Quebec Automobile Insurance Question-Answering With Retrieval-Augmented Generation

David Beauchemin[*†], **Zachary Gagnon**[*‡], and **Richard Khoury**[†]
Université Laval, Québec, Canada
Computer Science Department[†]
[†]{david.beauchemin, richard.khoury}@ift.ulaval.ca,
[‡]zachary.gagnon.1@ulaval.ca

Abstract

Large Language Models (LLMs) perform outstandingly in various downstream tasks, and the use of the Retrieval-Augmented Generation (RAG) architecture has been shown to improve performance for legal question answering (Nuruzzaman and Hussain, 2020; Louis et al., 2024). However, there are limited applications in insurance questions-answering, a specific type of legal document. This paper introduces two corpora: the Quebec Automobile Insurance Expertise Reference Corpus and a set of 82 Expert Answers to Layperson Automobile Insurance Questions. Our study leverages both corpora to automatically and manually assess a GPT4-o, a state-of-the-art LLM, to answer Quebec automobile insurance questions. Our results demonstrate that, on average, using our expertise reference corpus generates better responses on both automatic and manual evaluation metrics. However, they also highlight that LLM QA is unreliable enough for mass utilization in critical areas. Indeed, our results show that between 5% to 13% of answered questions include a false statement that could lead to customer misunderstanding.

1 Introduction

To protect their financial situation and property, vehicle owners and homeowners need to buy property damage insurance. However, most people have little to no proper knowledge of insurance products, and rely on insurance representatives to help them properly select and comprehend these products (RLRQ, 2004; AMF, 2019). As a result, in order to protect the public, insurance regulators, such as the "Autorité des marchés financiers" (AMF) in Quebec, make sure that insurance representatives are well-trained and educated, and that insurers properly inform their customers (AMF, 2024a).

However, customers are increasingly interested in buying insurance products online (Claire et al., 2018). This change impacts how an insurer can adequately inform their customer. Traditionally, customers buy products in person or through phone insurance representatives, which allows an insurance expert to help the customer understand the different products and buy the correct one. With an online sale, customers are left to gather information by themselves (AMF, 2018; Johnson, 2018). Moreover, insurance is regulated locally, which means that insurance products, coverages and laws are different from one jurisdiction to the next. Consequently, while many resources are available online, such as "infoassurance" (IBC and GAA, 2024), only the limited set of resources from one's own locality are applicable, and customers must take care not to get information from elsewhere.

The rapid progress in natural language processing and the growing availability of insurance data present unprecedented opportunities to bridge the gap between people and insurance knowledge. For instance, legal text summarization (Shukla et al., 2022) holds the potential to simplify complex legal documents for laypeople. Similarly, insurance question-answering (QA) could offer affordable, expert-like assistance to non-expert customers.

To this end, we present an end-to-end approach aimed at generating high-quality answers to Quebec automobile insurance questions. Our methodology harnesses the popular "Retrieval-Augmented Generation" (RAG) approach. The main contributions of this work are therefore:

1. The creation and release of a Quebec Automobile Insurance Expertise References Corpus[1];
2. The creation and release of a corpus of 82 Expert Answers to Laypeople Automobile Insurance Questions[2];

[*]Contributed equally to this work.

[1]https://github.com/GRAAL-Research/quebec-insurance-rag-corpora

[2]https://github.com/GRAAL-Research/quebec-insurance-rag-corpora

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 48–60

3. A set of experiments to assess the performance of GPT-4o, a state-of-the-art LLM, on our QA corpus, including a manual evaluation of the generated answers.

This paper is outlined as follows: first, we study the relevant questions-answering legal RAG research and its related corpora in Section 2. Then, we propose our corpora in Section 3, and in Section 4, Section 5 and Section 6 we present a set of experiments aimed at evaluating the performances of GPT-4o at answering Quebec automobile insurance questions. Finally, in Section 7, we conclude and discuss our future work.

2 Related Work

Legal-Domain QA RAG The advent of large language models (LLMs) has led to advances in many previously arduous tasks, such as in the application of the RAG concept in QA tasks, which has attracted a great deal of research interest in recent years (Pipitone and Alami, 2024). Answering legal questions has always been more complex due to the inherent difficulties of exploiting specialized texts that stem from handing specialized terminology (Wiratunga and Ram, 2011) and intricate sentence structures (Katz et al., 2023). Recently, Louis et al. (2024) has presented an end-to-end methodology to generate answers to any statutory law question leveraging a RAG architecture, along with a long-form legal question answering dataset comprising 1,868 expert-annotated legal questions in French. Likewise, the insurance sector, with its complex documents and nuanced information, could benefit from these advancements. Consequently, although research is mainly focused on the legal field, there is also a growing interest in the insurance sector, including for insurance RAG. Nuruzzaman and Hussain (2020) have presented a chatbot that generates accurate and contextual responses by identifying intentions and entities while ensuring semantic relevance and meaning of responses. It is trained on domain-specific datasets to understand insurance-specific terms and information. It notably uses RAG strategies to generate responses. Likewise, Na et al. (2022) focuses on a single-turn dialogue covering insurance QA on a Korean dataset to respond to insurance customers.

Legal and Insurance Corpora The number of datasets available in the legal and insurance domains has increased in recent years (Martinez-Gil, 2023; Cui et al., 2023). One example is CUAD (Hendrycks et al.), a dataset for legal contract review that includes 13,000 human annotations. The first insurance QA dataset was proposed by Feng et al. (2015), and consists in 16,889 question-answer pairs; they also conducted experiments to assess different approaches at answering insurance questions. More recently, Butler (2023) have proposed a corpus of 2,124 synthetic question-answer pairs concerning Australian law. The corpus was generated using GPT-4 and the Open Australian Legal Corpus, but the answers were not reviewed by an insurance expert. However, as of yet, no such corpus exists for automobile insurance questions.

3 Corpora

This section describes the two corpora we created for our work: our French corpus of automobile insurance expertise references documentation for the Province of Quebec (Canada), and our French corpus of 82 layperson questions about Quebec automotive insurance and their expert answers and annotations. First, we will describe our process for creating each corpus[3] and then present some key statistics.

