QA Analysis in Medical and Legal Domains: A Survey of Data Augmentation in Low-Resource Settings

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), but their success remains largely confined to high-resource, general-purpose domains. In contrast, applying LLMs to low-resource domains poses significant challenges due to limited training data, domain drift, and strict terminology constraints. This survey provides an overview of the current landscape in domain-specific, low-resource QA with LLMs. We begin by analyzing the coverage and representativeness of specialized-domain QA datasets against large-scale reference datasets what we refer to as *ParentQA*. Building on this analysis, we survey data-centric strategies to enhance input diversity, including data augmentation techniques. We further discuss evaluation metrics for specialized tasks and consider ethical concerns. By mapping current methodologies and outlining open research questions, this survey aims to guide future efforts in adapting LLMs for robust and responsible use in resource-constrained, domain-specific environments.

1 Introduction

Over the years, large language models (LLMs) (OpenAI et al., 2023; Gemini et al., 2024; DeepSeek-AI et al., 2025) have demonstrated remarkable performance across a variety of natural language processing (NLP) tasks. However, these advances remain largely confined to domains for which massive training corpora are available (Kaplan et al., 2020). In contrast, low-resource datasets (Ravichander et al., 2019; Möller et al., 2020) pose significant challenges for LLMs due to data scarcity and underrepresentation. The lack of sufficient quantity and quality of data leads to gaps in lexical coverage (Hangya et al., 2022), cultural knowledge (Li et al., 2024), and syntactic nuances (Lucas et al., 2024). Consequently, LLM performance in low-resource settings is markedly inferior to that observed with well-resourced datasets. This disparity strongly limits AI progress in the affected domains.

This survey article highlights the methods and evaluations employed in low-resource and specialized domains. We argue that the diversity and quality of datasets are more important than the accumulation of large volumes of mediocre data. This perspective is supported by studies showing that the quality of training data has a significant impact on language model performance, especially in low-resource environments (Micallef et al., 2022; Sajith and Kathala, 2024). To mitigate data scarcity, data augmentation has emerged as an effective solution (Seo et al., 2024), allowing the generation of additional examples to enhance model robustness.

Natural language processing (NLP) encompasses a broad range of tasks, such as text summarization, topic modeling, and text generation (Wikipedia LLMs, 2025). In this study, we focus explicitly on the question answering (QA) task, as it represents a particularly dynamic research area, especially in low-resource contexts. In domain-specific applications notably in the private sector and independent research settings, QA systems and chatbots (Afzal et al., 2024; Megahed et al., 2024) are commonly used to facilitate user interaction with datasets and to evaluate model capabilities. Moreover, with the advent of large language models, QA systems can be adapted to perform other NLP tasks through data restructuring and model fine-tuning. Nonetheless, despite these advances, domain-specific applications continue to face major challenges in low-resource environments.

2 Problem Statement

Overview Low-resource environments for Large Language Models (LLMs) are contexts in which essential resources such as large and diverse corpora, annotated datasets, domain expertise, or data availability are severely limited or entirely absent. These constraints go well beyond the challenges typically associated with low-resource languages. Even in high-resource languages like English, many specialized domains, such as certain branches of medicine or scientific research, suffer from a chronic lack of data (Seo et al., 2024). Since LLMs are primarily pretrained on large, generic corpora, they often fail to generalize to tasks that require fine-grained and domain-specific knowledge. For example, in the biomedical field, although there is a large volume of general medical text, datasets focused on rare diseases or specific clinical trials remain scarce or even nonexistent, which leads to distributional shifts and reduced model performance (Chen et al., 2024b).

These limitations pose major challenges for question-answering (QA) systems in low-resource domains. QA systems require not only extensive lexical coverage but also precise factual knowledge, domain-specific reasoning abilities, and the capacity to extract or infer information from context. When specialized corpora are scarce, QA models struggle to learn the terminology, background knowledge, and inference patterns necessary to produce accurate and relevant answers. Furthermore, in the absence of expert-designed annotations, it becomes difficult to adapt models to handle specialized question types, which increases the hallucination rate and reduces the reliability of responses. Although there is no universally recognized threshold to define a low-resource environment, we consider a dataset to fall into this category when it is not commonly used for the pretraining of large language models, particularly in the case of datasets absent from standard benchmarks.

