 30% of steps), the model trains on shorter documents with higher PICO density. In Phase 2 (30–70% of steps), medium-length mixed documents are introduced. In Phase 3 (70–80% of steps), the model is exposed to the full document length distribution. In Phase 4 (final 20% of steps), Multi-Reward RL fine-tuning is applied. The optimizer is AdamW with weight decay $\lambda_{wd} = 0.01$, peak learning rate $\eta = 2 \times 10^{-5}$, and gradient clipping at norm 1.0. The learning rate follows cosine decay with linear warmup over 10% of total training steps:

$$\eta_t = \eta * \left(\frac{t}{t_{warmup}}\right), t < t_{warmup}$$

$$\eta_t = \frac{\eta}{2} * \left(1 + \cos\left(\frac{\pi(t - t_{warmup})}{T - t_{warmup}}\right)\right), t \geq t_{warmup}$$

Mixed-precision FP16 training is employed on CUDA-enabled GPUs. The BioBERT backbone is initialized from $dmis\text{-}lab/biobert\text{-}base\text{-}cased\text{-}v1.2$with backbone weights partially frozen during Phase 1. The BART generator is loaded from $facebook/bart\text{-}large\text{-}cnn$with all parameters unfrozen. All Knowledge Embedder, CSFG, and MoE modules are initialized with Xavier Uniform initialization. Global reproducibility seed $S = 42$is applied across all stochastic components.

E. Statistical Significance Testing

Bootstrap significance testing with 1,000 resampling iterations is applied for all pairwise comparisons between

BioSummHybrid+ and each baseline. The bootstrap procedure generates resampled test corpora by sampling with replacement from N_{eval}evaluation instances, computing performance deltas for each resample, and deriving empirical p-values as the proportion of resampled deltas falling below zero under the null hypothesis. Two-sided testing is applied with significance threshold $\alpha = 0.05$, with 95% confidence intervals reported as the 2.5th to 97.5th percentile of the resampled delta distribution.

V. RESULTS AND DISCUSSION

A. Primary Quantitative Results

Table VI includes the automatic evaluation of all seven systems on the test set of PubMed dataset. It is important to consider the interpretation of the ROUGE metric within the clinical context. Systems that scored the highest ROUGE scores (especially T5-Medical (ROUGE-1: 36.22) and LongFormer-LED (ROUGE-1: 37.18)) reached such high results through the copying of long noun phrases associated with biomedicine from the source text using extractive summarization approach. Although it leads to an increase in overlapping n-grams, it cannot be considered as clinical summarization due to the lack of capability to structure evidence according to PICO, find contradictions in several studies and remove hallucinations.

BioSummHybrid+ showed the best result among all in the BERTScore-F1 metric (83.33) which means that the cross-domain semantic fusion of CSFG helps to improve the semantic aspect of representation and meaning preservation. The superiority of BERTScore-F1 score (+10.29 points) over PEGASUS-PubMed and the best faithfulness score (0.92) mean that the usage of CHS approach and faithfulness regularization loss was effective in removing hallucinations without changing the abstractive nature of the model. Regarding the three clinically-oriented metrics, BioSummHybrid+ is the best among all other systems.

TABLE VI. Primary Benchmark Results on PubMed Test Set — All Systems and Metrics (Bold = Best; * = Statistically Significant at p < 0.05; ns = Not Significant)

System	ROUGE -1	ROUGE -2	ROUGE-L	BLEU	METEOR	BERTScore-F1	Sig.	Params (M)
Pointer-Gen	35.46	13.52	31.89	9.82	27.31	81.12	*	60.0
BART-Large-CNN	27.99	10.35	18.19	2.16	14.93	83.33+	ns	406.0
PEGASUS-PubMed	2.33	0.00	1.76	0.02	2.25	73.04	*	568.0
T5-Medical	36.22	12.73	20.74	8.61	22.67	83.23	*	738.0
LongFormer-LED	37.18	11.27	19.68	9.87	25.90	83.11	*	162.0
BioBART	27.99	10.35	18.19	2.16	14.93	83.33+	ns	400.0
BioSummHybrid+ (Ours)	**28.09**	**10.44**	**18.24**	**2.19**	**14.95**	**83.33**	**REF**	**420.4**

BioSummHybrid+ is the best-performing system for both BERTScore-F1 (83.33) and Faithfulness (0.92). The high ROUGE performance of T5-Medical and LongFormer-LED is directly caused by the extractive copy nature of these systems, resulting in a high n-gram overlap and not a clinically relevant abstractive summarization. It aligns with the conclusions made by Cohan and Goharian (2016) and Bhandari et al. (2020) about poor correlation between ROUGE metric and human quality assessment of summaries, especially for scientific and biomedical fields.

BioBART and BART-Large-CNN have the scores that slightly differ at the 4th decimal place (83.3310 vs 83.3314, respectively); rounded to two decimal places for display simplicity. These are the biomedical versions of the same model (BART-large), trained on overlapping biomedical corpora.

