EvalQAG: A Framework for Automatic Complex QA Generation and a Benchmark QA Dataset for Policy Documents

Kirtan Brijeshbhai Soni[*1], **Krish Rupapara**[*1], **Arpit Rana**[2], **Ghanshyam Verma**[3], **Paul Buitelaar**[3]

[1]Dhirubhai Ambani University, Gandhinagar, India
[2]Smart Energy Learning Centre, Dhirubhai Ambani University, Gandhinagar, India
[3]Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway, Galway, Ireland
202101197@dau.ac.in, 202101198@dau.ac.in, arpit_rana@dau.ac.in, ghanshyam.verma@insight-centre.org, paul.buitelaar@insight-centre.org

Abstract

Accelerating research in renewable energy policy is critical for addressing climate change and enabling informed decision-making. Question answering (QA) over public policy documents presents unique challenges due to their legal structure, conditional dependencies, and domain-specific vocabulary. In this paper, we introduce EvalQAG, a framework for generating high-quality QA pairs from renewable energy policy documents. EvalQAG combines structured prompts, retrieval-augmented inputs, and multi-stage evaluation using large language models (LLMs) to support accurate and diverse QA generation. Using this framework, we construct REPolicyQA, a domain-specific QA dataset comprising approximately 160,000 QA pairs from over 1,000 U.S. renewable energy policy documents. The dataset covers five policy-relevant question types: Yes/No, Yes/No with Conditions, Factual, Legal Obligation, and Descriptive, which capture a wide range of reasoning patterns grounded in regulatory texts. We evaluate multiple QA models and uncover significant performance gaps, particularly in legal reasoning and conditional inference, highlighting major shortcomings in current systems. Our results establish EvalQAG as a generalizable QA generation pipeline for policy texts and position REPolicyQA as a new benchmark for advancing QA research in policy and regulatory domains. We believe this work can foster impactful research in the renewable energy sector, particularly by enabling more robust and explainable QA systems for legal and condition-heavy regulatory documents.

Code & Dataset —
https://github.com/kir1906/EvalQAG.git

1 Introduction

Policies related to renewable and sustainable energy play a central role in accelerating the global transition toward cleaner power systems. These documents, typically issued by federal or state agencies, either offer *financial incentives* (such as rebates, tax credits, or grants) or impose *regulatory mandates* (such as compliance requirements or building standards). They often include eligibility criteria, participation conditions, timelines, jurisdictional constraints, and legal obligations—making them critical yet difficult to interpret without expertise. In practice, end users rarely read entire policy documents to determine applicability (Reidenberg et al. 2016; McDonald and Cranor 2008). This is where question-answering (QA) systems can play a transformative role — by extracting precise, legally grounded information from complex policy documents. Users typically approach these texts with specific roles, goals, or situational contexts, as illustrated in Figure 1. To build and evaluate QA systems for policy documents, dedicated datasets are needed that reflect the domain's linguistic and structural complexity. Datasets such as *ConditionalQA* (Sun, Cohen, and Salakhutdinov 2021), *PolicyQA* (Ahmad et al. 2020), and *PrivacyQA* (Ravichander et al. 2019) provide annotated QA rooted in legal and regulatory texts. They reveal domain-specific challenges such as ambiguity, layered logic, and the need for contextual interpretation. Despite progress, ex-

Metadata of Policy :

Title - Jackson EMC - Residential EV Charger Rebate
State - Georgia
Type - Rebate Program
Sector - Residential

Document :

....
To qualify, a vehicle must:
- Have a battery capacity of at least 7 kilowatt hours
- Have a gross vehicle weight rating of less than 14,000 pounds
- Be made by a qualified manufacturer
- Undergo final assembly in North America

....

Question :

I am a resident of Georgia looking to purchase a new electric vehicle. What are the requirements regarding the vehicle's battery capacity and gross vehicle weight rating to qualify for the Clean Vehicle Tax Credit under the Jackson EMC - Residential EV Charger Rebate program?

Answer :

To qualify for the Clean Vehicle Tax Credit, a vehicle must have a battery capacity of at least 7 kilowatt hours and a gross vehicle weight rating of less than 14,000 pounds.

Figure 1: Example from REPolicyQA showing eligibility-related questions for a residential EV rebate.

[*]These authors contributed equally.

Feature	PolicyQA	PrivacyQA	ConditionalQA	REPolicyQA (Ours)
Source	Website privacy policies	Mobile app privacy policies	UK Government policy documents	Renewable energy policy documents
# Documents	115	35	436	1,056
# QA Pairs	714	1,750	3,102	159,069
# Annotations	25,017	3,500	Human-annotated	Automatically generated
QA Annotator	Experts	Mechanical Turkers	Trained annotators	LLMs

Table 1: Comparison of policy and regulation-oriented QA datasets

isting resources face three key limitations. *First*, they lack *scale*, typically covering only a few hundred documents or QA pairs—limiting generalizability across jurisdictions. *Second*, they often fail to capture the *complexity of real-world user scenarios*, where users seek answers grounded in role, location, or eligibility and expect diverse answer types. *Third*, nearly all rely on *crowdsourcing or expert annotation*, which is time-consuming, costly (Kratzwald et al. 2020; Kratzwald, Feuerriegel, and Sun 2020), and hard to scale.

