Computational Comparative Analysis of Global Water Legislation: An NLP and LLM-Based Framework for Cross-Jurisdictional Policy Assessment

by

Adilkhan Alikhanov

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

February 2026

© Nazarbayev University 2026. All rights reserved.

Author .....
Department of Data Science
February 2026

Certified by .....
Siamac Fazli
Associate Professor
Thesis Supervisor

Accepted by .....

Dr. Elizabeth Arkhangelsky

Dean, School of Engineering and Digital Sciences

Abstract

The research outlined within this dissertation provides an approach to analyzing international water legislation by using a computational pipeline to process water legislation from 165 different countries written in over 35 different languages and represented by over 10 different writing systems. The computational pipeline included seven steps: extracting the text from documents, translating that extracted text into English, evaluating the quality of those translations based on multiple metrics, utilizing a large language model to extract legal information from the translated text, calculating the similarities between each piece of legislation utilizing embedded representations of the text, and finally clustering these similar pieces of legislation together to identify patterns of similarity among them. This computational pipeline shows how automated methods may provide an extension to the existing manual comparative tradition in water law research, allowing researchers to analyze large amounts of data that would be impossible to compare manually. Important findings from this project were: (1) that the quality of the translation was sufficient enough to allow for meaningful comparison in the majority of the sample set (based on COMET reference-free quality estimation the mean score was 0.83); however, it was determined that there existed a phenomenon referred to as "contextual flattening," where low resource languages had been reduced to a flat context that did not take advantage of the linguistic complexity present in the original language; (2) that the large language model-based extraction pipeline was able to extract all relevant information regarding three dimensions of water law policy—groundwater regulation, river basin management, and polluter-pays principle—with 100% compliance with the schema; (3) that cluster analysis revealed five distinct typologies of water law that corresponded with some extent to traditional classifications of legal families but also indicated cross-traditional convergence in basin-based governance practices; and (4) that the polluter-pays principle was found to be the most frequently used mechanism of implementation although it was never explicitly mentioned in any of the examined country profiles. The methodology presented in this dissertation will serve as the basis for future research involving the use of computational comparative law in areas outside of the water sector.

Keywords: computational comparative law, water legislation, natural language processing, large language models, machine translation evaluation, hierarchical clustering, text embeddings, policy analysis

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Global Water Governance Challenge

Water is arguably the single most important resource for human existence, yet water is available to us in limited quantities and we continue to experience the greatest challenge of the twenty-first century regarding its governance and management. According to the United Nations, there are currently 2.2 billion people in the world who do not have access to clean drinking water and 4.2 billion people (more than half of the total world population) without access to safely managed sanitation [1]. Climate Change has increased pressure on both the amount of freshwater available to us and its quality. Changes in precipitation, melting glaciers and increasing ocean levels are changing the way the hydrologic cycle works and threatens both the amount and quality of freshwater on every continent [2]. Freshwater demands are expected to increase by 20–25% globally by 2050, due to population growth, urbanization and expanding agriculture; however, at the same time, the constraints on the supply side are also tightening [3].

The sheer magnitude of the crisis is further emphasized by the interrelationships between water scarcity and other development challenges. Approximately 70 percent of global freshwater usage is attributed to agriculture and competition between agricultural, industrial and domestic uses is growing in virtually all large river basins [1]. Over 40 percent of irrigated agriculture around the globe depends on groundwater; however, in many areas groundwater is being used at unsustainable rates, causing decline in aquifer levels in some areas of South Asia, the Middle East and North Africa [2]. Transboundary water resources—water resources shared by two or more sovereign states in 310 international river basins and 468 transboundary aquifer systems—add a geopolitical component to the governance challenge, requiring legal frameworks capable of operating within multiple jurisdictional boundaries. The combination of these stresses make the study of how governments create and enforce laws and regulations governing their water resources not just an academic exercise but an issue of immediate

practical necessity.

In light of this context, the importance of national legislation in managing and regulating water resources becomes evident. Water laws create the legal framework by which societies allocate water rights, regulate water extraction and use, protect water quality, manage river basins and aquifers and resolve disputes between competing users. It is widely recognized that effective water legislation is a necessary condition for achieving Sustainable Development Goal 6 (clean water and sanitation for all) and for improving resilience to the water-related effects of climate change [1]. The quality, comprehensiveness and enforceability of national water laws vary greatly across the world's more than 190 sovereign states and reflect differences in legal traditions, institutional capacity, hydrologic conditions and development priorities.

The consequences of poor water governance are well documented. In areas where water legislation is weak or not enforced, unregulated groundwater extraction has resulted in the depletion of aquifers, the sinking of land and the salination of freshwater supplies. Similarly, in areas where there are no effective provisions for controlling pollution, industrial and agricultural discharges have degraded surface water and groundwater quality, resulting in significant costs for downstream users and ecosystems that are often difficult to repair. Further, in areas where there are no river basin management frameworks in place, upstream-downstream conflicts over water allocation are not resolved, and result in a destabilization of both economic development and social stability. Therefore, in order to identify both best practices and governance deficits, it is imperative to understand the global landscape of water legislation—which countries provide a comprehensive regulatory structure, which countries only address specific aspects of governance, and what type of legislative approach is typically associated with a given legal tradition or hydrologic context.

Despite the importance of water legislation being widely recognized, our knowledge of how various countries regulate their water resources is still incomplete and disjointed. Comparative water law research—initiated by Dante Caponera at the Food and Agriculture Organization of the United Nations (FAO) and continued by researchers such as Stefano Burchi—has generated substantial qualitative analysis of water legislation in select jurisdictions [4, 5]. However, comparative water law research has been limited by the inherent characteristics of manual analysis. Typically, comparative water law research focuses on less than 20 countries, is restricted to languages available to the researcher, and requires significant amounts of expert time to complete. As such, there has never been a comprehensive global-level comparison of water legislation. This leaves many fundamental questions unanswered: Do national water laws group together into recognizable categories? Have the diffusion of governance paradigms (such as Integrated Water Resources Management) across the globe resulted in similar forms of legislation? How do colonial legal heritage, hydrologic conditions and development

priorities influence the content of water legislation in different regions?

The unanswered questions reflect a deeper issue; there are insufficient methodologies for comparative water law scholars to address this large-scale challenge. To accomplish this we need to move away from the limitation of a manual or small sample comparison to something that will be more systematic. This next section addresses the reasons behind how difficult it has been to scale up comparative water law, and the potential of recently developed technology to provide a new way to approach this challenge.

1.2 The Challenge of Comparative Water Law at Scale

The aspiration to compare water legislation across all the world's jurisdictions confronts formidable practical obstacles. More than 190 sovereign states maintain independent legal systems, each producing water legislation in its own official language or languages. The corpus of global water law spans more than 35 languages and at least 10 distinct script systems, including Latin, Arabic, Cyrillic, Georgian (Mkhedruli), Ethiopic (Ge'ez), Lao, Thaana, Hebrew, Greek, and Hangul. These languages are embedded in diverse legal traditions—civil law, common law, Islamic law, and customary law—each of which employs distinct drafting conventions, terminological frameworks, and structural organisations for legislative texts.

The manual approach to comparative water law, for all its scholarly rigour, cannot scale to meet this challenge. A single expert might reasonably compare the water legislation of 10–20 countries within a common legal tradition and language family. Extending this comparison to 100 or more countries, spanning dozens of languages and multiple legal traditions, would require a team of multilingual legal experts working for years—an investment that no research institution or international organisation has been willing or able to make. The challenge is compounded by the heterogeneity of document formats in which legislative texts are published: some countries maintain well-organised digital legal databases with machine-readable text, while others publish their legislation only as scanned PDF documents, sometimes of poor quality, requiring optical character recognition before any analysis can begin. The availability of digital legislative texts also varies dramatically across regions, with Sub-Saharan Africa, Central Asia, and the Pacific Islands presenting particular gaps. The result is that global water governance is studied primarily through case studies, regional surveys, and thematic reviews, with no systematic computational approach to the full breadth of the world's water legislation.

Recent advances in natural language processing (NLP), machine translation, and large language models (LLMs) offer a potential path through this impasse. Cloud-based document processing systems can extract text from PDFs in any language and script.

Neural machine translation systems, while imperfect, can produce serviceable translations across hundreds of language pairs. Large language models can be prompted to extract structured information from unstructured legal texts with a degree of accuracy that, while not yet matching expert human performance, is sufficient to support large-scale exploratory analysis. Embedding models can represent legal texts as dense numerical vectors, enabling the computation of semantic similarity across documents regardless of their original language. And hierarchical clustering methods can identify patterns of similarity and difference in high-dimensional data without requiring prior assumptions about the number or nature of the groups.

The convergence of these technologies creates, for the first time, the technical conditions under which a genuinely global comparison of water legislation becomes computationally tractable. Each individual technology addresses a specific barrier that has constrained manual comparative law: document AI overcomes the format heterogeneity of legislative texts (scanned PDFs, born-digital documents, multi-column layouts); machine translation overcomes the language barrier that has restricted comparative studies to researchers' personal linguistic repertoires; LLM-based extraction overcomes the laboriousness of manually identifying and cataloguing legal provisions across hundreds of documents; and embedding-based similarity analysis overcomes the difficulty of quantifying conceptual proximity between legal texts drafted in different traditions and vocabularies.

No prior work has attempted to integrate these technologies into a unified framework for global-scale comparative legal analysis. The present thesis does exactly this, developing and applying a computational pipeline that processes water legislation from 164 countries across more than 35 languages to produce the first comprehensive computational comparison of national water law.

It is now possible to outline what this research will answer with regard to the governance challenges identified, and the computational tools used to address those challenges. The following research questions reflect this layered approach by moving from technical feasibility (i.e., do the tools work?) to substantive findings (i.e., what does the output tell us about governance?) to methodological self-assessment (i.e., how effective were the tools in producing useful data for analysis?).

1.3 Research Questions

There are four main research questions in the thesis.

RQ1: *Can a computational pipeline reliably process, translate, and compare water legislation from 164+ countries across 35+ languages?* This question addresses the technical feasibility and reliability of the pipeline.

RQ2: *What global patterns emerge in groundwater management, river basin governance, and polluter-pays principle adoption across national water legislation?* This question addresses the substantive legal findings of the study.

RQ3: *Do national water laws cluster into identifiable typological groups based on their policy content?* This question is concerned with classifying national water laws into typological groups based on their policy content.

RQ4: *How does machine translation quality affect the reliability of cross-lingual legal comparison?* This question addresses the methodological risk associated with machine translation in the study.

These four research questions have been designed to build upon one another. The answer to RQ1 will determine if the computational infrastructure is sufficient to support substantive legal analysis. RQ2 will use the same infrastructure to provide empirical findings regarding the content of global water legislation. RQ3 will move from descriptive findings to analytical classification by determining if the observed diversity can be reduced to meaningful typological groupings. RQ4 will address the methodological limitations of the study in order to establish the evidence base for calibrating confidence in the findings of RQ2 and RQ3. Together, these four questions define a research program that spans from technical feasibility through substantive legal discovery to methodological self-assessment.

1.4 Contributions

The thesis makes four principal contributions to the fields of computational legal analysis and comparative water law:

1. **First global-scale computational comparative water law framework.** The thesis introduces the first fully computational comparative water law framework that encompasses 164 countries and over 35 source languages. The computational comparative framework described in this thesis is intended to complement rather than replace the manual comparative tradition established by Caponera and Burchi.

2. **Multi-metric translation fidelity framework.** The thesis develops a multi-metric framework for measuring the quality of machine translated legal texts, combining COMET-based reference-free quality estimation with embedding-based semantic fidelity analysis and manual verification. This framework quantifies the degree of information loss introduced by machine translation and identifies a phenomenon termed *contextual flattening*, where translations are linguistically fluent but semantically impoverished.

3. **LLM-based legal extraction pipeline.** The thesis develops a legal extraction pipeline using a large language model that can extract structured legal features from water legislation across diverse legal traditions and languages with sufficient accuracy to support quantitative comparative analysis.

4. **Empirical water law typology through clustering.** The thesis develops an empirically grounded typology of water law through the application of hierarchical clustering to similarity matrices derived from embeddings, identifying recurring patterns of legislative attention to groundwater management, river basin governance, and pollution control that provide a new lens for understanding global water governance.

1.5 Thesis Organisation

The thesis is organized in the following way. Chapter 2 discusses the literature that forms the basis of this thesis, focusing on comparative water law, legal NLP, machine translation evaluation, embedding-based document similarity, and hierarchical clustering, and identifies the research gap at the intersection of these literatures. Chapter 3 explains the construction of the global water legislation corpus, including document sourcing, text extraction, translation, and quality verification, and sets out the empirical foundation for the analysis. Chapter 4 explains the computational methodology in detail, relating each pipeline stage to the research questions it addresses, from LLM-based legal information extraction through embedding-based similarity analysis to hierarchical clustering. Chapter 5 presents the empirical results regarding the three policy dimensions—groundwater regulation, river basin management, and the polluter-pays principle—and describes the clustering analysis that produces five distinct water law typologies. Chapter 6 discusses the findings in the context of comparative water law scholarship, examines the methodological limitations of the study particularly in relation to the quality of machine translation, and considers implications for water governance policy. Chapter 7 summarizes the contributions of the thesis, identifies the constraints on interpreting them, and sets out the most promising directions for further research.

Chapter 2

Literature Review

This Chapter outlines the theoretical bases that support a computationally-based approach to comparing water laws of 164 countries using five separate but interconnected areas: comparative water law; Natural Language Processing (NLP) of legal texts; Machine Translation Quality Evaluation; Document Similarity via Embedding-Based Methods; and Hierarchical Clustering. Each of the five areas have developed independently of one another and there is no prior work that has combined them to form a single, computational framework for comparing water legislations of all countries at a global level. The remainder of this Chapter traces the development of each area; identifies key methods used and their limitations; and articulates the specific gaps in current approaches that the Thesis will address.

The five components are logically related to each other rather than being independently related to each other as separate components. Comparative water law defines the necessity for interjurisdictional analysis at scale in terms of substantive content. Legal NLP offers methods for extracting information from heterogeneous legislation. Machine translation evaluation deals with the language barriers present within a multi-jurisdictional data set. Embedding-based similarity allows for quantitative comparisons across jurisdictions. Hierarchical clustering generates typology by which patterns in legislative can be found. However, it is important to note that there are known constraints or limits associated with the concepts underpinning these relationships. In comparative legal research, legal family categories (civil law, common law and Islamic law) have been criticized as too simplistic of conceptualizations of more complex hybrid systems. Similarly, the IWRM framework, although adopted in many jurisdictions has been criticized for its imposition of a one size fits all approach to governance across different institutional contexts. Additionally, the use of LLMs as an extraction tool carries the potential for hallucinations and the possibility that output that is structurally correct may still not represent semantic accuracy.

2.1 Comparative Water Law

Comparative water law has a long history as a scholarly discipline based on recognition that water is a shared and essential resource requiring legal frameworks that extend beyond national borders. While the body of work in comparative water law is broad, the foundational work is attributed to Dante Caponera who was employed by the Food and Agriculture Organization of the United Nations (FAO) and produced the most comprehensive treatment of water law across jurisdictions. His seminal book, *Principles of Water Law and Administration: National and International*, first published in 1992 and revised posthumously by Marcella Nanni in a second edition [4] has become the definitive source for understanding how different countries regulate water resources. Caponera's contributions can be identified in three ways: establishing a taxonomy of water rights regimes; documenting the historical development of water law from Roman and Islamic traditions to modern codification; and compiling legislative provisions for dozens of countries. Importantly, Caponera showed that regardless of the vast differences in legal traditions—civil law, common law, Islamic law and customary law—there are a number of recurring regulatory issues that appear universally, including the allocation of water rights, the protection of water quality, and the management of shared water courses.

Building on the work of Caponera, Stefano Burchi has made important contributions to the comparative analysis of water legislation as part of the work of the FAO and the International Association for Water Law (AIDA). Through the FAO Legislative Study series, Burchi has written about the legal frameworks for groundwater governance [6]; the regulation of water rights and permits [7]; and the transposition of international water law principles into national legislation [5]. Burchi's research is notable for its focus on the implementation gap—the difference between legislative provisions on paper and their actual implementation in practice. In addition to his research on comparative water law, Burchi has assisted in the development of model water codes and legislative guidelines that have influenced water law reform in developing countries [8].

An important institutional resource for scholars of comparative water law is the FAO FAOLEX database (previously LEGIS) which contains legislative texts and policy documents related to food, agriculture and natural resources, including water, from countries world-wide [9]. The FAOLEX database has provided a major repository for researchers undertaking comparative studies of water legislation across jurisdictions, and it has facilitated cross-country comparisons of water legislation at the same time as providing the opportunity for the type of large-scale cross-country comparisons undertaken by this Thesis.

Comparing water legislation across legal traditions is known to be difficult. For

example, civil law systems, which exist in much of continental Europe, Latin America, and much of Africa, typically contain water regulations within comprehensive codified frameworks. Common law systems, existing in the UK, USA, Australia and former British colonies, rely more heavily on riparian doctrines, prior appropriation principles, and case law. Islamic water law, which is derived from Sharia principles and the concept of *shafa*, gives fundamental rights to water access while regulating the distribution of water through community-based mechanisms [4]. Customary water law, which is still operational in much of Sub-Saharan Africa and the Pacific Islands, governs water access through un-written norms enforced by traditional authorities [6]. The fact that the legal traditions differ in substantive and structural terms—and in terms of vocabulary and philosophy—makes the task of manually comparing the content of water legislation from around the world very difficult.

There have been two conceptual frameworks that have had a particular influence on structuring modern water legislation around the world and therefore structuring comparative study. The Dublin Principles, formulated at the International Conference on Water and the Environment in 1992, outlined four guiding principles: that fresh-water is a finite and vulnerable resource; that water development and management should be participatory; that women play a central role in water provision and safeguarding; and that water has an economic value and should be recognized as an economic good [10]. The Dublin Principles stimulated the shift towards the paradigm of Integrated Water Resources Management (IWRM) which advocates for the coordinated development and management of water, land and related resources to maximize economic and social welfare while ensuring environmental sustainability [11]. IWRM has been widely adopted as a guiding framework for water law reform, especially in developing countries and its influence is reflected in legislative provisions that require river basin organisations, stakeholder involvement and environmental flows [12].

While there has been a considerable amount of scholarly literature on comparative water law, this field of study remains almost exclusively a manual endeavour. Scholars such as Caponera, Burchi and others have based their analyses on intensive study of water legislation from multiple legal traditions, and have made comparisons between the various pieces of legislation in a piecemeal way. There is no previous work that has attempted to use computational or automated methods for comparative water law at a global scale. Given that there are approximately 190 sovereign states, each with its own water legislation—often in languages other than English and with different legal traditions—the manual approach is inherently limited in scope, reproducibility and scalability. The present Thesis aims to address this limitation directly by establishing a computational pipeline that enables the processing, translation, analysis and comparison of water legislation from 164 countries.

It is simply a matter of fact that comparative water law lacks computational ap-

proaches - it is not for a lack of available tools. In the last ten years, the area of Natural Language Processing (NLP) has produced many techniques which in theory could be used for the type of large scale comparisons of legal documents that would be impossible for the manual scholar to accomplish. This section reviews some of this recent development with an emphasis on Legal NLP and how Large Language Models can be employed as a tool for extracting structured information from legislative texts.

2.2 Natural Language Processing in Legal Analysis

Natural language processing (NLP) has emerged as a transformative tool for legal analysis, giving rise to the subfield variously termed "Legal NLP," "Legal AI," or "computational legal analysis." Legal texts present unique challenges for NLP systems: they employ archaic and highly formalised language, domain-specific terminology with precise meanings that may diverge from ordinary usage, complex syntactic structures including nested subordinate clauses and cross-references, and culturally embedded concepts that resist straightforward translation [13, 14]. These characteristics have motivated the development of specialised models, datasets, and evaluation benchmarks for the legal domain.

