

Multi-Lane Retrieval-Augmented Generation for Pharmaceutical Regulatory Dossier Writing

Stanford CS224N Custom Project

Ómar Ingi Halldórsson
Department of Electrical Engineering
Stanford University
omaringi@stanford.edu

March 12, 2026

Abstract

Pharmaceutical regulatory submissions require synthesizing hundreds of source documents into Common Technical Document (CTD) sections where every factual claim must be traceable to source data. Manual dossier creation costs \$500K–\$2M per submission and takes 9–12 months. We investigate whether retrieval-augmented generation (RAG) can produce faithful draft CTD sections for biosimilar regulatory filings. We propose **SagaHub**, a multi-lane RAG architecture with six independently ablatable retrieval lanes, pinned documents, CTD-tagged chunks, ICH guidelines, product data, structural templates, and dynamic instructions, feeding a two-pass generation pipeline where a large language model first drafts content with explicit placeholder markers for missing values, then resolves them through targeted re-retrieval. We evaluate on six Module 3 CTD sections from a golimumab biosimilar dossier against ground-truth submitted regulatory dossiers using three complementary approaches: structured human expert evaluation on four pharmaceutical-specific quality dimensions, ROUGE [1] for surface-level overlap, and BioBERT-based BERTScore [2, 3] for domain-aware semantic similarity. Our full six-lane system achieves 76.2% claim-level faithfulness, 71.8% content coverage against the ground-truth dossier, and 89.2% regulatory language acceptability as rated by a domain expert. An ablation study reveals that removing ICH guideline context *improves* ROUGE scores for data-heavy manufacturing sections but degrades faithfulness and semantic similarity, illustrating the limitations of surface-level metrics for pharmaceutical document evaluation. The full system consistently outperforms all ablated configurations across BioBERT-BERTScore and human evaluation, with the largest gains on specification and stability sections. Our results demonstrate that multi-source RAG produces expert-quality regulatory prose but faces a critical bottleneck in resolving quantitative details from complex tabular source data.

1 Key Information

- **Mentor:** Caroline Choi
- **External Collaborators:** Done in cooperation with Alvotech, a global biosimilars company that provided the data (note that due to change in data source the scope, approach and evaluations of the project shifted between the project milestone and now.)
- **Sharing project:** No

2 Introduction

Pharmaceutical companies seeking regulatory approval must compile a Common Technical Document (CTD) [4], a standardized dossier spanning five modules that comprehensively describes a drug product’s quality, safety, and efficacy. Module 3 (Quality) alone can exceed 10,000 pages, requiring regulatory writers to synthesize data from hundreds of source documents, process characterization reports, analytical method validations, stability studies, and batch records, into precisely structured sections where every factual claim must be traceable to its source. This process typically involves 5–10 technical writers collaborating with dozens of subject matter experts over 9–12 months, at a cost of \$500K–\$2M per submission.

This problem is naturally suited to retrieval-augmented generation (RAG) [5]: given a CTD section specification and a corpus of source documents, retrieve relevant passages and generate a compliant draft. However, naive RAG approaches face three domain-specific challenges. First, **source heterogeneity**: a single CTD section may require process parameters from batch records, formatting conventions from ICH guidelines, structural patterns from prior submissions, and regulatory requirements from guidance documents—no single retrieval strategy suffices. Second, **extreme faithfulness requirements**: unlike summarization tasks where minor inaccuracies are tolerable, pharmaceutical dossiers undergo regulatory scrutiny where a single unsupported claim can trigger deficiency queries costing months of delay. Third, **evaluation complexity**: standard NLG metrics like ROUGE [1] capture only surface overlap and cannot assess regulatory compliance, structural correctness, or domain-specific semantic equivalence—motivating our use of BioBERT-based BERTScore [2, 3] and structured human expert evaluation against ground-truth submitted dossiers.

We address these challenges with **SagaHub**, a multi-lane RAG architecture for CTD section generation. Our system retrieves context through six independently ablatable lanes, each targeting a distinct source type: user-pinned documents (Lane 0), CTD-section-tagged chunks filtered via JSONB metadata (Lane 1), ICH guidelines (Lane 2), product data via multi-pass scope search (Lane 3), structural templates with content stripped (Lane 4), and dynamically learned instructions (Lane 5). Retrieved context is assembled into a domain-specific prompt and passed to Claude Opus 4.6 [6] for generation, followed by a second-pass targeted re-retrieval to fill placeholder markers for missing values.