Research Questions We also aim to explore several research questions. First, it is essential to identify effective strategies to increase the quantity and quality of domain-specific data using LLMs, particularly in areas where such data is scarce. Second, we seek to understand which approaches can enhance the adaptation of LLMs to domain-specific tasks. Third, it is necessary to establish robust evaluation frameworks and metrics to accurately assess model performance in these contexts. Finally, to consider the ethical, privacy, and fairness implications when deploying LLMs in specialized domains. Accordingly, we formulate the following research questions:

- **Q1**: How can domain-specific data be effectively expanded using LLMs?
- **Q2**: Which approaches improve the adaptation of LLMs to domain-specific tasks?
- **Q3**: How can the performance of LLMs be evaluated in low-resource settings?
- **Q4**: What ethical, privacy, and fairness considerations must be addressed?

3 Related Work

Ding et al. (2024a) propose a domain analysis along two axes data and learning. They define four "data perspectives" (creation, annotation, reformulation, co-annotation) and present various learning paradigms ranging from supervised fine-tuning to alignment-based learning. They also illustrate concrete applications, such as Dr. LLaMA for medical question answering (where ChatGPT or GPT-4 rewrite or generate new question–answer pairs) and the selective masking strategy of DALE. Chai et al. (2025) complement this approach with a clear technical taxonomy, encompassing simple methods, prompt-based techniques, information retrieval based approaches, and hybrid methods. However, neither of these studies offers a systematic comparison of the different paradigms applied to the specific constraints of low-resource biomedical or legal domains, such as privacy requirements or distributional shifts.

Our survey builds on these contributions by focusing specifically on data augmentation for question answering in low-resource biomedical and legal contexts. Using targeted datasets, we evaluate how well different augmentation techniques address the unique constraints of these domains. Rather than proposing a new theoretical framework, our contribution lies in a detailed, data-driven comparison that highlights the practical relevance of each approach in sensitive settings.

4 Literature Review and Analysis

4.1 Article Identification Methodology and Analysis

Article Identification To conduct our analysis, we aim to identify under-represented dataset subsets within their respective domains. We focus specifically on datasets in the biomedical and legal fields, as these two areas have been extensively studied in the large language model (LLM) research community. Although a substantial body

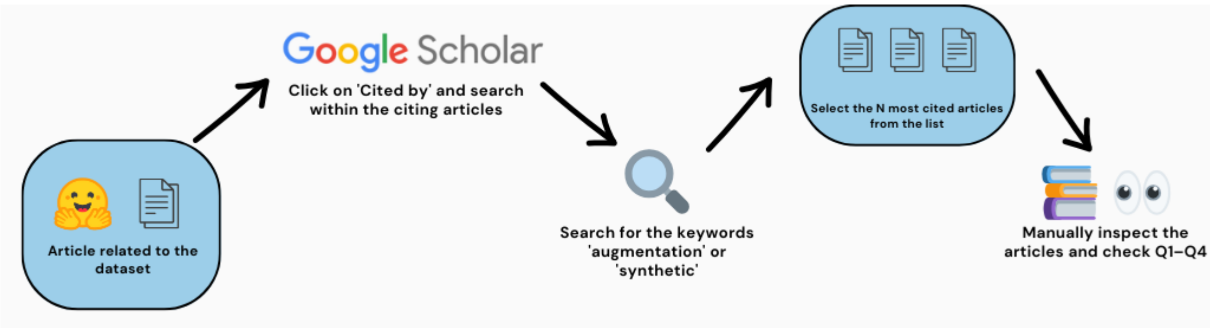


Figure 1: Workflow for identifying relevant papers on dataset augmentation

of literature exists for these domains, it remains difficult to locate publicly available low-resource datasets, often due to privacy concerns, access restrictions, or the absence of standardized repositories. Consequently, for each domain, we restrict our analysis to three or four dataset types that are accessible and sufficiently documented to permit analysis.

As illustrated in Figure 1, we implemented a structured workflow to identify research on dataset augmentation and synthetic data generation. To explore this issue systematically, we performed a literature review focusing on augmentation techniques and synthetic data generation applied to our selected datasets.

Using Google Scholar, we searched for articles containing either the keyword *augmentation* or the keyword *synthetic*, written in English, then filtered them to retain only those related to natural language processing (NLP). These two keywords were chosen to broadly cover the relevant literature on data augmentation, and Google Scholar's full-text indexing allowed us to identify works where these terms appear beyond the title or abstract. This approach facilitated the identification of potentially relevant contributions. We then selected up to N research articles each dataset, with $N \leq 3$[1], excluding review articles and those that mention augmentation techniques only in their related work sections. Review articles were excluded because, although they provide useful overviews, they generally do not present detailed methodological analyses or empirical results specific to the datasets under study. This filtering based on publication type enabled us to concentrate on the most influential and technically substantial contributions to data augmentation methodologies.

