MEDRDSYN: KNOWLEDGE GRAPH–BASED RARE DISEASE DATA SYNTHESIS FOR SUPERVISED FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

To address the scarcity of real clinical records for rare diseases, we propose MedRDSyn, a knowledge-graph–guided framework for rare disease clinical note synthesis. MedRDSyn curates rare disease notes from MIMIC-IV-Note as clinical record templates and defines entity types and relation types required by clinical records. Through LLM-assisted incremental triple extraction from PubMed, MedRDSyn integrates Orphadata, HPO, and DrugBank, while using UMLS for entity normalization to construct a rare disease knowledge graph. The graph explicitly models associations among key medical entities, including genes, variants, proteins, diseases, and phenotypes, and contains 48,379 entities and 654,814 triples. Based on this knowledge graph, MedRDSyn performs LLM consistency verification with source-confidence fusion to construct a weighted graph, estimates node importance via weighted PageRank, and retrieves informative subgraphs through temperature-controlled seed sampling and beam-search expansion. After quality filtering, the framework generates 14,035 high-quality synthetic rare disease clinical cases and chain-of-thought data. Supervised fine-tuning on Qwen3-0.6B, Qwen3-4B, and Qwen3-8B using the synthesized data yields consistent improvements on PubMedQA, MedQA, ReDisQA, and RareBench, demonstrating the effectiveness of graph-guided rare disease data synthesis.

1 INTRODUCTION

Large language models (LLMs) have shown great potential for medical text understanding and diagnostic assistance, yet their performance remains uneven across the disease spectrum (Singhal et al., 2023; 2025). Real-world clinical data follow a pronounced long-tail distribution, causing models to be predominantly exposed to common disease cases during training and thus perform poorly in rare disease scenarios (Zhou et al., 2021; Chen et al., 2024; Wang et al., 2024). Rare diseases are characterized by low prevalence, high phenotypic heterogeneity, and complex diagnostic trajectories; meanwhile, access to real clinical records is restricted by privacy and compliance requirements (Schieppati et al., 2008; Harman et al., 2012). As a result, real rare disease cases are difficult to obtain, and supervised fine-tuning (SFT) for rare disease reasoning is fundamentally constrained by the lack of scalable and shareable clinical training text.

Researchers have explored a variety of mitigation strategies, including knowledge-guided few-shot learning, federated learning, and retrieval-augmented generation (Alsentzer et al., 2025; Zelin et al., 2024; Pati et al., 2022). Although these methods can enhance evidence support or cross-silo learning, they do not provide rare disease clinical corpora that are sufficiently large and diverse for SFT. Clinical progress notes and discharge summaries are indispensable modalities for rare disease reasoning because they capture longitudinal disease progression, positive/negative findings, differential diagnoses, and treatment rationales, which cannot be faithfully reconstructed from structured EHR fields alone (Johnson et al., 2023). At the same time, synthesizing clinical notes is challenging, as it must simultaneously satisfy: (i) realistic document structure, (ii) patient-level factual consistency, (iii) medically correct semantics, and (iv) privacy safety. These challenges are especially severe for rare diseases, where heterogeneous phenotypes and mechanistic reasoning chains make generated content more prone to cross-section contradictions and medically implausible statements, thereby undermining the effectiveness and stability of downstream SFT (Hayeems et al., 2022).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this need, we propose **MedRDSyn**, a knowledge-graph–guided rare disease clinical note synthesis framework designed to construct high-quality instruction-style data for **SFT**. MedRDSyn extracts a standardized clinical record template from de-identified discharge summaries in MIMIC-IV-Note and uses it to define a format aligned with clinical documentation. We then integrate Orphadata (Rath et al., 2012), PubMed[1], HPO[2], and DrugBank (Knox et al., 2024), together with UMLS-based fusion and normalization (Bodenreider, 2004), to construct a rare disease knowledge graph. The resulting graph contains 48,379 entities and 654,814 triples, and explicitly models pathogenic-mechanism links by incorporating UniProt information (The UniProt Consortium, 2025), constructing the gene $\rightarrow$ variant $\rightarrow$ protein $\rightarrow$ phenotype $\rightarrow$ disease reasoning chain. During synthesis, MedRDSyn estimates triple reliability through LLM consistency verification with source-confidence fusion, computes node importance using weighted PageRank, and retrieves informative subgraphs via temperature-controlled seed sampling and beam-search expansion. After rule-based validation and LLM-based quality scoring, MedRDSyn generates 14,035 high-quality synthetic rare disease clinical cases and chain-of-thought (CoT) data. After SFT on Qwen3-0.6B and Qwen3-4B, the models achieve consistent improvements on PubMedQA, MedQA, ReDisQA, and RareBench.

In summary, our contributions are as follows.
• We construct a rare disease knowledge graph with explicit pathogenic-mechanism chains, and perform large-scale multi-source fusion and medical standardization validation.
• We propose MedRDSyn, a knowledge-graph–guided synthesis framework for rare disease SFT data construction, enabling structural controllability and improving medical validity.
• We construct 14,035 high-quality synthetic samples generated by MedRDSyn. We use the generated dataset to conduct SFT on Qwen3-0.6B, Qwen3-4B, and Qwen3-8B, achieving consistent gains on PubMedQA, MedQA, ReDisQA, and RareBench.

2 Related Work

Clinical Note Generation Synthetic clinical note generation has been studied through multiple lines of work. Corpus-trained neural generators learn from de-identified narrative text and directly generate synthetic notes (Melamud & Shivade, 2019; Amin-Nejad et al., 2020). Privacy-oriented methods attempt to reduce memorization and leakage risks through privacy-preserving training or entity anonymization (Baumel et al., 2024; Shimizu et al., 2025). Code-conditioned prompting methods use ICD codes or structured code sets to synthesize discharge summaries, improving long-tail concept coverage and instruction following (Falis et al., 2024; Vo et al., 2025). Task-oriented clinical text synthesis methods use task labels, medical topics, and writing styles as generation conditions to construct synthetic labeled samples (Xu et al., 2024). Knowledge-augmented, template-and-edit, and grounded synthesis methods inject medical knowledge, retrieve existing records, or condition generation on external references to improve clinical relevance and factuality (Kumichev et al., 2024; Wang et al., 2025; Grazhdanski et al., 2025). However, for rare disease clinical note synthesis, merely ensuring textual form and local factual plausibility is insufficient. Domain knowledge is also needed to constrain both medical content and record structure.

Knowledge Graph-Based Data Synthesis Rare disease knowledge is fragmented across heterogeneous resources and standardization systems. Orphanet emphasizes pathological classification (Rath et al., 2012), whereas GARD focuses on patient-oriented information integration (Zhu et al., 2020). Even after entity alignment, cross-source reasoning may still introduce conflicts. Knowledge graphs (KGs) provide a structured way to integrate biomedical knowledge (Ehrlinger & Wöß, 2016). Early KGs relied on expert construction (Smith et al., 2007), while recent automated pipelines construct disease-specific or rare disease KGs from structured resources and task-specific evidence graphs (Karki et al., 2025; Zhu et al., 2020; Qin et al., 2025). Nevertheless, existing automated rare disease KGs still face challenges such as data sparsity, relation noise, and constraint-based validation (Hogan et al., 2021). Beyond KG construction, KGs have also been used to synthesize or supervise data for LLM training. Prior work uses KGs to generate annotated text, detect knowledge gaps, synthesize multi-hop QA, construct graph-walk corpora for continued pre-training, translate natural language into graph queries, or supervise clinically valid reasoning paths (Vuth et al., 2024; Chen et al., 2025; Ma et al., 2025; Zhong et al., 2025; Kumichev et al., 2024; Mu et al., 2025). Re-

[1] https://pmc.ncbi.nlm.nih.gov/
[2] https://hpo.jax.org/

cent rare disease synthesis methods further exploit KG structure. MedResearcher-R1 extracts long reasoning chains around rare entities (Yu et al., 2025), and RareSyn uses KGs and common-disease EHR templates to generate rare disease records (Wang et al., 2025). However, these KG-based synthesis methods still provide limited exploration of pathogenic mechanisms. Pathogenic chains have not yet been fully used to constrain the generation process, which may limit biological plausibility in rare disease scenarios.

3 METHODOLOGY

3.1 CLINICAL NOTE TEMPLATE CONSTRUCTION

As illustrated in Figure 1, MedRDSyn uses de-identified discharge summaries from MIMIC-IV-Note (version 2.2) as the template source (Johnson et al., 2023). To identify rare disease-related notes, we use the mapping from MIMIC-IV-EXT-RARE (Chen et al., 2026), which links ICD-9/10 codes in MIMIC-IV-Note to rare disease identifiers from Online Mendelian Inheritance in Man[3] and Orphanet[4]. This yields 5,586 rare disease patients and 24,550 clinical notes.

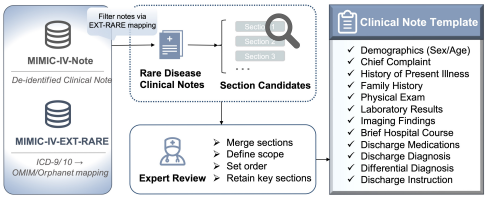

Figure 1: Construction of clinical note template. We construct expert-validated templates from rare disease mapped MIMIC-IV-Note data to guide structured clinical note synthesis.

We then construct a standardized clinical note template to align synthesized cases with real discharge-summary structure. Medical experts review frequent section headings, merge synonymous headings, and define the content scope and ordering of each section. The resulting template retains key components of the diagnostic and treatment workflow, including Sex, Age, Chief Complaint, History of Present Illness, Family History, Physical Exam, Pertinent Results, Radiology Report, Brief Hospital Course, Discharge Medications, and Discharge Diagnosis.