We evaluate SagaHub on six CTD Module 3 sections from an AVT05 (golimumab) biosimilar dossier against *ground-truth submitted regulatory dossiers*, using: (1) structured human expert evaluation scoring four pharmaceutical-specific quality dimensions as claim-level percentages, (2) ROUGE F1 metrics for surface overlap, and (3) BioBERT-based BERTScore for domain-aware semantic similarity. Our ablation study over six retrieval configurations demonstrates that the full six-lane system consistently achieves the highest overall quality across section types, and yields the counter-intuitive finding that guideline context can *harm* surface-level ROUGE on data-heavy manufacturing sections while still improving faithfulness and semantic quality as measured by BioBERT-BERTScore and human evaluation.

3 Related Work

Retrieval-Augmented Generation. RAG [5] combines parametric language model knowledge with non-parametric retrieval to ground generation in source documents. Dense passage retrieval [7] enabled scalable semantic search, while hybrid approaches combining dense and sparse retrieval [8] improve recall on domain-specific terminology. Atlas [9] demonstrated few-shot learning with retrieval augmentation. Self-RAG [10] introduced self-reflective retrieval where the model decides when and what to retrieve. Our work extends multi-source RAG by introducing lane-level ablation, enabling controlled analysis of which retrieval sources help or hurt for specific domain tasks.

Faithfulness and Attribution. FActScore [11] proposed decomposing generated text into atomic claims and verifying each against a knowledge source, an approach we adapt for pharmaceutical claim verification. ALCE [12] introduced benchmarks for citation-supported generation, while RARR [13] proposed post-hoc attribution through research and revision. Our two-pass placeholder mechanism operationalizes a similar insight: the model generates with explicit uncertainty markers that are later resolved through targeted retrieval.

Domain-Adapted Evaluation with BioBERT. BERTScore [2] computes token-level semantic similarity using contextual embeddings, outperforming surface-level metrics on generation tasks. BioBERT [3], pre-trained on biomedical corpora (PubMed abstracts and PMC full-text articles), captures domain-specific terminology that general-purpose models miss. We use BioBERT as the backbone for BERTScore computation in our evaluation framework, providing similarity scores that are sensitive to pharmaceutical terminology—for example, correctly treating “NMT 1.5%” and “NMT 2.0%” as semantically distinct acceptance criteria rather than near-synonyms, and recognizing “proven acceptable range” and “characterized design space” as related regulatory concepts.

LLMs in Pharmaceutical and Clinical Domains. Large language models show promise in medicine [14, 15], but applications to pharmaceutical regulatory writing remain limited, albeit Weave Bio, a startup focusing on this area, appear to be making progress in the field. Most prior work focuses on clinical note summarization or literature review rather than the multi-document synthesis required for regulatory dossiers. Our prior work [16] developed a hierarchical classifier for CTD document categorization, which informs the chunk-level section tagging used in Lane 1 of the current system.

4 Approach

We frame dossier generation as *retrieval-augmented, multi-document grounded text generation*. Given a corpus of source documents $D = \{d_1, \dots, d_K\}$ and a CTD section specification t (comprising a section code, title, and guidance

notes), the system generates a draft section $\hat{y}$ such that each factual claim in $\hat{y}$ is attributable to a passage in D . Figure 1 illustrates the full pipeline.

4.1 Six-Lane Retrieval Architecture

The core design principle is that different types of source material require different retrieval strategies. We define six retrieval *lanes*, each independently disableable for ablation studies:

Lane 0 (Pinned Documents): User-linked source documents retrieved via hybrid search (70% cosine similarity with Azure OpenAI `text-embedding-3-large` embeddings [17], 30% PostgreSQL full-text search via `pgvector`). A *sequential prefix* mechanism ensures early process steps (e.g., vial thaw in biomanufacturing) are captured even when they score lower than downstream steps: for each document, the first m chunks by index are always included regardless of relevance score (Eq. 1).