3.1 Corpus Creation

3.1.1 Quebec Automobile Insurance Expertise Reference Corpus

This corpus is composed of a set of documents extracted from seven official and reliable online sources about automobile insurance in Quebec. These sources have been selected in partnership with a Canadian insurance company. They have been divided into the following four categories:

- The **Laws** category includes two pieces of provincial legislation related to Quebec automobile insurance. The first one is the *Loi sur l'assurance automobile* (Quebec, 2024), which establishes the regulations governing insurers and insureds in Quebec. The second one is the *Code de sécurité routière* (Quebec, 2016), and it governs the use of all motorized vehicles and pedestrians on public roads to ensure safety.
- The **F.P.Q. 1** category includes the manually extracted Quebec mandatory-approved automobile insurance contracts (AMF; Beauchemin and Khoury, 2023). The F.P.Q. 1 is divided into civil

[3]We also discuss the risk of data leakage in our Limitations section.

liability and property damage. Optional coverages are described in endorsements. We have included one realistic synthetic contract that includes all available endorsements.
- The **Insurance Regulator Educative Resources** category includes informative resources from the AMF, Quebec's regulatory body for financial and insurance products and services (AMF, 2024a). We included its educational information related to automobile insurance for customers.
- The **Domain-Specific Educative Resources** category includes educative resources from four insurance domain organizations. They all propose various educational resources to the public through their online blog. The first, the *Chambre de l'Assurance de Dommages*, is the regulatory body that oversees the training and ethics of insurance agents, brokers and claims adjusters (ChAD, 2024). The second, the *Groupement des Assureurs Automobiles*, is the association of all home and car insurance insurers in Quebec. It oversees and develops various mechanisms to improve the property damage system (GAA, 2024). The third, *Éducaloi*, is a non-profit organization created by the Quebec Ministry of Justice that informs the public on legal matters, such as insurance products (Éducaloi, 2024). Lastly, *Infoassurances* is an insurance information website created by the Insurance Bureau of Canada and the *Groupement des Assureurs Automobiles* for the purpose of "properly informing customers about property insurance" (IBC and GAA, 2024).

We have selected 21 online documents from these sources that focus on the subject of "automobile insurance". The documents can be pieces of legislation, legal insurance documents, informative resources, or informative blog articles. The content of each document has been manually extracted and cleaned to remove trailing whitespace, along with paragraphs that are either "replaced" or "repealed" in a piece of legislature.

3.1.2 Corpus of Expert Answers to Laypeople's Automobile Insurance Questions

This corpus comprises a set of French questions and answers related to automotive insurance in Quebec. They were manually extracted from highly-reliable sources that were selected in partnership with a Canadian insurance company, like for the previous references corpus. Our selected sources are divided into the following four categories:

- The **Quebec Insurance Company FAQ** category includes question-answer pairs taken from the FAQ web pages of four insurers', namely, Beneva (Beneva, 2024), Desjardins Assurances (Desjardins Assurances, 2024), Belairdirect (Belairdirect, 2024) and Sonnet (Sonnet, 2024). These insurers have been selected based on two selection criteria. First, they must sell automotive insurance in Quebec. Second, the questions in their FAQ must not overlap with those of other selected insurers. For example, Intact Assurance's (Intact Insurance, 2024) FAQ is identical to Belairdirect's, since both companies belong to the same corporation[4], and therefore that insurer was excluded.
- The **Regulator** category includes insurance professional practice examination questions and answers from the regulator (AMF, 2024b).
- The **Domain-Specific Educative Resources** category includes question-answer pairs available through two educative resources and blogs from insurance sector organizations, namely the *Chambre de l'Assurance de Dommages* and *Infoassurances*. These two sources are also used as reference sources. We have carefully ensured no overlap between the extracted questions and the extracted reference content from these sources.
- The **Quebec Public Automobile Insurance Plan** category includes question-answer pairs from the Quebec government agency responsible for the automobile insurance plan that covers all bodily injuries (SAAQ, 2024).

We extracted 82 question-answer pairs from these sources, along with a category for each pair. Seven categories were extracted from the sources; each question is related to one of the following categories:

- **Legal Obligations** are questions related to the insuree's and insurer's legal obligation. For example, it could be a question about the minimum amount of civil liability insurance required.
- **Civil Liability Coverage** are questions related to civil liability coverage. This could be for example a question about how a civil liability claim works.
- **Property Coverage** are questions related to at-fault accidents and the scope of property damage

[4] https://www.intactfc.com/en

protection. For example, there could be a question about how to file an at-fault claim.

- **Endorsement** are questions related to any endorsements in insurance contracts. For example, it could be a question about the protections found in an endorsement.
- **Terms and Conditions** are questions related to an insurance contract's general terms and conditions. It could be a question about the consequence of a customer not paying their premium for instance.
- **General** are questions related to the general elements of the insurance sector. One example could be a question about why insurance companies use credit scores during the insurance proposal.
- **Public Automobile Insurance Plan** are questions related to bodily injury coverage offered by the public automobile insurance plan in Quebec. This for example could be a question about the program coverage and exclusions.

3.2 Corpora Analysis

Table 1 presents some key statistics of our French corpora and similar English insurance QA corpora introduced in Section 2. For the English insurance QA corpora, we have used their latest official version[5]. All statistics were computed using SpaCy's latest language-specific tokenizer (Honnibal et al., 2020). They exclude new lines (`\n`), whitespaces, punctuations and some special characters (<, >, | and $). Moreover, to evaluate the reading complexity level of the contracts, we compute readability scores using the frequently used Flesch-Kincaid formula (Flesch, 1948). It computes a score using a scale from 0 (hardest) to 100 (easier) to assess the readability level. We will first analyze our reference corpus and then compare our QA corpus with similar corpora using Table 1.

3.2.1 Our References Corpus Analysis

In Table 1 (left side), we see that all four sources share relatively similar statistics. Indeed, the average number of lexical words (LW), average sentence lengths (both), and average number of sentences are relatively similar. Moreover, since legal documents are known to be complex and lengthy and to use specialized vocabulary (Katz et al., 2023), we can see that the average number of tokens, lexical richness and average Flesch-Kincaid score are lower than the two other types of documents.