A major breakthrough in the development of legal NLP models came with the introduction of LegalBERT by Chalkidis et al. [13], which demonstrated the effectiveness of pre-training a BERT-based language model on a large dataset of legal text drawn from sources including contracts, court judgments and statutes from the US, EU and UK. Pre-training LegalBERT on legal text substantially improved the performance of LegalBERT on several downstream legal NLP tasks including text classification, named entity recognition and question answering compared to a BERT model pre-trained on a general corpus of text. LegalBERT demonstrated that training a model in a legal domain significantly improves its performance on tasks related to the legal domain, a finding that has since been supported by studies examining other models such as CaseLaw-BERT [15] and legal-domain versions of multilingual models.

Contract analysis and information extraction represent the most commercially developed applications of legal NLP. Systems for automated contract review extract key provisions—parties, dates, obligations, termination clauses—using a combination of rule-based patterns, named entity recognition (NER), and relation extraction [16]. The Contracts Understanding Atticus Dataset (CUAD), introduced by Hendrycks et al. [16], demonstrated that large language models could achieve accuracy approaching that of experienced lawyers in identifying 41 different types of legally important clause in contracts when the models were trained on a sufficiently large dataset of contracts.

Zero-shot and few-shot classification methods have opened new possibilities for legal text categorisation without the need for large annotated datasets. By leveraging pre-

trained language models' capacity to generalise from natural language task descriptions, researchers have applied zero-shot classification to legal topic identification, jurisdiction detection, and regulatory provision categorisation [17]. The LexGLUE benchmark, introduced by Chalkidis et al. [17], provides a standardised evaluation framework for legal NLP models across multiple classification tasks, facilitating systematic comparison of approaches.

Large language models (LLMs), particularly the GPT family developed by OpenAI, have transformed the field of legal NLP. GPT-4 and its successors have demonstrated remarkable capabilities in legal text summarisation, question answering, and structured information extraction [18]. Katz et al. [18] showed that GPT-4 could pass the Uniform Bar Examination, scoring in the 90th percentile, a result that underscored the models' capacity for legal reasoning. More relevant to the present thesis, LLMs have proven effective at extracting structured information from unstructured legal texts when guided by carefully designed prompts. Prompt engineering for legal information extraction involves specifying output schemas—typically JSON structures with predefined fields—providing examples of desired outputs, and incorporating domain-specific instructions that guide the model's attention to relevant legal concepts [19]. The use of structured JSON outputs with schema validation ensures that extracted information conforms to a consistent format amenable to downstream computational analysis.

To date, no previous study has systematically applied LLMs to the task of extracting structured policy features from water legislation across 164 countries written in over 35 languages, nor have existing studies of legal NLP focused primarily on contracts, court opinions and regulatory texts from a single jurisdiction or linguistic tradition. The primary objective of the present thesis is therefore to develop an LLM-based extraction system for legal information across three policy dimensions (the "AquaLex Scrutinizer") applicable to the full range of countries with water legislation.

The multilingual nature of this task introduces a dependency on machine translation, the quality of which must be rigorously evaluated.

2.3 Machine Translation Quality Evaluation

Machine translation (MT) quality evaluation is a mature but rapidly evolving field that has moved from purely lexical metrics toward learned evaluation models capable of assessing translation adequacy at the semantic level. The foundational automatic evaluation metric for machine translation is the Bilingual Evaluation Understudy (BLEU) score, introduced by Papineni et al. [20]. BLEU measures the overlap of n-grams between a machine-generated translation and one or more human reference translations, producing a score between 0 and 1. Despite its widespread adoption, BLEU has well-known limitations: it correlates poorly with human judgement for morphologically rich

languages, it penalises legitimate paraphrases, and it is insensitive to errors in meaning preservation that do not affect surface-level n-gram overlap [21]. These limitations are particularly acute for legal texts, where semantic fidelity is paramount and where multiple phrasings may convey identical legal obligations.

The shortcomings of surface-level metrics motivated the development of learned evaluation models that assess translation quality using neural representations. The Crosslingual Optimized Metric for Evaluation of Translation (COMET) framework, introduced by Rei et al. [22] and refined in subsequent work [23], represents the current state of the art in MT evaluation. COMET models are trained on human quality judgements from the Conference on Machine Translation (WMT) shared tasks and learn to predict human adequacy scores from source-translation-reference triplets. Crucially, the COMET framework also supports reference-free quality estimation through models such as `Unbabel/wmt22-cometkiwi-da` [24], which assess translation quality using only the source text and the translation, without requiring a human reference. This capability is essential for scenarios—such as the present thesis—where reference translations are unavailable for the vast majority of the documents under analysis.

The challenge of translating legal texts extends beyond general-purpose MT evaluation. Legal translation must preserve not only the surface meaning of individual sentences but also the precise legal concepts, obligations, and rights encoded in the source text [25]. Domain-specific terminology poses a particular challenge: terms such as "riparian rights," "prior appropriation," "usufruct," or "police des eaux" carry precise legal meanings that may lack direct equivalents in the target language. Legislative structure—including the hierarchical organisation of articles, paragraphs, and subparagraphs—must also be maintained to preserve the internal cross-references that are characteristic of legal drafting.

A further concern, identified in recent work on MT for specialised domains, is the phenomenon of "contextual flattening" or "semantic drift," whereby machine translation systems produce fluent but semantically impoverished translations that lose domain-specific nuance [26]. In the context of water legislation, this might manifest as the translation of a specific groundwater permit requirement into a vague statement about water management, preserving general meaning while losing the regulatory specificity that is the object of comparative analysis.

Despite the maturity of MT evaluation as a field, limited work has addressed the specific challenge of evaluating machine translation quality for legislative and legal documents across a large number of language pairs. The present thesis confronts this gap by deploying the COMET framework—specifically the `wmt22-cometkiwi-da` reference-free quality estimation model—to evaluate translations of water legislation from over 35 source languages into English, thereby establishing a quantitative basis for assessing the reliability of the translated corpus that underpins all subsequent analysis.

After the translation process has produced a set of translations, which were then evaluated for quality, the next question is how you are going to use those translations (and evaluate them) for comparison purposes. It would be impossible to read 164 different translated water law documents side by side for an evaluation of comparisons. Instead, there needs to be a way to represent each of these documents as numbers, so that they may be compared objectively using numerical methods rather than subjective evaluations. That is the purpose of text embeddings in the pipeline, and this is where we will review the literature on document similarity based on text embeddings.

2.4 Embedding-Based Document Similarity

Text embeddings—a way of representing text as dense, numerical vectors—have become the new standard for computing semantic similarity in Natural Language Processing (NLP). This area has been developing since early distributional semantics, to the current high-dimensional, contextual embeddings driving the latest similarity metrics.

Word2Vec, presented by Mikolov et al. [27], showed how shallow neural networks trained on word co-occurrence patterns can generate vector representations, where semantic relationships can be encoded as geometric relationships: the often cited example of $\text{vec}(\text{king}) - \text{vec}(\text{man}) + \text{vec}(\text{woman}) \approx \text{vec}(\text{queen})$ demonstrated how these representations can capture analogical reasoning. GloVe (Global Vectors for Word Representation), developed by Pennington et al. [28], achieved similar results using a global matrix factorization approach that combines the benefits of count-based and prediction-based methods. Although Word2Vec and GloVe generated static embeddings (a single vector per word, regardless of context), they provided the fundamental insight that semantic similarity can be operationalized as vector proximity in a learned embedding space.

BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin et al. [29], marked a paradigm shift toward contextual embeddings, where the representation of each token depends on the context in which the token appears. However, the BERT architecture was not optimized to generate sentence-level embeddings that would be useful for computing similarities. The limitations of BERT in generating sentence-level embeddings were alleviated by the development of Sentence-BERT (SBERT) [30], which fine-tuned BERT using a siamese network architecture on natural language inference and semantic textual similarity datasets. SBERT made it feasible to efficiently generate semantically meaningful sentence embeddings, enabling the effective computation of similarities among large document collections. Models available in the sentence-transformers library, including `all-mpnet-base-v2` and `all-MiniLM-L6-v2`, have become standard tools for semantic search and document similarity tasks.

Commercially available embedding models have further advanced the capabilities

of representation quality. The OpenAI `text-embedding-3-large` model generates embeddings with dimensionality of up to 3,072 and has achieved state of the art performance on standard retrieval and similarity benchmarks [31]. High dimensional embeddings provide a fine grain level of detail to represent subtle semantic distinctions that are important in many applications, including legal text comparison. In legal text comparison, differences in wording can result in significantly different interpretations of regulatory intent.

The standard metric for comparing embedding vectors is cosine similarity, defined as follows:

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{2.1}$$

where $\mathbf{a}$ and $\mathbf{b}$ are embedding vectors. The cosine similarity is bounded between -1 and 1 (or 0 and 1 for non-negative embeddings), is independent of the magnitude of the vectors, and provides an interpretable measure of semantic alignment. Due to these properties, cosine similarity is particularly suited for comparing documents of different lengths, as it captures directional similarities rather than absolute distances in the embedding space.

To enable efficient similarity searches over large collections of embeddings, the Facebook AI Similarity Search (FAISS) library [32] provides optimized implementations of approximate nearest neighbor algorithms that can scale up to billions of vectors. FAISS supports multiple index types, including flat (exact) indices, inverted file indices with product quantization, and hierarchical navigable small world graphs; thus, enabling tradeoffs between search accuracy and computational efficiency that are important when comparing large numbers of documents.

Similarity using embeddings has been extensively applied in document retrieval, semantic search, duplicate detection, and cross-language information retrieval [30]. However, there is no previous work applying embeddings to compare legal systems across jurisdictions—by using embeddings to quantify the degree of legislative attention each country directs toward specific dimensions of water governance.

The similarity scores computed using the embedding approach will provide input to the final analytical step: grouping countries into clusters based on their legislative profiles, which requires techniques from unsupervised machine learning.

2.5 Hierarchical Clustering in Text Analysis

Hierarchical clustering techniques offer a suitable framework for discovering structure in high-dimensional data without requiring specification of the number of groups prior to application. Among hierarchical agglomerative techniques, Ward's minimum variance criterion [33] has emerged as the most widely used technique for text and document

clustering. Ward's method merges clusters based on minimizing the total within cluster variance at each merge, producing compact, well-separated clusters that are particularly appropriate when the objective is to identify groups of similar entities—in this case, countries with similar legislative profiles.

Ward's method begins with each observation as a singleton cluster and proceeds by iteratively merging the pair of clusters whose merger results in the smallest increase in total within-cluster variance. The resultant hierarchy is naturally represented as a dendrogram, a tree-like diagram wherein the vertical axis represents the distance (or dissimilarity) at which merges occur. Dendrograms are useful for comparative legal analysis because they reveal not only cluster memberships but also the gradation of similarity between groups: two countries that merge at a low height are more legislatively similar than two that merge at a greater height.

The decision between hierarchical and partitional clustering techniques—most notably K-means [34]—depends upon relevant tradeoffs. K-means requires specifying the number of clusters beforehand, generates a flat partition rather than a hierarchy, and can be sensitive to initial conditions. Hierarchical clustering generates a complete nested hierarchy that contains all possible levels of clustering, allowing any number of clusters to be extracted by cutting the dendrogram at a suitable level. For exploratory analysis of legislative similarity, where the number of "legal families" is not known a priori and where the nested structure of similarity is itself informative, hierarchical clustering is the preferred choice.

Cluster quality may be evaluated using the silhouette score [35], which measures the cohesion of each observation within the assigned cluster relative to the separation of the observation from the nearest adjacent cluster. Silhouette scores range from -1 to 1, where values close to 1 indicate observations that are well clustered, values near zero indicate observations that reside on cluster boundaries, and negative values indicate potential misassignments. The mean silhouette value across all observations provides a global measure of clustering quality that can be used to evaluate differences in the number of clusters identified and/or differences in the linkage criteria utilized.

Hierarchical clustering has been successfully applied in political science and comparative politics to identify groupings of countries based on policy characteristics. Research has utilized clustering to identify welfare state regimes [36], varieties of capitalism [37], and environmental policy types [38]. In the legal domain, clustering methods have been applied to case law to identify citation patterns among judges and to group court decisions according to legal topic [39]. However, no previous study has applied hierarchical clustering to compare water legislation at a global scale to identify typologies of water law—distinct patterns of legislative attention to dimensions of water governance that recur across countries.

This thesis addresses this gap by applying Ward's method to cosine similarity ma-

trices derived from embedding-based comparisons of water legislation across 164 countries, generating dendrograms that illustrate the hierarchical structure of global water law and identifying clusters corresponding to distinct approaches to groundwater management, river basin governance, and pollution control.

The next part of this paper will show how the five strands of methods used throughout this thesis relate to each other and are integrated as an overall research strategy, because it is through their combination that this research contributes. This section also identifies the contributions made by this thesis to each of the individual strands of methods literature.

Each of these three areas has shown some degree of omission or lack of research development, so they represent potential areas for research development, and therefore constitute a number of possible research questions.

2.6 Research Gap and Contributions

The preceding sections have surveyed five bodies of scholarship, each mature in its own right yet largely disconnected from the others. Comparative water law, pioneered by Caponera and advanced by Burchi and the FAO, has produced rich qualitative accounts of water legislation across countries but has relied exclusively on manual methods that cannot scale to comprehensive global coverage. Legal NLP has developed powerful tools for text classification, information extraction, and structured output generation, but these tools have not been applied to water legislation or to the comparative analysis of legislative provisions across jurisdictions and languages. Machine translation evaluation has advanced from surface-level metrics to learned quality estimation models, yet limited attention has been paid to the specific challenges of translating and evaluating legal texts across the diversity of languages in which water legislation is drafted. Embedding-based similarity methods offer a principled means of quantifying semantic proximity between documents, but they have not been deployed for cross-jurisdictional legal comparison. Finally, hierarchical clustering provides a natural framework for identifying typologies in multidimensional data, but it has not been applied to the problem of classifying countries by their water law profiles.

This thesis is positioned at the intersection of these five fields and makes four principal contributions:

1. **Global-scale computational comparative law.** This thesis introduces, to the best of our knowledge, the first fully computational pipeline for comparative legal analysis of water legislation at the global scale, encompassing 164 countries and over 35 source languages. This pipeline demonstrates that automated methods can complement and extend the manual comparative tradition established by

Caponera and Burchi.

2. **Multi-metric translation fidelity framework.** The thesis develops a systematic approach to evaluating the reliability of machine-translated legal texts using COMET-based reference-free quality estimation, establishing quantitative confidence levels for the translated corpus that underpins all subsequent analysis.

3. **LLM-based legal information extraction pipeline.** The "AquaLex Scrutinizer," powered by GPT-4.1-mini with domain-specific prompts and JSON schema validation, demonstrates that large language models can extract structured legal features from water legislation across diverse legal traditions and languages with sufficient accuracy to support quantitative comparative analysis.

4. **Water law typology through clustering.** By applying Ward's hierarchical clustering to embedding-based similarity matrices, the thesis identifies empirically grounded water law typologies—recurring patterns of legislative attention to groundwater management, river basin governance, and pollution control—that provide a new lens for understanding global water governance.

Together, these contributions establish a methodological framework that bridges the gap between the qualitative tradition of comparative water law and the quantitative capabilities of modern NLP, machine learning, and data science. The following chapter describes the construction of the global water legislation corpus—the empirical foundation upon which this framework operates—before Chapter 4 presents the computational methodology in detail.

Chapter 3

Data and Corpus Construction

In addition to an overview of how the corpus was constructed, this chapter discusses the empirical foundation for all the comparative analysis that follows. The corpus contains primary water legislation from 165 countries, in their native languages, converted to machine-readable text, where applicable, translated into English, and validated using multiple metrics.

This chapter will discuss the method employed to source the legislation, the composition of the corpus, the extraction of text from the legislative documents, the pipeline used to translate the text, the methods used to assess the quality of the translations, a description of the methodology used to manually verify the translations, and a summary of the completed corpus.

3.1 Document Sourcing Strategy

The objective of constructing the corpus was to collect the primary national water legislation of all sovereign states for which such legislation can be found in a digitally accessible format. To achieve this goal three main sources were used:

1. **FAO FAOLEX Database.** The FAOLEX database of the Food and Agriculture Organization [9] was the primary source for the legislative texts. The FAOLEX database is the world's largest repository of national legislation, regulations and policies related to food, agriculture and natural resources including water. The database provides either full-text documents or references to official government legal databases for the majority of the world's countries.

2. **Official Government Legal Databases.** Where FAOLEX provided metadata or links to the legislation for countries for which there was no direct link to the primary legislation, the primary legislation was obtained directly from national legislative portals, official gazettes and government legal repositories. This was

especially important for countries that had recently passed or updated water legislation that was not yet reflected in the FAOLEX database.

3. **International Legal Repositories.** Other supplementary sources include the International Water Law Project, the ECOLEX database (a collaborative project of IUCN, UNEP, and FAO) and the regional legal databases maintained by organisations such as the African Union and the Organisation of American States.

For each country, the criteria for selecting the legislation was to select the most complete available water law. Most often this would be the primary water resources act, water code or equivalent framework legislation for that country. When a country's water governance is distributed over several statutes (e.g., separate laws for groundwater, surface water, and water quality), the most complete single statute was selected. The target number of countries was 165, which corresponds to all sovereign states that have identifiable national water legislation in a digital format.

There were three practical challenges encountered during the sourcing process. The first challenge was the variability in the availability of digital legislative texts around the globe. While most European and Anglophone countries have good quality online legal databases for their legislation, in Sub-Saharan Africa, Central Asia, and the Pacific Islands many countries publish their legislation primarily in print and only very few digital versions exist. The second challenge was identifying the "primary" water legislation. In federal states like Australia, India and the United States water governance is divided between national and subnational governments and therefore requires decisions to be made regarding which level of legislation should be analyzed. For consistency, national-level legislation was given priority in all cases. The third challenge was that some countries do not have standalone water legislation but instead include water governance provisions in broader environmental codes, natural resource laws or public health statutes. In these instances, the most relevant document regarding water governance for that country was selected, with the understanding that this may not represent the entire scope of the country's water governance framework.

No digitally accessible national water legislation could be located for 27 countries. These exclusions are mostly confined to small island developing states (especially in the Pacific and Caribbean, where water governance may occur via ministerial decrees or customary practices, and not formal legislation), conflict-affected states (in which the governmental legal framework has been impacted) and states with customary water governance systems (in which unwritten customs provide the basis for allocating water rights and not formal legislation). Excluded countries are not randomly distributed by geographic area; Sub-Saharan Africa and the Pacific Islands contain the greatest number of excluded countries, while Europe and the Americas are represented with almost complete coverage. Thus, the final corpus of 165 countries represents roughly

85% of the world's sovereign states and includes virtually all of the world's population, and thus provides nearly comprehensive—if not exhaustive—coverage of the world's water legislation.

3.2 Corpus Composition

3.2.1 Language Distribution

As previously stated, the corpus encompasses over 35 languages, and uses ten different script systems to reflect the linguistic diversity of global water governance. Table 3.1 summarises the language distribution by script system and processing tier.

Table 3.1: Language and script distribution of the corpus. Tier 1: Latin-script European languages; Tier 2: Latin-script with extended character sets; Tier 3: non-Latin scripts.

Script System	**Languages**	**Tier**
Latin	English, French, Spanish, Portuguese, Italian, German, Dutch, Danish, Icelandic, Catalan, Indonesian, Somali, Azerbaijani	1
Latin (ext.)	Polish, Czech, Slovak, Hungarian, Croatian, Slovenian, Estonian, Lithuanian, Latvian, Romanian, Vietnamese, Turkish, Finnish	2
Cyrillic	Russian, Bulgarian, Serbian	3
Arabic	Arabic, Persian (Farsi), Dari	3
Greek	Greek	3
Georgian	Georgian	3
Ethiopic	Amharic	3
Lao	Lao	3
Thaana	Dhivehi	3
Hangul	Korean	3
Hebrew	Hebrew	3

The overwhelming use of Latin-script languages reflects the historical and administrative influence of European languages in national legislation around the world: French is the legislative language of much of Western and Central Africa, Spanish is the legislative language of Latin America, and Portuguese is the legislative language of lusophone Africa. The 68 documents written in English from countries including the United Kingdom, the United States, Australia, India, Kenya, South Africa, and a host of Caribbean and Pacific Island nations did not require translation and were processed directly. The English-language subset of the corpus is geographically diverse, containing former British colonies in every region: Sub-Saharan Africa (Kenya, Tanzania, Uganda, Zimbabwe, Zambia, Botswana, Namibia, and others), Southern and

Southeastern Asia (India, Pakistan, Bangladesh, Sri Lanka, Myanmar), the Caribbean (Jamaica, Barbados, Belize, Trinidad and Tobago), and the Pacific (Fiji, Samoa, Tonga, Vanuatu, Papua New Guinea). This geographical diversity of the English-language subset is useful analytically, as it allows for a comparison of whether patterns seen in the translated subset are similar to those seen in the untranslated segment of the corpus.