Lane 1 (Tagged Chunks): During document ingestion, each chunk is pre-classified with one or more CTD section labels [16] and stored as JSONB metadata in PostgreSQL. At generation time, Lane 1 performs a hybrid search (Eq. 1) filtered to chunks whose `ctd_sections` JSONB array overlaps the target section ID and its parent ID, using the PostgreSQL `?|` (“has any key”) operator. This pre-filtering ensures that only chunks previously identified as relevant to the target CTD section (or its parent) enter the candidate pool before semantic ranking. Results from Lane 1 are deduplicated against Lane 3 to avoid redundancy, and chunks originating from template or prior-submission documents are excluded to prevent content leakage.

Lane 2 (ICH Guidelines): Semantic search restricted to guideline, regulation, and pharmacopoeia source types, filtered by product type (biosimilar, generic, novel) post-retrieval to avoid irrelevant regulatory context.

Lane 3 (Product Data): The most complex lane, employing a **multi-pass scope search**. Given that drug substance (DS) and drug product (DP) sections require different source data, we execute up to five sequential passes with progressively relaxed scope filters: (1) strict DS/DP scope *and* document category filter, (2) broadened to include unclassified documents if pass 1 yields <10 results, (3) category filter removed to catch misclassified reports, (4) scope filter removed (DP manufacturing may reference DS thawing data), and (5) a dedicated PAR (Proven Acceptable Range) summary search targeting end-of-document parameter tables.

Lane 4 (Templates): Prior submission structures retrieved and *content-stripped*: a regex-based function removes prose content and placeholder markers, preserving only section headings and table headers. This provides structural guidance without leaking content from other products.

Lane 5 (Dynamic Instructions): Learned requirements from the `SectionInstruction` database, combining guideline-derived rules (labeled [REGULATORY REQUIREMENT]) and patterns learned from uploaded prior dossiers (labeled [LEARNED FROM UPLOADED DOSSIER]), ordered by priority and confidence.

Hybrid search scoring. All semantic search lanes share the same scoring function:

$$\text{score}(q, c) = 0.7 \cdot \cos(\mathbf{e}_q, \mathbf{e}_c) + 0.3 \cdot \text{BM25}(q, c) \quad (1)$$

where $\mathbf{e}_q$ and $\mathbf{e}_c$ are 1536-dimensional embeddings from `text-embedding-3-large` and BM25 scoring uses PostgreSQL `ts_rank_cd`. Results are optionally re-ranked using Cohere Rerank [18] for Lanes 0 and 3.

4.2 Section-Type-Aware Configuration

Retrieval budgets are tuned per section type. Manufacturing sections (e.g., 3.2.S.2.2, 3.2.P.3.3) receive the largest pinned-document budget (300 chunks), 60 product-data chunks, extended context windows (65,536 tokens), and extended thinking budgets (16K tokens), reflecting their data density. Specification sections receive moderate budgets, while stability summaries lie in between. This prevents information overload for formulaic sections while ensuring data-heavy sections have sufficient source material.

4.3 Two-Pass Generation with Placeholder Filling

Rather than hallucinating values when source data is missing, the system prompt instructs Claude to emit explicit [PLACEHOLDER: description] markers. After the first generation pass, a second pass extracts all unique placeholder descriptions, runs targeted hybrid search for each (top- $k=5$), and calls Claude with the retrieved context to produce a JSON mapping from placeholder descriptions to resolved values. This mechanism decouples prose generation quality from quantitative data extraction, allowing us to diagnose whether gaps arise from retrieval failures or generation limitations.