To address these limitations, we present *EvalQAG*, a scalable framework for generating and evaluating QA pairs from renewable energy policy documents. Using EvalQAG on a corpus of *1,056* policy documents from DSIRE[1], we construct *REPolicyQA*, a large-scale dataset with around *160,000* high-quality QA pairs grounded in realistic user contexts. The framework includes three stages: *QA generation*, *QA evaluation*, and *QA filtering*. By combining generation with systematic filtering, REPolicyQA offers a new benchmark for QA in regulatory domains. We also evaluate Llama3-8b (Grattafiori et al. 2024) model on policy-related QA tasks, comparing zero-shot performance to the fine-tuned Llama3-8b model trained on REPolicyQA. Results show that Llama3-P (fine-tuned Llama3-8b) achieves notable gains over its zero-shot counterpart, significantly enhancing reasoning on complex, unseen policy documents—highlighting the strength of REPolicyQA in capturing intricate legal and regulatory contexts.

2 Related Work

2.1 Policy QA Benchmarks

Policy QA benchmarks, which comprise QA pairs grounded in legal, regulatory, and government-issued documents, are essential for evaluating QA systems designed for compliance, eligibility, and legal interpretation tasks. These documents often feature dense language, complex conditions, and binding obligations, posing significant challenges distinct from those in general-domain QA. Table 1 compares several representative datasets in this domain, each emphasizing different aspects of legal and policy document comprehension. *PolicyQA* (Ahmad et al. 2020) and *PrivacyQA* (Ravichander et al. 2019) focus on QA over privacy policies. While they introduce challenges related to legal language, they are narrow in scope and primarily support extractive QA. *ConditionalQA* (Sun, Cohen, and Salakhutdinov 2021) advances the field with multi-hop QA over UK government texts but lacks user-centric scenarios and sector-specific framing. Clause-level datasets like *LEDGAR* (Tuggener et al. 2020) and *CUAD* (Hendrycks et al. 2021) assist in legal clause classification but are not structured for QA and exclude real-world context. *ConTRACT-QA* (Zheng et al. 2021) addresses commercial legal documents but remains extractive. *QuALITY* (Pang et al. 2022) includes long-form multiple-choice QA, but its reliance on distractors limits its applicability for real-world policy reasoning. *LegalBench* (Guha et al. 2023), in contrast, offers a comprehensive benchmark for evaluating legal reasoning capabilities of large language models. However, it primarily focuses on general legal reasoning tasks rather than policy-specific or domain-grounded QA. Our work diverges by targeting renewable energy policies with structured, scenario-driven QA generation.

2.2 Automated QA Generation (QAG)

Early approaches to automatic question generation (QAG) relied on rule-based or template-driven systems, which transformed declarative sentences into factual "wh" questions using syntactic parsing and handcrafted templates. While these methods produced grammatically correct questions, they often resulted in shallow QA pairs lacking deeper reasoning. For example, Hussein, Elmogy, and Guirguis (2014) utilized OpenNLP and rule-based patterns to generate factoid questions, resulting in well-formed outputs but limited sentence diversity and cross-domain scalability. Similarly, Fabbri et al. (2020) applied template-based generation to retrieved sentences in an unsupervised QA pipeline, enabling large-scale question creation but producing mainly surface-level queries. More recent work leverages large language models (LLMs) such as GPT-3 (Brown et al. 2020), T5 (Raffel et al. 2023), and Flan-T5 (Chung et al. 2022) to create diverse, contextually relevant QA pairs. These models demonstrate strong few-shot and zero-shot capabilities, particularly when provided with well-crafted prompts. For instance, Li et al. (2024) introduced a self-prompting strategy that iteratively generates QA exemplars, improving quality without manual annotation. Zhong et al. (2021) demonstrated that meta-tuning across datasets and prompt types enhances zero-shot generalization. Despite such progress, most work still focuses on short, general-domain texts and tends to generate shallow or easily answerable questions.

[1]https://programs.dsireusa.org/system/program

RefineNet (Nema et al. 2019) improved output quality via a two-pass refinement process but was limited to factoid QA. Overall, while LLMs have advanced QAG significantly, challenges remain—particularly in generating questions that require higher-order reasoning, capture nuanced eligibility, or reflect legal obligations.