The linguistic diversity of the corpus creates several challenges for computational processing. First, the degree of morphological complexity of the various languages is enormous: agglutinative languages such as Finnish, Hungarian and Turkish aggregate multiple grammatical elements into single words, creating vastly different token distributions compared to analytic languages such as English and French. Second, right-to-left (RTL) script languages—Arabic, Hebrew, and Thaana—require special treatment when extracting text to ensure proper reading order, particularly in documents that contain both RTL text and left-to-right numbers and Latin-script terms. Third, certain scripts create unique obstacles for optical character recognition: Georgian Mkhedruli and Ethiopic (Ge'ez) script have limited representation in the typical OCR training datasets, whereas Arabic script is compounded by the fact that contextual letter forms change based upon their position within a word. As a result of these obstacles, a multi-method extraction architecture was developed as discussed in Chapter 4.

3.2.2 Document Types and Temporal Range

The above-mentioned linguistic complexities affect both the processing of documents and the interpretation of them. Differing legal systems contain differing methods of addressing water governance through legislation; and the above-referenced corpus reflects this difference. The legislative instruments contained within the corpus vary from comprehensive water codes which compile all relevant water law into a single document, to narrowly scoped "framework" acts, to broad environmental statutes which address water as one of multiple environmental components. Since the format of an act or statute adopted by a country will often shape the scope and degree of specificity of the provisions that an extraction pipeline may be able to extract from the corpus, this variation in form is worthy of examination. As a result of the fact that civil-law countries typically have a codification of all aspects of water resources management into a single statute (i.e., a comprehensive water code), France's Code de l'eau, Romania's Water Law and Bulgaria's Water Act are examples of civil-law countries' use of a comprehensive water code. In contrast, water resources acts in common-law countries provide a framework for the governance of water, though they may not codify water governance within a larger codified framework. Thus, water resources acts are typical of the governance of water in common-law countries such as Australia, Kenya and Zimbabwe. Environmental codes with water-specific titles (e.g., Italy and Sweden) rep-

resent the increasing recognition of water as an environmental as well as an economic resource and integrate water governance into a broader statutory framework governing environmental protection. Specialised water-management statutes (e.g., Japan's River Act, which is focused solely on river and flood management, or Canada's Water Act, which focuses on water quality) represent a more limited approach to the governance of water where various components of water governance are governed separately under multiple statutes.

Document sizes vary significantly in terms of length. Document sizes range from 3 pages (Oman) to greater than 230 pages (Bulgaria's Water Act at 236 pages, Poland's Water Law Act at 211 pages, Romania's Water Law at 211 pages). The variability in document length correlates roughly with the scope of the statute: specialised or framework statutes are much shorter than comprehensive water codes.

The time frame covered by the corpus spans more than eight decades. Costa Rica's *Ley de Aguas* of 1942 represents the earliest legislation included in the corpus, while Vietnam's Law on Water of 2023 is the most recently enacted. Most of the legislation in the corpus was enacted during the 1990s and 2000s, which was a decade of widespread water law reform inspired by the Dublin Principles (1992), the development of IWRM as a governance paradigm, and the influence of the EU Water Framework Directive (2000) on both European and non-European water legislation [12]. A secondary peak of legislation was enacted in the 2010s and early 2020s and reflects a more recent wave of water law reform that has been inspired by concerns about climate adaptation, the adoption of the Sustainable Development Goals (SDGs) in 2015, and the increasing recognition of groundwater depletion as a governance priority. The large temporal range of the corpus means that some of the entries in the corpus include legislative frameworks that have subsequently been amended or repealed, a limitation that is inherent in any corpus-based approach and will be addressed further in Section 3.7.

3.3 Text Extraction Results

3.3.1 Corpus Manifest

An automated manifest generator created a corpus manifest—a systematic catalog of the source documents in the non-English subset of the corpus—that characterised 104 source documents in their original languages. The documents were characterised by country code, language, script type, page count, file size, text layer presence, layout heuristics (single-column versus multi-column), right-to-left (RTL) flag, and OCR requirement indicators. All 104 documents were designated as successful in the manifest generation process, and the country codes were validated against the `pycountry` library.

The manifest indicated that there were 14 Arabic-script documents (including Ara-

bic, Persian, and Dari), 8 Cyrillic-script documents, 2 Greek-script documents, and individual documents in Hangul, Lao, Ethiopic, Hebrew, Georgian, and Thaana scripts. The remaining documents were written in Latin script with varying levels of diacritics. The two documents that were in DOCX format (Turkmenistan and Ukraine) instead of PDF necessitated separate extraction paths through the `python-docx` library.

3.3.2 Extraction Quality

Automated assessments of the quality of the extraction of the 104 documents across all script systems yielded the following results. Of the 104 documents, 102 (98.1%) received an "Excellent" rating, indicating that the extraction was successful and no problems were identified. The other two documents, Serbia's and Montenegro's water laws, received "Acceptable" ratings because of the low level of Cyrillic content detection (0.0%) that resulted from the source documents being written in Latin-script variants of the Serbian and Montenegrin languages rather than Cyrillic script. Therefore, the results of the extraction quality assessment by script system demonstrated consistent performance across all script families. All 74 Latin-script documents, all 14 Arabic-script documents, the 2 Greek documents, and the single documents in Hangul, Lao, Ethiopic, Hebrew, Georgian, and Thaana scripts passed validation. Specifically, Google Cloud Document AI performed well with regard to complex scripts, correctly handling right-to-left Arabic text, Georgian Mkhedruli, Ethiopic syllabary, and Lao script without needing to configure the script specifically.

3.4 Translation Pipeline Results

Approximately 96 of the 165 country-level documents in the corpus were not in English and thus needed to be translated into English. The remaining approximately 69 documents were in English and therefore did not require translation. The translations were performed using the Google Cloud Translate API v2 [40].

The translation pipeline employed a segment-by-segment approach, breaking down each document into sections (or "chunks") of 4,000 characters to balance the need to preserve context with the API payload limits. Each chunk was submitted with automatic source language detection, and the translated chunks were then combined into the final English text. A retry mechanism (up to three retries per chunk with 10-second delays) was implemented to handle transient API failures. Additionally, the design of the pipeline was resumable to enable partial processing of the pipeline in one run followed by continuation of the pipeline in subsequent runs.

The pipeline successfully translated all of the non-English documents, creating complete English-language versions of the entire corpus. Handling typologically diverse

languages—from highly morphological agglutinative languages (Hungarian, Finnish, Turkish) to tonal languages (Vietnamese, Lao) to right-to-left semitic languages (Hebrew, Arabic)—was facilitated by the extensive language coverage of Google's neural machine translation models. However, the quality of the translations varied greatly across languages, as discussed below.

3.5 Translation Quality Results

The translation quality of the corpus was assessed using the multi-metric fidelity framework discussed in Chapter 4. Since all downstream analyses depend upon the fidelity of the translations, assessing the quality of the translations is important for establishing the evidential basis of the research question. Below, a summary of the main findings is provided; details regarding the results of the assessments, including the per-language COMET scores, the semantic fidelity comparisons, and the identification of translation-induced semantic loss are presented in Chapter 5.

3.5.1 COMET Score Overview

Reference-free evaluation of the quality of the translations was conducted using the COMET model `Unbabel/wmt22-cometkiwi-da` [24] for all 39 source languages in the translated portion of the corpus. The COMET scores obtained for the translations ranged from 0.681 (Tigrinya, $n = 23$ segments) to 0.866 (Romania, $n = 192$ segments), with a mean of approximately 0.83 for the corpus as a whole. The mean COMET score of approximately 0.83 exceeds the commonly accepted threshold of 0.80 for sufficient translation quality, providing quantifiable evidence of the adequacy of the translated corpus as a whole.

3.5.2 Language Performance Tiers

Based on the results of the COMET evaluations, three performance tiers were identified:

- **High quality** (COMET ≥ 0.85): Romanian (0.866), Catalan (0.857), Estonian (0.856), Korean (0.852), Georgian (0.852), and Serbian (0.850) had high quality translations, indicating that the translations for these languages were of demonstrably high quality.
- **Adequate quality** ($0.80 \leq$ COMET < 0.85): The middle of the distribution, including 28 languages ranging from Italian (0.848) to German (0.811), including major world languages such as French (0.834), Spanish (0.831), Russian (0.826), and Portuguese (0.822), had adequate translations. While the translations in

this tier were acceptable for use in downstream analyses, there were some quality concerns related to specific segments.

- **Lower quality** (COMET < 0.80): Arabic (0.794), Polish (0.788), Somali (0.751), Dari (0.701), and Tigrinya (0.681) had lower translation quality with significant variation in translation quality among the segments of each of these languages.

3.5.3 The Contextual Flattening Discovery

In addition to the reference-free quality assessment of the translations based on COMET scores, a comparative semantic fidelity analysis based on embeddings was conducted and resulted in the identification of a phenomenon referred to as *contextual flattening*. Some translations exhibited adequate or even high COMET scores, while simultaneously demonstrating extremely low cosine similarity between the original-language and translated-language text embeddings. For instance, Georgia's legislation received a COMET score of 0.852 (within the top five), but a cosine similarity of only 0.240. Similarly, the cosine similarity for Amharic was found to be only 0.117, despite receiving a COMET score of 0.841. The difference between the two metrics provides useful information: COMET assesses whether a translation appears to read as fluent and adequate English, whereas embedding-based cosine similarity assesses whether the semantic content of the original document is retained in the translation. Therefore, a translation can meet the first criterion and fail the second if it replaces specific legal and regulatory terminology with more general language. The existence of contextual flattening in machine-translated documents highlights a systematic risk for computational legal analysis, which is explored further in Chapter 6.

It was recognized that while automated metrics were necessary, they would never provide sufficient insight into how automated translation is able to cause translation-induced meaning loss. Therefore, another level of human assessment was required in order to determine if there was a correlation between the results of the human reviewers and the automatic quality score. The differences between the COMET scores and the similarity in embeddings between the source and target language for certain languages clearly showed that no single metric should ever be relied upon without further verification. Therefore, a structured manual review process was developed for the purpose of (1) verifying the accuracy of the text extraction process for different script families, and (2) to verify if the automated quality scores accurately reflected the judgment of a human reviewer. To strike a balance between being thorough enough to provide useful data to the project team and the limited time available for the project, the protocol described here was limited to thirty documents in the corpus instead of using the entire corpus.

3.6 Manual Verification Protocol

A manual verification protocol of thirty legislative documents was conducted to test the validity of automated quality metrics for extracted text from the corpus and to validate the quality of the extracted text across all of the major script families in the corpus. Each document was selected to represent one of the main script families, and the verification method was adjusted based upon the type of script:

- **Latin-script documents** (19 documents): Visual scans were performed on the nineteen documents in Spanish (Costa Rica, Mexico, Chile, Honduras, Nicaragua, Bolivia, Venezuela, Uruguay, Paraguay, El Salvador, Ecuador, Colombia, Equatorial Guinea), French (France, Luxembourg, Mali, Côte d'Ivoire), German (Germany), Hungarian (Hungary), Polish (Poland), Slovak (Slovakia), Slovenian (Slovenia), and Catalan (Andorra) to verify whether the extracted text accurately represented the source PDFs. Verification included side-by-side comparisons of the source PDF and extracted text, confirmation of the page number match for the first and last paragraphs, and a check for missing sections.

- **Arabic-script documents** (4 documents): Right-to-left visual verification was performed on the four documents from Saudi Arabia, Afghanistan, Oman, and Sudan. Verification included confirming right-to-left text flow in the extracted text; comparing the first and last lines of the PDF and extracted text; checking whether numbers and dates corresponded (these are script-invariant); and verifying whether the document structure had been maintained.

- **Non-Latin complex scripts** (7 documents): Integrity of Unicode characters for Lao (Laos), Russian/Cyrillic (Turkmenistan), Greek (Greece), Ethiopic (Ethiopia), Hebrew (Israel), Georgian (Georgia), and Thaana (Maldives) were verified. Verification confirmed that there was no corruption of script-specific Unicode characters; that there were no excessive or insufficient characters in relation to the length of the document; and that there were no spurious Latin-character artifacts in predominantly non-Latin texts.

The manual verification protocol confirmed that the automated extraction pipeline generated accurate representations of the text across all of the primary script families in the corpus. No examples of missing sections, corrupted characters, or text order reversal were found among the 30 documents manually verified. In addition, article numbering and hierarchical document structure (chapters, sections, articles, paragraphs) were preserved in all cases so that the cross-references typical of legislative drafting could be interpreted in the extracted text. Two Arabic-script documents were noted to contain

minor artifacts (excessive whitespace between words and occasional insertion of line breaks in the middle of sentences), however, the artifacts were not impactful to the semantics of the extracted text and were addressed during the text cleaning phase. Additionally, for the Latin-script documents, verification confirmed that diacritical marks—essential for distinguishing legal terms in languages such as French, Spanish, and Romanian—were accurately preserved throughout the extraction process.

In addition to validating the quality of the extracted text, the manual verification also provided an independent validation of the quality scores estimated by COMET. Side-by-side comparison of the source and translated text for the Latin-script documents confirmed that legal concepts, article numbering, and regulatory structures were preserved in translation. For the other non-Latin complex scripts, the verification focused on the integrity of the characters (rather than translation accuracy) and confirmed that Unicode characters were correctly preserved without corruption or substitution. The protocol identified the same high-risk languages (Georgian, Amharic, Tigrinya) that were identified as needing caution when interpreting the results of downstream analysis using the embedding-based semantic fidelity analysis.

To ensure that the 30 documents selected represented all of the script families and language levels in the corpus, the selection of the 30 documents was designed to include representation of all of the major script families and language levels. While a larger verification sample would increase the statistical certainty of the results, the 30-document protocol was designed to trade off the thoroughness of the protocol with the practical limitations of the time required for the manual review. The agreement between the results of the manual verification protocol and the quality metrics computed automatically by the system provides evidence that the automatic assessment of quality of translation and extraction can serve as a reliable surrogate for quality assessments of the system for the entire corpus.

3.7 Final Corpus Summary

The corpus of legislative texts includes complete text extraction and translation for 165 countries, making it the largest and most comprehensive digital corpus of national water legislation compiled to date. Table 3.2 summarises the key characteristics of the corpus.

The corpus of legislative texts represents a significant advancement over previous collections of water legislation. The FAO's FAOLEX database provides access to individual legislative texts, but the database does not comprise a structured, machine-readable corpus of legislative texts that is amenable to computational analysis. Regional collections of water legislation—such as those developed by the European Commission for EU member states, or by the African Union for water policy for the

Table 3.2: Summary characteristics of the global water legislation corpus.

Characteristic	**Value**
Total countries with legislation	165
Countries with English-language originals	68
Countries requiring translation	∼96
Source languages	35+
Script systems	10+
Temporal range	1942–2023
Extraction quality: "Excellent"	102/104 (98.1%)
Extraction quality: "Acceptable"	2/104 (1.9%)
Mean COMET score (translated subset)	∼0.83
COMET score range	0.681–0.866
Manual verification documents	30
Manual verification script types	9

continent—cover the geographic areas served by the regional organisation, but they do not cover the globe. Prior academic comparative studies have assembled collections of ten to twenty legislative texts for purposes of manual comparison [4, 5]. The corpus described here contains 165 countries with complete text extraction and translation, spans more than 35 languages and 10 script families, and has documented quality metrics for each step of processing. To the best of our knowledge, no prior corpus of similar scope, linguistic diversity, or documented quality has existed in the field of comparative water law. As such, the corpus will function as a base for future work in computational comparative water law research.

The data quality control measures taken during the development of the corpus—quality scoring of extraction quality during extraction, multi-metric assessment of translation quality, and manual verification of quality across the various script families—have created a foundation of documented reliability for the downstream analytical tasks described in Chapters 4 and 5. Known quality limitations—specifically, the lower quality of translations for Tigrinya, Dari, Somali, and the context flattening effect on Georgian, Amharic, and Lao—have been documented and the implications of the limitations for the interpretation of results are discussed in Chapter 6.

It is important to note that the corpus reflects a single point in time regarding the state of national water legislation. National water legislation is not static: countries regularly modify, repeal, or substitute their water laws as a result of changes in governance paradigms, climate pressures, and obligations under international agreements. The time span of the corpus—legislation from 1942 (Costa Rica) to 2023 (Vietnam)—means that some entries in the corpus represent historical legislative frameworks that have since been superseded. Temporal heterogeneity is an intrinsic constraint to any large-scale legislative corpus and is further discussed in Chapter 6.

Now that the corpus has been constructed, its quality has been validated, and the limitations of the corpus have been documented, the next chapter describes the computational methodology used to transform the raw legislative texts in the corpus into structured, comparable forms that enable answers to the four research questions posed in Chapter 1.

Chapter 4

Methodology

The next step in the process was the computational analysis of the legal text to determine whether the legal instruments contained in the corpus could be used to inform the four research questions. The pipeline for the computational analysis of the legal text comprised seven sequential steps: (1) corpus creation and manifest generation; (2) extraction of the legal text from various formats of legal documents; (3) machine translation of non-English versions of the legal text to create a monolingual dataset in English; (4) assessment of the quality of the translations using a multi-metric framework for evaluating translation fidelity; (5) use of large language models to extract legal information relevant to the three dimensions of policy; (6) analysis of the similarities among the legal texts using embeddings and ranking of similarities using concordance; and (7) hierarchical clustering and visualisation of the similarities among the legal texts.

4.1 Overview of the Methodological Framework

In addition to the substantive contributions of the pipeline to answering the four research questions, the ability to carry out the seven stages of the pipeline constituted the feasibility study for the project, as stated in RQ1.

While it would be impossible to undertake comparative legal analysis on this scale manually, the computational approach made possible by the pipeline provided a solution to the problem of the need for large amounts of multilingual legal expertise in order to carry out such an analysis globally. Traditionally, comparative water law scholarship (e.g., Caponera, 1991; FAO Legislative Studies Series) has been restricted to qualitative studies of no more than about twenty jurisdictions and has been limited to languages accessible to the scholars. However, the pipeline provided a means of carrying out a uniform global analysis of water law in Arabic, Cyrillic, Georgian, Ethiopic, Lao, Thaana and many Latin-script languages, which would have required decades of multilingual legal expertise to accomplish manually.