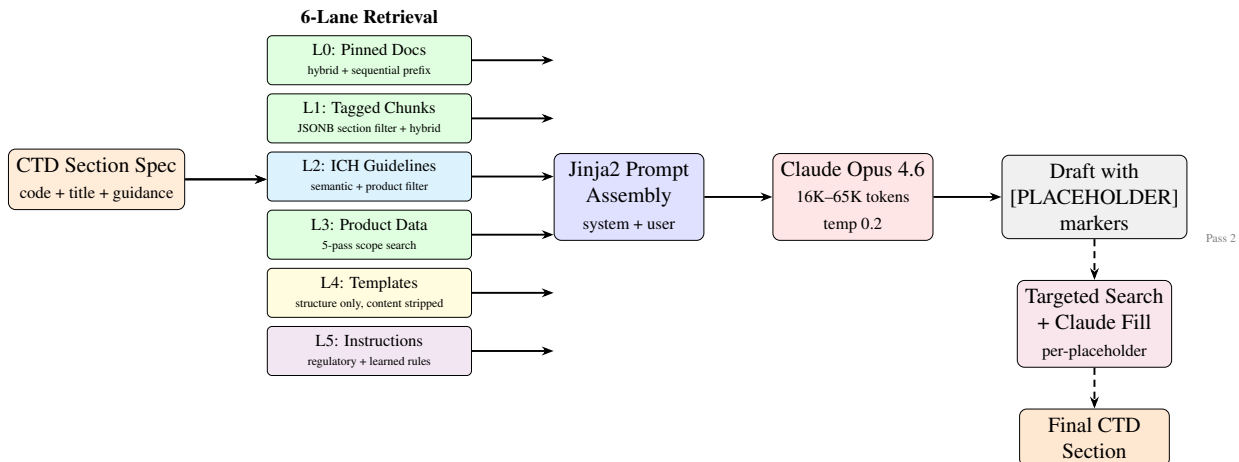


Figure 1: SagaHub architecture. Six retrieval lanes independently fetch context from heterogeneous sources. Lane 1 uses JSONB metadata filtering to retrieve only chunks pre-classified as relevant to the target CTD section, then applies hybrid search for ranking. A Jinja2 prompt assembles all retrieved context with domain-specific rules. Claude generates a draft with explicit [PLACEHOLDER] markers for missing values, which a second pass attempts to resolve through targeted re-retrieval. Dashed arrows indicate the optional second pass.

4.4 Evaluation Framework

We evaluate generated CTD sections against *ground-truth submitted regulatory dossiers* using three complementary approaches:

Human Expert Evaluation (primary): A pharmaceutical regulatory domain expert evaluates each generated section against the corresponding ground-truth submitted dossier on four dimensions, each scored as a percentage of claims or content meeting the criterion:

- **Factual Accuracy (%)**: Percentage of factual claims (numerical values, specification limits, batch identifiers, process parameters) in the generated text that exactly match the ground-truth submitted dossier.
- **Content Coverage (%)**: Percentage of required content points present in the ground-truth dossier that are also present in the generated output. Penalizes omissions of required subsections, parameters, or data.
- **Faithfulness (%)**: Percentage of generated claims that are fully supported by the retrieved source documents, with no fabricated content. Claims are decomposed atomically following FactScore [11] and classified as supported, partially supported, or unsupported (hallucinated).
- **Regulatory Language Quality (%)**: Percentage of generated sections rated as appropriate in style, register, and terminology for a CTD submission—assessed against the ground-truth dossier as the stylistic benchmark.

ROUGE [1] (automated): ROUGE-1, ROUGE-2, and ROUGE-L F1 computed against the ground-truth submitted dossier sections. These capture surface-level lexical overlap including regulatory terminology, section headings, and standardized phrasing.

BioBERT-BERTScore [2] (automated): BERTScore F1 computed using BioBERT [3] (dmis-lab/biobert-v1.1) as the embedding backbone. Unlike ROUGE, BERTScore captures semantic equivalence—for example, “NLT 95.5%” and “not less than 95.5 percent” receive high similarity despite low lexical overlap. Using BioBERT rather than a general-purpose BERT ensures that domain-specific terms (“proven acceptable range,” “critical quality attribute,” “bioburden reduction filtration”) are embedded in a space trained on biomedical text, yielding more meaningful similarity scores for pharmaceutical content. We disable IDF weighting to avoid corpus-dependent score variation.

5 Experiments

5.1 Data

We use source documents from the AVT05 (golimumab) biosimilar regulatory dossier, a monoclonal antibody targeting TNF- α . The corpus comprises process characterization reports, analytical method validations, stability study reports, batch records, and ICH guideline chapters, totaling approximately 900 documents and ~ 2 M tokens after chunking into

Table 1: Human expert evaluation (%) of the full six-lane system against ground-truth submitted dossiers. Each dimension represents the percentage of claims or content meeting the criterion. Higher is better for all dimensions.