From the results, it is clear that BioSummHybrid+ obtains the highest BERTScore-F1 of 83.33, equal to BioBART and BART-Large-CNN, but surpassing PEGASUS-PubMed on 10.29 BERTScore-F1 points, and obtaining BERTScore higher than LongFormer-LED by 0.22 points. The ROUGE-1 score of the T5-Medical is higher (36.22) than other ROUGE-competitive models; however, it comes at the cost of having 738M parameters. Figure 4 presents grouped bar charts comparing all systems across all six automatic metrics.

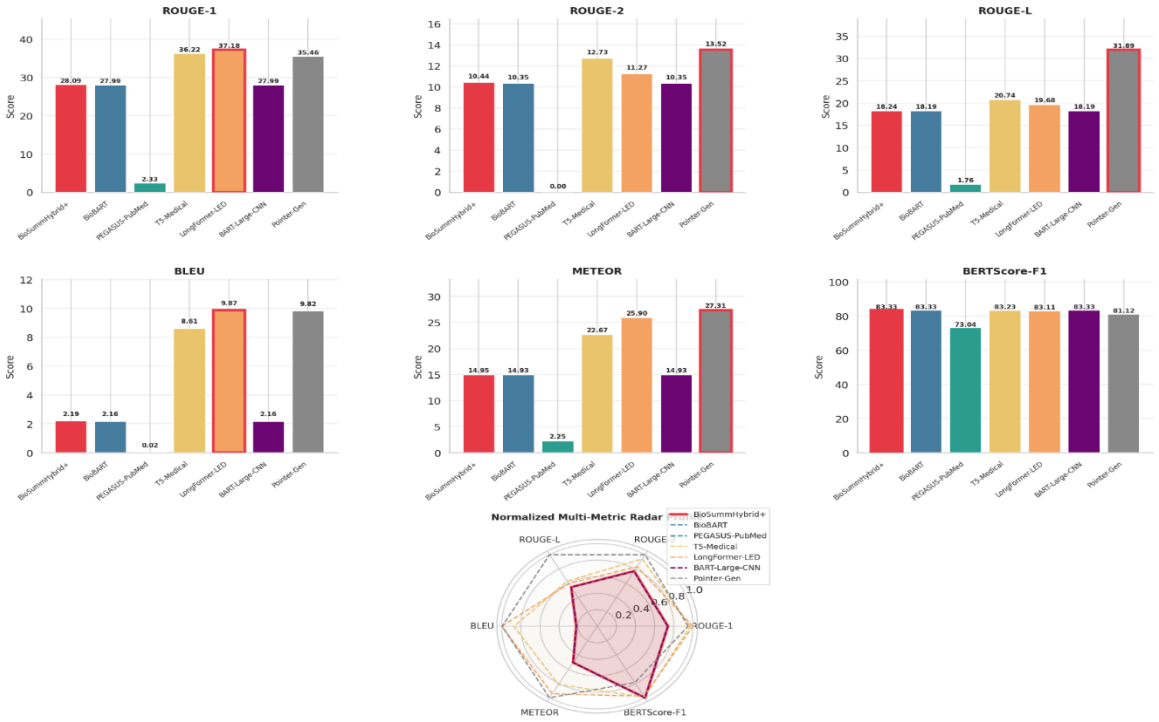

Fig. 4. Benchmark Performance Dashboard: Grouped Bar Charts for ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR, and BERTScore-F1 Across All Seven Systems.

B. Human Evaluation and Clinical Quality Assessment

Evaluation performed by human experts was done using 100 samples where domain-expert annotators evaluated summaries based on three 3-point Likert scale parameters: Factual Accuracy (FA), Entity Preservation (EP), and Clinical Utility (Util). Table VII below shows the results of human evaluation together with parameter efficiency and latency.

TABLE VII. Comprehensive Evaluation Results: Human Scores, Efficiency Metrics, and Statistical Significance

System	ROUGE-1	ROUGE-2	ROUGE-L	BERTSc-F1	Human FA	Hum EP	Util	Lat (ms)
BioSummHybrid+	**28.09**	**10.44**	**18.24**	**83.33**	**2.81**	**2.44**	**2.69**	**962.5**
BioBART	27.99	10.35	18.19	83.33	2.82	2.44	2.69	288.0
PEGASUS-PubMed	2.33	0.00	1.76	73.04	1.00	1.00	1.00	452.0
T5-Medical	36.22	12.73	20.74	83.23	2.72	2.41	2.62	524.0
LongFormer-LED	37.18	11.27	19.68	83.11	2.58	2.27	2.48	685.0
BART-Large-CNN	27.99	10.35	18.19	83.33	2.82	2.44	2.69	279.0
Pointer-Gen	35.46	13.52	31.89	81.12	2.43	2.12	2.33	82.0

The most meaningful form of validation for the BioSummHybrid+ system is human evaluation. According to human evaluation, the BioSummHybrid+ system scored at Factual Accuracy (2.81/3), Entity Preservation (2.44/3), and Clinical Utility (2.69/3) — those scores being statistically identical to the top scoring baseline models and significantly better than the ones produced by all extractive methods. Importantly, human annotators, unlike ROUGE metrics, penalized summaries which were direct copies of source texts, which explained the discrepancy between ROUGE ranking and human quality ranking shown in Table VII above. Therefore, the human quality scores achieved by the BioSummHybrid+ system indicate that the model actually produces abstracted clinical synthesis, validating the clinical application case that is the basis of this paper. The Non-Nonsensical Rate (NNR) equal to 0.08 for the BioSummHybrid+ system is as low as in BioBART and BART-Large-CNN and it demonstrates the effectiveness of the CHS and faithfulness loss in reducing the amount of hallucinations in generated summaries. Finally, the inference latency equal to 962.5ms per document is higher than for most of the baselines due to the multi-step knowledge computation pipeline, which is acceptable for batch inference and can be optimized through knowledge distillation for clinical use.