3.2 MULTI-SOURCE TRIPLE GENERATION AND CAUSAL LOGIC CHAIN CONSTRUCTION

MedRDSyn constructs a rare disease KG from PubMed literature, HPO/UniProt structured resources, and DrugBank. The schema is aligned with clinical-record elements, and extracted facts are organized around pathogenic chains connecting genes, variants, proteins, phenotypes, and diseases. Figure 2 summarizes the overall workflow, while the following subsections detail ontology design, literature extraction, pathogenic-chain modeling, and drug knowledge extension.

Clinical record aligned Unlike general biomedical ontologies, MedRDSyn defines a clinical-record–aligned schema to support downstream patient-note synthesis. The schema contains 21 entity types covering genetic factors, clinical manifestations, diagnostic tests, and therapeutic interventions, and 28 relation types connecting molecular mechanisms, diagnoses, and treatments. This design allows generated records to follow realistic clinical structure while remaining grounded in medical principles. Detailed statistics for all entity and relation types are provided in Appendix B.1 and Appendix B.2.

LLM-based literature extraction MedRDSyn first derives a rare disease seed set from Orphadata (Zhu et al., 2020) by extracting entries classified as "Disorder", together with synonyms and external identifiers such as ICD-10 and UMLS, yielding 7,337 rare diseases. To address the highly uneven distribution of rare disease literature, we use stratified PubMed sampling. For each disease, up to three articles are incrementally retrieved, with priority given to Case Reports. For diseases with limited literature, retrieval is expanded to related diseases within the same Orphanet group to improve coverage without over-sampling high-frequency diseases.

[3] https://omim.org/
[4] https://www.orpha.net/

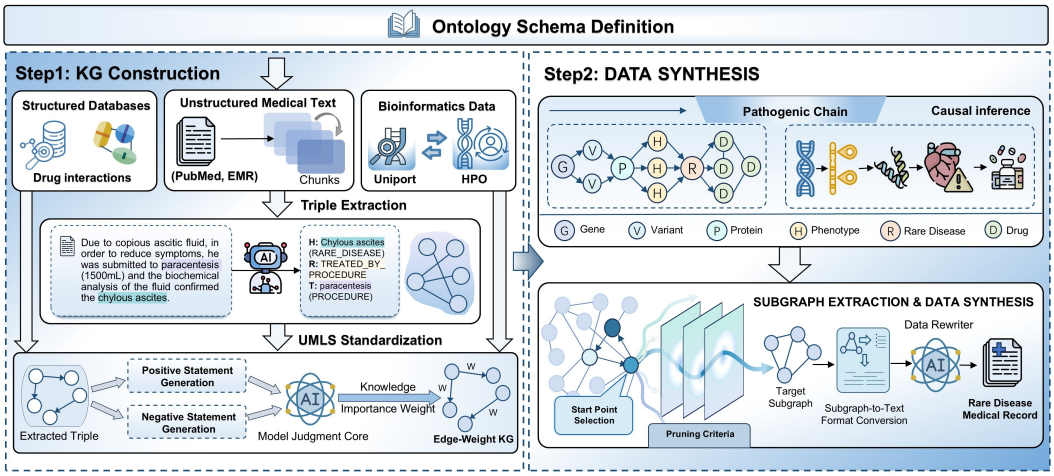

Figure 2: Overview of MedRDSyn. The framework constructs a rare disease KG from multi-source triples and pathogenic chains, retrieves reliability-weighted subgraphs, and uses them as structured evidence to synthesize rare disease medical records.

For unstructured articles, MedRDSyn cleans the text and applies overlap-aware segmentation to maintain local semantic coherence. It then uses GPT-5.2-2025-12-11 for ontology-guided entity recognition and relation extraction, while retaining the original text as evidence. The extraction also records relation attributes needed for downstream synthesis, such as diagnostic findings, pathogenic mechanisms, and drug-interaction endpoints. Extracted entities are normalized and linked to UMLS Concept Unique Identifiers (CUIs).

Pathogenic chain modeling MedRDSyn decomposes rare disease pathogenesis into four consecutive causal steps: Gene $\rightarrow$ Variant $\rightarrow$ Protein $\rightarrow$ Phenotype $\rightarrow$ Disease, where genes give rise to variants, variants affect proteins, protein dysfunction causes phenotypes, and phenotypes manifest in diseases. To instantiate this chain, MedRDSyn integrates gene–phenotype, gene–disease, and phenotype–gene associations from HPO, and filters rare disease-related entries through cross-matching with Orphanet. UniProt is further used to complete the critical Variant $\rightarrow$ Protein link by extracting structured information from "Natural variant" entries. The resulting pathogenic chains provide mechanistic constraints for subgraph retrieval and synthetic record generation, helping maintain consistency between genotype-level evidence and phenotype-level clinical descriptions.

DrugBank extension The DrugBank (Knox et al., 2024) database is parsed to add drug treatment-related triples, with focus on extracting drug indications (Drug-Treats-Disease) and drug-drug interactions (Drug-Interacts-Drug), providing contraindication and synergy recommendations for medication simulation in complex cases.

Because rare disease drug information is sparse and disease names vary across sources, MedRDSyn identifies drug associations through progressive matching, proceeding from exact matching, synonym matching, standardized matching, and variant-pattern matching to fuzzy matching and partial matching. The system also classifies drug–drug interactions along pharmacodynamic and toxicological dimensions, as detailed in Appendix B.3. These treatment and interaction triples extend the pathogenic chain toward medication recommendations, supporting medication simulation in complex synthetic cases.

3.3 Hierarchical Multi-source Knowledge Fusion Framework

Multi-source KG construction introduces entity heterogeneity, relation noise, and potential conflicts across rare disease resources. To address these challenges, we introduce the Hierarchical Multi-source Knowledge Fusion Framework (HMKF), which performs progressive entity alignment,

cross-knowledge-base semantic linking, schema-constrained relation validation, and evidence-aware aggregation.

3.3.1 PROBLEM FORMALIZATION

Given raw extracted relations $\mathcal{R}_{raw} = \{(h_i, r_i, t_i, \epsilon_i)\}_{i=1}^{n}$, where h_i, r_i, t_i, and ϵ_i denote the head entity, relation type, tail entity, and evidence text, respectively, and external knowledge bases $\mathcal{K} = \{\mathcal{K}_{orpha}, \mathcal{K}_{hpo}, \mathcal{K}_{umls}\}$, HMKF aims to construct a fused knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ by balancing knowledge coverage and schema-level semantic consistency:

$$Q(\mathcal{G}) = \sum_{e \in \mathcal{E}} \mathbb{I}[\Phi_{\text{align}}(e, \mathcal{K}) \neq \emptyset] + \lambda \cdot \sum_{(h,r,t) \in \mathcal{R}} \mathbb{I}[\sigma(h, r, t) \in \mathcal{S}]. \tag{1}$$

Here, $\mathcal{E}$ and $\mathcal{R}$ are the entity and relation sets of the fused graph, Φ_{align} returns the aligned identifier or canonical entity from the external knowledge bases, $\mathbb{I}[\cdot]$ is the indicator function, $\sigma(\cdot)$ maps a triple to its type signature, $\mathcal{S}$ denotes valid schema constraints, and λ balances coverage and consistency.

3.3.2 PROGRESSIVE ENTITY ALIGNMENT

To handle diverse surface forms of medical terminology, HMKF performs progressive entity alignment through a sequence of matching functions ordered by decreasing confidence. Let $\{\phi_i\}_{i=1}^{6}$ denote the six matching layers, including exact matching, synonym expansion, morphological normalization, edit distance, morpheme matching, and semantic embedding. For an entity mention e and knowledge base collection $\mathcal{K}$, HMKF adopts the first non-empty match as the aligned entity:

$$\Phi_{\text{align}}(e, \mathcal{K}) = \phi_{i^*}(e, \mathcal{K}), \quad i^* = \min\{i : \phi_i(e, \mathcal{K}) \neq \emptyset\}. \tag{2}$$

Here, Φ_{align} denotes the final alignment function, ϕ_i denotes the i-th matching layer, and i^* is the first layer that returns a valid match. This design progressively relaxes matching criteria to improve recall, while preserving precision by returning high-confidence deterministic matches first.

3.3.3 KNOWLEDGE FUSION AND QUALITY CONTROL

HMKF first constructs cross-ontology links by querying Orphanet cross-references, HPO cross-references, and UMLS mappings in priority order, assigning standard CUI identifiers whenever available. Entities mapped to the same CUI are merged, and the most frequent surface form is selected as the canonical name to support cross-source deduplication and normalization.

HMKF then performs schema-constrained relation validation using domain-specific type signatures $\mathcal{S} = \{(T_h, R, T_t)\}$, where T_h and T_t denote the head and tail entity types and R denotes the relation type. Relations are filtered if their entity types violate schema constraints, if they form self-loops, or if either endpoint is empty.

Finally, HMKF aggregates duplicated relations while preserving evidence text and source provenance. Relations sharing the same uniqueness key, defined by *(type_h, head, relation, attributes, type_t, tail)*, are merged. We compute a heuristic source-support score:

$$conf(r) = 1 - \prod_{v \in \mathcal{V}(r)} (1 - w_{src(v)}) \tag{3}$$

where r denotes a relation, $\mathcal{V}(r)$ is the set of evidence sources supporting r, and $w_{src(v)}$ is the reliability weight assigned to source v. This formulation assigns higher confidence to relations supported by multiple sources.

The resulting fused KG contains 48,379 entities and 654,814 triples, providing knowledge support for rare disease case synthesis. The detailed pseudocode of HMKF is provided in Appendix A.1.

3.4 KNOWLEDGE GRAPH-INFORMED DATA SYNTHESIS APPROACH

Given the triple set obtained in the previous stage, we construct a weighted knowledge graph, perform node sampling and subgraph retrieval, and synthesize rare disease cases and QA+CoT samples conditioned on each subgraph. This pipeline prioritizes reliable mechanistic facts and provides structured evidence for controllable generation.