3.2.2 Question-Answering Corpora Comparison

We can see in Table 1 (right side) that our QA corpus shares similar patterns to the other corpora. Indeed, for all corpora, the questions use less than half the vocabulary size as the answers and are half as long in terms of tokens, LW, number of sentences, and average sentence length as the answers. They are also easier to read than the answers based on the Flesch-Kincaid score. However, ours is significantly smaller compared to other similar corpora due to its nature. Indeed, the other two similar corpora focus on the broader insurance domain. For example, Insurance QA includes questions about all types of insurance (property, life, and health) throughout the USA. In contrast, our corpus focuses on a single insurance product for a single province in Canada.

4 QA Methodology

This section details our methodology for leveraging a large language model (LLM) to answer insurance questions. Our choice of architecture is similar to Louis et al. (2024), Ajmi (2024), and Wiratunga et al. (2024). We use a RAG architecture to inject domain expertise into an LLM generation for QA. Like the previous authors, our RAG architecture is inspired by the concept of "advanced RAG" (Gao et al., 2023), an architecture that adds a pre- and post-retrieval steps to the traditional processing. Our architecture was built using `LangChain` (Chase, 2022), a Python framework that consolidates the various components of the RAG architecture. As illustrated in Figure 1, first, a retriever selects a small subset of insurance documents from our reference corpus (red), some relevant to the question and some not. Then, a generator conditions its answer on the subset of articles returned by the retriever (blue). We describe these two components in details in the following subsections.

4.1 Retriever

The function of our retriever component is to extract from our reference corpus portions of texts, such as sentences or paragraphs, that are relevant to a question and to present them at the forefront of the returned results. It is a two-step operation consisting of pre-processing and retrieval steps.

[5] https://github.com/shuzi/insuranceQA, https://huggingface.co/datasets/umarbutler/open-australian-legal-qa

	References Corpus					Our QA Corpus		Australian Legal QA		Insurance QA	
	Laws	F.P.Q. 1	Regulator	Sector	Avg	Questions	Answers	Questions	Answers	Questions	Answers
Number of QA pair	N/A	N/A	N/A	N/A	N/A	82		2,124		16,889	
Vocabulary size	4,638	1,751	1,038	1,029	6,201	367	950	6,657	13,583	3,658	19,355
Avg number of tokens	89.41	87.43	115.52	109.83	90.60	14.45	57.98	26.03	85.99	7.36	99.98
Avg number of LW	38.95	42.14	51.2	49.71	40.20	6.68	25.67	14.57	44.6	4.03	45.91
Avg number of sentence	4.13	8.12	7.33	7.12	4.96	1.24	3.00	1.41	3.12	1.00	5.42
Avg sentence length (tokens)	21.37	11.81	17.05	16.67	19.56	12.22	20.86	21.09	31.22	7.32	19.57
Avg sentence length (LW)	9.25	5.89	7.61	7.51	8.61	5.59	9.34	11.74	16.52	4.02	9.12
Lexical richness	0.11	0.18	0.37	0.36	0.10	0.48	0.37	0.21	0.14	0.05	0.02
Avg Flesch-Kincaid score	46.67	56.45	61.41	65.6	49.31	73.66	60.19	55.8	46.1	71.25	66.78

Table 1: Aggregate statistics of our French corpora and similar English insurance QA corpora introduced in Section 2. "Avg" stands for average, "LW" for lexical words.

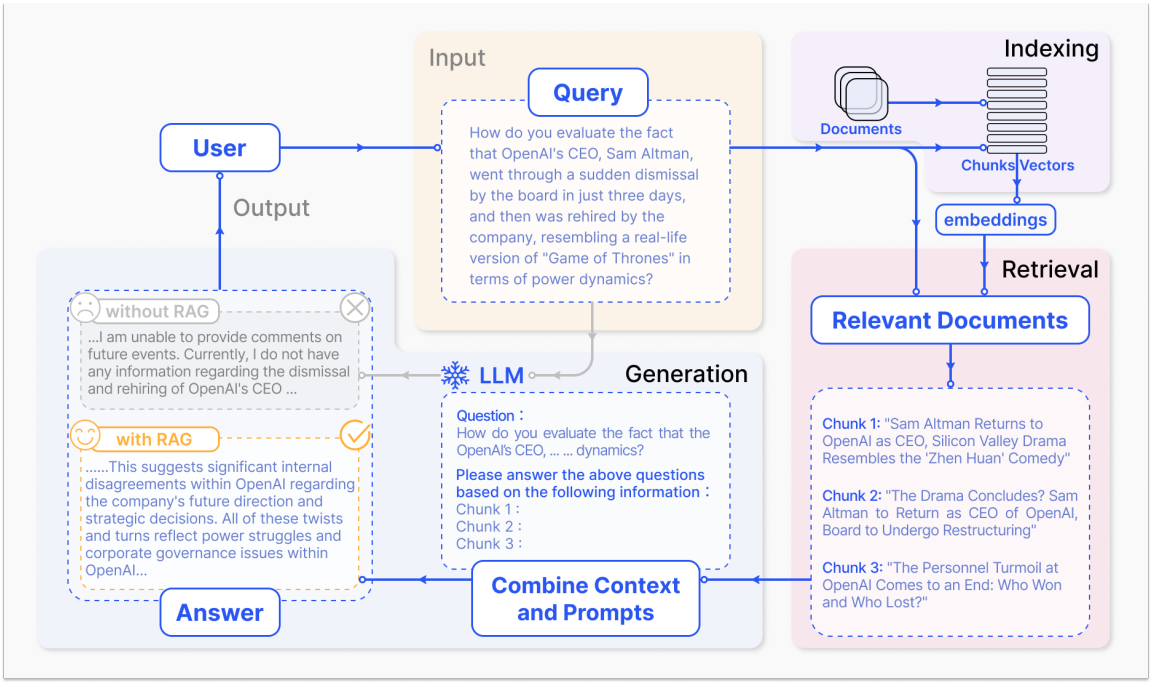

Figure 1: A representative instance of the 3-steps RAG process applied to question answering. 1) Indexing: Documents are split into chunks, and encoded into vectors in a vector database. 2) Retrieval: Retrieve the Top k chunks most relevant to the question based on semantic similarity. 3) Generation: Input the original question and the retrieved chunks together into LLM to generate the final answer. The illustration is taken from Gao et al. (2023).