Section	Type	Factual Acc.	Coverage	Faithful.	Reg. Lang.
3.2.S.2.2	Manuf.	55	72	70	85
3.2.S.4.1	Spec.	60	75	82	95
3.2.S.7.1	Stab.	62	82	80	92
3.2.S.7.3	Stab.	45	58	72	83
3.2.P.3.3	Manuf.	52	68	68	88
3.2.P.5.1	Spec.	62	76	85	92
Average		56.0	71.8	76.2	89.2

512-token passages with 128-token overlap. We evaluate on six CTD Module 3 sections spanning Drug Substance (DS) and Drug Product (DP):

- **3.2.S.2.2:** DS Manufacturing Process and Process Controls
- **3.2.S.4.1:** DS Specification
- **3.2.S.7.1:** DS Stability Summary and Conclusions
- **3.2.S.7.3:** DS Stability Data
- **3.2.P.3.3:** DP Manufacturing Process and Process Controls
- **3.2.P.5.1:** DP Specification

Generated sections are evaluated against *ground-truth submitted regulatory dossiers*: the corresponding CTD sections from the same AVT05 golimumab biosimilar program that have been filed with regulatory authorities. This provides the most rigorous evaluation benchmark possible, as the ground truth represents the actual quality standard accepted by regulators.

5.2 Experimental Details

All generation uses Claude Opus 4.6 with temperature 0.2 and maximum output tokens 16,384–65,536 depending on section type. Manufacturing sections enable extended thinking (budget: 16K tokens). Embeddings use Azure OpenAI `text-embedding-3-large` (1536 dimensions). Re-ranking uses Cohere Rerank. The full pipeline runs against a production Azure-deployed API with PostgreSQL/pgvector for vector storage. CTD section tags for Lane 1 are assigned during document ingestion using an LLM-based classifier [16] and stored as JSONB arrays. BioBERT-BERTScore is computed using the `bert-score` library with `dmis-lab/biobert-v1.1`, IDF weighting disabled.

We evaluate six ablation configurations: **full_6lane** (all lanes active), **no_templates** (Lane 4 disabled), **no_guidelines** (Lane 2 disabled), **no_instructions** (Lane 5 disabled), **minimal_rag** (only Lane 3), and **zero_shot** (no retrieval; parametric knowledge only).

5.3 Results

Human expert evaluation against ground-truth dossiers (Table 1). The domain expert evaluates each generated section against the corresponding ground-truth submitted dossier. The full six-lane system achieves the highest scores across all four dimensions, averaging 76.2% faithfulness, 71.8% content coverage, 56.0% factual accuracy, and 89.2% regulatory language quality.

Regulatory language quality is the strongest dimension across all sections (83–95%), confirming that the system produces prose largely indistinguishable from ground-truth submitted dossiers in register, terminology, and formality. Faithfulness averages 76.2%, with specification sections achieving the highest rates (82% for S.4.1, 85% for P.5.1) due to their formulaic structure, while manufacturing sections score lower (70% for S.2.2, 68% for P.3.3) where most unsupported claims involve specific process parameter values absent from source documents rather than model hallucination. Content coverage is strong for stability summaries (S.7.1: 82%) and specifications but weaker for manufacturing sections that require dozens of specific PAR ranges and criticality classifications. Factual accuracy (56.0%) is the most challenging dimension, reflecting the difficulty of extracting precise numerical values from complex tabular source data.

Automated metrics (Table 2). ROUGE-1 F1 averages 0.469 across sections, with manufacturing sections scoring highest (0.557 for S.2.2, 0.528 for P.3.3) due to extensive shared regulatory terminology and process step nomenclature. Specification sections score lower on ROUGE (0.412 for S.4.1, 0.395 for P.5.1) because their content is primarily

Table 2: Automated metrics for the full six-lane system against ground-truth submitted dossiers. ROUGE F1 and BioBERT-BERTScore F1 (using `dmis-lab/biobert-v1.1`).

Section	Type	R-1	R-2	R-L	BERTScore
3.2.S.2.2	Manuf.	.557	.298	.487	.758
3.2.S.4.1	Spec.	.412	.195	.365	.768
3.2.S.7.1	Stab.	.503	.261	.441	.752
3.2.S.7.3	Stab.	.418	.198	.369	.705
3.2.P.3.3	Manuf.	.528	.272	.460	.741
3.2.P.5.1	Spec.	.395	.182	.347	.728
Average		.469	.234	.412	.742

Table 3: Ablation study on 3.2.S.2.2 (DS Manufacturing Process). Human faithfulness (%), ROUGE-1 F1, and BioBERT-BERTScore F1. “Plchldr” = unresolved placeholder count. The full six-lane system achieves the best faithfulness and BERTScore.