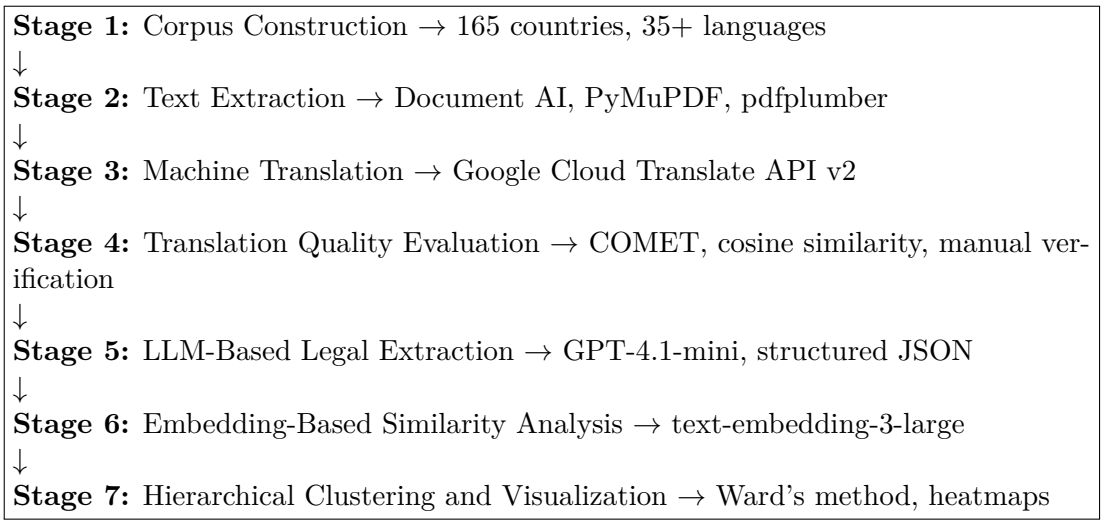

Figure 4.1: Overview of the seven-stage computational pipeline for comparative water legislation analysis.

4.2 Corpus Construction

This chapter does not contain the detailed description of how the corpus was constructed (sourcing strategy, language, document type, time period) as was done in Chapter 3. Instead it describes only those specific automated pipeline steps used to convert raw legislative documents into a structured and processable form. This distinction is made because the actual construction of the data set via the manifest generation, extraction routing, and quality validation phases constitutes an additional separate methodological contribution of this research and can be separated from the descriptive attributes of the resultant data set.

4.2.1 Document Sourcing Strategy

The corpus was constructed by gathering the primary national water legislation of 165 countries. The source documents were gathered from official government legal databases, the FAO Legal Database (FAOLEX), and national legislative portals. In each country, the most comprehensive water legislation available was chosen—usually the main water resources act, water code or similar framework legislation. The corpus contains documents with lengths ranging from 3 pages (small island states) to over 200 pages (comprehensive water codes), with publication dates ranging from 1942 (Costa Rica) to 2023 (Vietnam).

Each of the documents was collected in its original language, therefore the corpus represents a collection of more than 35 languages and ten different script systems. The language representation of the corpus corresponds to the political geography and linguistic legacy of global water governance: Latin-script languages (French, Spanish, Portuguese, German, etc.) represent the majority of the world's national legislation,

followed by Cyrillic-script languages (Russian, Bulgarian, Serbian), Arabic-script languages (Arabic, Persian, Dari), and a much smaller tail of less common scripts like Georgian, Ethiopic (Amharic, Tigrinya), Lao, Thaana (Dhivehi), Hangul (Korean), and Hebrew.

4.2.2 Corpus Manifest Generation

An automated manifest generator was used to generate a systematic list of the corpus by parsing the document filenames and analyzing some of the basic file attributes without having to analyze the content of the documents. For each document, the automated manifest generator generated the following metadata fields:

- **Country identification**: ISO 3166 alpha-2 country code, verified against the `pycountry` library, with additional fallbacks for non-standard codes.
- **Language and script classification**: The language names were standardized from the raw filename identifiers (ISO 639-3 codes, full names, and common abbreviations) and assigned to one of three levels of complexity:
  - *Tier 1 (Easy)*: Simple Latin-script European languages (English, Spanish, French, German, etc.).
  - *Tier 2 (Harder)*: Extended Latin-script languages (Polish, Czech, Hungarian, Vietnamese, Turkish, etc.).
  - *Tier 3 (Hardest)*: Scripts outside of the Latin family (Arabic, Persian, Hebrew, Georgian, Ethiopic, Korean, Lao, and Cyrillic).
- **Document properties**: Number of pages, file size, presence of a text layer, and layout attributes extracted by using the lightweight probing feature of PyMuPDF.
- **Layout heuristics**: Ratio of aspect to calculate the presence of landscape orientation, ratio of density of characters to detect the possibility of multi-column layouts, and ratio of file size per page to indicate whether an OCR operation might be necessary.
- **RTL detection**: Flag indicating whether the script is written from right-to-left (Arabic and Hebrew families).

The automated manifest generator served as the routing mechanism for the subsequent extraction steps and allowed the automated selection of the best extraction route for each document, depending on the language, script, layout, and OCR requirements of the document.

4.3 Text Extraction

4.3.1 Multi-Method Extraction Architecture

Because the corpus included both born-digital PDFs with embedded text layers, and scanned PDFs requiring OCR, and because the layout of the documents varied widely (multi-column layouts, Cyrillic, etc.), a multi-tool extraction strategy was developed with intelligent routing and automatic fallback.

Primary Extraction: Google Cloud Document AI

Google Cloud Document AI [41] was utilized as the primary extraction tool since it offers leading OCR and text extraction functionality with wide multilingual support. Based on the length of the document, two extraction modes were used:

- **Synchronous processing**: Documents of 15 pages or fewer were processed via the synchronous API endpoint, resulting in instant extraction results.
- **Batch processing**: Longer documents were uploaded to Google Cloud Storage (GCS) and processed in batches via the asynchronous batch-processing API, with the results retrieved once the batch had completed processing.

Fallback Methods

When Google Cloud Document AI was unavailable or did not produce acceptable results, two local extraction tools were used as fallbacks:

- **PyMuPDF (fitz)**: Basic text extraction via the `page.get_text()` method, useful for extracting text from born-digital PDFs with embedded text layers. PyMuPDF was the fastest method, but did not include OCR functionality for scanned PDFs.
- **pdfplumber**: Another extraction tool that performed better with multi-column layouts and tables. When the layout detector identified a multi-column document, pdfplumber was used instead of the standard PyMuPDF extraction method.

For the approximately 25 documents in DOCX format (the minority of the corpus), the `python-docx` library was used for direct extraction of the text.

4.3.2 Layout Detection

A custom `LayoutDetector` class was developed to identify documents with multi-column layouts, which requires special treatment to preserve reading order. The `LayoutDetector` worked as follows:

1. All text blocks on each page were extracted using PyMuPDF's dictionary-mode text extraction (`page.get_text("dict")`).
2. The left edge x-coordinates of all text blocks were collected and rounded to the nearest 50 pixels for clustering.
3. Clusters of positions with more than 15% of the total number of text blocks were considered "significant" column positions.
4. If two significant position clusters existed that were separated by more than 30% of the page width, the document was considered to be a two-column document. If there were three or more clusters, the document was considered to contain multiple columns.
5. The predominant layout type on the first three pages of a document determined the document's overall classification.

Document classifications of multi-column were passed to pdfplumber for extraction, while documents classified as single-column were extracted using the faster standard method of PyMuPDF.

4.3.3 Diagnostic Testing

Before implementing a production extraction strategy, a diagnostic test phase was undertaken on eight exemplar documents representing the full scope of the corpus: simple Latin-script PDFs (French, Spanish); multi-column layouts (German, Mexican); RTL scripts with embedded text layers (Iraqi Arabic); RTL scripts that required OCR (Egyptian Arabic); complex-script documents (Lao); Cyrillic multi-column layouts (Russian). Each of the documents was extracted using all three extraction methods (PyMuPDF standard, pdfplumber, and layout-aware extraction), and the results were compared on the basis of number of characters, number of words, extraction time, and the degree of similarity between the extracted texts using `difflib.SequenceMatcher`.

4.3.4 Text Cleaning

The extracted texts underwent a conservative cleansing process intended to remove artifacts without removing legally significant content:

1. **Control character removal**: While preserving newlines, tabs, and carriage returns, all non-printing control characters were removed.

2. **Whitespace normalization**: Trailing whitespace was removed from each line; multiple consecutive blank lines were reduced to a maximum of two blank lines; and all spaces within lines were reduced to single spaces.

3. **Hyphenation repair**: Broken words at line ends with hyphens (common artifacts of PDF formatting) were rejoined using the regular expression `(\w)-\s*\n\s*(\w)`.

4. **Page number removal**: Independent page numbers (such as single integer values, bracketed integer values, or "Page N" labels) were removed.

For each of the extracted documents, a confidence score was calculated based on the extraction success (40%), validation status (30%), adequate character count (15%), and the proportion of cleaned text removed (15%). Any document with a confidence score below 0.80 was flagged for manual examination. The final output was composed of clean `.txt` files divided into two directories: `english/` for the approximately 68 documents originally in English, and `to_translate/` for the roughly 96 documents in other languages.

After completing the steps of data extraction and cleaning, all documents within the corpus had been transformed into plain text files so they could move on to the next process. Approximately ninety-six non-English documents sat in the translation queue waiting to have their language converted from whatever they were originally written in to English so there would be a means by which valid cross country comparisons could be made. This challenge is certainly not trivial; the task of translating legal texts from over thirty five languages, where each has its unique syntax and domain specific terms, demanded a pipeline capable of processing large amounts of material (both in number and type) without generating too much semantic drift.

4.4 Translation Pipeline

Google Cloud Translate API v2 [40] was used to translate all legislative text that was not written in English into English. A chunking strategy was used to divide the larger documents into manageable sections while still meeting the limitations of the API:

1. All documents were divided into chunks of 4,000 characters. The character limit of 4,000 was determined to allow for enough context for each chunk to have some value, but small enough so as to keep within the API limits.

2. After division into chunks, each chunk was sent through the translation API with English (`en`) designated as the target language. The source language was automatically detected by the API based on the text it received and utilized Google's neural machine translation models.

3. In addition to submitting each chunk to the translation API, a retry mechanism was established to address possible temporary API errors. Up to three attempts at translating each chunk were made with a ten-second delay between attempts.

4. Once translated, the individual chunks were then concatenated back together to create the complete English version of each document.

5. A resumable design allowed previously completed documents to be skipped when the pipeline ran again, allowing for incrementally processing the data set.

More than thirty-five different source languages were successfully translated through the pipeline, many with typologically complex morphology and syntax. Non-Latin scripts such as Arabic right-to-left text, Cyrillic, Lao syllabic script, Georgian Mkhedruli, Ethiopic (Ge'ez) script, and Thaana presented particular challenges for maintaining translation quality in the pipeline, which are discussed below.

4.5 Translation Quality Evaluation

The use of machine translation in the translation of legal documents poses a significant threat of semantic distortion, especially for languages typologically far removed from English. To assess and minimize this risk, a multi-metric evaluation framework for assessing the fidelity of machine translation was created. The evaluation framework contained both automatic and manual assessments of translation quality; therefore, it represents a methodology contribution of the thesis.

4.5.1 COMET Reference-Free Quality Estimation

An automated assessment of the quality of machine translation was accomplished utilizing COMET (Crosslingual Optimized Metric for Evaluation of Translation) [22], specifically the `Unbabel/wmt22-cometkiwi-da` model [24]. This reference-free quality estimation model uses only the source text and the machine translation output to assess the quality of a machine translation, thus avoiding the need for a reference human translation, which would have been prohibitive given the volume of the corpus, where reference translations were unavailable.

The steps followed to perform the COMET analysis were as follows:

1. Source-text/translated-text pairs were created from the original-language and translated-language versions of the legislative documents.

2. Each pair was formatted as a dictionary with fields for `src` (source) and `mt` (machine translation), which is the format that is expected by the COMET framework.

3. The `wmt22-cometkiwi-da` model was loaded, and applied in batches of eight pairs, with GPU acceleration used where available and CPU fallback used otherwise.

4. The COMET score for each pair was aggregated by source language to produce a language-level quality estimate for the machine translations.

The COMET scores produced across the corpus ranged from 0.681 (Tigrinya) to 0.866 (Romanian), with an average of about 0.83—the generally recognized threshold of 0.80 for acceptable translation quality. The top-scoring Latin-script European languages were Romanian (0.866), Catalan (0.857), and Estonian (0.856). Low-resource languages with complex writing systems had lower scores (Tigrinya 0.681, Dari 0.701, Somali 0.751).

4.5.2 Embedding-Based Semantic Fidelity

To provide a complementary metric to the COMET score, cosine similarity was calculated between the text embeddings of the original-language legislative documents and their English translations. This type of measure assesses whether the semantic content of the translation is preserved compared to the source, regardless of the linguistic quality of the translation.

It was found in this study that there exists a problem we term *contextual flattening*: even though certain translations were deemed to have sufficient COMET scores (i.e., they were considered to be linguistically fluent and adequate), these translations exhibited very low cosine similarities with respect to the original-language text, indicating that the translation (although sufficiently competent in terms of grammar) had lost much of the domain-specific semantic detail that was contained in the original-language text. Examples of such languages include Georgian, with a cosine similarity of only 0.240, although the COMET scores were satisfactory, and Amharic, with a cosine similarity of 0.117. These examples illustrate how well-trained neural machine translation systems can produce fluent English translations that, however, lose much of the unique legal and cultural context of the original-language text.

This discovery has important implications regarding the interpretation of the results from the downstream analyses: in cases where legislation has undergone significant

contextual flattening, the similarity analyses will likely reveal less variability among those countries, which is explicitly noted in the Results Discussion section.

4.5.3 Manual Verification Protocol

A manual verification protocol was devised and executed over thirty documents representing nine different script types to verify the automated quality metrics. Documents were selected to represent the most common script families in the corpus:

- **Latin script**: A visual scan verification of the alignment between the original source and the translation.
- **Arabic and Hebrew (RTL)**: A visual right-to-left verification of directionality artifacts in the extracted text with attention to correct directionality.
- **Cyrillic**: A character-level verification of proper Unicode handling and consistent transliteration.
- **Georgian, Ethiopic, Lao, Thaana, Greek**: Integrity checks of the Unicode of script-specific characters to confirm that the characters were not corrupted during either the extraction or translation processes.

The manual verification confirmed that the automated quality metrics provided a good proxy for the quality of machine translation for most of the language families represented in the corpus, while also identifying the specific high-risk languages (Georgian, Amharic, Tigrinya) where caution should be exercised when interpreting the results of downstream analyses.

4.6 LLM-Based Legal Information Extraction

The primary analytical task—extracting structured legal provisions from unstructured legislative text—was accomplished using two complementary LLM-based information extraction strategies. The three policy areas that have been selected to be extracted—groundwater regulations, river basin management, and the “polluter pays” principle—are those which most clearly illustrate the repeated issues concerning cross-boundary jurisdictional regulation commonly reported in literature on Integrated Water Resources Management (IWRM) and the comparative law literature on water law developed primarily by Caponera and Burchi [4, 5]. All three areas were supported by the 1992 Dublin Principles which provided an international normative basis for connecting all three areas with respect to groundwater resources as a limited resource that requires controlled access; river basins as the appropriate geographic area for governance; and

economic tools, including the polluter pays principle as necessary tools for sustainable management. In combination, the three areas allow for a multi-dimensional analysis of how legislative approaches toward these areas vary among different jurisdictions based upon resource access, spatial governance and economic accountability.

4.6.1 Lightweight Topic-Specific Extraction

The first strategy employed a lightweight, topic-specific extraction strategy for extracting provisions from legislative text that relate to individual policy dimensions.

Model and Configuration

The extraction model used was OpenAI GPT-4.1-mini [42], which offered a good balance of capability and cost-effectiveness, as well as a large context window. The above-described GPT-4.1-mini was chosen for three key reasons. First, it had a sufficient length of context window to allow the processing of long legislative documents without having to do too much chunking. Second, it produced structured JSON output via function calls which allowed us to enforce the schema of the extraction strictly. Third, it cost significantly less than full scale models (i.e., GPT-4o) on a per token basis allowing us to be able to run our extraction process on each of the 165 legislative documents. A comparison of extraction results from various LLM vendors will be an area of research for future work in Chapter 7. Each legislative document was split into chunks of 15,000 characters (with 500-character overlaps to maintain the context from the preceding and succeeding chunks); each chunk was then separately inputted into the model.

Prompt Design

The system prompt defined the LLM as "a dedicated legal AI assistant" and explicitly stated the following operational constraints:

1. **Exclusive topic focus**: The model was told to consider only the specified policy dimension (for example, "River Basin Management" or "Water Pollution Control") and its direct synonyms.

2. **Evidence-based extraction**: The model was restricted from making assumptions not directly supported by the text itself.

3. **Chain-of-thought reasoning**: The prompt described a reasoning process: scan the text for the presence of keywords, find the exact sentence(s) containing relevant provisions, eliminate duplicate provisions, and format the output.

4. **Strict JSON schema**: The output was limited to a JSON object with four attributes: `is_relevant` (Boolean), `key_excerpts_original` (exact quote of key excerpts), `summary_english` (English summary), and `confidence_score_percent` (Integer, 0–100).

Two topic-specific extraction jobs were established: one for river basin management (using keywords: "river basin management", "river basin plan", "river basin councils", "catchment management", "watershed management", "basin planning"), and another for pollution control (using keywords: "polluter-pays principle", "pollution limits", "pollution prevention", "cost recovery from polluter", "environmental liability"). Although researchers select their own keywords which can introduce an element of bias into the analysis; the keywords selected in this research, whether for the Topic-Specific or Comprehensive Extraction Strategies, have been derived from the well-established terminology in IWRM, the classification terms for water legislation found in the FAO FAOLEX Database [9], and the Comparative Water Law literature of Caponera and Burchi [4, 5]. These are all representative of the principal legal concepts identified by other researchers' comparative studies on water governance. Alternative keyword selections have not been tested within this study (see Section 7.3).

Caching and Reliability

Each response generated by the LLM model to each chunk was saved to disk with filenames that included the name of the source file, topic, and chunk number (e.g., `France_river_basin_chunk_001.json`). Thus, if the pipeline needed to restart after a failure, the cached responses could be reused, reducing the need for duplicate API calls.

4.6.2 Comprehensive "AquaLex Scrutinizer" Extraction

The second strategy employed a comprehensively, multi-topic extraction strategy to extract provisions from the entire legislative document in a single run.

System Prompt Engineering

The LLM was instantiated as "AquaLex Scrutinizer," a persona of the AI system, with an exhaustive extraction requirement. The prompt for the system (approximately 1,500 words) outlined the following requirements for the extraction:

1. **Exhaustive extraction**: The model was required to identify every single provision, clause, article, section, or paragraph that was relevant to any of the three policy dimensions, including direct keyword matches, synonymous terms, and conceptually similar terms.

2. **Multi-group relevance handling**: If a provision was relevant to more than one policy dimension, the model was to reproduce the provision in the output for each of the relevant policy dimensions, along with the corresponding keyword annotations and summaries.

3. **Precise section identification**: Each provision that was identified was to be annotated with its official legal identifier (e.g., "Article 15, Paragraph 3") or, if no formal identifier was available, a placeholder description.

4. **Verbatim text preservation**: The field for extracted text was to contain the full verbatim text of the legal provision, without any summarizing or paraphrasing.

The four stipulations above collectively guaranteed that the extraction was both inclusive and could be traced back to the text from which the legislation was derived. If, for example, the model did not preserve the provisions verbatim, there would be no way to determine if a flagged provision actually existed within the source legislation or if it was a result of the model's own interpretation. Once the extraction constraints were defined, the next issue was how to organize the extracted data into a structured manner that could be uniformly analyzed and compared among all 165 countries.

Output Schema

The output was structured as a JSON object that has three top-level keys for the three different policy dimension types:

```
{
  "groundwater": [
    {
      "section_id": "Article 15, Paragraph 3",
      "text": "<verbatim legal text>",
      "summary": "<1-2 sentence summary>",
      "keywords": ["groundwater permitting", ...]
    }, ...
  ],
  "river_basin": [ ... ],
  "polluter_pays": [ ... ]
}
```

Each group of keywords was mapped to a defined list of permitted keywords. There were 11 keywords in the `groundwater` group (e.g., "groundwater ownership", "groundwater permitting", "groundwater monitoring", "aquifer management", "well drilling

regulations", "transboundary aquifers", "groundwater recharge"). There were nine keywords in the `river_basin` group (e.g., "river basin management", "river basin council", "watershed management", "integrated water resources management (IWRM) at basin scale", "transboundary river agreements"). There were ten keywords in the `polluter_pays` group (e.g., "polluter-pays principle", "pollution charges", "discharge fees", "environmental liability for pollution", "remediation costs recovery").