4.1.1 Pre-Processing

During the pre-processing step, all our documents in the reference corpus go through a two-step pre-processing stage to prepare our document for our retrieval algorithm. The first step is to split the document into smaller chunks of text (i.e. chunking). Based on the best practices for RAG in Wang et al. (2024), we use a fixed chunk size of 500 characters which gives optimal performance for document search since it standardizes their size for better similarity search results. Moreover, legal documents are similar to the financial reports of Yepes et al. (2024) because they use a standard structure to present their content. For example, laws are divided in chapters composed of articles relevant to their subject, which are in turn composed of sub-articles related to the main article. We thus process the documents using a parent-child split function to capture this structure. However, the complete chunk is also supplied for generation when the similarity function is performed during retrieval on the child-split. Namely, if a sub-article is extracted as a relevant text, the main article's text chunk will be provided, not only the sub-article.

The second step is to encode all chunks into dense embedding representation for retrieval. To do so, we use `text-embedding-ada-002` (Greene et al., 2022), a 1,536 dimension multilingual all-purpose embeddings model proposed by OpenAI. This embedding model has proven successful in the insurance field (Mohanan, 2024).

4.1.2 Retrieval

The retrieval step seeks to retrieve a subset of articles using an algorithm that leverages dense word embeddings (i.e. `text-embedding-ada-002`) for retrieval. Our retrieval process is a 3-step process that uses the question as a query. First, the question is encoded using the retriever embedding model. Second, using our dense retriever, we retrieve the `top-5` relevant documents from the reference corpus using cosine similarity to measure the semantic similarity between the query and each document. Third, we merge all relevant reference documents using a context compressor (Cheng et al., 2024). This compressor calls an LLM for each reference using the extracted document (context), the user query, and a formatted prompt that specifies the compressor's task. With this prompt, the LLM is asked to return only the relevant part of the context given the query and, if needed, reformulate the context in certain difficult-to-understand cases, as is sometimes the case with technical legal texts. The compressor reduces the context size, thus keeping the prompt size within an acceptable range, in order to prevent certain issues. Indeed, it is known that excessively large prompts can degrade the quality of answer generation (Levy et al., 2024). Moreover, a lost-in-the-middle effect can cause a language model to omit information contained in the middle portion of the prompt (Liu et al., 2024). This approach helps merge content from different sources to create a better-contextualized reference document for an LLM to generate an answer (Wang et al., 2024).

4.2 Generation

Our generator's goal is to formulate an exhaustive and concise answer to an automotive insurance question based on the information provided by the retrieval process. Our generator uses `GPT-4o`, the latest OpenAI LLM model. The prompt is constructed using the question and the context obtained from the retrieved reference documents, along with specific task instructions designed to guide the LLM in formulating a comprehensive and accurate answer.

As shown in Figure 2 and Figure 3, we have used two prompts for our experiment. The first (Figure 2) is a zero-shot prompt where the LLM is simply asked to answer the question. The second (Figure 3 is a domain-specific prompt that gives additional information to support the LLM. In each prompt, `{input}` corresponds to the question, and `{context}` to the retrieved references.

Répondez à la question suivante EN FRANÇAIS.

Voici la question: {input}.

(a) Basic zero-shot prompt adapted from Kew et al. (2023) followed by the input question to respond to.

Answer the following question IN FRENCH.

Here's the question: {input}.

(b) Translation of the prompt presented in Figure 2a.

Figure 2: Zero-shot prompt used for text generation. Blue boxes contain the task instructions. Yellow boxes contain the prefix for the model to continue.

Vous êtes un expert en assurances automobile dans le domaine de l'assurance de dommages. Vous répondez à des questions EN FRANÇAIS liés à l'assurance automobile AU QUÉBEC. Vous utilisez le contexte fourni ci-bas. Répondez EN PHRASES COMPLÈTES et soyez concis.

Voici la question: {input};
Voici le contexte : {contexte}.

(a) Domain-specific prompt with prompt engineering (i.e. role, task, domain of application) adapted from Kew et al. (2023) followed by the input question to respond to.

You are an automobile insurance expert in the property and casualty insurance field. You are answering questions in FRENCH related to automobile insurance in QUEBEC. Use the context provided below. Answer in FULL PHRASES and be concise.

Here's the question: {input};
Here's the context: {context}.

(b) Translation of the prompt presented in Figure 3a.

Figure 3: Prompt used for text generation. Blue boxes contain the task instructions. Yellow boxes contain the prefix for the model to continue.

5 Experiments

The goal of our experiments is to assess whether an LLM can adequately answer technical questions with complex answers, namely Quebec insurance questions, with or without a RAG architecture. To achieve this, we conduct experiments to automatically and manually evaluate six approaches.

5.1 Experimentation Setup

Baseline For our baseline, we use our zero-shot prompt to assess GPT-4o out-of-the-box capabilities to answer Quebec insurance questions. We label it `Zero-shot` in our result tables.

RAG Architecture Approaches For our other five experiments, we use our RAG architecture described in Section 4 and the domain-specific prompt, with an increasing number of reference

sources. Namely, we start with an approach that uses no references. The difference between this approach and the baseline is only prompt engineering. Then, we incrementally add in reference sources. The next approach only uses the **Laws** source, the following adds the **F.P.Q. 1**, then we add the AMF reference, and finally we add the educative resources to use all four references. We label these five approaches, `No references`, `Laws`, `Laws, F.P.Q. 1`, `Laws, F.P.Q. 1, AMF` and `All references` respectively.