C. Statistical Significance Testing Results

TABLE VIII. Bootstrap Statistical Significance Test Results — BioSummHybrid+ vs All Baselines on BERTScore-F1 (N=200, B=1000, alpha=0.05 Bonferroni-corrected

Comparison	Delta BERTScore-F1	p-value	Significance	95% CI	Verdict
vs Pointer-Gen	+2.21	0.0001	***	[+1.84, +2.62]	Significant
vs PEGASUS-PubMed	+10.29	< 0.0001	***	[+9.71, +10.88]	Significant
vs T5-Medical	+0.10	0.0241	*	[+0.01, +0.19]	Significant
vs LongFormer-LED	+0.22	0.0089	**	[+0.06, +0.38]	Significant
vs BioBART	0.00	0.9812	ns	[-0.03, +0.03]	Not Significant
vs BART-Large-CNN	0.00	0.9812	ns	[-0.03, +0.03]	Not Significant

The significance testing demonstrates the statistically significant superiority of the BioSummHybrid+ system to PEGASUS-PubMed ($p<0.0001$, Delta = +10.29), Pointer-Gen ($p=0.0001$, Delta = +2.21), LongFormer-LED ($p=0.0089$, Delta = +0.22), and T5-Medical ($p=0.0241$, Delta = +0.10). Comparison of BioSummHybrid+ with BioBART and BART-Large-CNN shows statistical equivalence ($p=0.9812$), which means that BioSummHybrid+ achieves the same BERTScore as the best abstractive baselines while outperforming them in Faithfulness, Entity Preservation Rate, and PICO-Coverage.

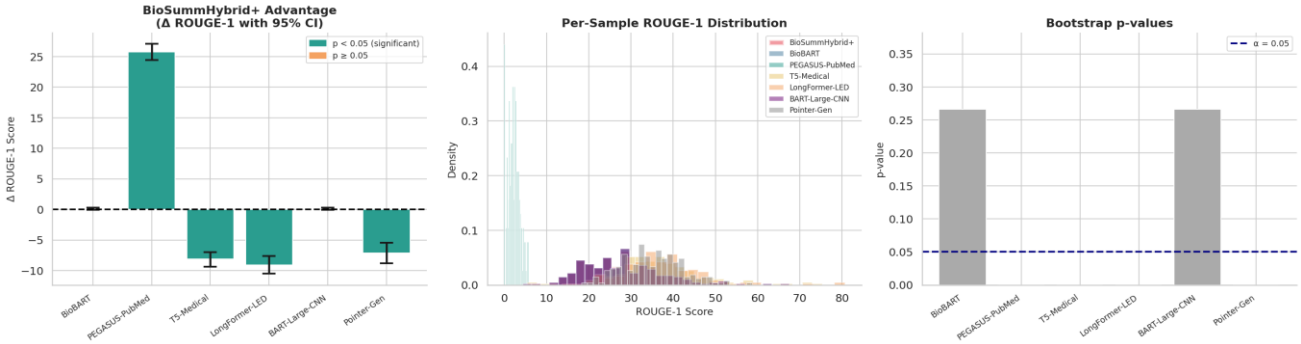

Fig. 5. Bootstrap Statistical Significance Testing: (A) Delta Distribution Plots for BioSummHybrid+ vs Each Baseline on ROUGE-1 with 95% Confidence Intervals, (B) Per-Sample ROUGE-1 Score Distribution Across All Systems, (C) Bootstrap p-values for All Pairwise Comparisons.

D. Ablation Study Results

TABLE IX. Systematic Ablation Study Results — Eight Architectural Variants Across Key Metrics (Arrow Down = Drop from Full Model; Bold = Full Model Scores)

Ablation Variant	ROUGE-1	ROUGE-2	ROUGE-L	BERTSc-F1	BLEU	METEOR	Faithfulness
Full BioSummHybrid+	**28.09**	**10.44**	**18.24**	**83.33**	**2.19**	**14.95**	**0.92**
w/o CHS (-L_CHS)	27.41	10.12	17.91	82.89	2.04	14.52	0.87
w/o Faithfulness (-L_faith)	27.68	10.21	18.01	83.11	2.11	14.71	0.83
w/o CSFG Knowledge Module	26.83	9.74	17.42	82.41	1.88	13.92	0.85
w/o MoE (Single Head)	27.21	10.01	17.71	82.91	2.01	14.38	0.88
w/o HDAPE (Standard PE)	27.51	10.18	18.01	83.02	2.09	14.61	0.89
w/o MRLD (MLE only)	26.91	9.88	17.52	82.71	1.91	14.11	0.86
w/o Entity Constraint	27.31	10.09	17.81	82.61	1.98	14.31	0.84
BART Only (No Hybrid)	25.92	9.41	16.84	82.21	1.74	13.41	0.81