Token Management and Rate Limiting

Token counting was performed using the `tiktoken` library with the encoding specific to the GPT-4.1-mini model. Two token limits were enforced:

- **Tokens-per-minute (TPM) limit**: The total token count (document tokens plus approximately 40,000-token prompt overhead) was used to enforce the account-level rate limit of 400,000 TPM. Any documents whose total token count would exceed the TPM limit were truncated to meet the limit, with a log entry made for each truncated document.
- **Model context window**: Any documents whose total token count exceeded the model context window limit of 1,000,000 tokens were truncated to fit, with a log entry made for each truncated document.

Schema Validation

A five-point schema validation was applied to every LLM response prior to it being accepted:

1. The output is a valid JSON dictionary.
2. The three top-level keys (`groundwater`, `river_basin`, `polluter_pays`) are present in the output.
3. Each value is a list of dictionaries.
4. Each dictionary contains the four required fields: `section_id`, `text`, `summary`, `keywords`.
5. The `keywords` field is a list of strings and the other string fields (`section_id`, `text`, `summary`) are of the proper type.

Any responses that failed schema validation were written out in detail, including the debugging information, and retried with exponential back-off (base delay of 20 seconds, jittered by a factor of 1.5–2.5 per attempt, capped at 5 minutes), for a maximum of five attempts per document. Any context-length error received from the API resulted in the document being skipped immediately, and no additional retries were made.

Output

The extraction pipeline produced 165 country-level JSON files, with each containing structured provisions for the three policy areas. Successfully validated extractions were cached to disk, thus enabling idempotent re-execution of the pipeline.

4.7 Embedding-Based Similarity Analysis

The similarity analysis stage quantitatively assessed the degree of policy convergence among countries by computing the semantic similarity of legislative content and policy-specific keyword vectors via dense text embeddings.

4.7.1 Embedding Model

All of the text embeddings were generated using OpenAI's `text-embedding-3-large` model [31], which generates 3,072-dimensional dense vector representations. This model was chosen because it has demonstrated state-of-the-art performance on semantic textual similarity benchmarks and can capture domain-specific nuances in legal and policy texts. Texts greater than the model's 8,190-token input limit were truncated to the input limit after tokenization using the `tiktoken` library.

4.7.2 Country-Level Similarity Scoring

For each country and for each policy area, the following steps were taken:

1. The English summaries generated by the LLM in Stage 5 were collected.
2. Embeddings were generated for each summary and for a set of topic-specific keywords.
3. A cosine similarity matrix was generated for the embeddings of the summaries and keyword embeddings using `sklearn.metrics.pairwise.cosine_similarity`.
4. The summary with the highest maximum similarity to any keyword was identified as the "best-matching" summary.
5. The final similarity scores for each country were determined as both the mean and maximum cosine similarity between the best-matching summary embedding and all keyword embeddings.

The topic-specific keyword sets were defined as follows:

- **River basin management**: “river basin management”, “river basin plan”, “river basin council”, “river basin organization”, “catchment management”, “watershed management”, “basin planning”.
- **Pollution control**: “polluter-pays principle”, “pollution limits”, “pollution prevention”, “cost recovery from polluter”, “environmental liability”, “pollution charges”, “pollutant discharge”.

4.7.3 Country Name Resolution

Since source filenames utilized diverse naming conventions for country names, a GPT-4.1-mini-based name resolution process was implemented to standardize English country names from raw filename stems. The resolved country names were saved to a persistent JSON file to ensure consistent country names across multiple executions of the pipeline and to reduce redundant API requests.

4.7.4 Country-to-Country Similarity

In addition to country–keyword similarity, country-by-country similarity matrices were generated. For each policy area, the aggregated legal text for each country was concatenated and embedded as a single vector. The resulting vector matrix (a matrix where one row corresponds to a country) was then used to generate a full pairwise cosine similarity matrix, yielding a $C \times C$ matrix (where C is the number of countries) for each of the three policy areas, as well as a single matrix representing the overall average similarity among countries.

4.7.5 Concordance Ranking

A concordance ranking procedure was specifically applied to the groundwater policy area to provide detailed annotations of the extracted provisions. Each extracted groundwater provision was annotated with its most semantically similar keyword from a set of seven groundwater-specific keywords:

1. “Groundwater rights”
2. “Groundwater monitoring”
3. “Underground water abstraction”
4. “Groundwater management”
5. “Groundwater permits”

6. "Transboundary aquifers"

7. "Aquifer recharge"

The concordance ranking procedure involved generating embeddings for all keyword terms and for all extracted legal text sections, calculating the cosine similarity between each text section embedding and each keyword embedding, and annotating each section with the keyword with the greatest similarity to the section (`top_keyword`) and the similarity value (`top_similarity`). This annotation enables further research on the thematic composition of groundwater legislation at the provision level.

4.8 Hierarchical Clustering and Visualization

4.8.1 Clustering Methodology

Country policies were clustered on the basis of their policy similarity profiles using two separate methodologies:

Hierarchical Clustering with Ward's Method

Hierarchical agglomeration was performed using Ward's minimum variance method [33] on the cosine distance matrices derived from the country-to-country similarity matrices. The distance metric was specified as $d_{ij} = 1 - s_{ij}$, where s_{ij} represents the cosine similarity between countries i and j. The resulting condensed distance matrix (the upper triangle) was then passed to the `scipy.cluster.hierarchy.linkage` function with `method='ward'`, which minimizes the total within-cluster variance at each merge step. Dendrograms were generated for each of the three policy areas and for the overall average similarity, demonstrating the hierarchical relationships among countries' legal approaches to water resource management.

K-Means Clustering

K-means clustering was applied to the feature vectors derived from the extraction results, using five clusters ($k = 5$) based upon exploratory analysis and the interpretability of the resulting cluster groups. The number of clusters, or k, was determined by evaluating various cluster evaluation criteria and methodologies, such as the "elbow" method and silhouette scores for each potential k, which ranged from $k = 2$ to $k = 10$ (see Figure 5.5 in Chapter 5). Other values were considered; at $k = 3$ the resulting clusters had been merged too much so that some relevant differences between the groundwater focused framework and basin-focused framework could no longer be discerned. At $k = 7$ there were too many small, distinct groups and they did not have

a consistent policy profile. Ultimately, the use of $k = 5$ resulted in an interpretable clustering scheme consisting of five clusters with policy profiles consistent with the variety of water governance approaches that are discussed in the literature comparing different legal systems. The feature vectors for each country were generated from the section counts for each policy area and boolean coverage flags, and were normalized before clustering.

4.8.2 Visualization Suite

A comprehensive visualization suite was developed to effectively communicate the results of the clustering and the patterns of similarity observed:

- **Clustered heatmaps**: Cosine similarity matrices with the rows and columns reordered to reflect the hierarchical clustering. These were generated using `seaborn.clustermap` with Ward's method and Euclidean distance on the similarity profiles. Custom colormaps were used to differentiate the individual policy areas from the overall average.
- **Dendrograms**: Independent dendrograms for each of the three policy areas and the overall average, with the country labels rotated for legibility and dynamic scaling of figure size.
- **Country keyword heatmaps**: Heatmaps illustrating the frequency of occurrence of specific keywords across all 164 countries, allowing researchers to identify those policy concepts that have been legislatively codified most frequently.
- **Radar charts**: "Water Policy DNA" profiles for individual countries or clusters, demonstrating the relative intensity of legislation across the three policy areas.
- **Bar charts**: Frequency distributions of keyword occurrences illustrating the most common policy concepts across the entire corpus.

4.8.3 Anomaly Detection and Data Quality

While conducting the similarity analysis, a perfect similarity score of 1.000 was found between North Korea (KP) and Papua New Guinea (PG); however, this finding is unlikely since the legal systems and languages of these countries are so dissimilar. Investigation revealed that the anomaly was caused by an embedding-quality issue at the text-data level: both countries produced identical or very similar zero-content embeddings, suggesting that they contained little or no extractable legal text. North Korea was subsequently removed from the study, but Papua New Guinea remained in the study and was validated to contain substantial amounts of legal content. Therefore,

the final clean dataset for the earlier NLP pipeline consisted of 55 countries with valid data quality, whereas the comprehensive LLM pipeline retained 164 countries after removal of the single outlier.

4.9 Summary

This chapter describes how the methodology presented in the chapter transforms a highly heterogeneous corpus of 165 national water laws—with over 35 languages and 10 script systems—into comparable and structured representations of water law through a seven-stage computational pipeline. The methodology contributes to the literature in the following ways: (1) a multi-method text extraction architecture with smart routing based on document characteristics; (2) a multi-metric translation quality assessment framework with reference-free quality assessment (COMET), semantic fidelity assessment (embedding-based similarity analysis), and manual evaluation, with the identification of translation-induced semantic drift; (3) a dual LLM-based legal extraction strategy with strict schema validation and caching; and (4) an embedding-based similarity analysis pipeline with concordance ranking that supports both macro-level country clustering and micro-level provision annotation. The results of applying the methodology to the corpus of global water legislation are presented in Chapter 5, where the results of the pipeline are analyzed across all three policy areas—groundwater regulation, river basin management, and the polluter-pays principle—and the results of the clustering analysis demonstrate empirically supported typologies of water law.

Chapter 5

Results

This chapter presents the results of applying the seven-stage computational pipeline described in Chapter 4 to the global water legislation corpus. The presentation proceeds from upstream validation—translation quality and extraction coverage—through the substantive findings for each policy dimension, to the cross-topic analysis and clustering results that constitute the central empirical contributions of the thesis.

5.1 Translation Quality Assessment

RQ4 asked how machine translation quality affected the reliability of cross-lingual legal comparisons. A multi-metric translation fidelity framework, described in Section 4.5, was used to assess translation quality of all non-English documents within the corpus. Two complementary metrics—COMET reference-free quality estimation and embedding-based cosine similarity—were calculated for each of the 39 source languages represented in the translated portion of the corpus. These results provide the quantitative evidence needed to evaluate the reliability of the translated corpus on which all subsequent analyses depend.

5.1.1 COMET Scores by Language

Table 5.1 lists the COMET scores for the top five and bottom five languages. The COMET scores for all 39 languages ranged from 0.681 (Tigrinya, $n = 23$ segments) to 0.866 (Romanian, $n = 192$ segments) with a mean for the entire corpus of about 0.83.

COMET results indicate a typologically graded distribution of language scores. As previously noted, Latin-script European languages scored best because they are typically the ones with the largest amounts of data in machine translation systems. Six languages met or exceeded the conventional threshold for high-quality translation of 0.85: Romanian (0.866), Catalan (0.857), Estonian (0.856), Korean (0.852), Georgian (0.852), and Serbian (0.850). The bulk of the distribution—25 languages

Table 5.1: COMET reference-free quality estimation scores: top five and bottom five languages. The *count* column indicates the number of translated text segments evaluated per language.

Rank	Language	Mean COMET	Std. Dev.	Count
Top five				
1	Romanian	0.866	0.026	192
2	Catalan	0.857	0.023	3
3	Estonian	0.856	0.029	140
4	Korean	0.852	0.013	4
5	Georgian	0.852	0.022	33
Bottom five				
35	Arabic	0.794	0.035	198
36	Polish	0.788	0.038	214
37	Somali	0.751	0.061	55
38	Dari	0.701	0.102	9
39	Tigrinya	0.681	0.141	23

including Italian (0.848), Finnish (0.847), Indonesian (0.844), Greek (0.843), and 24 other languages—fall into the range between 0.80 and 0.85, and thus represent acceptable translation quality for further analysis. Below or just at the 0.80 threshold were Arabic (0.794), Polish (0.788), and Somali (0.751). Dari (0.701) and Tigrinya (0.681) had much lower scores with higher standard deviation (0.102 and 0.141 respectively), indicating variable translation quality across the different text segments.

The mean of approximately 0.83 for the entire corpus is greater than the generally accepted threshold of 0.80 for acceptable translation quality, providing a quantitative basis for confidence in the entire translated corpus. However, results for Dari and Tigrinya indicate that downstream findings for countries whose legislation has been translated from Dari and Tigrinya (Afghanistan and Eritrea, respectively) should be viewed with increased caution.

5.1.2 Semantic Fidelity and Contextual Flattening

Unlike the COMET results, the embedding-based cosine similarity analysis comparing the semantic representation of original-language and translated-language texts found many patterns that differ significantly from those found in the COMET results. Table 5.2 shows the joint COMET and cosine similarity scores for some languages selected to illustrate the phenomenon of *contextual flattening*.

One of the most striking results is the group of languages that have adequately high COMET scores yet very low embedding-based cosine similarity. Georgia's legislation, for example, had a COMET score of 0.852, which placed it in the top five languages for translation quality; however, its cosine similarity between the original-language

Table 5.2: Translation fidelity comparison: COMET scores versus embedding-based cosine similarity for selected languages illustrating semantic drift. Languages are ordered by cosine similarity to highlight the discrepancy with COMET scores.

Language	**COMET Mean**	**Cos. Sim. Mean**	**Cos. Sim. Std.**	**Count**
High fidelity (concordant metrics)				
French	0.834	0.825	0.071	444
Spanish	0.831	0.820	0.051	700
Italian	0.848	0.806	0.048	75
Catalan	0.857	0.804	0.062	3
Contextual flattening (discordant metrics)				
Georgian	0.852	0.240	0.087	33
Somali	0.751	0.235	0.095	55
Lao	0.831	0.208	0.107	39
Amharic	0.841	0.117	—	1
Moderate fidelity (partially discordant)				
Uzbek	0.814	0.313	0.131	29
Dari	0.701	0.407	0.096	9
Tigrinya	0.681	0.484	0.276	23

and translated-language embeddings was only 0.240. Amharic also demonstrated a similar difference: a COMET score of 0.841 and a cosine similarity of only 0.117. Lao demonstrated the same type of difference (COMET 0.831, cosine similarity 0.208).

As such, the results demonstrate that a neural machine translation system may generate translations that are grammatically correct and linguistically accurate (as measured by COMET) but lose the relevant domain-specific semantics (as measured by the embedding-based cosine similarity). We interpret this result as *contextual flattening*: the translation system produces fluent English but replaces the specific legal and regulatory terms of the original legislation with more general language resulting in a semantically-poorer output. Therefore, we believe that countries whose legislation has been translated from Georgian (Georgia), Amharic (Ethiopia), or Lao (Laos) will show less differentiation in the similarity analysis, since their translated texts may have converged to a generic "water management" semantic profile instead of the specific regulatory provisions of the original legislation.

On the contrary, languages with both high COMET scores and high cosine similarity—French (0.825), Spanish (0.820), Italian (0.806), and Portuguese (0.800)—have strong semantic preservation. This is likely due to both the typological proximity of these languages to English and the availability of large amounts of training data for these language pairs in current neural machine translation models.

5.1.3 Text Extraction Quality

Automated quality assessment of the text extraction stage assessed the quality of 104 documents across 10 script systems. Out of these 104 documents, 102 documents (98.1%) received an "Excellent" rating, indicating successful extraction with no detected issues. Two documents—the Serbian and Montenegrin water laws—received "Acceptable" ratings due to low Cyrillic content detection (0.0%), which was traced to the use of Latin-script variants of the Serbian and Montenegrin languages in the source documents rather than to extraction failure. No documents received "Poor" or "Missing" ratings.

In addition, the quality assessment by script system verified robust performance across all script families: all 74 Latin-script documents, all 14 Arabic-script documents, both Greek documents, and the single documents in Hangul, Lao, Ethiopic, Hebrew, Georgian, and Thaana scripts all successfully validated. On the other hand, only the Cyrillic category had difficulties, with 2 of 8 documents having problems, and both problems were related to the source document containing Latin-script versions of the languages instead of Cyrillic.

The quality of these extracted values shows that the initial pre-processing (document) stage did not add much noise to the overall process. As the text was successfully extracted from each of the ten script families, the next question is how reliable was the LLM-based extraction phase — which transformed raw legislative text into structured policy information? This question is addressed in the next section.

5.2 LLM Extraction Coverage

Prior to reporting the findings in this section, it would be appropriate to provide a brief explanation of how the number of countries referred to herein have been counted. The LLM extraction pipeline has extracted and processed water legislation for 165 countries. Following identification of North Korea as a non-standard embedding outlier (see Section 5.8) the total number of countries within the dataset was decreased to 164. In addition, only the first 65 countries identified with original-language source legislation were used to evaluate and cluster keyword frequencies and perform clustering since direct keyword matching will typically yield better results when performed against original language text versus translated versions. Three additional English-language source documents were removed from this subset prior to extracting the data due to poor data quality. An early NLP-based embedding pipeline provided additional quality filtering resulting in a final set of 55 countries whose data quality had been validated for that specific analysis.

The LLM-based Legal Information Extraction Pipeline produced 165 country-level

JSON files; each contained structured provisions for the three policy dimensions, i.e., groundwater regulation, river basin management and polluter-pays principle. Every extraction output was subjected to the five-point schema validation described in Section 4.6. Structural validation checked the presence of the three required top-level keys, correctness of the data type for every field and non-empty keyword list for extracted provisions. Every one of the 165 outputs passed the structural validation check, thus validating that there were no extraction failures due to invalid structure. It should be noted that structural validation confirms schema compliance only; it does not assess the factual accuracy of the extracted legal content.

The LLM-based extraction successfully extracted information about all 165 of the countries. The results for keyword frequencies and clusters that are in Sections 5.3–5.5, however, used only a 65-country subset that was previously created by an NLP-based extraction pipeline to provide validated keyword data and similarity matrices for those countries. The extraction pipeline provided the following coverage for the three policy dimensions within the 65-country subset:

- **Groundwater regulation**: 62 of 65 countries (95.4%) had at least one extracted provision.
- **River basin management**: 60 of 65 countries (92.3%) had at least one extracted provision.
- **Polluter-pays principle**: 60 of 65 countries (92.3%) had at least one extracted provision.

There were only two countries that did not have any groundwater regulations and five countries that did not have any river basin or pollution regulations. Manual inspection validated that these countries' water legislation either did not provide any relevant provisions to either the relevant policy dimension or they had very limited provisions concerning the respective policy dimension (i.e., general environmental provisions).

To help the reader better understand the results presented below, it is important to be aware of the two techniques for extracting data that have been employed in Sections 5.3–5.5. Keyword frequency counts in Sections 5.3.2, 5.4.2 and 5.5.2 were produced by a lightweight topical specific extraction process that utilized an exact keyword match against the legal texts. Provision counts extracted using the large language model (LLM) in Sections 5.3.4, 5.4.3 and 5.5.3 were produced by the comprehensive AquaLex Scrutinizer pipeline which determined whether a provision was relevant based upon semantic relevance as opposed to an exact phrase match. Thus, these two processes can be viewed as mutually exclusive; i.e., they will extract provisions that contain either explicit regulatory language terms, or alternative/broader

language, therefore providing two differing numbers of provisions based upon the two extraction techniques being utilized.