5.2 Evaluation

5.2.1 Automatic Evaluation

Following Chen et al. (2019), we evaluate the accuracy of machine-generated answers compare to reference answers using three N-grams based metrics: BLEU-{1, 4, AVG} (Papineni et al., 2002), ROUGE-{1, 2, L} (F1-Score) (Lin and Och, 2004) and METEOR (Banerjee and Lavie, 2005) scores. We also use two deep similarity metrics that measure the similarity between a machine-generated text and a reference document to compute "how semantically related are those two documents" using words embedding: BERTScore (Zhang et al., 2019), and MeaningBERT (Beauchemin et al., 2023). Each metric uses a slightly different approach to compute this similarity. The first feeds the machine- and human-generated documents separately through a BERT model, then computes a token-by-token alignment between the documents using pairwise cosine similarity. The second, MeaningBERT, uses a fine-tuned pre-trained BERT model train to predict how similar two documents are; the model aims to maximize its correlation with human evaluation. We report the results averaged over five restarts from different random seeds.

5.2.2 Manual Evaluation

To discern the strengths and shortcomings of our generator with or without using an RAG architecture, we conduct a detailed manual analysis of all question-answer pairs. Inspired by Chartier et al. (2024) and Baray et al. (2024), we, in partnership with our insurance partner, have developed an evaluation guide with an exam-like setup to evaluate each pair.Based on the expert answers, we defined a set of criteria, or key elements that a machine-generated answer must include. To evaluate each criterion, we developed a grading scale inspired by the one used by Chartier et al. (2024) and Baray et al. (2024); this scale is presented in Table 2. In case of a false answer to a criterion, we penalize the score with a negative point since an erroneous answer could mislead the customer or hinder their understanding of an insurance product. On the other hand, a complete answer to a criterion results in the maximum score of 2 points. In total, 288 criteria have been extracted from the human answers. On average, each question has 3.51 criteria with a standard deviation of 1.75. The maximum grade a system can receive is $288 \times 2 = 576$ points, and the lowest is -288 points when a system always gives a false answer.

Grade	Description
-1	The system gives a false answer for the criterion i. For example, an answer states that civil liability covers property damage on the insured car if the owner is responsible, which is false.
0	The system does not give a proper answer to the criterion i or give an answer at all.
1	The system gives an incomplete answer to the criterion i.
2	The system gives a complete answer to the criterion i.

Table 2: Our evaluation grading scale to evaluates a machine-generated answer using a set of criteria.

Since we ran 5 runs of each setup with random restarts, we randomly select one of the five for our manual evaluation. One of the authors, with ten years of experience in Quebec Insurance, conducted the evaluation. Appendix A presents the evaluation interface used for our evaluation (in French). During the evaluation, the evaluator is randomly presented with a randomly-selected generated answer from one of the six experimental setups, and he does not know which approach he evaluate.

6 Results

In this section, we present and discuss both our quantitative and qualitative results. We also have conducted an ablation study that also use each source individually in Appendix B.

6.1 Quantitative Results

The left-hand side of Table 3 presents the results of the automatic metrics averaged over the five random restarts, with **bolded** value indicating the best score. First, we observe that, for all automatic metrics, on average, the `All references` approach outperforms other methods. Moreover, this method's BLEU, ROUGE, and METEOR scores

indicate that it gives answers using a vocabulary similar to that of humans in the ground truth. These scores are 40% to 300% higher than the zero-shot baseline approach. It shows that using all our references greatly improves the LLM's ability to answer questions properly. However, surprisingly, the second best approach is the `No reference` approach, which outperforms approaches using the same prompt along with a subset of our references. We hypothesize that using Laws and other juridical documents confuses the LLM and generates longer text that are penalized by automatic N-grams metrics. We will explore and discuss this in the following section.

A second observation is that the approach with the highest variation in performance over the five restarts is `All references`. Indeed, this approach's standard deviation is the highest of all setups, and is nearly three times higher than the lowest one. It indicates that using this approach can also yield suboptimal generations.

Finally, to further assess our approaches' performance, we report the two best approaches z-test significance test in Table 4. Our null hypothesis is that the pair of approaches have equal performances, meaning that values smaller or greater than $|1.96|$ allow us to reject the null hypothesis with $\alpha = 0.05$. A positive value means that the `No references` model (left) performs significantly better than the `All references` (right), and a negative value means the opposite. We can see that for most metrics, `All references` has a significantly better performance compared to `No references`; we can conclude that `All references` is better than `No references`.

6.2 Qualitative Results

The right-hand side of Table 3 also presents the manual grading obtained using our evaluation guide, with **bolded** value indicating the best score. Once again, we observe that `All references` approach outperforms other methods, achieving a score nearly double that of the baseline method.

Moreover, `No references` scores are higher than approaches that use a portion of the references corpus.This seems to indicate that responses from partial references are not just longer but are also incomplete. Indeed, we observed that using legal documents generates longer responses, but the generated answers tend to be of lower quality. For example, to the question "What is the recommended amount of civil liability insurance I should carry when driving outside Quebec?" (translated), the `Laws` model answers with the definition of civil liability instead of responding with the recommended amount of 2 million dollars. In contrast, the `No references` approach answers with the correct amount. It is likely due to data leakage: GPT-4o might have been trained using some of our references and memorized the correct answer. By forcing a different context from incomplete references, the LLM seems to forget or overwite that information.

An interesting situation occurred with the question "I was injured in a car accident. What should I do?" (translated). All evaluated generations take the questions literally and assume the driver has just been injured, and thus propose steps to secure the insuree such as "call an ambulance". In contrast, the ground truth specified the administrative steps to proceed with a bodily injury claim. It shows that, in this case, LLM cannot infer the actual context of the question.

Another interesting situation is whether or not the model abstains from answering in cases where the context is unknown or the information to respond to the question is unavailable for the model. In none of the cases we examined, the model abstained from answering the question. It always strived to be as helpful as possible. However, while sufficient, our prompt could be enhanced to further boost performance. By adapting it to generate better responses and prevent the model from responding when uncertain, we hypothesize that we could improve its performance by improving the prompt.

Moreover, in many cases, without specific references to Quebec insurance specifications, the response contained French insurance information. For example, the `No References` model responded to many questions with specific details of automobile insurance with France-based examples such as civil liability coverage. This pattern disappeared with the addition of the references.