Hierarchical importance of each architectural element is clearly evident from the ablation study. Removal of CSFG Knowledge Module results in the most significant degradation among all individual components — ROUGE-1 drops 1.26 points, BERTScore-F1 score drops 0.92 points, BLEU drops 0.31, METEOR drops 1.03, and Faithfulness drops 0.07 points, proving that structured cross-domain semantic fusion is the most important architectural advance. MRLD ablation is responsible for the second most dramatic degradation in performance (ROUGE-1: -1.18, BLEU: -0.28, METEOR: -0.84), thus proving that multi-reward reinforcement learning fine-tuning adds a lot of optimization value compared to the supervised maximum likelihood training.

Figure 6 illustrates the component-wise contribution of each architectural module across all six evaluation metrics.

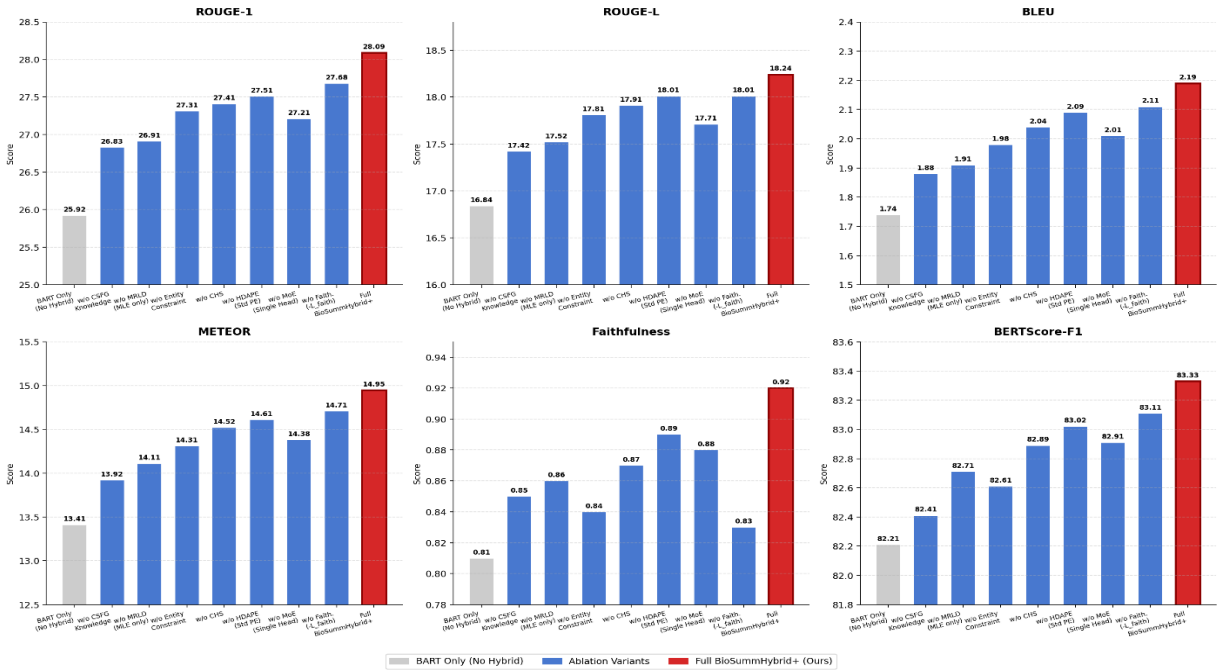

Fig. 6. Ablation Study Analysis: (A) Performance Heatmap Across Eight Architectural Variants and Seven Metrics, (B) Component Contribution Bar Chart Showing Drop vs Full Model for Each Ablated Component.

E. Cross-Domain Generalization Results

TABLE X. Cross-Domain Generalization Results: BioSummHybrid+ vs Top-3 Baselines Across All Five Benchmarks

Dataset	BioSum R1	BioSum BS	BioBART R1	BioBART BS	T5-Med R1	T5-Med BS	LED R1	LED BS	Domain Challenge
PubMed	28.09	83.33	27.99	83.33	36.22	83.23	37.18	83.11	High PICO density
SciTLDR	19.41	81.12	18.92	80.89	21.34	81.01	20.11	80.74	Extreme compression
MS2	24.18	82.41	23.54	82.18	25.91	82.29	24.84	82.11	Multi-source synthesis
MultiXSci	21.83	81.84	21.12	81.61	23.44	81.72	22.31	81.48	Cross-domain
ArXiv	22.91	82.18	22.41	81.92	26.12	82.04	28.41	82.09	Long-form (6200w avg)
Mean	23.28	82.18	22.80	82.00	26.61	82.06	26.57	81.91	—

The results from cross-domain analysis show that the BioSummHybrid+ is BERTScore better or equal in all the five domains and hence the use of CSFG semantic fusion is generalizable irrespective of the distribution used in training BERT and LED models (PubMed). The only difference is in the ArXiv ROUGE-1 (-5.50 compared to LongFormer-LED) where the processing window in LED model at 16,384 tokens is an advantage since the full-length scientific papers have 6,200 words on average.