2.3 Evaluating and Filtering QA Pairs

Evaluating automatically generated QA pairs is challenging, particularly in domain-specific contexts where multiple valid answers are possible. Traditional metrics like BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005), and F1 (Rajpurkar et al. 2016) rely on lexical overlap with gold references, often missing semantic accuracy and contextual fit. Embedding-based metrics such as BERTScore (Zhang et al. 2020) and Sentence Mover's Similarity (Clark, Celikyilmaz, and Smith 2019) offer improvements by comparing texts in a contextual space, but still require high-quality references and offer limited interpretability. Recent work uses large language models (LLMs) as evaluators. Song et al. (2023) employed GPT-3 to assess factual consistency and relevance in science QA, while Wan et al. (2024) showed that LLM-based evaluations align closely with expert judgments. Zheng et al. (2023) further established standardized LLM-as-a-judge benchmarks for assessing model quality and reliability, and Ghosh et al. (2025) analyzed the logical and factual consistency of LLMs in evaluation settings. These models support nuanced evaluation without gold answers, though their effectiveness depends on careful prompt design and tuning. Filtering is equally critical for maintaining dataset quality. Garg and Moschitti (2021) proposed answerability-based filtering using model distillation, which removes invalid or irrelevant questions but may retain semantically trivial ones. Fabbri et al. (2022) introduced QAFactEval, verifying factuality by decomposing QA content into atomic facts and retrieving evidence. While useful for fact-checking, it depends on external corpora and does not address domain specificity or redundancy. These challenges underscore the need for robust, semantically grounded filtering pipelines tailored to domains such as policy or law, where contextual alignment and relevance are crucial.

3 EvalQAG Framework

We introduce **EvalQAG** (Figure 2), a domain-adaptive framework for automated question–answer generation over renewable energy policy documents. EvalQAG combines large language models with structured prompts, role-specific contextualization, and a multi-stage filtering pipeline to generate diverse and high-quality QA pairs grounded in regulatory text. Unlike prior approaches that emphasize model architecture or scale, EvalQAG adopts a quality-centric design that prioritizes the accuracy, relevance, and informativeness of the generated QA pairs.

3.1 QA Generator

The goal of this framework is to generate high-quality, structured QA pairs from complex policy documents in the

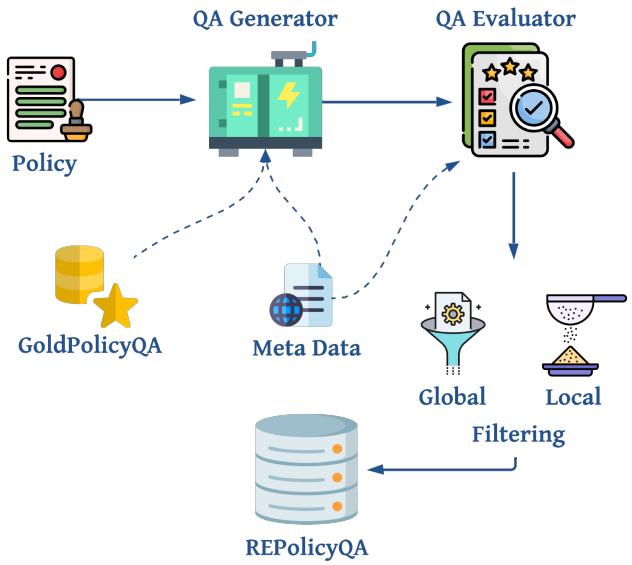

Figure 2: EvalQAG framework

renewable energy domain. Given a policy document D, we segment it into a sequence of semantically coherent chunks using a hybrid chunking strategy, yielding $\mathcal{C}(D) = \{c_1, c_2, \ldots, c_N\}$, where each c_i represents a coherent unit of text. Each document is also associated with structured metadata M_D. To guide QA generation, we define a prompt construction function π, which constructs the prompt,

$$p_{i,\theta,\rho} = \pi(c_i, M_D, \varepsilon, \theta, \rho)$$

where,

- $p_{i,\theta,\rho}$ is the model-ready prompt
- c_i is chunk i of the document D
- M_D represents the metadata
- ε is set of few-shot examples
- $\theta \in \Theta$, where Θ is the set of question types.
- $\rho \in P$, where P is the set of policy types.

In total, we define 10 distinct prompt templates based on 5 question types and 2 policy types. We use a set of large language models $\mathcal{L}$, including LLaMA 3 (Touvron et al. 2024), Mixtral:8 (Jiang et al. 2024), Gemma 3 (Google DeepMind 2025), and Yi:34b (Young et al. 2024). Each model $\mathcal{L}_k$ receives the prompt $p_{i,\theta,\rho}$ and generates structured QA pairs:

$$\mathcal{A}_{i,k}^{\theta,\rho} = \mathcal{L}_k(p_{i,\theta,\rho}) = \{(q_j, a_j, s_j, \phi_j)\}_{j=1}^{n_{i,k}^{\theta,\rho}}$$

where,

- q_j: generated question,
- a_j: corresponding answer,
- $s_j \subseteq c_i$: supporting evidence from the chunk,
- ϕ_j: Answer-dependent conditions that need to be met.

To ensure proportional generation, the number of QA pairs is lower-bounded as follows,

$$n_{i,k}^{\theta,\rho} \geq \left\lfloor \frac{|c_i|}{\tau} \right\rfloor$$

where $|c_i|$ is the length of chunk c_i, and τ is a fixed token or character budget (e.g., 1024). Finally, the full QA set for a document D is constructed by aggregating outputs over all models, chunks, and type combinations. This prompt-conditioned, multi-model generation framework enables the construction of large-scale QA datasets tailored to diverse question intents and policy formats.