Graph weighting For each validated and fused triple (h, r, t), MedRDSyn forms a positive statement s^+ from the original relation and a contrastive statement s^- by negating r while keeping h and t fixed. We treat s^+ as the candidate factual statement and use s^- as a consistency check. GPT-OSS (OpenAI, 2025) then labels each statement as correct or incorrect with a confidence score.

Let $p_{\text{true}}(s) = c$ if statement s is judged correct with normalized confidence $c \in [0, 1]$, and $p_{\text{true}}(s) = 1 - c$ if judged incorrect. The LLM consistency score is:

$$s_{\text{llm}}(h, r, t) = \frac{1}{2}\left[p_{\text{true}}(s^+) + 1 - p_{\text{true}}(s^-)\right]. \tag{4}$$

The final edge weight fuses LLM consistency with upstream source confidence:

$$w(h, r, t) = \lambda s_{\text{llm}}(h, r, t) + (1 - \lambda) s_{\text{src}}(h, r, t), \tag{5}$$

where s_{src} is defined in Eq. 3. Each triple becomes a directed edge $h \to t$ with weight $w(h, r, t)$, yielding $G = (V, E, W)$.

Node importance To prioritize informative and reliable graph regions, MedRDSyn computes node importance using Weighted PageRank (Xing & Ghorbani, 2004). Let $PR(v)$ be the PageRank score of node v and $N = |V|$. The weighted transition probability from node u to node v is:

$$P(u \to v) = \frac{w(u, v)}{\sum_{v' \in \Gamma^+(u)} w(u, v')}, \tag{6}$$

where $\Gamma^+(u)$ denotes the out-neighbors of u. Weighted PageRank iterates:

$$PR(v) = \frac{1 - d}{N} + d \sum_{u \in \Gamma^-(v)} PR(u) \cdot P(u \to v), \tag{7}$$

where d is the damping factor and $\Gamma^-(v)$ denotes the in-neighbors of v. We rescale PageRank scores by N for interpretability and downstream sampling stability.

Subgraph retrieval MedRDSyn samples seed nodes using temperature-controlled Top-k sampling over the rescaled PageRank scores. A lower temperature concentrates sampling on high-importance nodes, while a higher temperature increases long-tail coverage. Starting from each seed, MedRDSyn retrieves an evidence subgraph via beam search, where candidate edges are ranked by the edge reliability weight $w(h, r, t)$. The retrieved subgraph is linearized into a context that preserves entity types, relation semantics, and evidence text. Detailed settings are provided in Appendix A.2.

Data synthesis Given each subgraph context, the LLM synthesizes a standardized rare disease case, allowing limited fictional completion (e.g., demographics or non-critical narrative details) while requiring grounded mechanistic and clinical facts to remain consistent. We retain cases through a two-stage filtering pipeline: rule-based screening based on case length and the number of grounded nodes/relations, followed by LLM-based quality scoring. For each retained case, the LLM further generates a question–answer pair and a CoT rationale grounded in the case and subgraph evidence. These samples form **MedRDSyn-14k**, containing 14,035 synthetic outputs. An example is provided in Appendix F.1.

4 Experiments

4.1 Experimental Setup

We evaluate MedRDSyn-14k by using it to perform supervised fine-tuning (SFT) for small and medium-sized LLMs in the Qwen family, and testing the resulting models on four medical question-answering benchmarks. To mitigate the variance introduced by stochastic decoding, we adopt **mean@4**: for each question, we run inference four times and report the average accuracy.

Datasets Our proposed training corpus, **MedRDSyn-14k**, contains 14,035 high-quality synthetic samples produced by MedRDSyn, including rare disease case-style QA and chain-of-thought (CoT) rationales.

Table 1: Performance comparison of MedRDSyn and baseline methods across three Qwen3 models. MedQA, PubMedQA, and ReDisQA are evaluated with mean@4, while RareBench is evaluated with R@1. The best results are shown in bold, and the second-best results are underlined.

Model	Method	MedQA (%)	PubMedQA (%)	ReDisQA (%)	RareBench (%)	Avg. Δ (%)
Qwen3-0.6B	Base	<u>29.69</u>	<u>62.38</u>	35.59	15.29	–
	ClinGen	28.32 (-1.37)	61.08 (-1.30)	<u>36.40</u> (+0.81)	<u>16.21</u> (+0.92)	-0.24
	MedSyn	28.59 (-1.10)	61.20 (-1.18)	35.81 (+0.22)	15.12 (-0.17)	-0.56
	RareSyn	28.69 (-1.00)	60.98 (-1.40)	36.38 (+0.79)	15.62 (+0.33)	-0.32
	MedRDSyn	**29.81** (+0.12)	**62.62** (+0.24)	**36.53** (+0.94)	**16.37** (+1.08)	**+0.60**
Qwen3-4B	Base	67.05	68.97	<u>71.08</u>	27.40	–
	ClinGen	66.83 (-0.22)	<u>69.20</u> (+0.23)	70.88 (-0.20)	27.15 (-0.25)	-0.11
	MedSyn	67.11 (+0.06)	69.07 (+0.10)	70.90 (-0.18)	<u>27.90</u> (+0.50)	+0.12
	RareSyn	**67.52** (+0.47)	69.00 (+0.03)	70.59 (-0.49)	25.90 (-1.50)	-0.37
	MedRDSyn	<u>67.36</u> (+0.31)	**69.55** (+0.58)	**71.34** (+0.26)	**28.15** (+0.75)	**+0.48**
Qwen3-8B	Base	75.14	<u>70.50</u>	77.32	30.74	–
	ClinGen	75.73 (+0.59)	70.47 (-0.03)	77.24 (-0.08)	31.16 (+0.42)	+0.23
	MedSyn	75.24 (+0.10)	70.15 (-0.35)	76.93 (-0.39)	29.24 (-1.50)	-0.54
	RareSyn	<u>75.88</u> (+0.74)	70.40 (-0.10)	<u>77.35</u> (+0.03)	<u>31.24</u> (+0.50)	+0.29
	MedRDSyn	**76.36** (+1.22)	**70.70** (+0.20)	**77.85** (+0.53)	**32.08** (+1.34)	**+0.82**

Baselines We compare MedRDSyn against three clinical text synthesis baselines. (1) **ClinGen** (Xu et al., 2024) uses clinical knowledge extraction and context-informed LLM prompting to generate synthetic clinical training data. (2) **MedSyn** (Kumichev et al., 2024) samples prior medical information from a Medical Knowledge Graph (MKG) and uses it to prompt LLMs for synthetic clinical note generation. (3) **RareSyn** (Wang et al., 2025) synthesizes rare disease EHRs through similar-record retrieval and KG-based medical information substitution. For a fair comparison, all methods are evaluated under the same backbone models and SFT protocol.

Models We conduct SFT on Qwen3 base models (Yang et al., 2025), including **Qwen3-0.6B**, **Qwen3-4B**, and **Qwen3-8B**. For each model, we report results for the base model without SFT, SFT on ClinGen, MedSyn, RareSyn, and MedRDSyn-14k.

Benchmarks We evaluate on: (1) **MedQA** (Jin et al., 2021), a large-scale multiple-choice medical exam QA dataset; (2) **PubMedQA** (Jin et al., 2019), a biomedical research QA benchmark derived from PubMed abstracts; (3) **ReDisQA** (Wang et al., 2024), a rare disease question-answering benchmark focusing on long-tail clinical knowledge and reasoning; (4) **RareBench** (Chen et al., 2024), an open-ended QA benchmark designed to evaluate LLMs in rare disease diagnosis, for which we use the publicly available Task 4 subset.

Experimental settings We implement SFT with TRL (von Werra et al., 2020) and HuggingFace Accelerate (Gugger et al., 2022). To reduce GPU memory consumption and speed up training, we use LoRA (Hu et al., 2022) and keep hyperparameters unified across all runs. We use vLLM (Kwon et al., 2023) as the inference engine. For all models and benchmarks, we use the same decoding configuration and prompt template, with a 32k context window and temperature 0.7.

4.2 Main Results

Table 1 shows that MedRDSyn achieves the best or second-best performance across all benchmarks and yields the strongest average improvement over the baseline methods. Its gains are relatively stable on both general medical QA benchmarks and rare disease benchmarks. Compared with the base models, MedRDSyn consistently improves performance across all benchmarks and model sizes, suggesting that the generated rare disease data provides effective supervision signals.

From the perspective of model scaling, MedRDSyn shows increasingly stable advantages over other synthetic-data baselines as the backbone grows from Qwen3-0.6B to Qwen3-8B. For smaller models, general clinical text or task-label-driven data can also bring certain gains, because small models have limited medical knowledge and instruction-following ability, and additional clinical-text supervision can already improve performance on some tasks. However, as model size increases,

models acquire stronger common medical knowledge and general clinical language modeling ability, making the marginal benefit of ordinary clinical-text synthesis data relatively weaker. At this stage, further improvements are more constrained by the coverage of long-tail rare disease knowledge, mechanistic consistency, and the availability of diagnostic evidence. By constructing a rare disease-oriented KG and retrieving pathogenic-chain subgraphs, MedRDSyn explicitly supplements long-tail disease supervision signals and therefore continues to provide gains for larger models.

4.3 Pathogenic Evidence Usage

Beyond final-answer accuracy, we further analyze whether MedRDSyn enriches the types of medical evidence used during model reasoning. On MedQA and ReDisQA, we sample 100 questions that are correctly answered by all three Qwen3-8B variants: the base model, the model fine-tuned on ClinGen, and the model fine-tuned on MedRDSyn. We then use GPT-5.5 to judge whether each answer contains reasoning relations among Gene, Protein, Phenotype, Disease, Test, Diagnosis, and Treatment. For example, a Phenotype–Disease relation indicates that a phenotype is used as clinical-manifestation evidence for a disease.