In Table 1, we adopt a structured approach to analyze each of the four research questions in the biomedical and legal domains. This framework enables a systematic examination of augmentation techniques applied to various low-resource datasets. We selected three to four datasets per domain. By mapping augmentation approaches to different dataset types, our study offers insights for researchers aiming to improve the performance of large language models (LLMs) in low-resource environments.

4.2 Embedding Model Selection

To analyze text distributions in embedding space, we selected specialized models for each domain based on the MTEB Leaderboard rankings[2], limiting our choices to models of up to 1 billion parameters to control computational costs. The selected models are available in Table 2.

4.3 Biomedical Domain

4.3.1 Overview of Selected Datasets

The biomedical domain remains one of the most critical for AI applications, given its potential to transform diagnosis, treatment planning, and patient management. Despite these promises, this field faces severe data limitations or inaccessibility outside of hospital settings. Although medical data can take many forms such as images, videos, and other modalities. We restrict this study to textual data to maintain a coherent scope.

Applying our methodology, we selected four low-resource medical QA datasets for in-depth analysis. To assess their representativeness, we compared them against `MedMCQA` (Pal et al., 2022), a large-scale dataset of 160,869 instances covering various medical subdomains. We refer to this ref-

[1] Some datasets are recent and still have few specialized methods.

[2] `https://huggingface.co/spaces/mteb/leaderboard`

Domain	Papers Citing Datasets	Q1	Q2	Q3	Q4
Medical	(Möller et al., 2020), COVID-QA	–	✓	✓	–
	↪ (Reddy et al., 2020)	✓	✓	✓	–
	↪ (Siriwardhana et al., 2023)	✓	✓	✓	–
	↪ (Samuel et al., 2024)	✓	✓	✓	–
	(Wang et al., 2024), ReDis-QA	✓	✓	✓	–
	↪ (Li et al., 2025)	✓	✓	–	✓
	↪ (Wang et al., 2025a)	✓	✓	✓	✓
	(Arias-Duart et al., 2025), CareQA	✓	✓	✓	–
	↪ (Wang et al., 2025b)	✓	✓	✓	✓
	(Chen et al., 2024a), Medbullets	–	–	✓	✓
	↪ (Kim et al., 2025)	✓	✓	✓	✓
	↪ (Wang et al., 2025b)	✓	✓	✓	✓
	↪ (Wang et al., 2025a)	✓	✓	✓	✓
Legal	(Ravichander et al., 2019), PrivacyQA	–	–	✓	–
	↪ (Vold and Conrad, 2021)	–	✓	✓	–
	↪ (Parvez et al., 2023)	✓	✓	✓	✓
	↪ (Nayak et al., 2024)	✓	✓	✓	–
	(Ahmad et al., 2020), PolicyQA	–	–	✓	–
	(Lin et al., 2022), TruthfulQA	–	–	✓	✓
	↪ (Wang et al., 2023)	–	✓	✓	✓
	↪ (Kim et al., 2023)	✓	✓	✓	✓
	↪ (Ding et al., 2024b)	✓	✓	✓	✓

Table 1: Overview of the intersection between each research question (Q1 to Q4) and the articles describing corpora in the two studied domains. A check mark ✓ indicates that the question is addressed, a dash indicates that it is not, and arrows ↪ denote the reuse of these datasets for various data augmentation methods.

erence corpus as `ParentQA`. The four specialized datasets are:

- **COVID-QA** (Möller et al., 2020): 2,019 expert-annotated question–answer pairs on COVID-19, using a SQuAD-inspired annotation protocol.
- **ReDis-QA** (Wang et al., 2024): 975 high-quality question–answer pairs covering 205 rare diseases.
- **MedBullets** (Chen et al., 2024a): 616 real clinical cases designed to evaluate reasoning and decision-making in complex clinical scenarios.
- **CareQA** (Arias-Duart et al., 2025): 2,769 instances annotated with both open- and closed-ended questions spanning medicine, nursing, biology, chemistry, psychology, and pharmacology.

4.3.2 Diversity Analysis

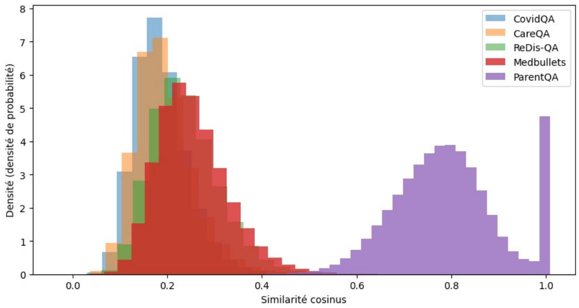
To assess lexical and semantic diversity of the low-resource medical QA corpora relative to the large-scale `ParentQA`, we conducted two complementary analyses: (i) lexical statistics including OOV rates and Shannon entropy (Table 3), and (ii) semantic similarity and OOV overlap analysis (Figure 2).