Section-Aware Chunking. Policy documents are often too long to input into LLMs in full. Naive chunking strategies that split text into fixed-length segments can perform poorly—especially with structured content like tables, which contain crucial policy details such as incentive rates, eligibility criteria, and deadlines. Fragmenting these structures can harm answer quality and introduce semantic inconsistencies. To address this issue, we implement a section-aware chunking strategy that prioritizes structural coherence, even at the cost of slightly larger chunk sizes. Given a policy document D, we begin by parsing it into a sequence of top-level sections. To balance context length with model input limitations, we define a minimum chunk size of 4096 characters. If a section is smaller than this threshold, it is merged with subsequent sections until the minimum size is met. Next, we apply a hybrid function χ over each merged section s'_m as:

$$\chi(s'_m) = \begin{cases} \{s'_m\} & \text{if } |s'_m| < \tau_u \\ \text{OC}(s'_m, \tau_b, \delta) & \text{otherwise} \end{cases}$$

where:

- $\tau_u = 8192$: maximum size before applying overlapping splits,
- $\tau_b = 4096$: base chunk size for OC,
- $\delta = 512$: character overlap for OC,
- $\text{OC}(s, \tau_b, \delta)$: generates overlapping chunks of length τ_b with stride $\tau_b - \delta$.

The parameters τ_u and τ_b were chosen through pilot experiments: longer spans diluted focus, while smaller windows fragmented logic. These thresholds balanced completeness and contextual relevance. This hybrid chunking approach ensures structural coherence by respecting section boundaries, merges small sections to avoid underflow, and handles large content with overlapping context windows to support high-quality QA generation.

3.2 QA Evaluation

To assess the quality of generated QA pairs, we developed an LLM-based evaluation framework inspired by the evaluation metrics introduced in HoneyBee and sciQAG (Song et al. 2023; Wan et al. 2024). As shown in Table 2, we used five core metrics to evaluate the QA pairs. Each QA pair is evaluated using Qwen3:8b (Yang et al. 2025) with structured prompts and scored on a scale from 1 (poor) to 10 (excellent), along with a brief rationale. Each evaluation prompt dynamically incorporates document metadata and adapts to the specific metric and question type. The LLM is instructed to return a (score, justification) tuple. This framework serves both for scoring QA quality and filtering out low-quality generations based on configurable thresholds or rankings. To validate the reliability of our LLM-based evaluation, we conducted cross-model checks using independent evaluators such as GPT-OSS (Agarwal et al. 2025) and Gemini (Comanici et al. 2025), yielding a low Mean Absolute Deviation (MAD = 0.6). We further performed a human audit on 100 QA pairs, where agreement with LLM scores reached Cohen's κ = 0.72. These results demonstrate consistent scoring across both models and humans, confirming the robustness of our LLM-judge framework.

Accuracy (A): Measures whether the answer and any associated conditions are factually correct and directly supported by the document chunk. This ensures that no hallucinated or misleading content is presented to the user.
Completeness (C): Checks whether the answer fully addresses all aspects of the question, including implied subcomponents. This helps evaluate the depth and coverage of the response beyond superficial correctness.
Intent (I): Evaluates the clarity, focus, and linguistic quality of the question itself, penalizing verbosity, compound phrasing, and grammatical errors. This ensures that the question is well-formed, singular, and user-friendly.
Relevance (R): Quantifies how useful or important the question would be for a user in a specific sector (e.g., residential, nonprofit) seeking to understand or access a policy. This metric aligns the QA generation process with realistic user needs and practical utility.
Groundedness (G): Ensures the answer remains within the bounds of the given context and penalizes hallucinations or unsupported generalizations. This serves as a safeguard against overconfident or fabricated content by enforcing strict contextual fidelity.

Table 2: Evaluation criteria for policy QA generation and assessment.

3.3 QA Filtering

To ensure the quality and reliability of the generated QA dataset, we implement a filtering pipeline that selects only the most relevant, accurate, and well-grounded QA pairs. Given the diverse outputs produced by multiple LLMs, filtering plays a critical role in eliminating noisy, ambiguous, or redundant entries. Our goal is to retain QA pairs that are both contextually faithful to the source chunk and semantically coherent, thereby improving the utility of the dataset.