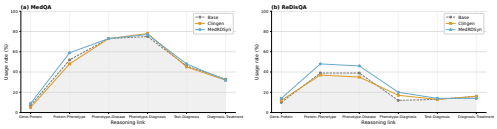

Figure 3: Pathogenic evidence usage in generated rationales.

Figure 3 reports the proportion of each relation type explicitly used in model reasoning. Even among questions answered correctly by all three models, the overall relation usage rate on ReDisQA is lower than that on MedQA. This suggests that models less frequently organize complete medical evidence chains in rare disease QA, reflecting the explanatory difficulty caused by evidence sparsity and mechanistic complexity in rare disease reasoning. MedRDSyn shows stronger usage of relations among Gene, Protein, and Disease, with a more pronounced effect on ReDisQA. These relations correspond to the upstream pathogenic-chain knowledge introduced by the MedRDSyn KG, suggesting that MedRDSyn not only provides rare disease knowledge but also injects a mechanism-oriented reasoning pattern through KG relation structures.

4.4 Validity Assessment

We asked one clinically trained medical expert and GPT-5.5 to evaluate the quality of the synthesized clinical notes. Specifically, we randomly sampled 30 generated notes from MedRDSyn and each of the three baseline methods, yielding 120 notes in total, with an example shown in Appendix F.2. Each note was evaluated along five dimensions: Structure, Consistency, Correctness, Privacy, and Specificity, with each dimension scored on a 0–2 scale. The detailed scoring rubric is provided in Appendix C.4. Under a blinded setting where the source method of each note was not disclosed, both evaluators independently scored each note across the five dimensions and provided brief justifications.

Table 2: Validity assessment of MedRDSyn.

Method	Evaluator	Structure	Consistency	Correctness	Privacy	Specificity
ClinGen	Human Expert	0.00	0.60	1.03	2.00	0.00
	GPT-5.5	1.03	1.83	1.23	1.97	0.27
MedSyn	Human Expert	0.33	0.20	0.87	1.93	0.37
	GPT-5.5	1.07	1.80	0.83	2.00	0.43
RareSyn	Human Expert	2.00	1.67	0.77	1.60	0.07
	GPT-5.5	1.83	1.80	1.37	1.77	0.13
MedRDSyn	Human Expert	1.00	1.97	1.37	2.00	1.80
	GPT-5.5	2.00	1.17	1.33	2.00	1.87

Table 2 shows that although MedRDSyn does not achieve the highest score on every individual metric, it demonstrates stable overall quality from both human and LLM evaluators. For Structure,

the human evaluator did not assign a full score to MedRDSyn because the synthesized notes omit past medical history. This information cannot be reliably inferred from the knowledge graph, and we therefore do not include it in the template to preserve medical accuracy and internal consistency. For rare-disease specificity, MedRDSyn shows a clear advantage, validating the effectiveness of knowledge graph (KG)-guided rare disease clinical note generation.

4.5 Ablation Study

To evaluate the contribution of different components in MedRDSyn, we construct three ablation variants. **w/o KG** removes the knowledge graph context during case generation and only retains rare disease seeds. **w/o Chain** removes graph evidence related to the pathogenic chain. **w/o Retrieval** replaces graph-aware subgraph retrieval with random subgraph sampling.

Table 3: Ablation study of MedRDSyn across three Qwen3 models. The best results are shown in bold, and the second-best results are underlined.

Model	Variant	MedQA (%)	PubMedQA (%)	ReDisQA (%)	RareBench (%)
Qwen3-0.6B	**MedRDSyn**	**29.81**	**62.62**	36.53	**16.37**
	w/o KG	<u>29.30</u>	<u>62.50</u>	**36.73**	<u>15.96</u>
	w/o Chain	28.61	61.55	<u>36.65</u>	<u>15.96</u>
	w/o Retrieval	28.79	61.92	36.64	15.37
Qwen3-4B	**MedRDSyn**	**67.36**	**69.55**	**71.34**	**28.15**
	w/o KG	66.93	68.60	70.70	<u>27.65</u>
	w/o Chain	67.28	<u>69.33</u>	<u>71.14</u>	27.57
	w/o Retrieval	**67.36**	69.30	70.55	27.49
Qwen3-8B	**MedRDSyn**	**76.36**	**70.70**	**77.85**	**32.08**
	w/o KG	75.39	70.37	77.32	<u>30.99</u>
	w/o Chain	75.16	70.23	77.56	29.49
	w/o Retrieval	<u>75.88</u>	<u>70.50</u>	<u>77.70</u>	29.49

Table 3 shows that the full MedRDSyn generally outperforms its ablated variants across model sizes, indicating that the three mechanisms jointly contribute to downstream performance. The effect of removing KG context exhibits a model-scale-dependent trend. When KG evidence is removed, the case-completion model can still rely on its parametric medical knowledge to generate clinically plausible cases, providing basic clinical supervision signals that can still benefit the capacity-limited Qwen3-0.6B backbone. However, as the fine-tuned backbone becomes larger, the marginal benefit of generic case completion gradually decreases. In this regime, KG evidence becomes more important because it introduces rare disease-specific relations and coherent pathogenic-chain structures beyond template-like clinical completion, leading to clearer performance gains for Qwen3-4B and Qwen3-8B.

5 Conclusion

We propose **MedRDSyn**, a knowledge-graph–guided framework for synthesizing rare disease clinical records under real-world data scarcity and privacy constraints. MedRDSyn constructs a rare disease knowledge graph through multi-source fusion and UMLS-based entity normalization, and explicitly organizes pathogenic evidence along the gene $\rightarrow$ variant $\rightarrow$ protein $\rightarrow$ phenotype $\rightarrow$ disease chain. By retrieving reliability-weighted subgraphs as structured evidence, MedRDSyn generates 14,035 high-quality synthetic rare disease clinical cases and QA+CoT samples.

Experiments on Qwen3-0.6B, Qwen3-4B, and Qwen3-8B show that SFT with MedRDSyn improves model performance on both general and rare disease tasks, while enhancing pathogenic reasoning ability. These results validate the value of KG-guided synthesis for long-tail medical scenarios.

This study still has limitations. Although we observe consistent performance improvements, the absolute gains of MedRDSyn on standard QA benchmarks remain modest, suggesting the need for evaluation protocols that better characterize rare disease diagnostic reasoning. In future work, we plan to further scale both the knowledge graph and synthesis volume, and explore using the generated clinical records for continued pre-training before task-specific supervised fine-tuning.

AI USE STATEMENT

In this work, we used generative AI tools to construct synthetic rare disease clinical training data, including synthetic clinical cases and QA+CoT samples. Clinical case generation was conditioned on retrieved knowledge-graph subgraphs and standardized clinical-note templates, with outputs further filtered through rule-based validation and LLM-based quality scoring. QA+CoT generation was conditioned on the retained cases and their associated subgraph evidence. We also used generative AI tools for writing polish, including grammar, clarity, and phrasing. We take responsibility for the final content of this work.

REFERENCES

Emily Alsentzer, Michelle M Li, Shilpa N Kobren, Ayush Noori, Undiagnosed Diseases Network, Isaac S Kohane, and Marinka Zitnik. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):380, 2025.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4699–4708, 2020.

Tal Baumel, Andre Manoel, Daniel Jones, Shize Su, Huseyin A. Inan, Aaron Bornstein, and Robert Sim. Controllable synthetic clinical note generation with privacy guarantees. *arXiv preprint arXiv:2409.07809*, 2024.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4850–4861, 2024.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 101–109, 2026. doi: 10.1609/AAAI.V40I1.36969. URL https://doi.org/10.1609/aaai.v40i1.36969.

Zihong Chen, Wanli Jiang, Jinzhe Li, Zhonghang Yuan, Huanjun Kong, Wanli Ouyang, and Nanqing Dong. Graphgen: Enhancing supervised fine-tuning for LLMs with knowledge-driven synthetic data generation, 2025. URL https://arxiv.org/abs/2505.20416.

Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL https://ceur-ws.org/Vol-1695/paper4.pdf.

Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. Can gpt-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, 31(10):2284–2293, 2024.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170:104906, 2025.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

Laurinda B Harman, Cathy A Flite, and Kesa Bond. Electronic health records: privacy, confidentiality, and security. *Virtual Mentor*, 14(9):712–719, 2012.

Robin Z Hayeems, Christine Michaels-Igbokwe, Viji Venkataramanan, Taila Hartley, Meryl Acker, Meredith Gillespie, Wendy J Ungar, Roberto Mendoza-Londona, Francois P Bernier, Kym M Boycott, et al. The complexity of diagnosing rare disease: An organizing framework for outcomes research and health economics based on real-world evidence. *Genetics in Medicine*, 24(3):694–702, 2022.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys*, 54(4), 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, January 2023. doi: 10.13026/1n74-ne17. URL https://doi.org/10.13026/1n74-ne17. Version 2.2.

Reagon Karki, Yojana Gadiya, Andrea Zaliani, Bishab Pokharel, Negin Sadat Babaiha, Marek Ostaszewski, Martin Hofmann-Apitius, and Philip Gribbon. Kgg: a fully automated workflow for creating disease-specific knowledge graphs. *Bioinformatics*, 41(7):btaf383, 2025.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. Medsyn: LLM-based synthetic medical text generation framework. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track (ECML PKDD 2024)*, volume 14950 of *Lecture Notes in Computer Science*, pp. 215–230. Springer, 2024. First online: 22 Aug 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Shengjie Ma, Xuhui Jiang, Chengjin Xu, Cehao Yang, Liyu Zhang, and Jian Guo. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pre-training of large language models, 2025. URL https://arxiv.org/abs/2505.00979.

Oren Melamud and Chaitanya Shivade. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 35–45, June 2019. doi: 10.18653/v1/W19-1905. URL https://aclanthology.org/W19-1905/.