Lexical Statistics. Table 3 reports for each corpus the unique vocabulary size $|\mathcal{V}|$, the number of vocabulary not found in `ParentQA` (OOV), and the Shannon entropy

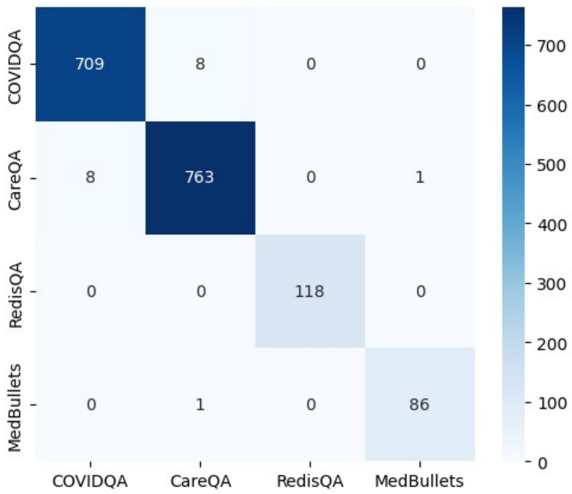
$$H = -\sum_{w \in \mathcal{V}} p(w)\, \log_2 p(w)\,,$$

computed from the empirical unigram distribution $p(w)$. Higher entropy indicates more balanced and extensive vocabulary usage; lower entropy signals concentration on a few frequent terms. All specialized corpora exhibit much smaller $|\mathcal{V}|$ and lower entropy than `ParentQA` (13.09 bits), reflecting their narrow scope and data scarcity. OOV counts range from 86 in `MedBullets` to 763 in `CareQA`, with examples like *creatininuria* and *endosymbionts* highlighting domain-specific terminology.

Semantic Similarity and Implications. Figure 2(a) displays the distribution of cosine similarities between sentence embeddings of each specialized corpus and those of `ParentQA` (embeddings generated by the model detailed in Table 2). `ParentQA` peaks at 1.0 (self-similarity) and centers around 0.75–0.85, indicating strong internal coherence. In contrast, the four low-resource corpora shift leftwards: `COVID-QA` peaks near 0.15, `CareQA` around 0.20, `ReDis-QA` at 0.22, and `MedBullets` at 0.27. Their flatter, wider curves reveal greater



(a) Cosine similarity between each specialized corpus and *ParentQA*



(b) Vocabulary overlap

Figure 2: Comparison of low-resource medical QA datasets to `ParentQA` in terms of (a) cosine similarity and (b) OOV vocabulary overlap.

internal heterogeneity in question phrasing. The large gap (0.50–0.60) relative to `ParentQA` highlights significant domain-induced divergence, both terminologically and syntactically. This "semantic distance" arises from specialized medical jargon (e.g., *furin*, *creatininuria*, *arrhythmiab*) and question structures unseen in generalist corpora.

Combined with low OOV overlap (Figure 2(b)) and reduced entropy (Table 3), these results confirm that each low-resource corpus is both lexically limited and semantically distant from `ParentQA`. These disparities call for domain-sensitive strategies such as targeted vocabulary augmentation, specialized pre-training, or robust adaptation techniques to overcome challenges in low-resource environments.

OOV Overlap. Figure 2(b) shows a heatmap of OOV term overlap between specialized corpora. The overlap is minimal (e.g., only 8 shared OOVs between `COVID-QA` and `CareQA`), indicating that each dataset introduces largely disjoint rare vocabulary. This low overlap underscores the difficulty of transferring lexical knowledge across specialized domains.

4.3.3 Positioning with Respect to the Research Questions

Among the methods examined, Q1 (*how to expand domain-specific data*) falls into two paradigms. On one hand, *few-shot* generation followed by filtering (e.g., *round-trip consistency*), as demonstrated in (Samuel et al., 2024) on CovidQA, enables rapid performance gains without requiring a massive pre-existing corpus. On the other hand, large-scale *chain-of-thought* pipelines combine reasoning extraction, synthesis, and document-based revision to generate hundreds of thousands or even billions of medical tokens, but they require extensive access to manuals, knowledge graphs, or clinical databases (Kim et al., 2025; Wang et al., 2025b).