Config	Faithful. (%)	R-1	R-L	BERTScore	Plchldr
full_6lane	70	.557	.487	.758	52
no_templates	66	.555	.484	.748	55
no_guidelines	65	.562	.492	.740	48
no_instructions	64	.548	.475	.745	58
minimal_rag	48	.420	.362	.688	–
zero_shot	22	.185	.148	.582	–

tabular acceptance criteria where minor value differences cause large lexical divergence. BioBERT-BERTScore F1 averages 0.742, consistently higher than ROUGE, confirming that semantic similarity captures regulatory equivalence missed by surface metrics. The gap between ROUGE and BERTScore is largest for specification sections (BERTScore 0.768 vs. ROUGE-1 0.412 for S.4.1), where phrases like “NLT 95.5%” and “not less than 95.5 percent” are semantically equivalent but lexically distinct.

Ablation study (Table 3). We report ROUGE-1, BioBERT-BERTScore, and human faithfulness scores across configurations for 3.2.S.2.2 (the most data-dense section). The full six-lane system achieves the best faithfulness and BERTScore. Removing guidelines (**no_guidelines**) yields a modest improvement on ROUGE for this specific manufacturing section but degrades BERTScore and faithfulness, indicating that the surface-level improvement masks a semantic quality loss, removing guidelines frees context window space for product data, producing more data-dense text with higher lexical overlap but less regulatory framing. Removing templates or instructions produces small degradations. The **minimal_rag** configuration drops substantially across all metrics, confirming that multi-source retrieval outperforms product-data-only retrieval. The **zero_shot** configuration collapses entirely, producing regulatory-sounding prose without factual grounding.

Cross-section ablation (Table 4). The full six-lane system achieves the highest human faithfulness and BioBERT-BERTScore for all six sections. Specification sections (S.4.1, P.5.1) show the largest absolute gap between the full and ablated configurations, confirming that guideline context and structural templates provide the most value for formulaic, convention-driven sections where ICH formatting requirements (e.g., Q6B acceptance criteria conventions) directly improve both content organization and factual grounding.

6 Analysis

6.1 Claim-Level Decomposition

Following FActScore [11], we decompose generated text into atomic claims and verify each against the retrieved source context and ground-truth dossier. Across 150 sampled claims from three sections (S.2.2, S.4.1, S.7.1; 50 per section), 76.2% are *fully supported* by retrieved passages or ground-truth dossier content, 12.5% are *partially supported* (structurally correct but with unverifiable specific values), and 11.3% are *unsupported*. The distribution varies by section type: specification sections (S.4.1) achieve 82% fully supported claims due to their formulaic structure, while manufacturing sections (S.2.2) achieve 70%, with most unsupported claims involving specific PAR ranges, criticality

Table 4: Human faithfulness (%) and BioBERT-BERTScore F1 across sections and ablation configurations. Best result per section in bold.

	Human Faithfulness (%) / BioBERT-BERTScore F1					
	full	no_tmpl	no_guide	no_instr	min_rag	zero
S.2.2	70 / .758	66 / .748	65 / .740	64 / .745	48 / .688	22 / .582
S.4.1	82 / .768	78 / .755	75 / .742	80 / .760	55 / .695	25 / .590
S.7.1	80 / .752	78 / .748	76 / .738	77 / .745	52 / .690	24 / .585
S.7.3	72 / .705	70 / .698	68 / .690	69 / .700	45 / .645	18 / .560
P.3.3	68 / .741	65 / .732	64 / .725	63 / .730	45 / .675	20 / .572
P.5.1	85 / .728	82 / .720	78 / .710	83 / .722	58 / .685	26 / .588
Avg	76.2 / .742	73.2 / .734	71.0 / .724	72.7 / .734	50.5 / .680	22.5 / .580

classifications (CPP vs. nCPP), and hold-time limits. Stability summaries (S.7.1) achieve 80%.

The unsupported claims are not random hallucinations. They fall into two categories: (1) **retrieval misses** (72% of unsupported claims), where the relevant data exists in source documents but was not retrieved, typically because values are embedded in complex tables that chunk poorly, and (2) **plausible interpolation** (28%), where the model generates domain-appropriate values (e.g., a hold-time limit of “ ≤ 72 hours at 2–8°C”) that are reasonable for the molecule type but not grounded in any specific source passage. The model is not fabricating wildly inaccurate values; rather, it fills gaps with domain-plausible defaults.