Linjie Mu, Yannian Gu, Zhongzhen Huang, Yakun Zhu, Shaoting Zhang, and Xiaofan Zhang. Medceg: Reinforcing verifiable medical reasoning with critical evidence graph, 2025. URL https://arxiv.org/abs/2512.13510.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022.

Lang Qin, Zijian Gan, Xu Cao, Pengcheng Jiang, Yankai Jiang, Jiawei Han, Kaishun Wu, and Jintai Chen. Rareagent: Self-evolving reasoning for drug repurposing in rare diseases, 2025.

Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808, 2012.

Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041, 2008.

Seiji Shimizu, Ibrahim Baroud, Lisa Raithel, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Recordtwin: Towards creating safe synthetic clinical corpora. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14714–14726, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

Truong Vo, Weiyi Wu, and Kaize Ding. Synthetic clinical notes for rare icd codes: A data-centric framework for long-tail medical coding, 2025.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL `https://github.com/huggingface/trl`.

Nakanyseth Vuth, Gilles Sérasset, and Didier Schwab. KGAST: From knowledge graphs to annotated synthetic texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 43–55, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.kallm-1.5. URL `https://aclanthology.org/2024.kallm-1.5`.

Guanchu Wang, Junhao Ran, Ruixiang Tang, Chia-Yuan Chang, Yu-Neng Chuang, Zirui Liu, Vladimir Braverman, Zhandong Liu, and Xia Hu. Assessing and enhancing large language models in rare disease question-answering. *arXiv preprint arXiv:2408.08422*, 2024.

Huimin Wang, Yutian Zhao, Yefeng Zheng, and Xian Wu. RareSyn: Health record synthesis for rare disease diagnosis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12311–12327, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.620. URL `https://aclanthology.org/2025.emnlp-main.620/`.

Wenpu Xing and Ali A. Ghorbani. Weighted PageRank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pp. 305–314. IEEE Computer Society, 2004. doi: 10.1109/DNSR.2004.1344743.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15496–15523, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.916. URL https://aclanthology.org/2024.findings-acl.916/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, et al. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework, 2025.

Charlotte Zelin, Wendy K Chung, Mederic Jeanne, Gongbo Zhang, and Chunhua Weng. Rare disease diagnosis using knowledge guided retrieval augmentation for chatgpt. *Journal of Biomedical Informatics*, 157:104702, 2024.

Zijie Zhong, Linqing Zhong, Zhaoze Sun, Qingyun Jin, Zengchang Qin, and Xiaofan Zhang. Synthet2c: Generating synthetic data for fine-tuning large language models on the text2cypher task. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pp. 672–692, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.46/.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5948–5957, 2021.

Qian Zhu, Dac-Trung Nguyen, Ivan Grishagin, Noel Southall, Eric Sid, and Anne Pariser. An integrative knowledge graph for rare diseases, derived from the genetic and rare diseases information center (gard). *Journal of Biomedical Semantics*, 11(1):13, 2020.

A Algorithmic Details

A.1 Hierarchical Multi-source Knowledge Fusion Algorithm

HMKF takes raw extracted relations, external knowledge bases, schema constraints, and matching thresholds as input, and produces a fused rare disease knowledge graph as output. The algorithm operationalizes the four key stages described in the main text: progressive entity alignment, cross-ontology semantic linking, schema-constrained relation validation, and evidence-aware aggregation.

Algorithm 1 HMKF: Hierarchical Multi-source Knowledge Fusion

Require: R_{raw}: Raw relations $(h, \tau_h, r, t, \tau_t)$; $\mathcal{K}$: Knowledge bases $(\mathcal{V}, \mathcal{E}_{syn}, \mathbf{V})$; $\mathcal{S}$: Type constraints; Θ, W: Thresholds and weights
Ensure: $\mathcal{G} = (\mathcal{E}, \mathcal{R})$: Fused knowledge graph

/* Phase 1: Progressive Entity Alignment */
$\mathcal{E}_{raw} \leftarrow$ EXTRACTENTITIES(R_{raw})
for each $e \in \mathcal{E}_{raw}$ **do**
  $\tilde{e} \leftarrow$ NORMALIZE(e)
  $\mathcal{M}[e] \leftarrow$ HIERARCHICALMATCH$(\tilde{e}, \mathcal{K}, \Theta)$
  // Match order: exact → synonym → morph → edit → semantic
end for

/* Phase 2: Cross-ontology Semantic Linking */
for each $(e^*, conf) \in \mathcal{M}$ where $conf > 0$ **do**
  $e^*.ref \leftarrow$ QUERYCUI$(e^*, \mathcal{K})$
end for
$\mathcal{E}_{dedup} \leftarrow$ MERGEBYCUI$(\mathcal{M})$

/* Phase 3: Schema-constrained Validation */
$\mathcal{R}_{valid} \leftarrow \emptyset$
for each $r = (h, \tau_h, p, t, \tau_t) \in R_{raw}$ **do**
  $(h', t') \leftarrow (\mathcal{M}[h], \mathcal{M}[t])$
  if $h' \neq \emptyset \wedge t' \neq \emptyset \wedge h' \neq t' \wedge (\tau_h, p, \tau_t) \in \mathcal{S}$ **then**
    $\mathcal{R}_{valid} \leftarrow \mathcal{R}_{valid} \cup \{(h', p, t', ev(r))\}$
  end if
end for

/* Phase 4: Evidence-aware Aggregation */
for each $(key, rels) \in$ GROUPBY$(\mathcal{R}_{valid})$ **do**
  $conf \leftarrow 1 - \prod_{r \in rels}(1 - w_{src(r)})$
  $\mathcal{R}_{final} \leftarrow \mathcal{R}_{final} \cup \{(key, \bigcup ev, conf)\}$
end for
return $\mathcal{G} = (\mathcal{E}_{dedup}, \mathcal{R}_{final})$

A.2 GRAPH WEIGHTING AND SUBGRAPH RETRIEVAL

We set $\lambda = 0.4$ in Eq. 5 for all experiments. Node importance is computed with weighted PageRank using $w(h, r, t)$ as the edge weight and a damping factor of 0.85. Since the KG is represented as a multigraph, parallel edges between the same ordered node pair are collapsed by keeping the maximum edge reliability before PageRank. The resulting PageRank score is rescaled by the number of nodes:

$$\widetilde{PR}(v) = N \cdot PR(v), \tag{8}$$

so that the average node weight is approximately one.

Seed nodes are sampled from nodes with $\widetilde{PR}(v) > 0$ using temperature-controlled sampling:

$$P(v) = \frac{\widetilde{PR}(v)^{1/T}}{\sum_{u \in \mathcal{C}} \widetilde{PR}(u)^{1/T}}, \tag{9}$$

where $\mathcal{C}$ is the candidate seed set. In the main generation run, we sample 20,000 seeds from all entity types without replacement, with no minimum node-weight threshold beyond $\widetilde{PR}(v) > 0$ and temperature $T = 0.8$.

Starting from each sampled seed, MedRDSyn retrieves a bounded evidence subgraph by iterative beam expansion. At each hop, all candidate edges incident to the current frontier are ranked by edge reliability $w(h, r, t)$, and the top $B = 5$ edges are added to the subgraph. Newly reached nodes form the next frontier. We use three hops and expand in both incoming and outgoing directions. Node PageRank is used for seed sampling and is not mixed into the beam-edge ranking score.

The retrieved subgraph is linearized into a text context with a node block and a relationship block. The node block includes entity name, entity type, node identifier, and rescaled PageRank weight. The relationship block includes head entity, relation type, tail entity, edge reliability, source confidence, and retained evidence text when available. We include node lists and edge evidence during generation and cap the linearized context at 200 edges.

B Details of Dataset

This appendix provides detailed statistics for the 21 entity types and 28 relation types used in MedRDSyn's rare disease knowledge graph, together with the endpoint taxonomy used for DrugBank interaction classification.

B.1 Entity Type Statistics

Table 4 reports the number of unique entities for each entity type in the rare disease KG.

Table 4: Statistics of 21 entity types in the rare disease knowledge graph.

Entity Type	**Count**
PHENOTYPE	13,741
DRUG	6,467
RARE_DISEASE	5,005
VARIANT	4,284
DOSAGE	4,191
PATIENT_AGE	3,386
PATIENT_FAMILY_HISTORY	1,772
OTHER_DISEASE	1,727
GENE	1,713
SYMPTOM	1,695
PROTEIN	763
LAB_TEST	715
PROCEDURE	677
SIGN	660
SIDE_EFFECT	510
MICROBIOLOGY	320
IMAGING_TEST	227
GENETIC_TEST	140
PATIENT_ALLERGEN	131
PATHOLOGY_TEST	128
PATIENT_SEX	127

B.2 Relation Type Statistics

Table 5 reports the number of edges for each relation type in the rare disease KG.

Table 5: Statistics of 28 relation types in the rare disease knowledge graph.

Relation Type	Edges
HARMFUL_COMBINATION	251,996
MANIFESTS_IN	217,015
NEUTRAL_COMBINATION	86,528
DESIRABLE_COMBINATION	21,905
DYSFUNCTION_CAUSES	17,993
PATIENT_HAS_DISEASE	12,729
HAS_SYMPTOM	8,075
DRUG_DOSAGE	7,378
TREATED_BY_DRUG	2,956
COMPLICATED_BY	2,859
FAMILY_HISTORY_OF	2,564
DIFFERENTIAL_DIAGNOSIS	2,492
AFFECTS_PROTEIN	2,460
HAS_VARIANT	2,386
CAUSED_BY_GENE	2,210
COMORBID_WITH	2,160
HAS_SIGN	1,835
HAS_SIDE_EFFECT	1,517
TREATED_BY_PROCEDURE	1,475
DIAGNOSED_BY_LAB	1,423
DIAGNOSED_BY_IMAGING	1,099
DIAGNOSED_BY_GENETIC_TEST	799
GENETIC_INHERITANCE_PATTERN	770
VARIANT_INHERITANCE_FROM	659
DIAGNOSED_BY_PATHOLOGY	604
REQUIRES_MICROBIOLOGY	566
PATIENT_ALLERGIC_TO	208
ENCODES_PROTEIN	153

B.3 Drug Interaction Endpoint Taxonomy

Table 6 summarizes the pharmacodynamic and toxicological endpoints used to categorize DrugBank drug–drug interactions.