For Q2 (*which approaches for LLM adaptation*), three main directions emerge. Fine-tuning on annotated corpora (e.g., RoBERTa + COVID-QA) provides consistent improvements starting from just a few thousand expert-labeled examples (Möller et al., 2020). *Chain-of-thought* instruction tuning improves accuracy across various medical benchmarks by explicitly incorporating reasoning during training (Kim et al., 2025). Finally, end-to-end or multi-phase RAG architectures combine tailored *retrieval* with reinforcement learning stages for more refined alignment with clinical criteria, but these models are heavily dependent on external knowledge and domain-specific metrics (Siriwardhana et al., 2023; Wang et al., 2025b).

Regarding Q3 (*evaluation and metrics*), generic *close-ended* indicators such as Exact Match, F1, and perplexity remain foundational across all domains (Möller et al., 2020; Samuel et al., 2024). Semantic-based measures (e.g., BERTScore, BLEURT) and automated judges like G-Eval (Chen et al., 2024a; Arias-Duart et al., 2025) provide deeper qualitative insights into generated responses, while human evaluation remains essential for verifying coherence and factual correctness in clinical contexts (Wang et al., 2025a).

Finally, for **Q4** *(ethical principles)*, most articles either omit these considerations or address them only superficially, highlighting a critical gap in healthcare applications, where patient safety, data confidentiality, and equitable access are paramount (Wang et al., 2025b,a). Given the potential risks of biased or inaccurate medical advice (Li et al., 2025), it is essential for future research to integrate *bias analysis, privacy-preserving protocols, and regulatory frameworks* into data augmentation strategies for biomedical low-resource settings.

Overall, two families of methods can be distinguished: on one hand, **generic methods** such as few-shot generation, chain-of-thought instruction tuning, and light fine-tuning on small annotated corpora, coupled with standard metrics like Exact Match, F1, and perplexity, offer quick implementation and performance gains of 5–10% with just a few dozen examples (Möller et al., 2020; Samuel et al., 2024; Chen et al., 2024a). On the other hand, **domain-specific methods** require access to specialized resources (manuals, knowledge graphs, expert annotations), careful prompt engineering, architectural modifications, and integration into complex fine-tuning pipelines. These methods are typically employed after applying generic techniques to establish a baseline and then further optimize performance by targeting domain-specific nuances. However, their increased effectiveness comes at the cost of reduced transferability, as they require prior adaptation.

4.4 Legal Domain

4.4.1 Overview of Selected Datasets

As the volume of legal cases increases, artificial intelligence plays a crucial role in reducing workloads, minimizing human errors, and accelerating judicial decisions while ensuring their consistency. By automating repetitive and time-consuming tasks such as document analysis and legal research, AI enables legal professionals to focus on strategic decision-making and nuanced case evaluations. Furthermore, predictive analysis helps anticipate outcomes, thus promoting transparency and consistency in judicial decisions (Lai et al., 2024).

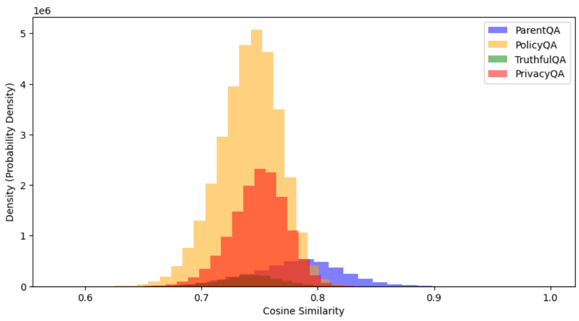
Applying our methodology to this domain, we identified three relevant legal QA datasets for in-depth analysis. We selected a single dataset as the `ParentQA` corpus: the legal subset of MMLU (Hendrycks et al., 2021), which includes the categories *international law*, *professional law*, and *US foreign policy*. These subsets, widely used for pretraining large language models, contain approximately 2 000 examples. The three specialized datasets selected for this study are as follows:

- **PolicyQA** (Ahmad et al., 2020): a reading comprehension dataset focused on website privacy policies, comprising over 17 000 question-passage-answer triplets aimed at concise responses.

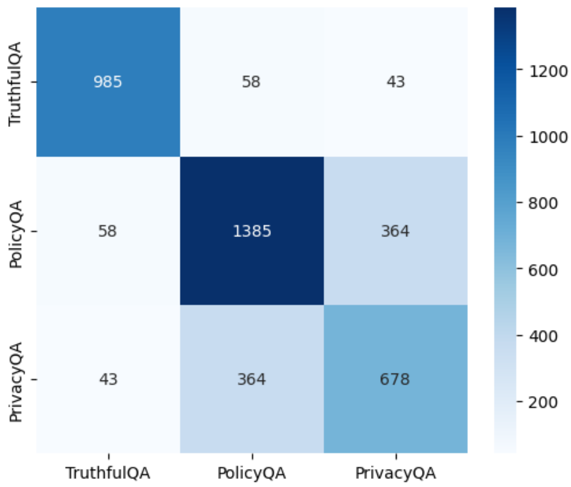
- **PrivacyQA** (Ravichander et al., 2019): a QA dataset on privacy policies, containing 1,750 questions and over 3 500 expert annotations, combining legal and computer science perspectives.