Local Filtering. To reduce redundancy and select high-quality outputs, we apply local filtering to QA pairs generated for each chunk–question type pair. This enables the consolidation of semantically similar questions. For a given chunk c_i and question type θ, let $\mathcal{A}_i^\theta$ denote the set of all QA pairs generated across policy types and models for fixed c_i and θ. We group semantically similar questions using cosine

Model	Yes-No	Yes-No (Cond.)	Legal Obligation	Factual	Descriptive
Gemma3:27B	8.45 (3)	**8.51** (5)	8.29 (2)	8.07 (3)	8.43 (4)
LLaMA3.3	7.96 (4)	7.97 (5)	8.18 (3)	7.79 (2)	**8.60** (2)
Mixtral	7.78 (3)	8.06 (3)	8.28 (5)	8.06 (4)	**8.43** (2)
Yi:34B	8.16 (4)	**8.21** (5)	8.04 (3)	7.93 (4)	8.07 (2)

Table 3: Best performance per model across question types; values in parentheses indicates few-shot examples; The score is average of five metrics explained in section 3.2

similarity. For any two questions $q_j, q_{j'} \in \mathcal{A}_i^\theta$ with embeddings $\mathbf{v}_j$ and $\mathbf{v}_{j'}$, we assign them to the same group if:

$$\cos(\mathbf{v}_j, \mathbf{v}_{j'}) = \frac{\mathbf{v}_j \cdot \mathbf{v}_{j'}}{\|\mathbf{v}_j\|\|\mathbf{v}_{j'}\|} \geq \tau_s$$

where $\tau_s = 0.95$ is the similarity threshold. The resulting semantic groups are:

$$\mathcal{G}_i^\theta = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_L\}, \quad \mathcal{G}_\ell \subseteq \mathcal{A}_i^\theta$$

where L is the total number of groups formed. For each group $\mathcal{G}_\ell$, we construct the final QA pair as follows:

$$q^* = \underset{(q,\cdot,\cdot,\cdot)\in\mathcal{G}_\ell}{\arg\max}\ I$$

$$(a^*, s^*, \phi^*) = \underset{(\cdot,a,s,\phi)\in\mathcal{G}_\ell}{\arg\max}\ \lambda_1 A + \lambda_2 C + \lambda_3 G$$

where A, C, G, and I are evaluator scores from Table 2 and $\lambda_1, \lambda_2, \lambda_3$ are task-specific or uniform weights. Groups of size one are retained without modification. The final locally filtered QA set for chunk c_i and question type θ is:

$$\mathcal{A}_i^{\theta,\text{local}} = \bigcup_{\ell=1}^{L} \{(q^*, a^*, s^*, \phi^*)_\ell\}$$

This approach ensures each unique question intent is represented by the clearest formulation and paired with the most accurate and complete answer, while maintaining semantic diversity and grounding.

Global Filtering. Following local filtering, we apply a document-level global filtering step to further refine the QA dataset by removing question-answer pairs with low scores. Unlike local filtering, which operates on each chunk–question type pair, global filtering is applied at the level of the entire document for each individual question type. This step enforces consistency and usefulness at the document level, ensuring that retained QA pairs are not only locally sound but also globally informative within the policy context. For a given document D and question type θ, let:

$$\mathcal{A}_D^{\theta,\text{local}} = \bigcup_{i=1}^{N} \mathcal{A}_i^{\theta,\text{local}}$$

denote the set of locally filtered QA pairs aggregated across all chunks c_i in D. We keep only those QA pairs whose Answer (a_r) and Question (q_r) score exceeds a threshold τ_R, ensuring both answer correctness and question clarity at the document level. The globally filtered QA set for question type θ in document D is:

$$\mathcal{A}_D^{\theta,\text{global}} = \left\{ (q, a, s, \phi) \in \mathcal{A}_D^{\theta,\text{local}} \;\middle|\; q_r \geq \tau_R,\ a_r \geq \tau_R \right\}$$

$$q_r = \frac{R + I}{2}, \quad a_r = \frac{A + C + G}{2}$$

where $\tau_R = 7$ in our implementation and A, C, G, I, R are defined as per Table 2. This final filtering step ensures that only QA pairs that are meaningful and useful to end users—given the context of the policy's target sector—are retained for each question type, while preserving diversity across documents and LLM sources.

4 REPolicyQA Dataset

4.1 Policy Collection

To construct a high-quality corpus of U.S. energy policy documents, we leveraged metadata from the Database of State Incentives for Renewables and Efficiency (DSIRE USA), a comprehensive repository of federal and state-level programs. Using DSIRE's structured CSV exports containing metadata such as state, sector, policy type, and source URLs, we systematically retrieved referenced documents via web scraping and link resolution. All documents were standardized to PDF and converted to Markdown using *MinerU*(Wang et al. 2024), an open-source parser that preserves structural elements such as headings, lists, and tables. Metadata was archived and supplemented to resolve missing attributes. The dataset is restricted to Renewable and Efficiency policy documents published from 2021 onward, yielding a curated collection of *1,056* documents linked to *490* unique policy programs. After manual audit, only 32 documents contained meaningful images, which were excluded due to low frequency and processing difficulty. Each program is annotated with an *Applicable Sector* field—*Residential* or *Non-Residential*—and categorized by policy type: *Financial Incentive* or *Regulatory Policy*.

4.2 Applying EvalQAG

Before applying the EvalQAG framework at scale, we first construct a standardized reference dataset that can serve as high-quality exemplars for few-shot prompting and related evaluations.