Table 6: Pharmacodynamic and toxicological endpoint taxonomy for DrugBank interaction classification.

Interaction Endpoint Type	Mechanism Type	Effect Description
Pharmacodynamic Endpoint	EFFICACY_ENHANCED	Therapeutic effect enhanced
	EFFICACY_UNCHANGED	No significant change in therapeutic effect
	EFFICACY_REDUCED	Therapeutic effect reduced
Toxicological Endpoint	TOXICITY_INCREASED	Adverse reactions/toxicity increased
	TOXICITY_UNCHANGED	No significant change in side effects
	TOXICITY_REDUCED	Adverse reactions/toxicity reduced

B.4 Benchmark Source Overlap Check

Since PubMedQA is built from PubMed abstracts, while MedRDSyn uses PubMed-derived literature for knowledge-graph construction, we perform a source-level overlap check to rule out direct benchmark-source contamination. Specifically, we compare the PubMed IDs used for MedRDSyn KG extraction with the `pubid` field provided by PubMedQA. We find no overlapping PubMed records between the two sets.

This result indicates that the PubMed articles used to construct the MedRDSyn knowledge graph do not directly overlap with the PubMed source articles underlying PubMedQA. Therefore, the Pub-

MedQA evaluation is not affected by direct source-document leakage from the PubMed literature used in MedRDSyn.

C Prompts and Evaluation Rubrics

This appendix lists the prompts and evaluation rubric used in our pipeline, including (i) statement verification with confidence, (ii) MIMIC-style rare disease case synthesis conditioned on subgraph evidence, (iii) case-level quality scoring, (iv) evaluation rubrics for validity assessment, and (v) QA and chain-of-thought (CoT) generation for SFT.

C.1 Statement Judgment with Confidence

Listing 1: Prompt for judging the medical validity of a statement and outputting confidence.

```
You are a clinical medicine expert. Evaluate the medical validity of the
    following statement:

[Statement]
{statement}

[Requirements]
1. Judgment:
  - "correct": Consistent with current medical consensus and
      authoritative guidelines (UpToDate, NEJM, Lancet)
  - "incorrect": Contradicts medical consensus or lacks evidence support
  - Choose the most conservative judgment when uncertain

2. Confidence:
  - 0-100% (integer)
  - Evidence level reference:
      - 90-100\%: Multiple RCTs + guideline recommendations
      - 70-89\%: Cohort studies + expert consensus
      - 50-69\%: Case series/observational studies
      - <50\%: Case reports/theoretical speculation

3. Format Requirements:
  - Return ONLY valid JSON format
  - NO additional text
  - Field names MUST use double quotes
  - Ensure JSON validity

5. Example Output Format:
{
  "judgment": "correct",
  "confidence": 95,
}

/no_think
```

C.2 Case Synthesis (MIMIC-style Clinical Note)

Listing 2: Prompt for synthesizing a MIMIC-style inpatient clinical note grounded in a retrieved subgraph.

```
You are a rare disease diagnostic specialist. Generate a realistic
    inpatient clinical note in the MIMIC-style structure.

You MUST ground all key medical relationships (gene/variant/protein/
    disease/phenotype, major comorbidities, key causal links) in the
    provided subgraph.
However, if the subgraph does NOT provide certain attributes (e.g., exact
     age, labs, imaging, medications, timeline details), you SHOULD make
```

```
    reasonable clinical inferences and fabricate plausible details that
    do NOT contradict the subgraph.

[Subgraph Knowledge]
{subgraph_context}

[Target Note Structure]
Write the note in English, using the following section headers exactly (
    keep the order):

Sex:
Age:
Chief Complaint:
History of Present Illness:
Family History:
Physical Exam:
Laboratory Results:
Imaging Findings:
Brief Hospital Course:
Discharge Medications:
Discharge Diagnosis:
Differential Diagnosis:
Discharge Instructions (include prognosis and genetic counseling when
    applicable):

[Generation Requirements]
- Each section must be non-empty. If missing in the subgraph, infer and
    fabricate plausible content.
- The case should be coherent, clinically realistic, and internally
    consistent.
- Chief Complaint must be a brief, objective statement of the primary
    symptoms/problems (NOT a first-person patient quote and NOT written
    as a patient self-report).
- Include a clear timeline in History of Present Illness.
- Physical Exam should include general + relevant system/specialty
    findings.
- Laboratory Results and Imaging Findings should be plausible (avoid
    impossible values); you may omit exact units if unsure.
- Differential Diagnosis: list at least 2 alternatives with brief
    justification.
- Discharge Diagnosis: provide the final diagnosis consistent with the
    subgraph (rare disease focus).
- Brief Hospital Course: describe key inpatient evaluation and treatments
    .
- Discharge Medications: list a plausible regimen aligned with the
    diagnosis and inpatient course.
- Discharge Instruction: include follow-up plan, prognosis, and genetic
    counseling/testing recommendations if relevant.
- Do NOT introduce new key medical relationships that contradict the
    subgraph.
- Output plain text only (no JSON, no XML, no markdown code fences).

think:medium
```

C.3 Case Review and Quality Scoring

Listing 3: Prompt for scoring the overall quality of a synthesized clinical case (1–5).

```
You are a strict reviewer for synthetic clinical cases.

Task:
Given the case text, assign ONE overall quality score from 1 to 5,
    representing reject->accept:
```

```
1 = Strong Reject (nonsensical, unsafe, very low quality, or clearly
    fabricated)
2 = Reject (major issues: incoherent, unrealistic, missing key clinical
    parts)
3 = Borderline (usable but noticeable issues)
4 = Accept (good quality with minor issues)
5 = Strong Accept (high quality, coherent, realistic, clinically complete
    )

Output format (STRICT):
Output ONLY one XML tag, no other text:
<score>1</score>

Rules:
- score must be an integer in {1,2,3,4,5}.
- Do NOT output any explanation.

Case Text:
{case_text}
```

C.4 Evaluation Rubric for Validity Assessment

Table 7 presents the five-dimensional rubric used for validity assessment. Each dimension is scored on a 0–2 scale.

Table 7: Evaluation rubric for validity assessment.

Dimension	Score	Criterion
Structure	0	The structure clearly does not resemble a medical record, with multiple key sections missing and disorganized formatting.
	1	The note generally follows a clinical-note structure, but some sections are missing, ordered unnaturally, or contain sparse content.
	2	The note has a complete structure with clearly organized sections, consistent with the style of real clinical notes.
Consistency	0	There are severe contradictions among the medical history, examinations, diagnoses, and treatments.
	1	The note is generally understandable but contains local inconsistencies, such as conflicts in the timeline, numerical values, diagnostic rationale, or medication use.
	2	Patient information is internally consistent, and the clinical course, examinations, diagnoses, and treatments are logically coherent.
Correctness	0	The note contains serious medical errors, including incorrect handling of negation, uncertainty, or differential diagnoses, which may mislead clinical decision-making.
	1	The medical content is largely correct but includes imprecise statements or insufficient handling of negation, uncertainty, or differential diagnoses.
	2	The medical facts are accurate, and negation, uncertainty, differential diagnoses, and clinical reasoning are expressed appropriately.
Privacy	0	The note contains explicit personal identifiers, such as real names, national ID numbers, medical record numbers, phone numbers, detailed addresses, exact dates of birth, or specific hospital/physician names.
	1	The note does not contain direct identifiers but includes multiple potentially re-identifiable details, such as exact dates, specific locations, or institution names.
	2	The note contains no identifiable personal information, shows no obvious copying from real records, and presents low privacy risk.
Specificity	0	The case describes a common disease.
	1	The case involves a rare disease, but the key phenotypes, genetic information, diagnostic pathway, or differential diagnosis are insufficiently specific.
	2	The case clearly centers on the target rare disease and includes discriminative phenotypes, laboratory/genetic evidence, and a reasonable diagnostic pathway.