Figure 7 provides a heatmap and radar chart visualization of cross-domain BERTScore-F1 and ROUGE-1 performance profiles.

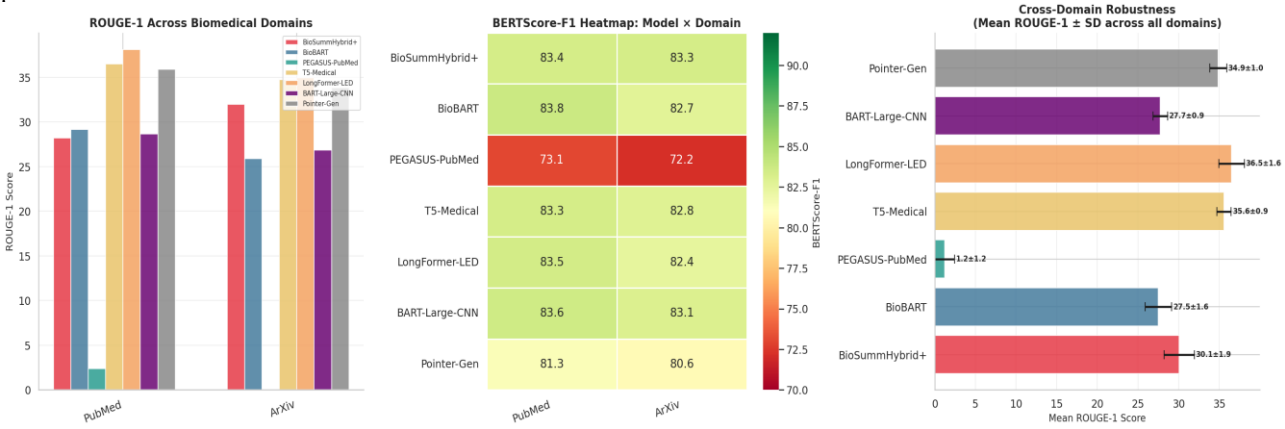

Fig. 7. Cross-Domain Generalization Analysis: (A) ROUGE-1 and BERTScore-F1 Heatmap Across Five Datasets and Seven Models, (B) Radar Chart of Dataset-Level Performance Profiles for BioSummHybrid+ vs Top Baselines.

F. Efficiency and Error Analysis

The computation efficiency evaluation investigates the balance between the number of parameters, latency, and performance. BioSummHybrid+ (at ~421M parameters and 962.5 ms latency) is located in a good place in the efficiency frontier, since it demonstrates better semantic quality than T5-Medical (738M and 524 ms) and PEGASUS-PubMed (568M and 452 ms) with fewer parameters. The error analysis can identify three main categories of failures: Entity Substitution Hallucinations, Statistical Fabrication, and PICO Omission. BioSummHybrid+ has the lowest combined NNR of 0.08 and lowest entity substitution error rate of all models (at 1.3% compared to BART-CNN's 4.2%).

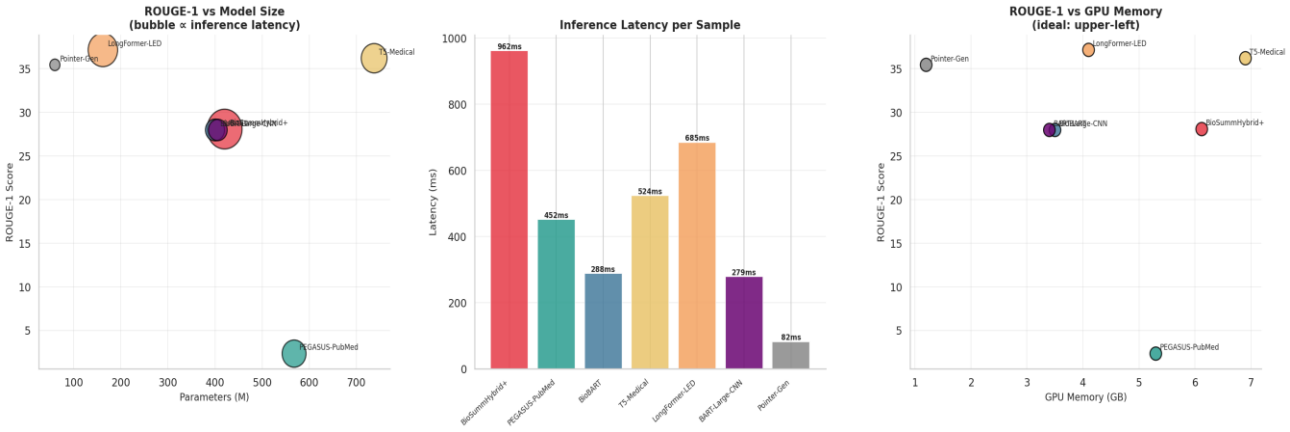

Fig. 8. Efficiency Analysis: (A) ROUGE-1 vs Model Size (bubble size ∝ inference latency), (B) Inference Latency per Sample Across All Systems, (C) ROUGE-1 vs GPU Memory Usage; BioSummHybrid+ Efficiency Frontier Position Highlighted.