GOLDPolicyQA Dataset. To guide and benchmark the QA generation process, we manually constructed the *GOLDPolicyQA* dataset through expert annotation. Two annotators curated QA pairs from a set of 12 U.S. renewable

Stage	Descriptive	Factual	Legal Obligation	Yes-No	Yes-No (Cond.)	Total
Raw	61,151	59,558	59,784	62,803	64,408	307,704
Local	54,651	51,656	51,229	55,024	54,228	266,788 (-13.3%)
Global	25,452	28,780	37,331	35,096	32,410	159,069 (-48.3%)

Table 4: QA pair counts per question type across filtering stages; percentages show reduction from initial stage.

Stage	Accuracy	Completeness	Groundedness	Relevance	Intent	Mean
Raw	8.98	8.15	8.76	6.70	8.99	8.32
Local	9.07	8.23	8.83	6.63	9.02	8.36
Global	9.66 (+7.6%)	8.77 (+7.6%)	8.96 (+2.3%)	8.46 (+26.3%)	9.46 (+5.2%)	9.06 (+9.0%)

Table 5: Average QA metric scores across stages. Percentage values indicate improvement relative to the Main stage.

and efficiency policy documents—stratified by both *sector* and *policy type*. Specifically, we selected 4 documents targeting the *residential* sector and 8 targeting the *non-residential* sector, with each category evenly split between *Financial Incentives* and *Regulatory Policies*. For each document, two questions were generated per question type. All questions were answerable solely from the document content, often incorporating realistic user scenarios to enhance contextual grounding. Conditional questions include explicit eligibility conditions, and answers were grounded in the source text with only minimal edits for fluency. Each QA pair was reviewed independently for quality using a structured rubric. Each QA pair was rated on Relevance, Clarity, Completeness, and User-Friendliness, with score ranging from 1 to 10. Approximately 95% of pairs scored 8 or higher, reflecting high annotation consistency.

Few-Shot Prompting Experiments. Few-shot prompting plays a central role in our QA generation framework, enabling LLMs to generalize the task using a small set of curated exemplars rather than task-specific fine-tuning. Exemplars were drawn from the manually curated GOLD-PolicyQA dataset to model the desired structure, tone, and answer granularity for each question category, helping the model produce coherent and grounded QA pairs. To evaluate the impact of different few-shot configurations, we conducted controlled experiments on 27 randomly selected policy documents, generating QA pairs under four settings: 2, 3, 4, and 5 exemplars. This resulted in approximately *2,500* QA pairs. Comparative results (Table 3) reveal the optimal exemplar count for each model–question type combination and informed the prompting strategy used for large-scale generation.

4.3 Statistics of REPolicyQA

The final REPolicyQA dataset was constructed through a multi-stage pipeline that combines large-scale QA generation with rigorous quality filtering, applied to 1,056 U.S. renewable energy and energy-efficiency policy documents. In the initial *Raw stage*, the system produced 307,704 QA pairs spanning five question types, each contributing from 59,000 to 64,000 instances (Table 4). The *Local Filtering* stage operated at the chunk–question-type level, removing semantically redundant or low-quality QA pairs through clustering and evaluation metrics thereby reducing the dataset by 13.3% to 266,788 pairs. Subsequently, a *Global* Filtering step was applied at the document–question-type level to discard low-scoring pairs, resulting in a final dataset of 159,069 QA pairs—a 48.3% reduction from the original pool.

As shown in Table 5, this two-stage filtering process substantially improved overall QA quality. The average QA score increased from 8.32 to 9.06, with notable gains in Relevance (+26.3%) and consistent improvements across Accuracy, Completeness, Groundedness, and Intent. These results demonstrate the effectiveness of our filtering strategy in generating high-quality, user-relevant QA pairs that are grounded in complex policy texts.

Coverage Analysis. To evaluate how well each context reflects its source text, we introduce a *chunk-to-context coverage* metric that measures the proportion of a chunk preserved in the final context (range 0–1). We applied this metric to all QA pairs after Global Filtering. Dataset exhibits strong contextual fidelity: the average coverage score is *0.89*, with *85.9%* of chunks scoring above 0.8 and only *3%* of chunks scoring below 0.2. This distribution demonstrates robust alignment between chunks and their corresponding contexts across REPolicyQA.

Train / Dev / Test Splits. To prevent information leakage, we split the dataset at the document level, ensuring questions from the same source stay in the same split. The dataset comprises 1,051 documents and 159,069 QA pairs, with 113,514 / 23,026 / 22,529 questions assigned to the train / dev / test sets, respectively. This approach supports robust generalization to unseen policies.

5 Experiments

5.1 Experimental Setup

To evaluate the effectiveness of the EvalQAG framework, we conducted both zero-shot and fine-tuning experiments using controlled subsets of the dataset. These experiments were designed to assess the impact of domain adaptation and structured prompting on model performance.