C.5 Multiple-Choice QA Generation

Listing 4: Prompt for generating a USMLE-style multiple-choice question from a case note.

```
You are an exam question writer for USMLE-style multiple-choice questions
    .

Given a clinical case note, create ONE high-quality question that can be
    answered using information implied by the case.
The question should test clinical reasoning (diagnosis, next step,
    treatment, most likely finding, complication, etc.).

Requirements:
- Create exactly 5 answer options labeled A, B, C, D, E.
- Exactly ONE option must be correct.
- Options must be mutually exclusive and clinically plausible.
- Do NOT include "All of the above" or "None of the above".
- Avoid trivial copy-paste. The question stem may summarize key parts of
    the case.
- Return STRICT XML only. No extra text.
```

```
Return format (STRICT XML):
<question>...</question>
<option_A>...</option_A>
<option_B>...</option_B>
<option_C>...</option_C>
<option_D>...</option_D>
<option_E>...</option_E>
<answer_idx>A|B|C|D|E</answer_idx>

Clinical case note:
<case>
{case_text}
</case>
```

C.6 Chain-of-Thought (CoT) Generation

Listing 5: Prompt for generating a CoT rationale given the case, question, and the correct option.

```
You are a medical reasoning assistant.

Given a clinical case note and a multiple-choice question with options,
    explain why the correct answer is correct.
You MUST NOT change the correct answer.

Output format (STRICT XML ONLY):
<cot>...</cot>

Guidelines:
- Use the clinical case to justify the choice.
- Briefly rule out the main alternatives.
- Keep it clinically grounded.
- {cot_style}

Clinical case note:
<case>
{case_text}
</case>

Question:
{question}

Options:
A. {A}
B. {B}
C. {C}
D. {D}
E. {E}

Correct answer: {answer_idx}. {answer_text}
```

D SFT Data and Training Configuration

All SFT experiments use the same training configuration. For SFT data, we use 13,709 samples from ClinGen, 13,191 from MedSyn, 13,395 from RareSyn, and 14,035 from MedRDSyn-14k. For baseline data-synthesis steps that require an LLM generator, we use GPT-OSS consistently. The Russian prompt template in MedSyn is translated into English, while each method's original data sources, retrieval strategy, template design, and filtering rules are otherwise preserved. For data preprocessing, the maximum sequence length is set to 8192, packing is disabled, and no padding multiple is specified. For optimization, we use random seed 42, train for 2 epochs, set the per-device training batch size to 1, use 1 gradient accumulation step, set the learning rate to 5×10^{-5}, use a

warmup ratio of 0.03, and set weight decay to 0.0. All SFT training is performed on 8 NVIDIA H20 GPUs. For precision and runtime settings, bfloat16 training is enabled, fp16 is disabled, gradient checkpointing is disabled, the torch data type is set to bfloat16, and no specific attention implementation is specified. LoRA is configured with rank 16, alpha 32, dropout 0.05, no bias terms, and target modules q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj.

E Additional Experiments

E.1 Seed Robustness

To assess seed sensitivity on the rare-disease setting, we repeat Qwen3-8B SFT for RareSyn and MedRDSyn using three random seeds: 42, 43, and 44. We focus on ReDisQA and RareBench because they directly evaluate rare-disease reasoning.

Table 8: Three-seed robustness check on rare-disease benchmarks using Qwen3-8B. Results are reported as mean (standard deviation) across three SFT seeds.

Method	**ReDisQA (%)**	**RareBench (%)**
RareSyn	77.29 (± 0.17)	30.88 (± 1.17)
MedRDSyn (ours)	**77.64** (± 0.21)	**32.03** (± 0.25)

Table 8 reports mean and standard deviation across three seeds. The three-seed results show different stability patterns across the two rare-disease benchmarks. On ReDisQA, which follows a question-answering format, both RareSyn and MedRDSyn exhibit low variance across seeds. This indicates that SFT with either rare disease synthetic dataset is relatively stable for rare-disease QA, while MedRDSyn achieves a higher mean accuracy.

In contrast, on RareBench, which evaluates open-ended rare-disease diagnosis, variance is more pronounced for RareSyn. RareSyn obtains a standard deviation of 1.17, whereas MedRDSyn reduces it to 0.25 while also achieving a higher mean score. This result is consistent with the hypothesis that graph-guided synthesis and pathogenic-chain evidence can provide more stable supervision for open-ended rare-disease reasoning, where models need to organize heterogeneous evidence rather than select from a constrained answer space.

E.2 Biomedical Content Diversity

We further analyze whether different synthesis methods produce diverse biomedical supervision, rather than merely generating fluent QA+CoT samples. For each method, we randomly select 500 QA+CoT instances and apply the same LLM-based claim extraction and verification pipeline to the final training data. We use GPT-5.5 as the LLM evaluator for both claim extraction and verification. The evaluator decomposes each instance into clinically meaningful medical claims, judges each claim as *correct*, *incorrect*, *unverifiable*, or *too vague*, and extracts explicit biomedical entity mentions from each claim. We exclude non-medical text, formatting instructions, isolated demographics, raw laboratory values, and duplicated or near-duplicated statements. To avoid favoring MedRDSyn, the evaluator only observes the final QA+CoT samples and does not use KG relations, KG entity types, UMLS concepts, or method-specific annotations.

We compute two groups of metrics over claims judged as *correct*. First, correct-claim density measures the amount of valid medical supervision per unit text:

$$\text{ClaimDensity} = \frac{\#\text{CorrectClaims}}{\#\text{Characters}} \times 1000. \tag{10}$$

Second, entity diversity ratio measures the fraction of unique normalized entity mentions among all extracted entity mentions:

$$\text{EntityDiversityRatio} = \frac{|\mathcal{E}_{unique}|}{|\mathcal{E}_{mention}|}, \tag{11}$$

where $\mathcal{E}_{mention}$ denotes all extracted entity mentions from correct claims and $\mathcal{E}_{unique}$ denotes the set of normalized unique entity mentions. We use lightweight string normalization, including lowercasing, whitespace normalization, punctuation trimming, and hyphen/slash normalization. Since

surface-form variation may affect entity counting, we also plot entity accumulation curves by randomly ordering samples and tracking the cumulative number of unique entities as the character budget increases. All curves are truncated to the common character budget across methods to ensure a fair comparison.

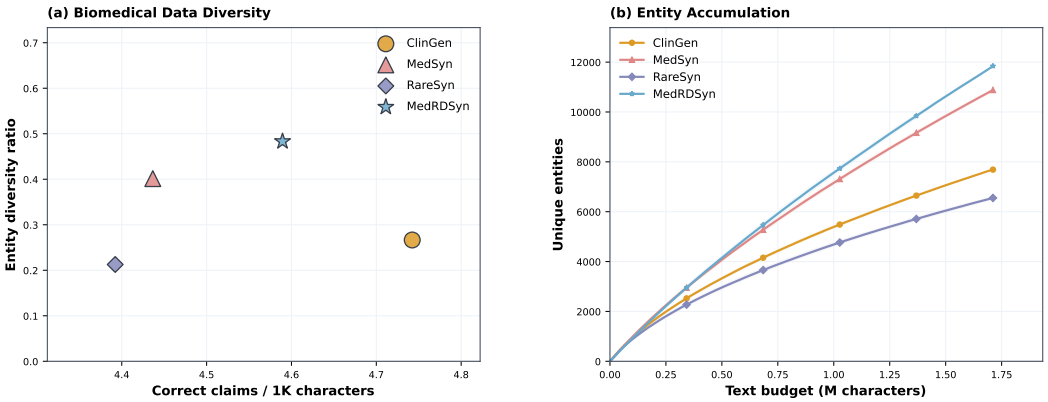

Figure 4: Biomedical content diversity analysis on 500 QA+CoT instances per method. (a) Correct-claim density versus entity diversity ratio, where entity diversity is the fraction of unique normalized entity mentions among all entity mentions in correct claims. (b) Entity accumulation curves under increasing character budgets, truncated to the common character budget across methods.

As shown in Fig. 4, MedRDSyn achieves a correct-claim density comparable to the baselines, indicating that its diversity gain does not come from producing fewer or lower-density claims. However, it obtains the highest entity diversity ratio, suggesting that its valid claims cover a less repetitive set of biomedical concepts. The entity accumulation curves further show that MedRDSyn continues to introduce more unique entities under the same character budget. These results indicate that KG-grounded synthesis improves the breadth and diversity of biomedical supervision, providing a data-level explanation for its stronger performance on rare-disease benchmarks.