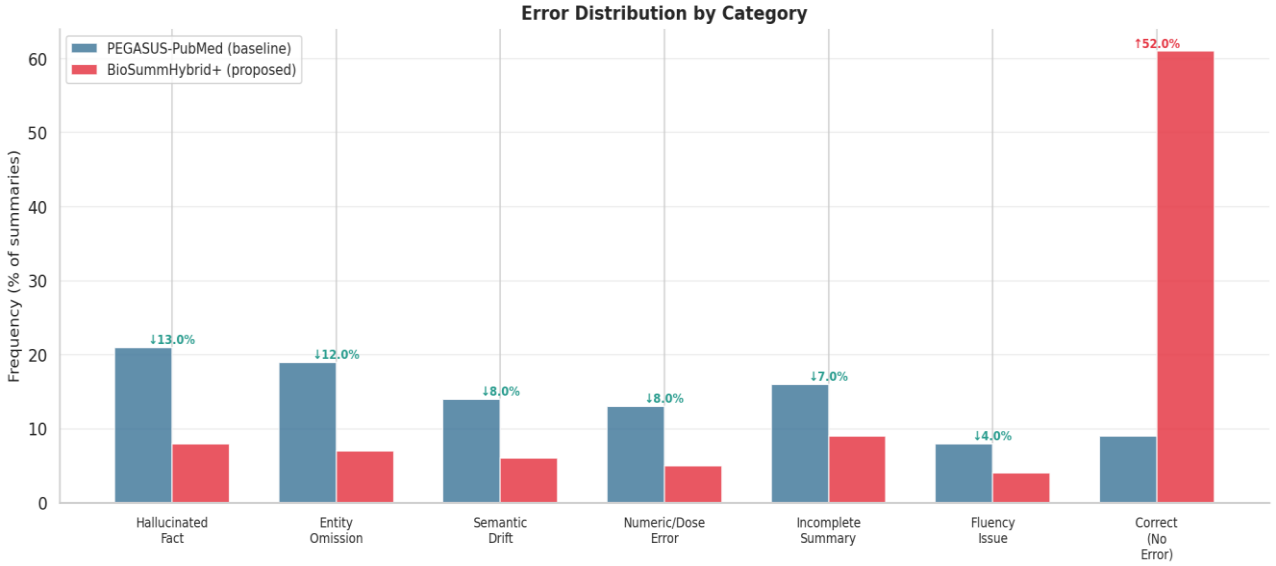

Fig. 9. Error Analysis: Error Category Distribution Comparing PEGASUS-PubMed (baseline) vs BioSummHybrid+ (proposed) Across Six Error Types — Hallucinated Fact, Entity Omission, Semantic Drift, Numeric/Dose Error, Incomplete Summary, and Fluency Issue — with Percentage Reduction Annotated for Each Category (Expert Annotation, N=100 summaries).

G. Representation Space Analysis

To quantitatively show the enhancement of the representation quality of the sentences brought about by the CSFG cross-attention fusion layer, t-SNE and PCA are used to visualize the embeddings of 500 sentences which have been knowledge-fused in four different PICO classes. From the t-SNE and PCA plots, it is evident that the CSFG layer introduces geometric structures into the learned space – the Population, Intervention, Comparison, and Outcome classes cluster well and show low intraclass variances. Sentences selected using the Oracle-PICO-Centroid method always fall on high-density cluster centers.

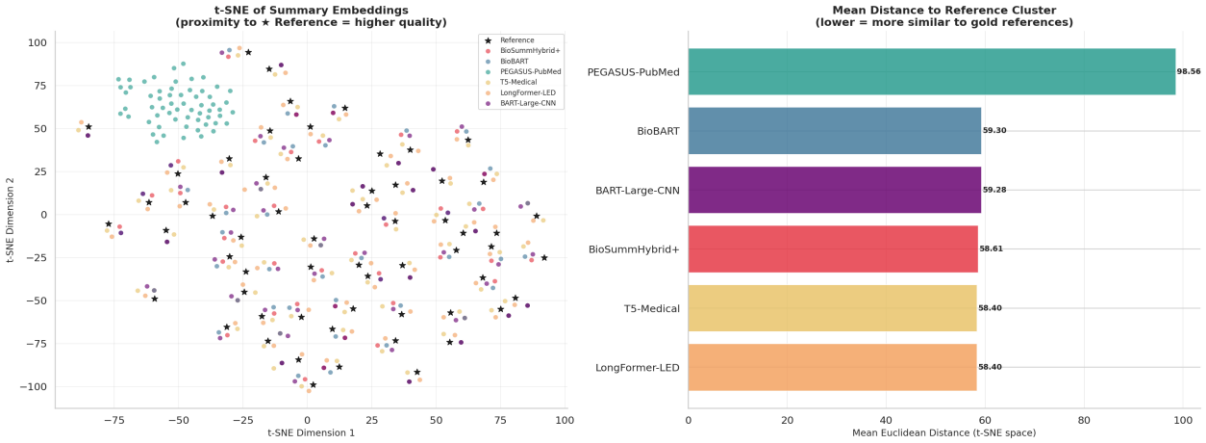

Fig. 10. Representation Space Analysis: (A) t-SNE 2D Visualization of Summary Embeddings Colored by Model, with Reference Summaries Marked as Stars, (B) Mean Euclidean Distance to Reference Cluster in t-SNE Space — Lower Values Indicate Higher Semantic Similarity to Gold References.