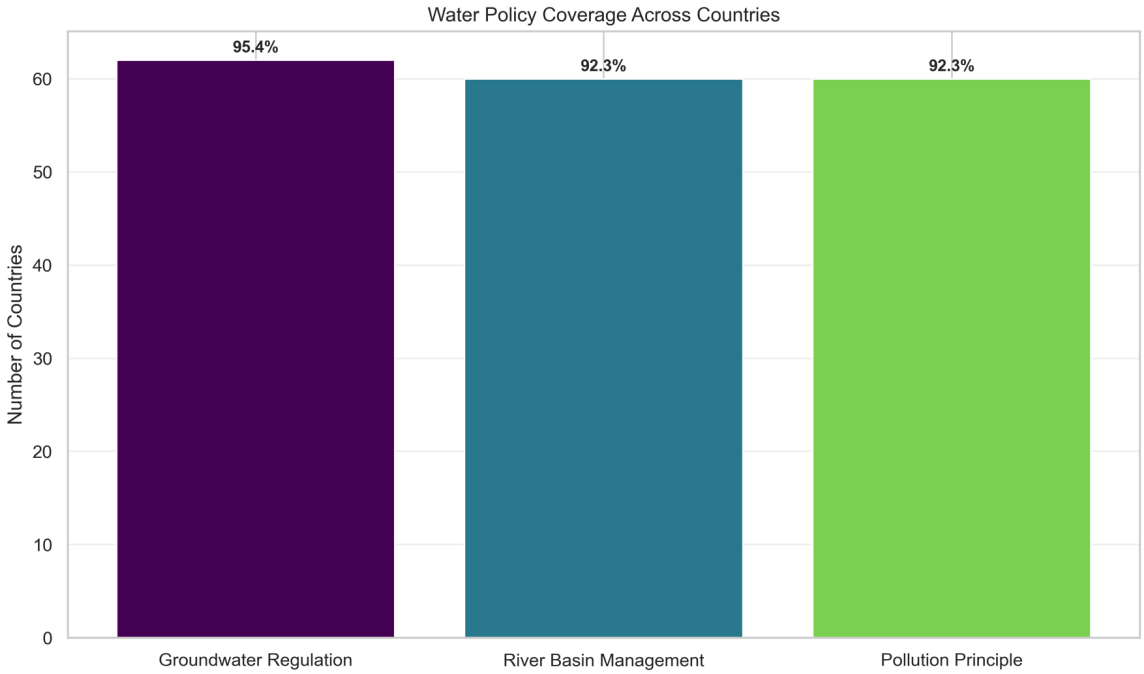

Figure 5.1: Policy coverage across the three policy dimensions for the 65-country analysis subset.

5.3 Groundwater Regulation: Global Patterns

Sections 5.3–5.5 of this dissertation addressed RQ2, which asked about the global patterns in groundwater management, river basin governance, and polluter-pays principle adoption across national water legislation.

5.3.1 Policy Coverage and Provision Density

Groundwater regulation was the most frequently addressed policy dimension in the corpus, as groundwater provisions were found in the water legislation of 62 of the 65 countries (95.4%) studied. The density of groundwater provisions varied greatly across countries, ranging from only one provision (e.g., Singapore, with only 1 extracted section) to over 200 provisions (North Macedonia, with 205 extracted sections).

5.3.2 Keyword Frequency Analysis

Keyword frequency analysis across the 65-country subset (Figure 5.2) showed that the term “aquifer” is the most common groundwater-related term, occurring 135 times

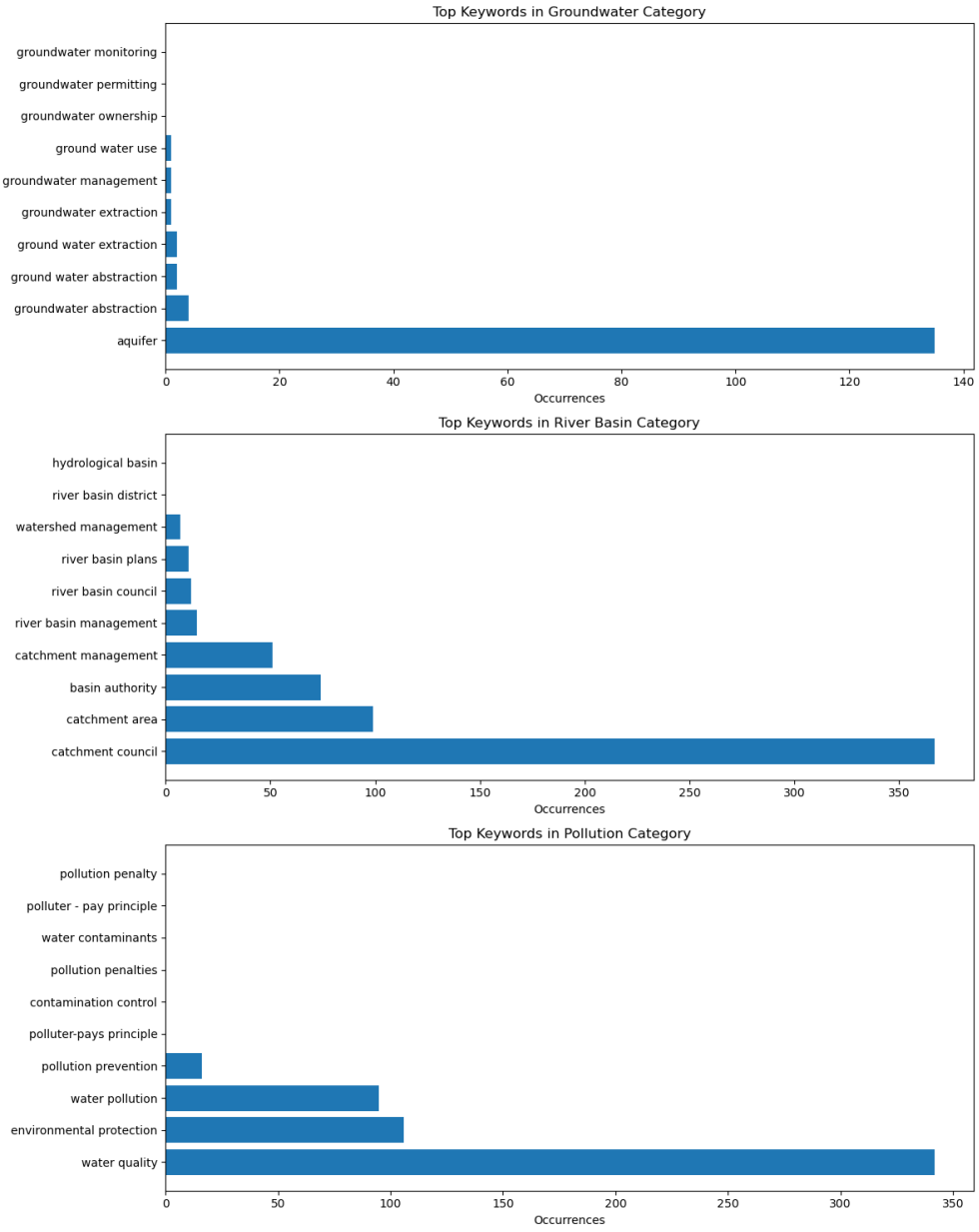

Figure 5.2: Top keywords by frequency across all three policy domains: groundwater (top), river basin (middle), and pollution (bottom). The dominance of “aquifer”, “catchment council”, and “water quality” reflects the terminology preferences of national water legislation in the corpus.

across all country profiles. All of the other groundwater-specific keywords occurred far less often, including: “groundwater abstraction” (4 occurrences), “ground water abstraction” (2 occurrences), “ground water extraction” (2 occurrences), “groundwater extraction” (1 occurrence), “groundwater management” (1 occurrence), and “ground water use” (1 occurrence). The keywords “groundwater ownership”, “groundwater permitting”, “groundwater monitoring”, “transboundary aquifers”, “groundwater rights”, and “groundwater governance” were not found with an exact match. The very low count of occurrences for the compound regulatory terms (such as groundwater abstraction) and the compound regulatory term “groundwater extraction” reflects that the exact match of a given keyword is being searched instead of the fact that these legal concepts are absent from the national legislation. This is further demonstrated by the results reported in the LLM-Extracted Provision Counts subsection below where the Legal Language Model was able to identify all the legal provisions concerning groundwater regulation regardless of what specific keywords were used.

This distribution shows that while national water legislation addresses groundwater provisions extensively, it rarely utilizes the specific technical terminology that would be associated with groundwater governance frameworks. The widespread use of “aquifer” (and the lack of use of compound regulatory terms like “groundwater permitting” or “groundwater monitoring”) suggest that these concepts are typically embedded in broader regulatory frameworks, rather than being referenced as standalone provisions.

5.3.3 Country-Level Groundwater Keyword Prevalence

Table 5.3 displays the countries with the largest numbers of groundwater keywords, based on exact keyword matching across the extracted legal texts.

Table 5.3: Top 10 countries by total groundwater keyword count. The dominant keyword is “aquifer” in all cases except where noted.

Rank	Country	Total Count	Primary Keyword
1	Namibia (NA)	22	aquifer
2	Belize (BLZ)	12	aquifer
3	Kenya (KE)	8	aquifer
4	Pakistan (PK)	7	aquifer, groundwater management
5	Lesotho (LSO)	6	aquifer
6	Zambia (ZM)	6	aquifer
7	Bangladesh (BGD)	6	aquifer
8	South Africa (ZA)	6	aquifer
9	Mauritius (MU)	6	aquifer
10	Australia (AU)	5	aquifer

It is noteworthy that the countries that ranked the highest in terms of groundwa-

ter keyword prevalence are sub-Saharan African countries (Namibia, Kenya, Lesotho, Zambia, and South Africa). This may indicate that groundwater is an especially important resource for arid and semi-arid climates, and therefore that national water legislation has placed more emphasis on the legislative framework governing aquifer management in those countries.

5.3.4 LLM-Extracted Provision Counts

The LLM-extracted provision counts represent another method of assessing how much legislative attention has been paid to groundwater through the number of provisions extracted per country. Within the 65-country subset, North Macedonia ranked first with 205 groundwater-related provisions, followed closely by Armenia (148), the Kyrgyz Republic (119), Kenya (114), South Africa (102), the UK (99), Albania (97), Namibia (88), Sweden (87), and Papua New Guinea (87). High LLM-provision counts are indicative of both the comprehensiveness of the underlying legislation, and the success of the LLM in extracting all relevant groundwater provisions, regardless of whether they utilized specific groundwater terminology.

5.4 River Basin Management

5.4.1 Policy Coverage

Provisions related to river basin management were identified in 60 of the 65 countries (92.3%) analyzed. The five countries without extracted river basin provisions were Barbados, Canada, Nepal, New Zealand, and Singapore. The former three countries lack basin-level governance structures in their water legislation, while the latter two countries utilize alternative institutional frameworks for water management, e.g., New Zealand's Resource Management Act uses a regional council model rather than basin-based governance.

5.4.2 Keyword Frequency Analysis

The river basin keyword analysis demonstrates a rich and varied terminology landscape. Unlike the groundwater dimension, where a single keyword dominated, river basin management terminology is spread out across multiple terms representing various governance traditions:

- **Catchment council**: 367 occurrences—the most frequent term, and representative of the prevalence of catchment-based governance in Commonwealth legal traditions.

- **Catchment area**: 99 occurrences—a geographic rather than institutional term.
- **Basin authority**: 74 occurrences—representative of the continental European and Francophone tradition of basin-level institutional governance.
- **Catchment management**: 51 occurrences—an operational term referring to active management interventions.
- **River basin management**: 15 occurrences—the EU Water Framework Directive terminology.
- **River basin council**: 12 occurrences—institutional governance at the basin level.
- **River basin plans**: 11 occurrences—planning instruments.
- **Watershed management**: 7 occurrences—primarily North American terminology.

As shown in Figure 5.2, the vast majority of countries (61 of 65) in the dataset utilize the term "catchment council". The significantly smaller number of countries utilizing "river basin management" (15 occurrences) highlights the fact that the 65-country English language subset is heavily represented by Commonwealth nations in Africa, the Caribbean and the Pacific. Those countries have implemented catchment-based governance systems in their water legislation that are modeled after British water law, whereas the "river basin" terminology has been promoted by the EU Water Framework Directive and the IWRM paradigm [12].

5.4.3 LLM-Extracted Provision Counts

The LLM-extraction identified the highest number of river basin provisions in Zimbabwe (185 sections), followed by Australia (165), Japan (139), Zambia (133), South Africa (106), Albania (95), North Macedonia (83), China (60), Norway (51), and Namibia (51). The presence of Zimbabwe, Australia and South Africa among the top-ranked countries aligns with the previously documented adoption of comprehensive basin-based water governance frameworks in these countries. Zimbabwe's Water Act of 1998 created catchment councils and sub-catchment councils as the primary institutional mechanism for water resources management, whereas the Australian Water Act of 2007 created the Murray-Darling Basin Authority with considerable basin planning authority [5].

5.5 Polluter-Pays Principle

5.5.1 Policy Coverage

Pollution-related provisions were identified in 60 of the 65 countries (92.3%) studied, identical to the coverage of river basin management. The five countries without extracted pollution provisions are characterized by either having very short or narrowly focused water legislation that does not address water quality or environmental liability.

5.5.2 Keyword Frequency Analysis

The pollution keyword analysis revealed a significant terminology finding: the specific phrase "polluter-pays principle" was not found anywhere in the 65 country profiles examined through exact keyword matching (0 occurrences). Although the keyword search used an exact string match on "polluter-pays principle," with its most common variations ("polluter pays," "polluter-pays," "the polluter shall pay"), it did not include semantic variations such as "liability for pollution" or "recovery of costs for environmental harm." However, the LLM extracted these forms of liability, using a form of contextual interpretation instead of a simple keyword match. Pollution-related governance could be identified even without using the exact term, by searching through the use of broader terms that relate to pollution:

- **Water quality**: 342 occurrences—the most frequently used pollution-related term; it represents the widespread use of water quality standards as the dominant regulatory mechanism.
- **Environmental protection**: 106 occurrences—a broad term encompassing pollution control in general environmental frameworks.
- **Water pollution**: 95 occurrences—direct reference to pollution as a regulatory concern.
- **Pollution prevention**: 16 occurrences—proactive measures against pollution.

The fact that the specific phrase "polluter-pays principle" could not be located in the corpus is an important finding. Although the principle has been generally acknowledged in international environmental law and included in various international agreements (e.g., OECD Recommendation on Guiding Principles Concerning International Economic Aspects of Environmental Policies; 1972 and Rio Declaration; Principle 16), none of the countries in this 65-country sample explicitly referenced the principle by name in their water legislation. However, some type of pollution related provision existed in 60 out of 65 countries. These include regulations for water quality standards,

discharge permits, pollution penalties, etc. The existence of these provisions are logically consistent with the reasoning behind the polluter pays principle. However, they do not represent evidence that the principle has been formally recognized as a policy norm or standard of conduct within those jurisdictions.

5.5.3 LLM-Extracted Provision Counts

In addition to identifying provisions by keyword matching, the LLM-extraction also identified pollution-related provisions by their semantic relevance. The LLM-extracted provision counts provide a more complete view of pollution-related legislative attention. Sweden ranked first with 109 pollution-related provisions, followed by North Macedonia (96), the Philippines (84), Albania (47), the Kyrgyz Republic (41), Mongolia (36), South Sudan (34), New Zealand (31), Belize (30), and Canada (29). The difference between the results of the exact keyword analysis (the "polluter-pays principle" matched zero times) and the results of the LLM-extraction (dozens of countries were identified as having substantial numbers of pollution-related provisions) demonstrate the value of semantic extraction compared to keyword matching for comparative legal analysis.

Together, both the keywords and the LLM extraction methods used to analyze the three policy areas indicate a common trend: Exact keyword matches find very little (a small portion) of the regulatory content found in national water legislation as compared to the LLM based method's ability to produce a much fuller picture of what has been legislated. The difference between these two methods is most evident when looking at the polluter-pays area, however this trend can be seen with some variation in all three of the policy areas. Since we have identified what the pipeline found in each of the policy areas separately; we are now going to see if there is a relationship between the three policy areas at the country level — that is, do countries enact laws evenly across the three policy areas of groundwater, river basins, and pollution or do countries focus more of their legislative efforts on one policy area over the other(s).

5.6 Cross-Topic Analysis

5.6.1 Policy Dimension Balance

The relative allocation of legislative attention across the three policy dimensions varies substantially across countries, revealing distinct regulatory profiles. To quantify this balance, the proportion of extracted provisions allocated to each dimension was computed for each country. Table 5.4 presents representative countries illustrating the range of policy profiles observed.

Three broad regulatory archetypes emerge from this analysis:

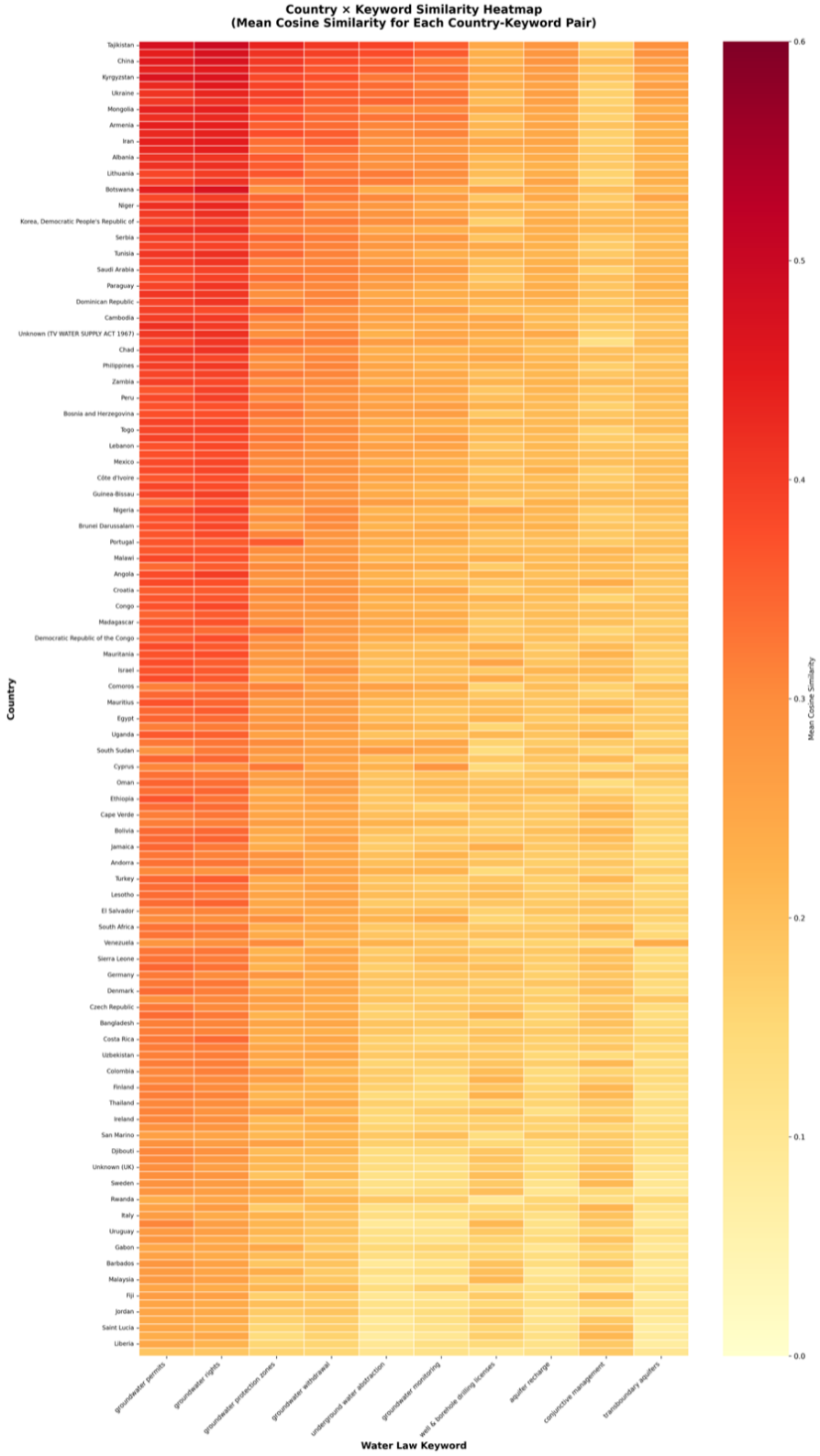

Figure 5.3: Country–keyword similarity heatmap for all 164 countries across 10 groundwater-related keywords, computed using sentence-transformer embeddings. Darker shading indicates higher cosine similarity.