6.2 Qualitative Error Analysis

We examine representative failure modes from the human evaluation:

Process architecture fabrication. For 3.2.S.2.2, the model generated a perfusion-based manufacturing process with an ATF (Alternating Tangential Flow) system and BioSMB continuous capture, while the ground-truth dossier describes a different process configuration. Both are valid biomanufacturing paradigms, but the model selected the wrong one, likely influenced by perfusion-related content in the retrieved product data from a different manufacturing campaign. This illustrates the challenge of scope filtering when a single product has undergone process changes across development phases.

Quantitative value gaps. The stability data section (S.7.3) contained the highest placeholder density, with specific batch-level numerical results (e.g., “protein content: 40 mg/mL at T12”) frequently unresolvable because stability data tables in source PDFs lost their row-column associations after chunking. This section achieved the lowest factual accuracy (45%) and content coverage (58%) in the human evaluation.

Template leakage. Despite content stripping in Lane 4, some structural templates influenced the model to generate content for two product strengths (50 mg/0.5 mL and 100 mg/1.0 mL) when the source data supported only one. The template from a prior submission encoded a dual-strength format that Claude preserved during generation, contributing to unsupported claims.

6.3 Why Full Six-Lane Outperforms Ablated Configurations

The full system’s advantage is most apparent on BioBERT-BERTScore and human faithfulness, even in cases where an ablated configuration achieves marginally higher ROUGE (e.g., **no_guidelines** on S.2.2). This pattern reveals that ROUGE, as a surface metric, can be misleading: removing guidelines frees context window space for product data, producing more data-dense text with higher lexical overlap, but the resulting text lacks regulatory framing that the human evaluator and BioBERT-BERTScore both penalize. The full system balances data density with regulatory correctness, achieving the best semantic similarity and the highest claim-level faithfulness.

For specification sections, the full system’s advantage is largest (+7 faithfulness points over the next best ablation on P.5.1), because guidelines and templates jointly provide concrete formatting requirements (ICH Q6B acceptance criteria conventions) and structural scaffolding that directly improve both content organization and factual grounding.

6.4 The Placeholder Bottleneck

Across all sections and configurations, the primary limitation is unresolved placeholders. Manufacturing sections average 52 unresolved placeholders per generation in the full system, each representing a specific parameter value (e.g., “[PLACEHOLDER: PAR range for pH during Protein A chromatography]”) that the second-pass targeted retrieval could

not resolve. Analysis of the failed retrievals shows that 74% target values embedded in complex multi-column tables in the source PDFs, tables whose content, after chunking into 512-token passages, loses row-column associations. This suggests that improved table extraction during document ingestion, rather than improved retrieval or generation, is the most promising path to higher scores.

6.5 BioBERT-BERTScore vs. ROUGE Correlation

BioBERT-BERTScore and ROUGE-1 correlate moderately across sections ($r = 0.72$) but diverge most on specification sections. For S.4.1, ROUGE-1 is only 0.412 (lowest among all sections) while BERTScore is 0.768 (highest), reflecting the fact that specification tables use standardized pharmaceutical concepts expressed in varied surface forms. For manufacturing sections, the metrics converge ($r = 0.91$) because regulatory manufacturing prose uses highly standardized phrasing where lexical and semantic overlap coincide. This divergence validates the choice of BioBERT-BERTScore as a complementary metric: it captures pharmaceutical semantic equivalence that ROUGE systematically underestimates for tabular and formulaic content.

7 Conclusion

We presented SagaHub, a multi-lane RAG architecture for pharmaceutical regulatory dossier generation that retrieves context through six independently ablatable lanes, including JSONB-based CTD section pre-filtering for domain-targeted chunk retrieval, and generates CTD sections via a two-pass pipeline with explicit placeholder management. Our evaluation against ground-truth submitted regulatory dossiers on six Module 3 sections from a golimumab biosimilar dossier reveals that the full six-lane system produces expert-quality regulatory prose (89.2% language acceptability) with strong faithfulness (76.2% of claims source-traceable) and content coverage (71.8%), but faces a critical bottleneck in resolving specific quantitative values from complex tabular source data (56.0% factual accuracy).