Finally, we can see that the zero-shot approach generates the lowest grade and the highest number of false answers. This highlights the risk of using an out-of-the-box LLM to generate technical answers with precise answer elements, as in our situation. It also highlights that using our RAG approach with our references corpus can lower this risk substantially.While the risk of false answers

	ROUGE			BERTScore	MeaningBERT	BLEU			METEOR	Exam	False
	1	2	L			Average	1	4		Score (%)	Statement
Zero-shot	0.27±0.10	0.09±0.06	0.16±0.06	66.93±4.33	71.42±11.21	4.10±4.25	21.28±10.63	1.39±2.63	24.23±7.61	27.43	34
No references	0.35±0.09	0.14±0.08	0.22±0.07	71.40±4.43	78.17±10.62	7.06±6.16	33.63±14.63	3.05±5.00	27.02±9.68	32.29	20
Laws	0.32±0.1	0.12±0.08	0.20±0.07	70.743±4.54	77.05±11.11	6.177±5.76	31.276±15.05	2.76±5.49	26.29±9.68	27.78	20
Laws, F.P.Q. 1	0.32±0.11	0.13±0.11	0.21±0.1	70.29±5.1	75.44±11.0	6.73±7.47	30.40±15.02	3.30±6.26	27.05±10.42	29.51	19
Laws, F.P.Q. 1, AMF	0.33±0.11	0.14±0.11	0.21±0.1	70.89±5.59	76.91±10.63	7.62±8.02	31.76±16.0	3.93±7.12	27.76±10.78	34.20	18
All references	**0.375±0.14**	**0.18±0.15**	**0.25±0.14**	**71.99±5.9**	**78.87±10.17**	**10.68±11.71**	**33.77±16.66**	**5.98±9.91**	**33.61±14.69**	**51.74**	14

Table 3: Automatic metrics (left) average and one standard deviation over the five restarts on our questions-answering corpus and manual (right) evaluation using our evaluation guide. The best score is **bolded**.

	ROUGE			BERTScore	MeaningBERT	BLEU			METEOR	Exam
	1	2	L			AVG	1	4		Score (%)
No references/All references	**-3.25**	**-3.94**	**-3.28**	-1.57	**-2.47**	**-3.60**	**-4.00**	**-3.49**	**-2.96**	**-2.52**

Table 4: Z-test significance test of our two bests approaches (**bold** value are rejected null hypothesis with $\alpha = 0.05$).

remain present, it is a better way for consumers get easier access to insurance expertise.

No analysis was done as to how the system performs when the question is out of the context of references – does it hallucinate an answer, does it abstain from answering? Would be important to classify what kinds of questions can be answered by the system in order to put guard-rails on it.

6.3 Discussions

Evaluation of RAG systems typically relies on automatic generation N-Grams metrics (Yu et al., 2024). As our results highlight, these metrics provide interesting insight into model performance. Such insight was used to steer the development of the solution. However, the legal field and documents are known to be lengthy and complex (Beauchemin et al., 2020; Beauchemin and Khoury, 2023). Thus, we are skeptical that only relying on this type of metrics is sufficient to develop robust systems; these metrics display an incomplete illustration of the system's response quality and cannot properly capture the legal and misinformation risks they pose to the public. Indeed, ROUGE and BLEU have been criticized for lacking semantic capabilities or correlating weakly with human judgment (Reiter, 2018; Tay et al., 2019; Beauchemin et al., 2023). Moreover, more recent approaches that leverage Transformer-based architecture, such as BertScore, have yet to be shown to achieve a strong correlation with human judgment (Beauchemin et al., 2023). For this reason, many RAG applications now focus on human evaluation (Yu et al., 2024). However, such an evaluation procedure is labour-intensive and costly, especially in specialized fields such as the legal domain. Our primary results show that one can use automatic metrics during development to steer one project. However, human evaluation should evaluate the final system qualitatively to assess a system's performance and risk properly, particularly in sensitive fields such as the legal domain.

7 Conclusion and Future Works

In conclusion, this paper introduced two new corpora: a Quebec automobile insurance expertise reference corpus and a corpus of expert answers to laypeople automobile insurance questions. To generate answers to the questions in our second corpus, we leverage an RAG architecture that uses our reference corpus. We experimented with six approaches: a zero-shot that did not use the RAG architecture, an RAG architecture without references, and four models that incrementally use more of our reference corpus. Our results demonstrate that, on average, using our complete reference corpus generates better responses based on both automatic and manual evaluation metrics. Our results show that between 5% to 13% of generated answers include a false statement that could mislead a customer, indicating that LLM-based technical and sensitive QA is not yet robust enough for mass utilization by the public.

In our future works, we plan to extend the references corpus to include AMF proprietary documents, such as their insurance representative training manual, and increase the number of expert-answered questions. Moreover, we would also like to experiment with other LLMs, and to conduct a real-life evaluation using real insurance customers. Finally, we plan to improve performance with prompt engineering and LLM fine-tuning.

Limitations

First, despite our efforts to make our systems more factually grounded using Quebec insurance references, our proposed framework remains at risk of generating hallucinations in its answers, as shown in Table 3.

Second, since our reference documents are available online, it is possible that GPT-4o and other LLMs could have been trained with some or all of our reference documents. Thus, the results we obtained may include some overfitting, which could make it difficult to generalize to unseen data.

Third, our study is limited to monolingual French documents and QA, and to a single application domain. Though we expect our results to be consistent in other languages and domains, we did not study that question.

Fourth, we acknowledge that our prompt might be considered too simplistic; our focus was not to rabbit-hole ourselves with prompt engineering but instead study the quantitative and qualitative capabilities of out-of-the-box solutions and minimalist RAG to assess the limitations of such technology.

Finally, consistent with prior studies (Krishna et al., 2021; Xu et al., 2023; Louis et al., 2024), we observe that conventional automatic metrics may not accurately mirror answer quality, leading to potential misinterpretations.