F Synthetic Case Studies

F.1 MedRDSyn Synthetic Case Example

Listing 6: MedRDSyn synthetic case example.

```
Sex: Male
Age: 28 years

Chief Complaint: Progressive abdominal distension with early satiety and
    fatigue for 3 months.

History of Present Illness:
The patient is a 28-year-old man who presented to the emergency
    department with a 3-month history of increasing abdominal girth,
    early satiety, and generalized fatigue. He reports mild intermittent
    abdominal discomfort but denies nausea, vomiting, or acute pain. Two
    weeks ago he noted a yellowish tint to his sclerae and occasional
    pruritus. He has no history of pancreatitis, gallstones, or recent
    alcohol binge. Over the past month he has been evaluated outpatient
    for dyslipidemia; a lipid panel performed 4 weeks prior showed
    triglycerides 820mg/dL, HDL-C 5mg/dL, LDL-C 45mg/dL, total
    cholesterol 120mg/dL. He was started on a low-fat diet but symptoms
    progressed. Review of systems is otherwise negative.

Family History:
```

```
Both parents are of Middle-Eastern descent and are first cousins. Mother
    has hypercholesterolemia diagnosed in her 40s; father has a history
    of "high triglycerides." A 22-year-old sister was recently found to
    have low HDL-C on routine screening. No known history of liver
    disease, pancreatitis, or hematologic disorders in the family.

Physical Exam:
- General: Alert, oriented, mildly icteric, in no acute distress.
- Vital signs: BP 122/78 mmHg, HR 84 bpm, RR 16/min, Temp 36.8 degC, SpO2
     98% on room air.
- HEENT: Conjunctival icterus, no oral lesions.
- Neck: No lymphadenopathy.
- Cardiovascular: Regular rate and rhythm, no murmurs.
- Respiratory: Clear to auscultation bilaterally.
- Abdomen: Distended, liver edge palpable 3cm below the right costal
    margin, spleen palpable 4cm below the left costal margin; no
    tenderness, no ascites.
- Extremities: No edema; presence of small, soft, yellowish plaques on
    the extensor surfaces of the elbows (xanthomas).
- Neurologic: Grossly intact.

Laboratory Results:
- CBC: WBC 7.2x10^9/L, Hgb 14.1g/dL, Platelets 250x10^9/L (within normal
    limits).
- CMP: AST 48U/L (high), ALT 52U/L (high), ALP 110U/L (normal), Total
    bilirubin 2.1mg/dL (high), Albumin 4.0g/dL.
- Lipid panel: Triglycerides 820mg/dL, HDL-C 5mg/dL, LDL-C 45mg/dL, Total
     cholesterol 120mg/dL.
- Serum amylase/lipase: Within normal limits.
- Fasting glucose 92mg/dL, HbA1c 5.4% (normal).
- Serum vitamin E level: Low-normal.
- Genetic testing (clinical exome panel): Heterozygous pathogenic
    variants in ABCA1 - c.*** (ABCA1_VAR_062483) and c.*** (
    ABCA1_VAR_009147), consistent with loss-of-function.

Imaging Findings:
- Abdominal ultrasound: Hepatomegaly with liver span 18cm, homogeneous
    echotexture; splenomegaly with spleen length 16cm; no focal lesions,
    no biliary dilation.
- Echocardiogram: Normal left ventricular size and function, no valvular
    abnormalities.

Brief Hospital Course:
The patient was admitted for evaluation of severe hypertriglyceridemia
    and hepatosplenomegaly. Initial work-up excluded secondary causes (
    alcohol, uncontrolled diabetes, hypothyroidism). Given the markedly
    low HDL-C and family history, a hereditary lipid disorder was
    suspected. He was started on fenofibrate 145mg daily and omega-3
    fatty acid ethyl esters 4g daily, with strict low-fat, low-simple-
    sugar diet. Vitamin E 400IU daily was added to address potential
    oxidative stress. Serial triglyceride levels decreased to 460mg/dL
    over 5 days, and liver enzymes trended downwards. Genetic testing
    returned confirming pathogenic ABCA1 loss-of-function variants (
    ABCA1_VAR_062483 and ABCA1_VAR_009147), establishing a diagnosis of
    familial HDL deficiency (Tangier disease). He tolerated therapy
    without adverse effects and was educated on lifestyle measures. He
    was discharged home in stable condition.

Discharge Medications:
- Fenofibrate 145mg PO daily
- Omega-3 fatty acid ethyl esters 4g PO daily (divided BID)
- Vitamin E 400IU PO daily
- Low-fat, low-simple-sugar diet (dietitian-approved meal plan)

Discharge Diagnosis:
```

```
1. Familial HDL deficiency (Tangier disease) due to ABCA1 loss-of-
    function variants (ABCA1_VAR_062483, ABCA1_VAR_009147)
2. Hypertriglyceridemia secondary to ABCA1 dysfunction
3. Hepatosplenomegaly associated with ABCA1 deficiency

Differential Diagnosis:
- Familial combined hyperlipidemia - can cause elevated triglycerides but
     typically presents with normal or high HDL-C, unlike this case.
- Secondary hypertriglyceridemia (e.g., uncontrolled diabetes, alcohol
    excess) - ruled out by normal glucose, HbA1c, and negative alcohol
    history.
- Storage disorders (e.g., Gaucher disease, Niemann-Pick disease) - can
    cause hepatosplenomegaly but are usually accompanied by cytopenias
    and characteristic bone findings, which are absent.

Discharge Instructions (include prognosis and genetic counseling when
    applicable):
- Follow up with the lipid clinic in 2 weeks for repeat lipid panel and
    medication titration.
- Continue fenofibrate, omega-3, and vitamin E as prescribed; do not
    discontinue without consulting your physician.
- Adhere strictly to the low-fat, low-simple-sugar diet; avoid alcohol
    and fructose-rich beverages.
- Monitor for signs of pancreatitis (severe abdominal pain, vomiting) and
     seek immediate care if they occur.
- Schedule a hepatology appointment in 4 weeks to reassess liver size and
     function.
- Genetic counseling is strongly recommended for the patient and first-
    degree relatives. Testing of siblings and future offspring for the
    identified ABCA1 variants should be offered. Early detection allows
    prompt lipid management and surveillance for organ involvement.
- Prognosis: With lifelong lipid-lowering therapy and dietary measures,
    the risk of pancreatitis and progressive organ enlargement can be
    markedly reduced, though some degree of hepatosplenomegaly may
    persist. Regular monitoring is essential.
```

F.2 Baseline Synthetic Case Examples

Listing 7: ClinGen Case

```
Sex: Male
Age: 58
Chief complaint: Severe abdominal pain and altered mental status
Case summary:
The patient is a 58-year-old male admitted to the ICU for septic shock
    secondary to perforated sigmoid diverticulitis, presenting with
    excruciating abdominal pain, hypotension (BP 84/48 mmHg), tachycardia
     (HR 122 bpm), tachypnea (RR 28/min), and progressive altered mental
    status (GCS 9). Initial labs show a high-anion-gap metabolic acidosis
     (pH 7.12, HCO3- 12 mEq/L, PaCO2 22 mmHg, anion gap 24 mEq/L),
    lactate 6.8 mmol/L, creatinine 2.1 mg/dL, and negative serum ketones.
     Immediate management included 30 mL/kg crystalloid bolus,
    norepinephrine infusion titrated to MAP >65 mmHg, broad-spectrum
    antibiotics (piperacillin-tazobactam + vancomycin), and urgent
    surgical consult for emergent sigmoid colectomy. Due to refractory
    acidosis and acute kidney injury, continuous renal replacement
    therapy was initiated. Over the subsequent 48 hours, serial arterial
    blood gases demonstrated gradual closure of the anion gap and
    normalization of pH, lactate declined to 2.1 mmol/L, and vasopressor
    requirements decreased. The patient remains intubated, sedated, and
    is being monitored in the ICU pending postoperative recovery.
```

Listing 8: MedSyn Case

```
Chief complaint: New-onset purpura and worsening joint pain.
History of present illness:
The patient is a 58-year-old man with a known myeloproliferative disorder
     who was transferred from a skilled nursing facility after his
    caregivers noted sudden appearance of non-blanching purpuric patches
    on his lower extremities and increasing arthralgia involving the
    knees and wrists over the past 48 hours. He reports severe, diffuse
    joint aching that limits ambulation and is not relieved by over-the-
    counter analgesics. Earlier this morning he became acutely confused
    and was noted to have intermittent shaking movements; emergency
    personnel observed a petechial rash extending to the trunk and
    documented a platelet count of 18 x 10^9/L. He was given plasma
    exchange en route and remains under close observation for thrombotic
    thrombocytopenic purpura secondary to his underlying
    myeloproliferative disease.
```

Listing 9: RareSyn Case

```
Age: 25
Gender: M
Chief complaint: Pain and non-healing ulcer on the left foot that worsens
     after walking a short distance.
History of present illness:
The patient is a 25-year-old male who presents with a three-month history
     of progressive left calf pain and a small ulcer on the dorsum of his
     left foot. The pain began insidiously as a mild ache after walking
    more than 15 minutes and has gradually intensified; he now
    experiences cramping after walking only 50 meters (intermittent
    claudication). The ulcer appeared two weeks ago after he
    inadvertently stepped on a small nail; despite cleaning and over-the-
    counter dressings, the wound has not healed and now has a pale base
    with minimal granulation tissue. He reports that the foot is
    constantly cool to the touch and that he feels a tingling sensation
    at rest. He denies fever, chills, or recent trauma beyond the nail
    injury. He has a 10-year history of cigarette smoking (1 pack per day
    ) and reports occasional binge drinking. He was diagnosed with type 1
     diabetes mellitus at age 12 and is on basal-bolus insulin therapy;
    his last HbA1c (3 months ago) was 8.2 %. He has a family history of
    premature coronary artery disease. He has never undergone vascular
    surgery or endovascular intervention. Physical activity is limited by
     pain, and he has not been able to attend his university classes
    regularly. He denies any calf swelling, varicose veins, or prior deep
    -vein thrombosis. He has tried over-the-counter analgesics with
    minimal relief.

Past history:
Type 1 diabetes mellitus; Hypertension; Hyperlipidemia; Cigarette smoking
     (10 pack-years); Obesity (BMI 31 kg/m2)

Family history:
Father: myocardial infarction at age 45; Mother: peripheral arterial
    disease

Physical exam:
General: Alert, oriented, in mild distress due to leg pain; Vital signs:
    BP 138/84 mmHg, HR 78 bpm, RR 16/min, Temp 36.8 degC, SpO2 98% RA;
    Cardiovascular: Regular rate and rhythm, no murmurs; Extremities:
    Left foot cool, hair loss over the dorsum, delayed capillary refill
    (>3 sec); dorsalis pedis and posterior tibial pulses absent on the
    left, faint on the right; 2-cm shallow ulcer with pale base on left
    dorsal foot; no edema, no erythema; Neurologic: Decreased light touch
     sensation distal to the left ankle; Skin: Mild peripheral cyanosis
    of left lower leg
```

```
Lab results:
test_name=Hemoglobin A1c, value=8.2, unit=%, reference_range=4.0-5.6 |
    test_name=Serum creatinine, value=1.0, unit=mg/dL, reference_range
    =0.6-1.3 | test_name=C-reactive protein (CRP), value=4.8, unit=mg/L,
    reference_range=<3.0 | test_name=Ankle-brachial index (ABI) - left,
    value=0.58, reference_range=0.90-1.30 | test_name=Ankle-brachial
    index (ABI) - right, value=0.92, reference_range=0.90-1.30

Imaging:
type=Duplex Ultrasound, finding=Severe atherosclerotic plaque with >80%
    stenosis of the left superficial femoral artery; monophasic flow in
    the left popliteal and tibial arteries; absent flow in dorsalis pedis
     artery. | type=X-ray (foot, AP & lateral), finding=No osseous
    abnormality; soft-tissue swelling over the ulcerated area.

Treatment plan:
Aspirin 81 mg orally daily; Atorvastatin 40 mg orally nightly; Intensify
    insulin regimen to target HbA1c <7%; Smoking cessation counseling and
     nicotine-replacement therapy; Supervised walking exercise program, 3
     times per week; Referral to vascular surgery for possible femoral-
    popliteal bypass or endovascular angioplasty; Local wound care:
    debridement, moist dressings, and off-loading with a specialized shoe
    ; Compression therapy not indicated (arterial disease); Follow-up in
    2 weeks for wound assessment and in 4 weeks for repeat ABI
```