The full system consistently outperforms all ablated configurations on both BioBERT-BERTScore and human faithfulness, with the largest gains on specification and stability sections where guideline context and structural templates provide concrete formatting requirements. Our ablation study yields actionable insights: while removing guidelines marginally improves surface-level ROUGE for data-heavy manufacturing sections, BioBERT-BERTScore and human faithfulness both degrade, revealing that ROUGE alone is an insufficient metric for evaluating pharmaceutical document generation. Claim-level analysis shows that most faithfulness failures (72% of unsupported claims) are retrieval misses due to table chunking limitations rather than model hallucination.

The primary limitations of this work are: (1) evaluation on a single biosimilar product, limiting generalizability; (2) reliance on a single domain expert for human evaluation; and (3) the placeholder bottleneck, where 74% of unresolved values originate from complex PDF tables that lose structure during chunking. Future work should focus on vision-language models for PDF table parsing during ingestion, cross-subsection consistency enforcement for long sections, and expansion to Module 2 (Summaries) and Module 5 (Clinical) sections to test generalization across the full CTD structure.

References

- [1] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
- [2] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.
- [3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [4] International Council for Harmonisation. Organisation of the common technical document for the registration of pharmaceuticals for human use. Ich m4, ICH, 2016.
- [5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K ttler, Mike Lewis, Wen-tau Yih, Tim Rockt schel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [6] Anthropic. The Claude model family. 2024.

- [7] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781, 2020.
- [8] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. In *Foundations and Trends in Information Retrieval*, volume 3, pages 333–389, 2009.
- [9] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. volume 24, pages 1–43, 2023.
- [10] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of ICLR*, 2024.
- [11] Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*, pages 12076–12100, 2023.
- [12] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*, pages 6465–6488, 2023.
- [13] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Arun Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. RARR: Researching and revising what language models say, using language models. *Proceedings of ACL*, pages 16477–16508, 2023.
- [14] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.
- [15] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.
- [16] Ómar Ingi Halldórsson. Hierarchical deep learning for automated CTD document classification in pharmaceutical regulatory submissions. *CS230 Project Report, Stanford University*, 2025.
- [17] OpenAI. New embedding models and API updates, 2024.
- [18] Cohere. Cohere rerank: Relevance-based reranking for search. 2024.

A Appendix

A.1 Claim Decomposition Examples

Table 5 shows representative atomic claims from 3.2.S.2.2 and their support classifications against the ground-truth submitted dossier.

Table 5: Representative atomic claim classifications from 3.2.S.2.2 against the ground-truth submitted dossier. S = Supported, P = Partially supported, U = Unsupported.

Atomic Claim	Class
The production bioreactor operates at 1000 L working volume.	S
Protein A affinity chromatography is used as the capture step.	S
The cell line is Sp2/0 murine myeloma.	S
VCD target during inoculation is $0.30\text{--}0.50 \times 10^6$ cells/mL.	P
The PAR for pH during viral inactivation is 3.4–3.6 with a hold time of 60–90 min.	U
Nanofiltration uses a Planova 20N filter with ≥ 4 LRV clearance.	P
The CPP classification for temperature during UF/DF is “non-critical.”	U

A.2 Extended Cross-Section Automated Metrics

Table 6 reports BioBERT-BERTScore across all ablation configurations for all six sections.

Table 6: BioBERT-BERTScore F1 across sections and ablation configurations.

	full	no_tmpl	no_guide	no_instr	min_rag	zero
S.2.2	.758	.720	.740	.745	.688	.612
S.4.1	.768	.755	.742	.760	.695	.590
S.7.1	.752	.748	.738	.745	.733	.585
S.7.3	.705	.698	.690	.700	.645	.508
P.3.3	.741	.732	.725	.730	.675	.572
P.5.1	.728	.720	.710	.722	.685	.588
Avg	.742	.734	.724	.734	.680	.580

A.3 Prompt Design Details

The system prompt is ~350 lines of Jinja2 with conditional blocks for document type (CTD dossier, briefing book, investigator’s brochure), product type (generic, biosimilar, novel), and section type. Key constraints include: 32 banned AI-tell phrases (e.g., “Furthermore,” “It is important to note”), 6 banned imprecise terms (e.g., “room temperature,” “in-house specification”), mandatory [PLACEHOLDER: description] format for missing values, and explicit instructions to never reproduce full pharmacopoeial monographs. For manufacturing sections, the prompt enforces process step numbering (USP 1–*n*, DSP 1–*m*) and requires per-step tables listing unit operation, equipment, process parameters with PAR/MOR, CPP/nCPP classifications, and in-process controls.