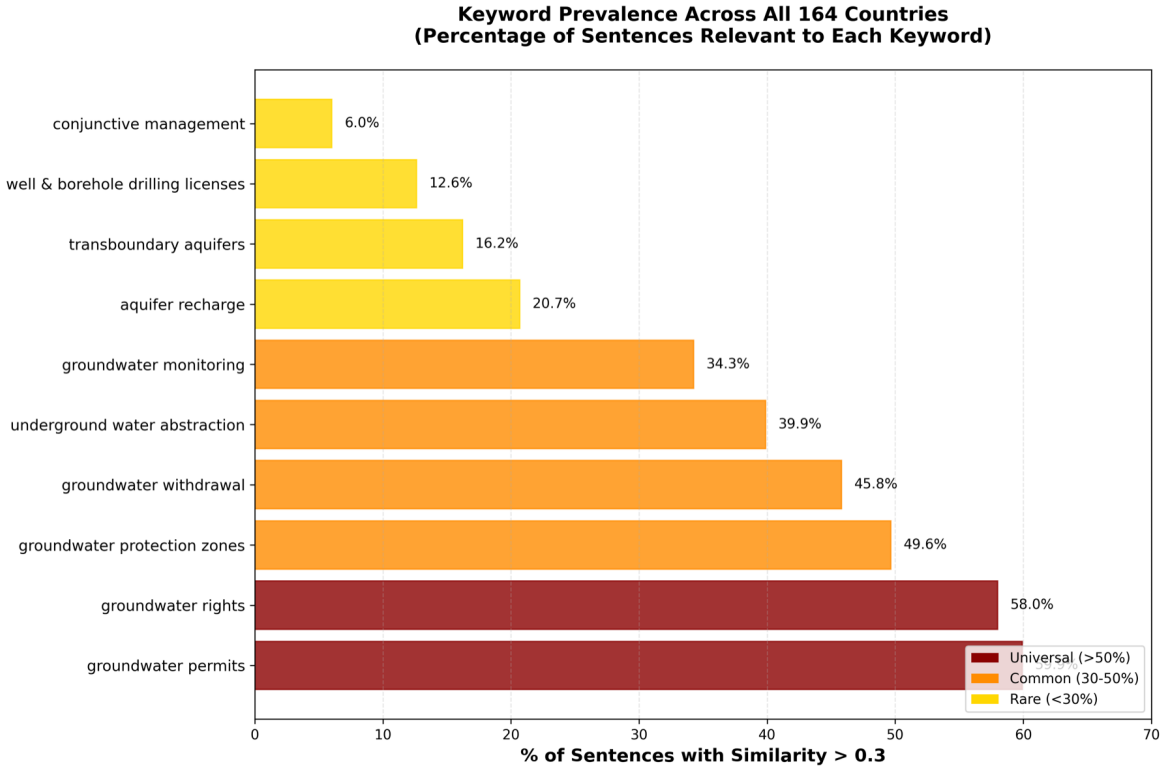

Figure 5.4: Keyword prevalence across 164 countries, showing the frequency of each groundwater-related keyword in the full corpus.

Table 5.4: Policy dimension profiles for selected countries, showing the percentage of extracted provisions allocated to each policy dimension. Countries are selected to illustrate the range of regulatory profiles.

Country	**Groundwater (%)**	**River Basin (%)**	**Pollution (%)**
Groundwater-dominant			
Botswana (BW)	92.3	1.5	6.2
Mauritius (MU)	89.5	9.3	1.2
Papua New Guinea (PG)	88.8	8.2	3.1
River-basin-dominant			
Japan (JP)	13.1	86.9	0.0
Australia (AU)	7.0	82.5	10.5
Zimbabwe (ZW)	27.6	70.9	1.5
Pollution-dominant			
Canada (CA)	3.3	0.0	96.7
New Zealand (NZ)	6.1	0.0	93.9
Philippines (PH)	11.5	1.0	87.5
Balanced			
Albania (AL)	40.6	39.7	19.7
Mongolia (MN)	52.0	24.0	24.0
South Sudan (SS)	46.3	28.7	25.0

1. **Groundwater-centric frameworks**: Groundwater-focused statutes are found in small island states and countries with limited surface water resources, such as Botswana (92.3% groundwater), Mauritius (89.5%), and Papua New Guinea (88.8%), which allocate the vast majority of their legislative provisions to groundwater management.

2. **Basin-governance frameworks**: Countries with strong river basin governance, such as Japan (86.9% river basin), Australia (82.5%), and Zimbabwe (70.9%), have a heavy emphasis on basin-level provisions, reflecting the adoption of basin-based governance models as the central organizing principle of their water legislation.

3. **Environmental-quality frameworks**: Pollution-focused statutes are often found in countries such as Canada (96.7% pollution), New Zealand (93.9%), and the Philippines (87.5%), which focus predominantly on pollution control and environmental quality.

A fourth type of country—such as Albania (40.6/39.7/19.7%), South Sudan (46.3/28.7/25.0%) and Mongolia (52.0/24.0/24.0%)—has an equal amount of provisions related to groundwater, basin management and pollution, suggesting comprehensive water governance frameworks that address all three dimensions with roughly comparable legislative attention.

5.6.2 Comprehensive vs. Partial Regulatory Frameworks

North Macedonia, with 384 total provisions (205 groundwater, 83 river basin, 96 pollution) ranks first in terms of legislative comprehensiveness; second place goes to South Africa (235); third to Zimbabwe (261); fourth to Australia (200) and fifth to Sweden (207). At the other extreme, Singapore (2) and Sri Lanka (2) rank last in terms of the number of provisions, indicating a lack of water legislation or that it is spread across multiple statutes that were not included in the study.

5.7 Clustering Results

Clustering provides insight into the grouping of countries with similar policies in the water sector. The clustering analysis addresses RQ3, which asks whether national water laws cluster into identifiable typological groups based on their policy content.

5.7.1 K-Means Clustering

Using K-means clustering with five clusters ($k = 5$, designated as Clusters 0–4), we analyzed the 65-country feature vector data generated from our LLM-extraction results.

Each feature vector consisted of the number of sections in each policy dimension, as well as boolean coverage flags. After normalization, the feature vectors were used for clustering.

The optimal value of k is determined by using both silhouette analysis and the elbow method. Figure 5.5 displays the results for k values between 2 and 10. The silhouette coefficient reaches its highest value at $k = 3$ (0.482) and again at $k = 5$ (0.476). The difference between these two peaks is 0.006. The elbow plot also displays an obvious change in slope between $k = 4$–5, after which there is a substantial decrease in the marginal amount of inertia reduced as k increases. Since $k = 3$ has the largest silhouette coefficient, we chose $k = 5$, as it allows us to differentiate types much more finely than $k = 3$. In addition, we believe that the additional type of statute will help capture the differences in the corpus (i.e., between those statutes that describe river-basin-dominant frameworks and those that are narrowly focused on specific statutes), which would otherwise have been included in the same cluster at smaller k values.

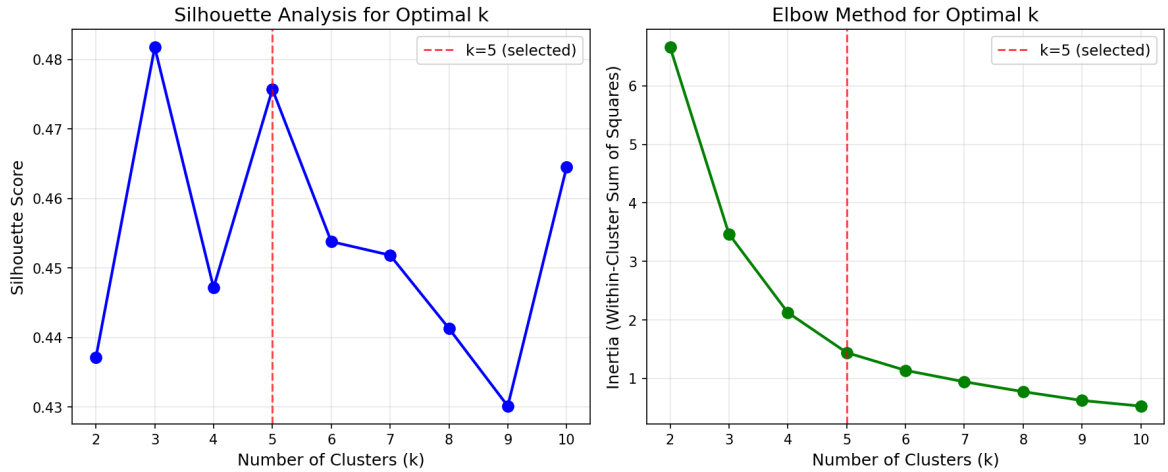

Figure 5.5: Silhouette analysis (left) and elbow method (right) for determining the optimal number of clusters. The silhouette coefficient peaks at $k = 3$ (0.482) and $k = 5$ (0.476); the elbow plot shows diminishing returns after $k = 5$.

Table 5.5 summarizes the characteristics of each cluster.

Cluster 0: Groundwater-Centric, Moderate Diversity

Cluster 0 represents a variety of groundwater-centric frameworks, but with moderate diversity. The majority of Commonwealth countries in sub-Saharan Africa (Kenya, Tanzania, Uganda, Botswana, Namibia, Malawi, Lesotho, Swaziland), the Caribbean (Jamaica, Belize, Dominica, Saint Lucia, Saint Kitts and Nevis, Saint Vincent and the Grenadines, Bahamas, Guyana), and the Pacific (Fiji, Tonga, Vanuatu, Samoa, Papua New Guinea) are represented in this cluster. They have a major focus on groundwater abstraction, well drilling and aquifer protection, due to their reliance on groundwater

Table 5.5: Summary of K-means clustering results ($k = 5$) for the 65-country subset. Cluster characterization is based on the dominant policy profile of member countries.

Cluster	**Size**	**Dominant Profile**	**Representative Countries**
0	38	Groundwater-centric, moderate diversity	Bangladesh, Belize, Botswana, Kenya, Jamaica, Mauritius, Namibia, Norway, Pakistan, Tanzania, Uganda, UK
1	9	Comprehensive, high provision count	Albania, Armenia, Kyrgyz Republic, North Macedonia, Mongolia, India, Philippines, Sweden, South Sudan
2	9	River-basin-dominant	Australia, Brazil, China, Grenada, Gambia, Japan, South Africa, Zambia, Zimbabwe
3	3	River basin/pollution, no groundwater	Great Britain (Water Industry Act), Sri Lanka, Myanmar
4	5	Minimal provisions, narrow scope	Barbados, Canada, Nepal, New Zealand, Singapore

resources. However, they tend to provide less detailed information regarding basin-level governance and pollution control. The average proportion of groundwater in this cluster is approximately 63%, with river basin provisions averaging about 22% and pollution provisions averaging around 15%.

The large size of cluster zero is also due to its spread through out many parts of the world; (the cluster contains 38 of the 65 countries within the sample), making it represent about 58% of all countries analyzed. The presence of such a high percentage of cluster zero within the sample is representative of the broad international application of the commonwealth law traditions in relation to legislation focused upon groundwater. Therefore, although there may appear to be an over representation of cluster zero within the data and therefore an issue with the clustering methodology, this can be explained by the broad geographical spread of the Commonwealth throughout Africa, the Caribbean, and the Pacific regions. However, the strong association of so many countries within cluster zero limits the ability to make fine grained distinctions amongst them. As a result, additional detail could potentially be provided regarding these countries' respective water law systems only through an increase in the number of features used to define each country's system or through performing an additional level of clustering to provide a clearer understanding of the various levels of similarity present amongst the members of cluster zero.

Cluster 1: Comprehensive, High Provision Count

This cluster (9 countries) contains countries with the most complete and comprehensive water legislation in the corpus. This includes the former Soviet republics and Yugoslavia, who traditionally write extensive, codified water legislation addressing all three policy areas equally. Countries like North Macedonia (384 total provisions), Armenia (214), the Kyrgyz Republic (204), and Albania (239) are examples of countries with this type of legislation. Other countries in this cluster include Sweden's Environmental Code (207 provisions) and South Sudan's Water Policy (136 provisions), which also have equally comprehensive legislation. The average number of provisions for this cluster is approximately 168, compared to approximately 31 for Cluster 0 and approximately 60 for Cluster 2.

The notable aspect of Cluster 1 is the relatively even distribution among the three policy areas. The average groundwater proportion in this cluster is approximately 44%, while the river basin proportion averages approximately 18% and the pollution proportion averages approximately 30%. The high pollution proportion is largely due to the fact that India's Water Act (75.9% pollution) and the Philippines' Water Act (87.5% pollution) are primarily pollution-control statutes, although they are part of this cluster.

Cluster 2: River-Basin-Dominant

This cluster (9 countries) is defined by the dominance of river basin management provisions. Countries such as Zimbabwe (70.9% river basin), Japan (86.9%), and Australia (82.5%) are the archetype, as they have implemented basin-based governance as the main organizing principle for their water legislation. South Africa (45.1% river basin, 43.4% groundwater) and Zambia (52.8% river basin, 42.5% groundwater) are members of this cluster, as they have developed comprehensive catchment management systems in addition to considerable groundwater provisions. The average river basin proportion in this cluster is approximately 63%.

Both Commonwealth (Australia, Zimbabwe, Zambia, South Africa) and non-Commonwealth (Japan, Brazil, China) countries are present in this cluster, indicating that basin-based governance is being implemented beyond colonial legal traditions. For instance, Brazil's National Water Policy of 1997 created river basin committees based on the French *agences de l'eau* model instead of the British catchment council tradition, but generates a similar legislative profile in this analysis.

Cluster 3: River Basin/Pollution, No Groundwater

The smallest cluster (3 countries) comprises Great Britain's Water Industry Act 1991, Sri Lanka's Water Act, and Myanmar's Conservation of Water Resources and Rivers

Law (2006). The common element in the three countries in this cluster is the absence of any groundwater provisions in their primary water legislation, with the majority of legislative attention focused on river basin management (59.5% for Great Britain, 100% for Sri Lanka, 76.9% for Myanmar) and pollution control (40.5% for Great Britain, 23.1% for Myanmar). This is because in these countries groundwater is regulated separately through other statutory instruments.

Cluster 4: Minimal Provisions, Narrow Scope

This cluster (5 countries) contains countries with very few extracted provisions: Barbados (12), Canada (30), Nepal (9), New Zealand (33), and Singapore (2). Although there is little overlap in the member countries, the countries in this cluster are similar in their narrow legislative focus. Specifically, the legislative focus of the countries in this cluster is generally limited to a subset of water governance issues, and therefore comprehensive water governance coverage is often distributed across multiple legislative instruments. For example, Canada's Water Act focuses nearly exclusively on pollution control (96.7%), while New Zealand's Resource Management Act focuses on pollution and environmental management as part of a broader resource consent regime.

The five K-means clusters shown above describe a useful way to partition data, but they show no information about how well-defined or compactly clustered the objects in each group are, nor do they indicate at what point clusters will merge as you increase the degree of dissimilarity. To investigate these issues, hierarchical agglomerative clustering was conducted on the same features and by using Ward's linkage method, resulting in dendrograms that illustrate the nested hierarchy of country relationships.

5.7.2 Hierarchical Clustering and Dendrogram Analysis

The dendrograms obtained through the use of Ward's method to perform hierarchical agglomerative clustering on the cosine similarity matrices for the three policy dimensions and the overall average provide visual representations of the nested hierarchies of legislative similarity among the countries included in the study. Figure 5.6 presents the groundwater dendrogram, and Figure 5.7 presents the overall average dendrogram.

In the groundwater dendrogram (Figure 5.6), we observe that the sub-Saharan African and Caribbean countries exhibit a tightly clustered group at relatively low merge heights, indicating that they share the same Commonwealth legislative tradition in their groundwater governance frameworks. A notable separation exists between the largest cluster of countries with moderate-to-high levels of groundwater similarity (approximately 1.0 merge height) and the smallest cluster of outlier countries—Canada, Fiji, Ireland, and Japan—that are merged at the highest level of the hierarchy, suggesting significant differences in their groundwater governance approaches.

The dendrogram for the overall average similarity across the three policy dimensions (Figure 5.7) provides the most detailed representation of the nested hierarchy of legislative similarities. At the highest level of the hierarchy, we observe a clear separation of comprehensive water governance frameworks (Clusters 1 and 2) from those that are more narrowly focused (Clusters 3 and 4), with Cluster 0 being located in an intermediate position in the hierarchy. The colour-coded branches reflect the five clusters that were identified by K-means, thereby confirming the correspondence between the partitional and hierarchical clustering analyses.

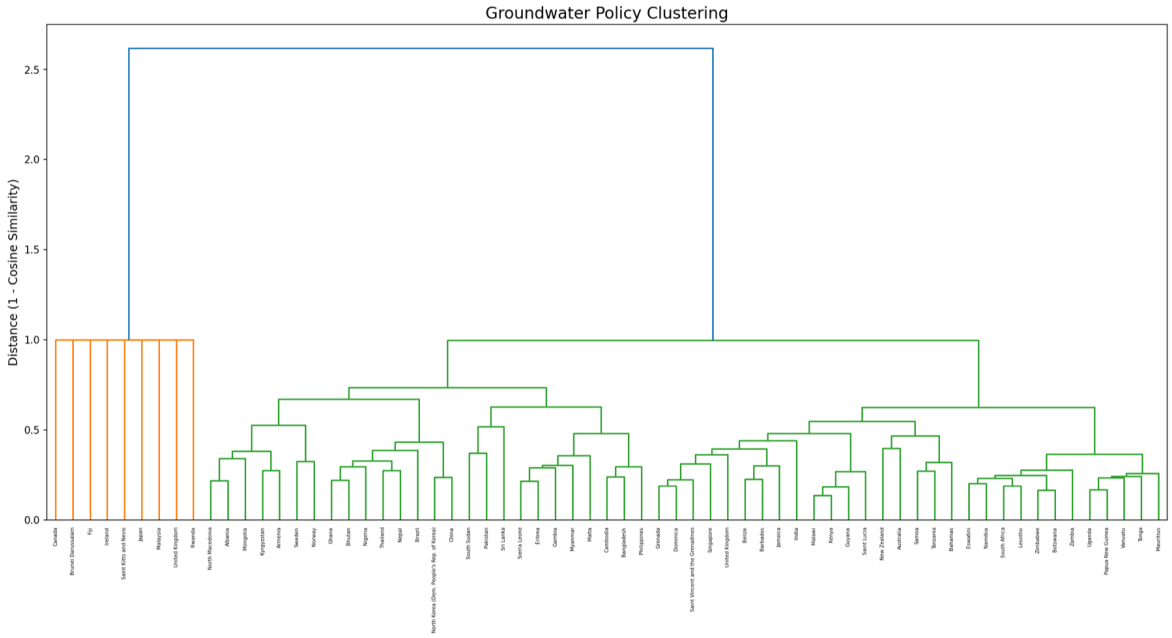

Figure 5.6: Hierarchical clustering dendrogram (Ward's method) for groundwater regulation similarity. Branch colours indicate cluster membership. The y-axis represents the cosine distance ($1 - \text{similarity}$) at which clusters merge.

We also developed a series of clustered heatmaps for each policy dimension, merging the pairwise cosine similarity matrices with the hierarchical clustering dendrograms to reorder the rows and columns of the heatmaps. In addition to observing strong diagonal dominance in the heatmaps for all three policy dimensions (indicating that the clustering algorithm produced internally coherent groups of countries), we observed three key patterns of similarity in the heatmaps: (i) distinct off-diagonal blocks of elevated similarity among cluster pairs (for example, the cluster of countries with river-basin-dominant frameworks and the cluster of countries with comprehensive frameworks); (ii) the similarity of countries in the same cluster; and (iii) topic-specific variation, with the similarity pattern for groundwater being the most uniform, and the similarity patterns for pollution being the most fragmented, indicating a wide range of regulatory approaches from command-and-control standards to market-based instruments.