Ethical considerations

As highlighted by Beauchemin et al. (2020), the premature deployment of legal NLP solutions, such as an insurance RAG system for the Quebec Insurance domain, poses a tangible risk to laypeople, who may uncritically rely on the answers it provides and thus inadvertently exacerbate their circumstances. Indeed, a layperson might use this kind of innovation as a viable source of information. Thus, the quality of the response needs to be as precise as possible. Furthermore, the use of AI in the legal field poses significant risks because of the presence of bias in corpora and the systems where many might be considered illegal (Bender et al., 2021; Beauchemin and Monty, 2022). We are committed to limiting the use of our dataset strictly to research purposes to ensure the responsible development of legal aid technologies and limit the risk of illegal, biased use.

Hardware & Librairies

Computations are performed on two 12 GO NVIDIA GTX 1080 TI and with proprietary OpenAI LLM and embeddings model using their API; one experimentation over the six approaches cost around 30 USD, the overall OpenAI budget was 1,050 USD.

Acknowledgements

This research was made possible thanks to the support of a Canadian insurance company, NSERC research grant RDCPJ 537198-18 and FRQNT doctoral research grant. We thank the reviewers for their comments regarding our work.

A Evaluation Annotation Interface

Figure 4 presents the evaluation interface used for our evaluation (in French). It is a custom adaptation of the Prodigy annotation tool (Montani and Honnibal, 2018).

B Ablation Study

Table 5 presents the ablation study based on the references used for the RAG, namely using only one source reference instead of the cumulative approach. Our results show that using the cumulative approach yields better results than using only one. We did not conduct the manual evaluation of our ablation study.

References

Ayyoub Ajmi. 2024. Revolutionizing Access to Justice: The Role of AI-Powered Chatbots and Retrieval-Augmented Generation in Legal Self-Help. In *The Brief*, volume 53-10.

Autorité des marchés financiers AMF. AMF approved forms. Accessed online (14-08-2024) https://lautorite.qc.ca/en/professionals/insurers/automobile-insurance/amf-approved-forms.

Autorité des marchés financiers AMF. 2018. *Mémoire présenté à la Commission des finances publiques sur le Projet de loi 141 : Loi visant principalement à améliorer l'encadrement du secteur financier, la protection des dépôts d'argent et le régime de fonctionnement des institutions financières*. Autorité des marchés financiers.

Autorité des marchés financiers AMF. 2019. *Québec Financial Education Strategy for 2019-2022 - Orientations and Action Plan*. Autorité des marchés financiers.

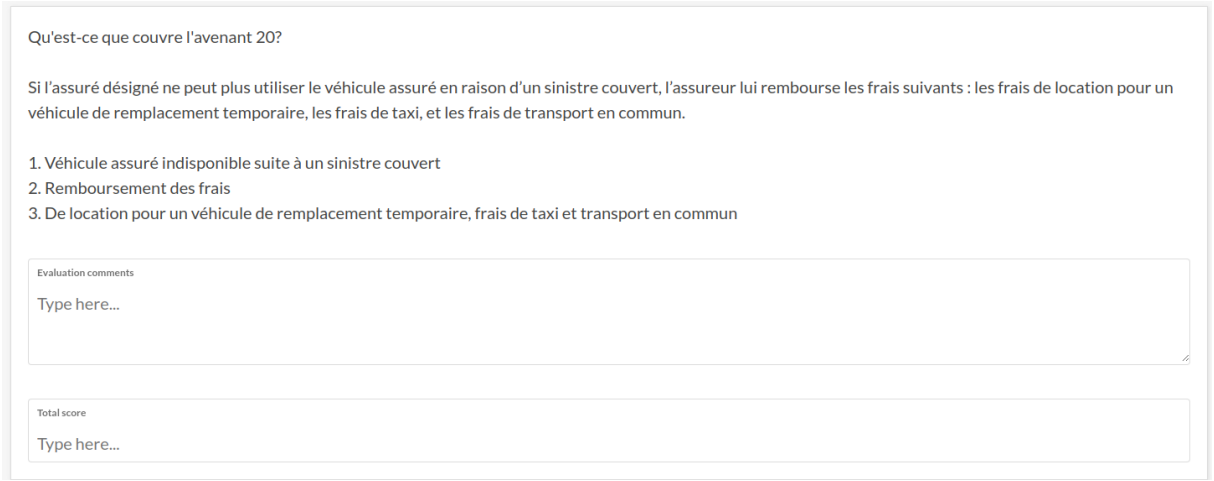

Figure 4: The Prodigy annotation interface (in French) used for evaluation.

	ROUGE			BERTScore	MeaningBERT	BLEU			METEOR
	1	2	L			Average	1	4	
All references	0.375±0.14	0.18±0.15	0.25±0.14	71.99±5.9	78.87±10.17	10.68±11.71	33.77±16.66	5.98±9.91	33.61±14.69
F.P.Q. 1	0.212±0.10	0.09±0.06	0.16±0.06	66.93±4.33	71.42±11.21	4.10±4.25	21.28±10.63	1.39±2.63	24.23±7.61
AMF	0.210±0.19	0.13±0.08	0.19±0.06	70.46±4.54	76.17±10.45	7.06±6.16	33.63±14.63	3.05±5.00	27.02±9.68
Educative Resources	0.240±0.08	0.13±0.07	0.19±0.09	71.26±4.34	77.55±10.22	7.36±5.16	33.17±12.62	3.44±5.01	27.38±9.68

Table 5: Automatic metrics average and one standard deviation over the five restarts on our questions-answering corpus of our ablation study.

Autorité des Marchés Financiers AMF. 2024a. Mission. Accessed online (14-08-2024) https://lautorite.qc.ca/en/general-public/about-the-amf/mission.

Autorité des Marchés Financiers AMF. 2024b. Practice Examination Questions. Accessed online (14-08-2024) https://lautorite.qc.ca/en/becoming-a-professional/damage-insurance/examinations/practice-examination-questions.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation With Improved Correlation With Human Judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jérôme Baray, Alain Decrop, and Gérard Cliquet. 2024. Modèle standardisé d'évaluation des ia génératives en soutien à la recherche marketing: Test de chatgpt. In *Colloque international de l'Association Tunisienne de Marketing*.