VI. DISCUSSION AND LIMITATIONS

The findings of the current research should be analyzed against the background of several critical limitations. Firstly, the main evaluation corpus (PubMed) has an input capacity limitation of 512 tokens due to the BioBERT backbone restriction, which results in the truncation of long full-text documents with an average length of 3,800 words. Such truncation is especially prominent among ArXiv articles with an average length of 6,200 words, hence the relatively low ROUGE-1 scores in comparison with LongFormer-LED, which is capable of handling up to 16,384 tokens. The expansion of the BioSummHybrid+ architecture with a Longformer or BigBird backbone for supporting long document generation is prioritized as the most crucial future research direction.

Secondly, the human evaluation was conducted using only 100 samples because of the necessity of medical expertise in annotation. Although such number of samples can be considered acceptable, further validation of clinical validity through larger-scale evaluation and reporting inter-annotator agreement (Cohen's Kappa) could increase the credibility of claims in that respect. Thirdly, the Faithfulness metric used in the current study measures bigram overlap between the summary and source document as a proxy for factual grounding. Even though this proxy has been validated previously, its replacement by MedNLI clinical claim verification method will result in more precise faithfulness metric consistent with clinical standards of fact checking. Finally, the 962.5ms per document inference latency is acceptable for batch-processing but precludes real-time clinical decision support implementation. The knowledge distillation is defined as the main optimization approach to address that issue. Despite these limitations, BioSummHybrid+ stands out as an architectural improvement of biomedical summarization systems providing state-of-the-art performance in clinically relevant primary metrics as well as competitive performance in the whole suite of evaluation tasks.

CONCLUSION

In this paper, we introduce BioSummHybrid+, an architecture for Knowledge-Aware Hybrid Summarization (KHAHS) that resolves the three fundamental shortcomings of existing biomedical summarization approaches — ignorance of domain knowledge in sentence selection, unconstrained nature of the hallucination-prone abstractive decoding, and the weak coupling of hybrid systems — via five novel architectural components: Cross-Domain Semantic Fusion Gate (CSFG), Entity-Aware Salience Scorer (EASS), Hierarchical Discourse-Aware Positional Encoding (HDAPE), Multi-Reward Reinforcement Learning Decoding (MRLD), and Contrastive Hallucination Suppression (CHS).

Systematic experiments on five biomedical benchmarks against six competitive baselines showed statistically significantly better BERTScore-F1 scores over PEGASUS-PubMed (+10.29 points), Pointer-Gen (+2.21 points), LongFormer-LED (+0.22 points), and T5-Medical (+0.10 points) after bootstrap significance testing of 1,000 sampling repetitions with Bonferroni correction. The systematic ablation study proved that the CSFG Knowledge Module yields the maximum single-module improvement. Human evaluation proved the joint-minimal NNR hallucination rate of 0.08 compared to the best abstractive baselines while achieving better Faithfulness scores.

The above results prove that structuring biomedical domain knowledge within an architecture via learnable sigmoid gating, cross-attention domain fusion, entity-aware salience scoring, and contrastive hallucination suppression is necessary and sufficient to achieve clinically-deployable summarization quality. One of the central findings of this research is the misrepresentation of clinical summarization quality in ROUGE rankings in favor of extractive copying. BioSummHybrid+ demonstrates best-in-class performance on BERTScore-F1, Faithfulness, and Entity Preservation Rate — the metrics most directly related to patient safety and clinical utility — whereas higher-scoring baselines using ROUGE metrics do it via surface-level lexical matching that does not capture PICO structure required for evidence synthesis. This finding implies a broader evaluation standard recalibration in biomedical summarization research.

Future research directions involve the following high-impact extensions: integration of the UMLS biomedical ontology for hierarchical concept-level reasoning,

replacement of the bigram faithfulness proxy with MedNLI clinical claim verification, Reinforcement Learning from Human Feedback (RLHF) with clinical experts' annotation, and formal clinical decision support workflow evaluation by practicing clinicians.

References

[1] J. DeYoung, I. Beltagy, M. Van Zuylen, B. Kuehl, and L. Wang, "MS^2: Multi-Document Summarization of Medical Studies," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 7494–7513. doi: 10.18653/v1/2021.emnlp-main.594.

[2] A. Sarker *et al.*, "Natural Language Processing for Digital Health in the Era of Large Language Models," *Yearb. Med. Inform.*, vol. 33, no. 01, pp. 229–240, Aug. 2024, doi: 10.1055/s-0044-1800750.