- **TruthfulQA** (Lin et al., 2022): a benchmark consisting of 817 adversarial questions across 38 categories, including a subset dedicated to legal questions, designed to evaluate the truthfulness of language model outputs.

4.4.2 Diversity Analysis



(a) Cosine similarity between each specialized corpus and *ParentQA*



(b) Vocabulary overlap

Figure 3: Comparison of legal datasets with `ParentQA` in terms of cosine similarity and vocabulary overlap

To measure both vocabulary range and semantic consistency across our three specialized QA sets versus `ParentQA`, we ran two complementary analyses: (i) lexical profiling via vocabulary size, OOV rates and Shannon entropy (Table 4); and (ii) internal semantic similarity distributions alongside OOV-overlap statistics (Figure 3).

Lexical Statistics. As Table 4 shows, all three specialized corpora possess drastically smaller vocabularies and lower entropy than `ParentQA` (11.50 bits). **PolicyQA** exhibits the smallest vocabulary (4 093 types) and lowest entropy (8.58 bits). **PrivacyQA** is richer (2 541 types, 9.11 bits), mixing policy-style prompts with occasional technical clarifications, while **TruthfulQA** despite only 2 616 types, yields surprisingly high entropy (10.51 bits). OOV counts against `ParentQA` mirror this pattern: PolicyQA's 1 385 unseen vocabulary (e.g. *cache*, *affiliate*) underscore domain-specific framing; TruthfulQA's 985 new terms (e.g. *cage*, *gasper*) reflect idiosyncratic references; PrivacyQA's 678 OOVs (e.g. *behavioral*, *latitude*) occupy a middle ground.

Semantic Similarity and Implications. Figure 3(a) reports histograms of all pairwise cosine similarities (sentence embeddings) within each corpus. `ParentQA` peaks sharply at ∼0.82, evidencing a large but internally consistent question pool. **PolicyQA** centers at ∼0.75 with a very narrow spread and the highest peak density, signifying highly repetitive structure across its many examples. **PrivacyQA** also peaks near 0.75 but with a modestly wider shoulder toward 0.65–0.70, indicating occasional outlier phrasings alongside core policy-style questions. By contrast, **TruthfulQA** peaks lower, around 0.72, and displays the broadest distribution (spanning 0.55–0.85), directly reflecting its adversarial design to cover diverse topics and linguistic traps.

OOV Overlap. Complementing these semantics, Figure 3(b) shows that OOV-sets are largely distinct: only 58 vocabulary overlap between TruthfulQA and PolicyQA, 43 between TruthfulQA and PrivacyQA, but 364 between PolicyQA and PrivacyQA highlighting their shared legal/policy jargon. Taken together, low entropy and high pairwise similarity in PolicyQA argue for template-like redundancy; TruthfulQA's entropy and spread warn of semantic unpredictability; and PrivacyQA sits in between.

4.4.3 Positioning with Respect to the Research Questions

Among the examined methods, Q1 (how to increase domain-specific data) involves generation and retrieval strategies: generation of semantically equivalent perturbations via paraphrasing with LLMs (Ding et al., 2024b), corpus synthesis through output comparison (Kim et al., 2023), example extraction using multi-retrievers (Parvez et al., 2023), and large-scale instruction generation from meta-templates (Nayak et al., 2024).

Regarding Q2 (approaches for adapting LLMs), the studies combine continual pretraining, fine-tuning, and reinforcement learning: PolicyQA fine-tunes a BERT model pretrained on a corpus of privacy policies to adapt it specifically to the task of extractive QA in this sensitive domain (Ahmad et al., 2020). Rowen activates a generic "retrieve-only-when-needed" mechanism (Ding et al., 2024b); ALMoST combines reward modeling, synthetic demonstrations, and RL (Kim et al., 2023); Citrus integrates CPT, SFT, and reflective RL for clinical tasks (Wang et al., 2025b); and (Vold and Conrad, 2021) demonstrates performance gains of +31% F1 and +41% MRR with RoBERTa fine-tuned on PrivacyQA.