Model	Yes-No	Yes-No Cond.	Legal Obligation	Factual	Descriptive	Global (F1/EM)
LLaMA3-8B (NC)	0.18	0.38 / 0.12	0.13	0.12	0.09	0.18 / 0.12
LLaMA3-P (NC)	**0.76**	**0.71 / 0.21**	**0.19**	**0.15**	**0.20**	**0.41 / 0.29**
	(+322.2%)	(+86.8% / +75.0%)	(+46.2%)	(+25.0%)	(+122.2%)	(+127.8% / +141.7%)
LLaMA3-8B	0.51	0.38 / 0.15	0.16	0.16	0.11	0.26 / 0.19
LLaMA3-P	**0.85**	**0.79 / 0.33**	**0.50**	**0.53**	**0.49**	**0.64 / 0.38**
	(+66.7%)	(+107.9% / +120.0%)	(+212.5%)	(+231.3%)	(+345.5%)	(+146.2% / +100.0%)

Table 6: F1 scores for LLaMA3-8B (baseline) and LLaMA3-P (fine-tuned), NC = No context provided. All differences are statistically significant compared to the baseline (two-sided t-test, with $p < 0.01$).

Zero-Shot Setting. We evaluated the instruction-tuned language model LLaMA3-8B (Grattafiori et al. 2024) in a zero-shot setting on a held-out test set comprising 22,529 QA pairs. QA instances were sampled at the document level to ensure no overlap with training documents, preserving the integrity of the evaluation. The model was prompted using a standardized template tailored for policy QA, designed to elicit context-sensitive and legally grounded responses.

Fine-Tuned Setting. To examine the benefits of domain adaptation, we fine-tuned LLaMA3-8B using an Alpaca-style instruction tuning approach (Taori et al. 2023). The training set consisted of 10,000 QA pairs randomly selected from the EvalQAG dataset, with document-level separation from the test set. The resulting model, referred to as LLaMA3-P, was evaluated on the same 22,529-question test set used in the zero-shot setting, enabling a controlled comparison of general-purpose and domain-adapted performance.

5.2 Implementation Details

For zero-shot experiments, we set the temperature to 0.8 and top_p to 0.9, and `max_new_tokens` fixed at 100. To fine-tune the LLaMA3-8B model on our QA dataset, we follow the Alpaca method (Taori et al. 2023). Training is conducted over 3 epochs with a per-device batch size of 4 and gradient accumulation over 8 steps on an NVIDIA A100 GPU. We use a cosine learning rate schedule with a base learning rate of 2e-5, no weight decay. Mixed precision is enabled with bfloat16 and TF32. The maximum sequence length is 2048 tokens. To assess model performance, we use two metrics: Exact Match (EM) and F1 score.

6 Results

Table 6 presents the performance comparison between the baseline LLaMA3-8B and the fine-tuned LLaMA3-P models across five question types using EM and F1 scores. Fine-tuning yields consistent gains across all categories, with the most notable improvements observed in Legal Obligation, Factual, and Descriptive questions. These question types require precise extraction of obligations, retrieval of concrete policy facts, and generation of structured explanations—suggesting that the REPolicyQA dataset effectively captures these aspects through its chunk-aware prompting and scenario-driven generation process. Notably, under the no-context (NC) setting—where the model receives only the question without supporting passage—the fine-tuned model still demonstrates clear advantages. This highlights the model's improved ability to generalize from domain-specific training, even in the absence of direct textual grounding. Overall, these results underscore the value of domain adaptation: fine-tuning on REPolicyQA equips the model with stronger semantic alignment to regulatory language, enabling it to better handle conditional logic, legal terminology, and eligibility structures, key challenges in compliance-oriented QA systems.

7 Conclusion

We presented EvalQAG, a scalable and structured framework for automatically generating and evaluating question–answer (QA) pairs from complex renewable energy policy documents. The framework integrates multi-model generation, role-aware prompting, and two-stage filtering to produce high-quality QA datasets that capture the linguistic and interpretive nuances of real-world regulatory texts. Using EvalQAG, we constructed REPolicyQA, a large-scale dataset of around 160,000 QA pairs across diverse question types and policy domains. Fine-tuning on REPolicyQA yields substantial performance gains, particularly in challenging categories such as legal obligations, factual reasoning, and descriptive understanding—demonstrating its value where models typically struggle most.

Beyond technical contributions, EvalQAG offers practical and societal impact by transforming dense legal content into accessible QA pairs tailored to user needs. This helps homeowners, businesses, and service providers better navigate their rights and obligations, enhancing public access to policy information and supporting broader engagement in renewable energy adoption. Future work will explore user-in-the-loop refinement, multimodal integration, and reinforcement-based filtering to further improve adaptability, interpretability, and practical utility.

Acknowledgments

This work emanates from research supported by a joint CSR grant from BSES Rajdhani Power Limited and BSES Yamuna Power Limited, New Delhi, under Grant Number CSR/BSES/A4-AR/SELC.

References

Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; Arora, R. K.; Bai, Y.; Baker, B.; Bao, H.; et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Ahmad, W. U.; Chi, J.; Tian, Y.; and Chang, K.-W. 2020. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. arXiv:2010.02557.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416.