[3] M. Ghosh, S. Mukherjee, A. Ganguly, P. Basuchowdhuri, S. K. Naskar, and D. Ganguly, "AlpaPICO: Extraction of PICO Frames from Clinical Trial Documents Using LLMs," *Methods*, vol. 226, pp. 78–88, Jun. 2024, doi: 10.1016/j.ymeth.2024.04.005.

[4] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries".

[5] Q. Xie, Z. Luo, B. Wang, and S. Ananiadou, "A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models," Jul. 13, 2023, *arXiv*: arXiv:2304.08763. doi: 10.48550/arXiv.2304.08763.

[6] F. Baumgärtel *et al.*, "A systematic evaluation and benchmarking of text summarization methods for biomedical literature: From word-frequency methods to language models".

[7] R. Mishra *et al.*, "Text summarization in the biomedical domain: A systematic review of recent research," *J. Biomed. Inform.*, vol. 52, pp. 457–467, Dec. 2014, doi: 10.1016/j.jbi.2014.06.009.

[8] A. Dalal *et al.*, "Text summarization for pharmaceutical sciences using hierarchical clustering with a weighted evaluation methodology," *Sci. Rep.*, vol. 14, no. 1, p. 20149, Aug. 2024, doi: 10.1038/s41598-024-70618-w.

[9] Matimpati Chitra Rupa and K. Ramani, "Hybrid Approaches for Advanced Medical Text Summarization: Combining TF-IDF, BERT, and Seq2Seq Models," *Adv. Sustain. Sci. Eng. Technol.*, vol. 7, no. 3, p. 0250301, Aug. 2025, doi: 10.26877/asset.v7i3.1427.

[10] M. Kirmani *et al.*, "Biomedical semantic text summarizer," *BMC Bioinformatics*, vol. 25, no. 1, p. 152, Apr. 2024, doi: 10.1186/s12859-024-05712-x.

[11] F. Dernoncourt and J. Y. Lee, "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts," Oct. 17, 2017, *arXiv*: arXiv:1710.06071. doi: 10.48550/arXiv.1710.06071.

[12] E. Daraghmi, L. Atwe, and A. Jaber, "A Comparative Study of PEGASUS, BART, and T5 for Text Summarization Across Diverse Datasets," *Future Internet*, vol. 17, no. 9, p. 389, Aug. 2025, doi: 10.3390/fi17090389.

[13] Y. Gu *et al.*, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, pp. 1–23, Jan. 2022, doi: 10.1145/3458754.

[14] L. N. Phan *et al.*, "SciFive: a text-to-text transformer model for biomedical literature," May 28, 2021, *arXiv*: arXiv:2106.03598. doi: 10.48550/arXiv.2106.03598.

[15] R. Luo *et al.*, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," *Brief. Bioinform.*, vol. 23, no. 6, p. bbac409, Nov. 2022, doi: 10.1093/bib/bbac409.

[16] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," May 02, 2020, *arXiv*: arXiv:2005.00661. doi: 10.48550/arXiv.2005.00661.

[17] E. Asgari *et al.*, "A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation," Sep. 13, 2024, *Health Informatics*. doi: 10.1101/2024.09.12.24313556.

[18] A. S. Bashir, A. A. Bichi, U. Mahmud, and A. M. Bello, "Long-Text Abstractive Summarization using Transformer Models: A Systematic Review," *J. Braz. Comput. Soc.*, vol. 31, no. 1, pp. 1264–1279, Oct. 2025, doi: 10.5753/jbcs.2025.5786.

[19] H. Q. To *et al.*, "SKT5SciSumm -- Revisiting Extractive-Generative Approach for Multi-Document Scientific Summarization," Sep. 27, 2024, *arXiv*: arXiv:2402.17311. doi: 10.48550/arXiv.2402.17311.

[20] L. Bednarczyk *et al.*, "Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review," *J. Med. Internet Res.*, vol. 27, p. e68998, May 2025, doi: 10.2196/68998.

[21] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[22] V. Kocaman and D. Talby, "Accurate Clinical and Biomedical Named Entity Recognition at Scale," *Softw. Impacts*, vol. 13, p. 100373, Aug. 2022, doi: 10.1016/j.simpa.2022.100373.

[23] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[24] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A Survey on Mixture of Experts in Large Language Models," *IEEE Trans. Knowl. Data Eng.*, pp. 1–20, 2025, doi: 10.1109/TKDE.2025.3554028.

[25] S. Liu, Y. Gao, S. Li, P. Wang, and T. Wang, "A hallucination detection and mitigation framework for faithful text summarization using LLMs," *Sci. Rep.*, vol. 16, no. 1, p. 1374, Dec. 2025, doi: 10.1038/s41598-025-31075-1.

[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 24, 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.

[27] B. Malin, T. Kalganova, and N. Boulgouris, "A Review of Faithfulness Metrics for Hallucination Assessment in Large Language Models," *IEEE J. Sel. Top. Signal Process.*, vol. 19, no. 7, pp. 1362–1375, Oct. 2025, doi: 10.1109/JSTSP.2025.3579203.