As for Q3 (evaluation and metrics), the studies use standard metrics adapted to each task: EM and F1 for extractive QA (Ahmad et al., 2020), and precision, recall, F1, and MRR for classification and ranking (Ravichander et al., 2019). These metrics are widely recognized for their robustness and ability to reflect performance in low-resource settings.

Finally, regarding Q4 (ethical principles), TruthfulQA warns against misinformation risks and the erosion of user trust caused by misleading answers, advocating for strong safeguards (Lin et al., 2022). ALMoST relies on the HHH benchmark (helpful, harmless, honest) to align models with human values and reduce harmful outputs (Kim et al., 2023). However, most studies do not comprehensively address ethical, privacy, or fairness concerns—yet these dimensions are essential for ensuring user trust, preventing algorithmic bias, and complying with regulations.

Generic approaches rely on paraphrasing, retrieval, and knowledge transfer mechanisms. They enable rapid prototyping and generalization across low-resource domains, but are limited by the consistency and depth of the base model (Kim et al., 2023; Ding et al., 2024a; Nayak et al., 2024). In contrast, domain-specific solutions leverage expert-curated corpora and workflows to achieve peak performance, at the cost of specialized data collection, domain expertise, and computational resources (Vold and Conrad, 2021; Wang et al., 2023). Therefore, it is advisable to start with minimal fine-tuning on a generic transformer, then progressively integrate architectural modules and targeted corpora to meet domain requirements and ensure ethical adoption.

Despite these advancements, a major challenge remains in the availability and structure of legal datasets. Many cases remain undocumented or inaccessible, exacerbating the inherent complexity of domain-specific language, frequent regulatory changes, and the need for high-quality annotated data (Abdallah et al., 2023). Furthermore, several legal subdomains remain largely unexplored in the context of LLMs including international trade agreements[3], space law[4], Antarctic Treaty law[5], and patent law in biotechnology and genetics[6], among others. The datasets available in these areas are still raw and unstructured, requiring significant preprocessing before they can be effectively leveraged for legal research or analysis.

5 Conclusion

In this paper, we presented an in-depth analysis of data augmentation strategies in low-resource settings, focusing on the biomedical and legal domains. We conducted our literature review by first identifying articles that describe relevant datasets, then analyzing papers on Google Scholar that propose data augmentation methods in relation to these datasets. We assessed their treatment of four key research questions: how to increase domain-specific data, which approaches to use for adapting LLMs, how to evaluate their performance, and what ethical implications should be considered. The review was supported by diversity analyses (cosine similarity and lexical overlap) to highlight differences between specialized datasets and their parent corpora, thereby revealing significant challenges related to data scarcity and specificity.

As a continuation of this work, a comparative empirical evaluation of different augmentation strategies applied to each dataset represents an important next step. This initial study also paves the way for identifying augmentation methods suited to low-resource contexts, aligned with the objectives of my thesis. I also plan to extend this analysis to other languages and specialized domains, in order to more thoroughly assess the robustness, generalizability, and limitations of the approaches studied.

[3] https://datatopics.worldbank.org/dta/table.html

[4] https://www.unoosa.org/oosa/en/ourwork/spacelaw/index.html

[5] https://www.ats.aq

[6] https://www.wipo.int/wipolex/en/

6 Limitations

Although this study offers insights into data augmentation and synthetic data generation for low-resource datasets, several limitations must be acknowledged.

Domain specificity This analysis is limited to the biomedical and legal fields. While these domains present diverse and complex challenges, extending the study to other sectors such as energy, finance, or the sciences could reveal additional nuances and enhance the broader applicability of augmentation techniques.

Keyword-based search constraints The literature search relied exclusively on the keywords *augmentation* and *synthetic*. This targeted approach may have excluded relevant works that use alternative terminology or methodologies, thus limiting the scope of our findings.

Parent dataset selection The parent dataset used in our analysis consists of a single large-scale collection, selected under the assumption that its diversity offers a robust reference point. However, incorporating additional and more diverse parent datasets would likely enhance the breadth and generalizability of our analysis.

Language bias We chose to use English-language datasets due to their accessibility and relative availability, which facilitated the identification of a broader literature base. However, this choice may introduce biases: LLMs trained primarily on English data tend to present Anglo-American perspectives as universal truths, thereby overlooking non-English viewpoints (Ramesh et al., 2023). This phenomenon can lead to systematic sampling bias and hinder faithful representation of the true diversity of subjects and opinions.

References

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1):127.

Anum Afzal, Alexander Kowsik, Rajna Fani, and Florian Matthes. 2024. Towards optimizing and evaluating a retrieval augmented qa chatbot using LLMs with human in the loop. *arXiv preprint arXiv:2407.05925*.