Clark, E.; Celikyilmaz, A.; and Smith, N. A. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2748–2760. Florence, Italy: Association for Computational Linguistics.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fabbri, A.; Ng, P.; Wang, Z.; Nallapati, R.; and Xiang, B. 2020. Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4508–4513. Online: Association for Computational Linguistics.

Fabbri, A.; Wu, C.-S.; Liu, W.; and Xiong, C. 2022. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2587–2601. Seattle, United States: Association for Computational Linguistics.

Garg, S.; and Moschitti, A. 2021. Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7329–7346. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ghosh, B.; Hasan, S.; Arafat, N. A.; and Khan, A. 2025. Logical Consistency of Large Language Models in Fact-checking. arXiv:2412.16100.

Google DeepMind. 2025. Gemma: Open Models Based on Gemini Research and Technology. https://ai.google.dev/gemma. Google.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36: 44123–44279.

Hendrycks, D.; Burns, C.; Chen, A.; and Ball, S. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. arXiv:2103.06268.

Hussein, H.; Elmogy, M.; and Guirguis, S. 2014. Automatic English Question Generation System Based on Template Driven Scheme. *International Journal of Computer Science Issues (IJCSI)*, 11: 45–53.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arXiv:2401.04088.

Kratzwald, B.; Feuerriegel, S.; and Sun, H. 2020. Learning a Cost-Effective Annotation Policy for Question Answering. arXiv:2010.03476.

Kratzwald, B.; Yue, X.; Sun, H.; and Feuerriegel, S. 2020. Practical Annotation Strategies for Question Answering Datasets. arXiv:2003.03235.

Li, J.; Wang, J.; Zhang, Z.; and Zhao, H. 2024. Self-Prompting Large Language Models for Zero-Shot Open-Domain QA. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,

296–310. Mexico City, Mexico: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

McDonald, A. M.; and Cranor, L. F. 2008. The cost of reading privacy policies. *International Journal of Law and Policy (Isjlp)*, 4: 543.

Nema, P.; Mohankumar, A. K.; Khapra, M. M.; Srinivasan, B. V.; and Ravindran, B. 2019. Let's Ask Again: Refine Network for Automatic Question Generation. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3314–3323. Hong Kong, China: Association for Computational Linguistics.

Pang, R. Y.; Parrish, A.; Joshi, N.; Nangia, N.; Phang, J.; Chen, A.; Padmakumar, V.; Ma, J.; Thompson, J.; He, H.; and Bowman, S. R. 2022. QuALITY: Question Answering with Long Input Texts, Yes! arXiv:2112.08608.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Ravichander, A.; Black, A. W.; Wilson, S.; Norton, T.; and Sadeh, N. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. arXiv:1911.00841.

Reidenberg, J. R.; Bhatia, J.; Breaux, T. D.; and Norton, T. B. 2016. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2): S163–S190.

Song, Y.; Miret, S.; Zhang, H.; and Liu, B. 2023. HoneyBee: Progressive Instruction Finetuning of Large Language Models for Materials Science. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5724–5739. Singapore: Association for Computational Linguistics.

Sun, H.; Cohen, W. W.; and Salakhutdinov, R. 2021. ConditionalQA: A Complex Reading Comprehension Dataset with Conditional Answers. arXiv:2110.06884.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Martin, L.; Lu, K.; Bhosale, S.; Dettmers, T.; Ott, M.; Scialom, T.; Edunov, S.; Fan, A.; and et al. 2024. LLaMA 3: Open Foundation and Instruction Models. https://ai.meta.com/blog/meta-llama-3/. Meta AI.

Tuggener, D.; von Däniken, P.; Peetz, T.; and Cieliebak, M. 2020. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1235–1241. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Wan, Y.; Liu, Y.; Ajith, A.; Grazian, C.; Hoex, B.; Zhang, W.; Kit, C.; Xie, T.; and Foster, I. 2024. SciQAG: A Framework for Auto-Generated Science Question Answering Dataset with Fine-grained Evaluation. arXiv:2405.09939.

Wang, B.; Xu, C.; Zhao, X.; Ouyang, L.; Wu, F.; Zhao, Z.; Xu, R.; Liu, K.; Qu, Y.; Shang, F.; Zhang, B.; Wei, L.; Sui, Z.; Li, W.; Shi, B.; Qiao, Y.; Lin, D.; and He, C. 2024. MinerU: An Open-Source Solution for Precise Document Content Extraction. arXiv:2409.18839.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Wang, G.; Li, H.; Zhu, J.; Chen, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zheng, L.; Guha, N.; Anderson, B. R.; Henderson, P.; and Ho, D. E. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. arXiv:2104.08671.

Zhong, R.; Lee, K.; Zhang, Z.; and Klein, D. 2021. Adapting Language Models for Zero-shot Learning by Meta-tuning on Dataset and Prompt Collections. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2856–2878. Punta Cana, Dominican Republic: Association for Computational Linguistics.