David Beauchemin, Nicolas Garneau, Eve Gaumond, Pierre-Luc Déziel, Richard Khoury, and Luc Lamontagne. 2020. Generating Intelligible Plumitifs Descriptions: Use Case Application with Ethical Considerations. In *Proceedings of the International Conference on Natural Language Generation*, pages 15–21.

David Beauchemin and Richard Khoury. 2023. RISC: Generating Realistic Synthetic Bilingual Insurance Contract. *Proceedings of the Canadian Conference on Artificial Intelligence*. Https://caiac.pubpub.org/pub/k18zu6c9.

David Beauchemin and Marie-Claire Monty. 2022. La discrimination en intelligence artificielle est-elle suffisamment encadrée ? Preprint.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: Assessing Meaning Preservation Between Sentences. *Frontiers in Artificial Intelligence*, 6.

Belairdirect. 2024. FAQ. Accessed online (14-08-2024) https://www.belairdirect.com/en/faq.html.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM conference on fairness, accountability, and transparency*, pages 610–623.

Beneva. 2024. FAQ - Car Insurance. Accessed online (14-08-2024) https://www.beneva.ca/en/car-insurance/help.

Umar Butler. 2023. Open Australian Legal QA.

Chambre de l'Assurance de Dommages ChAD. 2024. About Us. Accessed online (14-08-2024) https://chad.ca/en/about-us/.

Mathieu Alexandre Chartier, Nabil Dakkoune, Guillaume Bourgeois, and Stéphane Jean. 2024. Évaluation des capacités de réponse de larges modèles de langage (llm) pour des questions d'historiens. In *24ème conférence francophone sur l'Extraction et la Gestion des Connaissances*, 40, pages 155–166.

Harrison Chase. 2022. LangChain.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating Question Answering Evaluation. In *Proceedings of the workshop on machine reading for question answering*, pages 119–124.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token. *arXiv:2405.13792*.

Bourget Claire, Lacombe Marie-Eve, Godbout René, Lanctôt Sébastien, Rajaobelina Lova, Ducharme Guillaume, Lavoie Annie, and Maynard Marie-Guy. 2018. *Assurance de dommages à l'ère du numérique*. Centre facilitant la recherche et l'innovation dans les organisations.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges. *IEEE Access*.

Desjardins Assurances. 2024. Insurance FAQ. Accessed online (14-08-2024) https://www.desjardins.com/qc/en/insurance/faq.html.

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying Deep Learning to Answer Selection: A Study and an Open Task. In *IEEE workshop on automatic speech recognition and understanding*, pages 813–820. IEEE.

Rudolf Flesch. 1948. A Readability Formula in Practice. *Elementary English*, 25(6).

Groupement des Assureurs Automobiles GAA. 2024. About Us. Accessed online (14-08-2024) https://gaa.qc.ca/en/who-are-we/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997*.

Ryan Greene, Arvind Neelakantan, Lilian Weng, and Ted Sanders. 2022. New and Improved Embedding Model.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. SpaCy: Industrial-strength Natural Language Processing in Python.

Insurance Bureau of Canada IBC and Groupement des Assureurs Automobiles GAA. 2024. Infoassurance - About Us. Accessed online (14-08-2024) https://infoassurance.ca/en/utility-menu/about-us/.

Intact Insurance. 2024. FAQ. Accessed online (14-08-2024) https://www.intact.ca/en/faq.

Christopher Johnson. 2018. Projet de loi 141 et vente par internet: où en est le RCCAQ? https://www.rccaq.com/cgi/page.cgi/_article_fr.html/Categories/Dans_la_mire/Projet_de_loi_141_et_vente_par_internet_o_en_est_le_RCCAQ_.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural Language Processing in the Legal Domain. *Available at SSRN 4336224*.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. *arXiv:2310.15773*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957. Association for Computational Linguistics.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models.

Chin-Yew Lin and FJ Och. 2004. Looking for a Few Good Metrics: ROUGE and Its Evaluation. In *Ntcir workshop*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable Long-Form Legal Question Answering With Retrieval-Augmented Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Jorge Martinez-Gil. 2023. A Survey on Legal Question–Answering Systems. *Computer Science Review*, 48:100552.

Monisha Mohanan. 2024. *Competitive Analysis of Embedding Models in Retrieval-Augmented Generation for Indian Motor Vehicle Law Chat Bots*. Ph.D. thesis, Dublin Business School.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A Modern and Scriptable Annotation Tool for Creating Training Data for Machine Learning Models.

Seon-Ok Na, Young-Min Kim, and Seung-Hwan Cho. 2022. Insurance Question Answering via Single-turn Dialogue Modeling. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 35–41.

Mohammad Nuruzzaman and Omar Khadeer Hussain. 2020. IntelliBot: A Dialogue-Based Chatbot for the Insurance Industry. *Knowledge-Based Systems*, 196:105810.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain.

Quebec. 2016. Code de la sécurité routière 2016.

Quebec. 2024. Loi sur l'assurance automobile 2024.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Recueil des lois et des règlements du Québec RLRQ. 2004. Act Respecting the Regulation of the Financial Sector.

Société de l'Assurance Automobile du Québec SAAQ. 2024. Québec's Public Automobile Insurance Plan in Brief. Accessed online (14-08-2024) https://saaq.gouv.qc.ca/en/traffic-accident/public-automobile-insurance-plan/in-brief.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 1048–1064. Association for Computational Linguistics.

Sonnet. 2024. Frequently Asked Questions. Accessed online (14-08-2024) https://www.sonnet.ca/faqs.

Wenyi Tay, Aditya Joshi, Xiuzhen Jenny Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-Faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation. In *Proceedings of the Annual Workshop of the Australasian Language Technology Association*, pages 52–60.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024. Searching for Best Practices in Retrieval-Augmented Generation. *arXiv:2407.01219*.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Nirmalie Wiratunga and Ashwin Ram. 2011. *Case-Based Reasoning Research and Development*. Springer.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3225–3245. Association for Computational Linguistics.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Leah Li. 2024. Financial Report Chunking for Effective Retrieval Augmented Generation. *arXiv:2402.05131*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv:2405.07437*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Éducaloi. 2024. Governance. Accessed online (14-08-2024) https://educaloi.qc.ca/en/governance/.