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749, Online. Association for Computational Linguistics.

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare LLMs beyond question-answering. *arXiv preprint arXiv:2502.06666*.

Yaping Chai, Haoran Xie, and Joe S Qin. 2025. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *arXiv preprint arXiv:2501.18845*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024b. Rarebench: Can LLMs serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4850–4861.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024a. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024b. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv:2402.10612*.

Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, and 1 others. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700, Singapore. Association for Computational Linguistics.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and S Yu Philip. 2024. Large language models in law: A survey. *AI Open*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Jiaxi Li, Yiwei Wang, Kai Zhang, Yujun Cai, Bryan Hooi, Nanyun Peng, Kai-Wei Chang, and Jin Lu. 2025. Fact or guesswork? evaluating large language model's medical knowledge with structured one-hop judgment. *arXiv preprint arXiv:2502.14275*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of Guarani-Spanish neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6385–6397, Mexico City, Mexico. Association for Computational Linguistics.

Fadel M Megahed, Ying-Ju Chen, Inez M Zwetsloot, Sven Knoth, Douglas C Montgomery, and L Allison Jones-Farmer. 2024. Introducing chatsqc: Enhancing statistical quality control with augmented ai. *Journal of Quality Technology*, 56(5):474–497.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and bert models for maltese. *arXiv preprint arXiv:2205.10517*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Nihal V Nayak, Yiyang Nan, Avi Trost, and Stephen H Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. *arXiv preprint arXiv:2402.18334*.

Josh OpenAI, Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Md Rizwan Parvez, Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian, and Kai-Wei Chang. 2023. Retrieval enhanced data augmentation for question answering on privacy policies. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 201–210, Dubrovnik, Croatia. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond english: Gaps and challenges. *arXiv preprint arXiv:2302.12578*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avi Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2020. End-to-end qa on covid-19: domain adaptation with synthetic training. *arXiv preprint arXiv:2012.01414*.

Aryan Sajith and Krishna Chaitanya Rao Kathala. 2024. Is training data quality or quantity more impactful to small language model performance? *arXiv preprint arXiv:2411.15821*.

Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.

Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. 2024. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Andrew Vold and Jack G. Conrad. 2021. Using transformers to improve answer retrieval for legal questions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 245–249, New York, NY, USA. Association for Computing Machinery.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, and 1 others. 2025a. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*.

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

Guoxin Wang, Minyu Gao, Shuai Yang, Ya Zhang, Lizhi He, Liang Huang, Hanlin Xiao, Yexuan Zhang, Wanyue Li, Lu Chen, and 1 others. 2025b. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, Singapore. Association for Computational Linguistics.

Wikipedia LLMs. 2025. Language model benchmark — Wikipedia, the free encyclopedia. [Online; accessed 10-March-2025].

A Additional Analyses

Table 2 lists the embedding models we selected for the biomedical and legal domains, along with their embedding dimensions and GPU memory requirements.

Domain	Model	Dim.	GPU Mem. (GB)
Biomedical	`jasper_en_vision_language_v1`	8960	3.8
Legal	`inf-retriever-v1-1.5b`	1536	2.9

Table 2: Characteristics of the selected embedding models.

Table 3 and Table 4 report lexical statistics for the medical and legal evaluation corpora, respectively, including vocabulary size, out-of-vocabulary (OOV) counts relative to ParentQA, Shannon entropy, and example OOV.

Corpus	Vocab. Size	OOV Count	Entropy (bits)	Sample OOV
ParentQA	275 944	—	13.09	—
COVIDQA	6 062	709	11.13	*furin, endosymbionts, …*
CareQA	9 943	763	11.87	*creatininuria, cathodic, …*
ReDisQA	3 041	118	10.42	*arrhythmiab, ophthalmos, …*
MedBullets	4 280	86	9.97	*escherchia, nonrebreather, …*

Table 3: Lexical statistics of the evaluation corpora, including vocabulary size, OOV counts relative to ParentQA, Shannon entropy, and example OOV.

Corpus	Vocab. Size	OOV Count	Entropy (bits)	Sample OOV
ParentQA	11 692	—	11.50	—
PolicyQA	4 093	1 385	8.58	*cache, estimating, affiliate, …*
TruthfulQA	2 616	985	10.51	*cage, gasper, england, …*
PrivacyQA	2 541	678	9.11	*behavioral, cache, latitude, …*

Table 4: Lexical statistics of the evaluation corpora: vocabulary size, OOV counts relative to ParentQA, Shannon entropy, and example OOV.