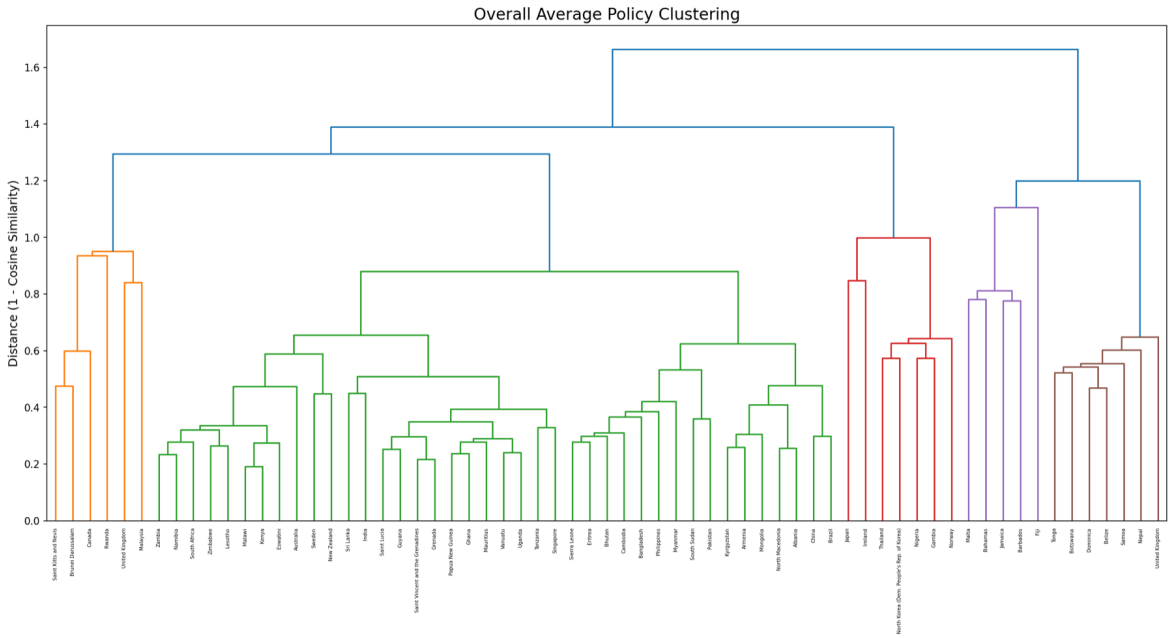

Figure 5.7: Hierarchical clustering dendrogram (Ward's method) for overall average policy similarity across all three dimensions. Branch colours indicate cluster membership identified by K-means ($k = 5$).

5.8 Anomaly Detection and Data Quality

Before we synthesize our results, we will describe the data quality controls used to establish the validity of the results presented above.

5.8.1 The North Korea–Papua New Guinea Anomaly

When computing the pairwise cosine similarity matrices, we discovered a perfect similarity score of 1.000 between North Korea (KP) and Papua New Guinea (PG). On a substantive basis, the result is implausible: North Korea's water legislation was originally written in Korean, and describes a centrally planned water management system; whereas Papua New Guinea's Water Resource Act governs a decentralized, common-law-based water governance regime in a former Australian territory. After investigation, we discovered that both countries yielded identical or near-identical embedding vectors, likely due to one or both of them producing very little extractable content that collapsed into a nearly zero embedding vector. As a consequence of the identical embeddings, we computed a trivially perfect cosine similarity.

5.8.2 Quality Control Actions

North Korea (KP) was removed from all subsequent analyses. Papua New Guinea (PG) remained in all subsequent analyses after we performed a manual validation to confirm that there existed substantive legal content in its extracted text and LLM output. We

did not identify any other country pairs that yielded perfect or near-perfect similarity scores, so the anomaly appears to be isolated rather than systematic.

5.8.3 Final Dataset Characteristics

Following the removal of North Korea, the clean dataset that was input into the earlier NLP pipeline consisted of 55 countries with validated data quality, and reflected additional exclusions that occurred during the keyword and semantic matching stages for countries that had insufficient extractable content. The comprehensive LLM pipeline retained 164 countries after removing North Korea from consideration. The mean pairwise cosine similarity of the clean dataset was about 0.42, indicating a high degree of diversity—on average, countries in the corpus are more dissimilar than similar in their legislative approaches to water governance, which is what we would expect based on the diversity of national legal traditions. To provide a little more background information for this discussion, the cosine similarity score can range from 0 to 1. The 0 value will indicate two sets of vector embeddings (in this case, the country and regulatory/legislative data) that have no overlap or shared understanding in terms of how they are used to describe semantics. Conversely, when the value reaches 1, it will be an indication that both sets of vector embeddings are essentially duplicates of one another. Therefore, given the average value of 0.42, it can be inferred that there is clearly some degree of commonality among the countries' regulatory vocabularies; however, they tend to emphasize different aspects of legislation across the three policy dimensions, which supports the diversity of legal frameworks and government approaches represented within the corpus.

5.8.4 Sensitivity to Translation Quality

The anomaly detection process also identified a possible relationship between translation quality and similarity analysis. Countries whose legislation was translated from languages that exhibit translation-induced semantic drift (Georgian, Amharic, Lao) may yield embedding vectors that are systematically biased toward a generic semantic profile, which could increase their similarity to other countries. Two factors mitigate the impact of this concern: (1) only a few countries are subject to this bias; (2) the semantic similarity analysis uses LLM-extracted summaries rather than raw translations, which provides an additional layer of normalization. Thus, although this relationship represents a methodological limitation that is addressed in greater detail in Chapter 6, it does not appear to have significantly impacted the results.

5.9 Summary of Results

The seven-stage computational pipeline applied to the global water legislation corpus has produced the following principal findings:

1. **For the majority of languages, the quality of the translations is sufficient to support downstream analysis.** The mean COMET score of 0.83 for the corpus exceeds the commonly accepted threshold for acceptable quality. However, for a small number of languages (Georgian, Amharic, Lao, Uzbek, Somali), we identified semantic compression—where translations are linguistically fluent but semantically impoverished—which requires careful interpretation of results for the affected countries.

2. **The LLM-based extraction achieved near-full coverage.** The extraction pipeline successfully processed 165 countries with 100% schema validation pass rates. The proportion of countries covered by the extraction pipeline for each of the three policy dimensions ranged from 92.3% (river basin management and polluter-pays principle) to 95.4% (groundwater regulation).

3. **The policy dimension for groundwater regulation is the most consistently addressed**, and "aquifer" is the dominant keyword across all 65 countries. The lack of compound regulatory terms for groundwater indicates that groundwater governance is generally embedded within larger water management frameworks, and is not typically governed as a separate regulatory domain.

4. **The terminology for river basin management reflects colonial and institutional influences.** The prevalence of "catchment council" (367 instances) relative to "river basin management" (15 instances) illustrates how the Commonwealth legal tradition influenced the English language portion of the corpus.

5. **The polluter-pays principle is generally implemented implicitly rather than explicitly.** Although we did not find a single instance of "polluter-pays principle", we did find numerous provisions governing water quality standards, pollution penalties, and environmental protection.

6. **We identified five distinct legislative clusters**, representing a spectrum of groundwater-centric frameworks (38 countries) to narrowly scoped statutes (5 countries), with comprehensive water codes (9 countries), river-basin-dominant frameworks (9 countries), and surface-water-only legislation (3 countries) positioned in an intermediate position in the spectrum.

7. **Automated quality assurance processes identified and resolved one anomaly** (the North Korea–Papua New Guinea perfect similarity), illustrating

the necessity of automated quality assurance in large-scale computational legal analysis.

These results raise several new research questions that go beyond the data themselves: Do these computational results differ from the results obtained by Caponera and Burchi using the expert-driven comparative water law methodology? What are the implications of translation-induced meaning loss for the reliability of cross-lingual legal analysis? What do the identified legislative typologies imply for water governance policy and reform? These questions are addressed in Chapter 6 by situating the results within the comparative water law literature, by assessing the methodological limitations of the study, and by exploring the practical implications of the results for water governance.

Chapter 6

Discussion

The authors discuss the results of the previous chapter in relation to the research questions provided in the introduction, in order to place the results in the context of comparative water law and computational legal analysis, as well as to examine the methodological constraints of this study.

6.1 Answering the Research Questions

6.1.1 RQ1: Feasibility of the Computational Pipeline

RQ1 examined whether an entire computational pipeline can effectively extract, translate, and compare the water laws of more than 164 countries written in more than 35 languages. The results indicate that the response is yes; however, there are certain conditions.

The computational pipeline was able to successfully extract, translate, and analyze water laws from 165 countries using 35+ languages and 10+ script systems. As a result, the computational pipeline achieved an 98.1% extraction rate, which the authors consider to be "excellent" (Section 5.1.3), and 100% of the extracted data were schema validated for LLM-based extraction (Section 5.2). Also, the COMET translation quality score of 0.83 is above the commonly accepted threshold of adequacy, and the extraction pipeline had policy coverage rates of 92–95% across the three policy dimensions. Thus, based on these results, the authors conclude that the use of computational comparative legal analysis for water laws is technically feasible at a global level.

However, the reliability of the computational pipeline is not consistent across all languages and scripts. Specifically, the authors note that low-resource language countries (i.e., Tigrinya, Dari, Somali) and the semantic drift experienced by Georgian, Amharic, and Lao introduce systematic biases into the computational pipeline's reliability. Therefore, the computational pipeline is best characterized as highly reliable

for the majority of countries in the dataset—specifically those falling within the high- and adequate-quality translation tiers—and provisionally reliable for the remaining countries, where results should be interpreted as indicative rather than as definitive.

6.1.2 RQ2: Global Patterns in Water Legislation

RQ2 focused on the substantive legal findings across the three policy dimensions. A number of global trends emerged from the research.

Groundwater regulation is the most consistently regulated policy dimension. Approximately 95.4% of countries included at least one provision related to groundwater. The authors noted that "aquifer" is the most frequently occurring term in the legislative texts of all countries studied (135 times), and that compound regulatory terms (e.g., "groundwater permitting," "groundwater monitoring") are rarely used in legislative texts. The authors suggest that this indicates that groundwater governance is usually integrated into larger water management frameworks and not treated as a separate regulatory domain. This has significant implications for how groundwater governance is measured globally. Studies seeking to measure groundwater regulation using terms like "groundwater permitting" or "groundwater monitoring" will significantly undercount the actual amount of groundwater regulation, since these concepts are legislated under the broad umbrella of water abstraction permits, resource management plans, and environmental protection provisions.

Countries in sub-Saharan Africa dominate the list of countries with the highest numbers of groundwater-related keywords. Countries like Namibia (22 mentions), Kenya (8), Lesotho (6), Zambia (6), and South Africa (6) have higher numbers of groundwater-related keywords due to the fact that these countries depend heavily on groundwater as the primary or sole source of reliable water supplies. The large amounts of groundwater-related regulations in these countries reflect the critical need for effective groundwater management in arid and semi-arid regions.

River basin management terminology shows the influence of legal tradition on governance vocabulary. The use of the term "catchment council" (367 times) far outnumbers the use of "river basin management" (15 times) in the English-language subset of the dataset, showing the strong influence of the Commonwealth legal tradition on the terminology of water governance. This finding demonstrates that the conceptual framework of water governance is influenced not just by hydrological and environmental factors, but also by the colonial and institutional legacy of legal systems.

The polluter-pays principle presents the most interesting terminological finding: zero verbatim mentions of the phrase despite its widespread endorsement in international environmental law. Instead, countries implemented the polluter-pays principle implicitly through various mechanisms (e.g., water quality standards, discharge per-

mits, pollution penalties, and remedial actions). This gap between international legal vocabulary and national legislative drafting has significant implications for the way international environmental law principles are seen to migrate into national law. The results suggest that the migration of international environmental law principles into national law is more likely to occur through substantive regulatory mechanisms, rather than through the use of the principles' standard terminology. Therefore, assessments of the degree to which countries adopt international environmental law principles (such as the polluter-pays principle) through keyword searches in legislative texts would significantly underestimate the actual level of adoption of these principles. The same methodological caveat could apply to other international environmental law principles such as the precautionary principle or the principle of common but differentiated responsibilities.

6.1.3 RQ3: Water Law Typologies

The third research question was whether the national water laws clustered into recognizable typological categories. Using K-means clustering ($k = 5$, Table 5.5), the authors identified five different types of legislative frameworks:

1. *Groundwater-centric frameworks* (38 countries, Cluster 0): mainly Commonwealth countries in sub-Saharan Africa, the Caribbean and the Pacific and representing a reliance on groundwater resources, and the influence of British water law traditions.

2. *Comprehensive water codes* (9 countries, Cluster 1): post-Soviet and post-Yugoslav countries, with exhaustive, codified water laws addressing all three policy dimensions in detail, along with complete environmental codes from other legal traditions.

3. *River-basin-dominant frameworks* (9 countries, Cluster 2): countries that have adopted basin-based governance as the main organizing principle of their water policies, encompassing both Commonwealth (Australia, Zimbabwe, South Africa) and non-Commonwealth (Japan, Brazil, China) countries.

4. *Surface-water-only legislation* (3 countries, Cluster 3): countries whose main water laws deal with surface water management and water industry regulation, without dealing with groundwater management.

5. *Narrowly scoped statutes* (5 countries, Cluster 4): countries with a very limited number of extracted provisions, and thus represent legislative frameworks where water governance is spread across multiple statutes.

These typologies are empirically grounded, but only partially correspond to the traditional classification of legal families. The fact that river-basin-dominant governance appears in both Commonwealth and civil-law countries is particularly interesting, suggesting that the adoption of basin-based governance is driven more by hydrological and institutional factors than by legal tradition itself.

Moreover, the results of the clustering demonstrate the limits of the conventional legal-family distinctions for explaining water-governance approaches. The distinction between civil-law and common-law countries is useful for explaining private law and procedural aspects of justice, but does not help explain water-governance approaches. Common-law countries appear in many different clusters (i.e., Cluster 0 for groundwater-centric, Cluster 2 for river-basin-dominant, Cluster 4 for limited-scope statutes), indicating that the regulatory content of water laws is determined by a variety of factors (hydrological characteristics, development needs, the influence of international water governance norms) that do not respect legal-family boundaries. These findings support the increasing recognition in comparative law literature that functional convergence is possible among legal traditions in cases where countries address similar governance problems [38].

6.1.4 RQ4: Translation Quality and Analytical Reliability

RQ4 examined the effect of machine translation quality on analytical reliability. The multi-metric translation-fidelity framework demonstrates that translation quality produces a quantifiable-but-manageable source of uncertainty (see Table 5.1 and Table 5.2) for the vast majority of the corpus. With respect to the 28 languages in the adequate-quality tier (COMET ≥ 0.80), the similarity analyses are likely to remain unscathed by translation artifacts. However, with regard to the five languages in the lower-quality tier (COMET < 0.80), the translation-induced noise may reduce the discriminatory capability of the similarity analyses, leading to the compression of the differences among the countries.

This phenomenon poses a qualitatively different risk: translations that are linguistically satisfactory but semantically impoverished may lead to the artificial inflation of the similarity scores of affected countries relative to those of other countries, and obscure the salient features of their water legislation. Although this risk is concentrated in a relatively small number of languages (Georgian, Amharic, Lao), it is a systematic constraint that cannot be adequately mitigated through automation alone.

With the pipeline's methodological reliability assessed, we now turn from discussion of validity in measurement to discussion of the substantive patterns that emerge from analysis results. Specifically, the clustering results raise questions that go beyond the data processing steps themselves and into comparative legal scholarship — specifically,

whether the groupings produced by the computational pipeline reflect known legal and institutional traditions or whether they reveal new relationships that expert-driven studies have not previously identified.

6.2 Legal Traditions and Policy Clustering

The three clusters generated through RQ3, and the nature of their relationships to each other, merit some discussion through the lens of comparative legal studies and the study of policy diffusion. Three patterns are particularly significant.

Colonial heritage and legislative similarity. The fact that there is a dense cluster of Commonwealth African, Caribbean and Pacific countries within Cluster 0 (groundwater-centric frameworks) indicates that colonial legal traditions continue to influence the development of national water laws and regulations, even after decades of independence. All of the countries listed above inherited British-influenced water laws that emphasized the need to protect groundwater abstraction rights and to establish permits for the abstraction of groundwater. Although all of the countries listed have implemented legislative reforms since gaining independence, they have retained many of the structural features of their colonial-era water law framework. Furthermore, the fact that all of the countries listed have used the term "catchment council" to describe their local institutions further illustrates the extent to which British water governance vocabulary has been transmitted to the legal drafting processes of Commonwealth countries.

Post-Soviet legal convergence. The fact that post-Soviet and post-Yugoslav countries are clustered together in Cluster 1 (comprehensive water codes) illustrates the similarities in institutional legacies of water governance under socialism. Socialist systems of government placed strong emphasis on centralised regulation of all aspects of water resources management through codified legislation. Therefore, although countries such as North Macedonia, Armenia, the Kyrgyz Republic and Albania have had differing post-independence experiences, they all inherited comprehensive water codes during the socialist era and have continued to maintain or expand the scope of codified regulation of water resources management through subsequent reforms. As a result, the water legislation of all four countries exhibit similar structural characteristics despite their different historical experiences.

Basin governance as a cross-tradition phenomenon. The fact that both Commonwealth (Australia, Zimbabwe, Zambia, South Africa) and non-Commonwealth (Japan, Brazil, China) countries are grouped together in Cluster 2 (river-basin-dominant frameworks) illustrates that the adoption of basin-governance is not limited to a particular legal tradition. Both the Australian Murray-Darling Basin Authority, Zimbabwe's catchment councils, Brazil's river basin committees (modeled on the

French *agences de l'eau*) and Japan's River Law all apply basin-level governance using different institutional mechanisms, but the legislatively similar results produced in the embedding-based analysis illustrate the commonality of basin-governance through IWRM and the Dublin Principles. This finding supports previous research on policy diffusion that has demonstrated that IWRM and the Dublin Principles have facilitated the global dissemination of basin-governance [12].

6.3 The Translation Quality Challenge

Given that there was considerable variability in the quality of translation of the corpus (COMET scores ranged from 0.681 to 0.866) a frank assessment of how this impacts upon the findings will follow.

In the case of most of the languages (COMET $\geq$ 0.80), the quality of translation was sufficient to enable meaningful comparative analysis of the legislation. High cosine similarities between the original and translated versions of the legislation for languages such as French (0.825), Spanish (0.820) and Italian (0.806) indicate that the semantic content of the legislation was largely preserved during translation. Therefore, the findings obtained from these languages—which accounted for the majority of the corpus—can be considered to be reasonably reliable.

However, in the case of languages that exhibited semantic drift, the situation was far more complicated. Although the COMET score for the Georgian legislation was high (0.852), the cosine similarity between the original and translated versions of the legislation was extremely low (0.240). This indicates that while the translated version of the Georgian legislation was linguistically competent, it lost much of the domain-specific legal content contained in the original. The practical consequence of this is that the positioning of Georgia's legislation in the clustering analysis may be distorted. Georgia's legislation may appear to be more similar to other countries' legislation than it actually is due to the fact that the translation flattened its unique regulatory features into generic water management language. A similar concern also applies to varying degrees to Ethiopia (Amharic, cosine similarity 0.117), Laos (Lao, cosine similarity 0.208), and Uzbekistan (Uzbek, cosine similarity 0.313).

As stated earlier, however, the similarity analysis was conducted on LLM-extracted summaries of the legislation rather than on the raw translations. The LLM summary generation process involves a second layer of normalisation that may help to offset the loss of semantic content that occurs as a result of translation-induced semantic drift, as the LLM generates summaries based on the legal content that it identifies in the translated text. It remains an open empirical question whether this layer of normalisation is sufficient to restore the ability of the LLM to discriminate between the legal content contained in the translated text and the legal content contained in

the original text.

6.4 LLM Extraction: Strengths and Limitations

The use of GPT-4.1-mini for legal information extraction represents both a strength and a vulnerability of the methodology.

Strengths. Advantages of the use of GPT-4.1-mini for legal information extraction include 100% schema validation pass rates across 165 country-level documents, demonstrating that structured JSON output with rigorous validation can yield consistently formatted data from highly heterogeneous legal texts. The dual extraction method employed here—topic-specific extraction for targeted retrieval and comprehensive AquaLex Scrutinizer extraction for exhaustive coverage—yielded complementary views that revealed patterns not apparent when viewed through the lens of either method alone. The contrast between the lack of exact matches for the "polluter-pays principle" (which were found via exact keyword searches) and the number of pollution-related provisions identified in each country via semantic LLM extraction illustrate the benefits of LLM-based extraction relative to traditional keyword-based methods for comparative legal analysis.

Limitations. Disadvantages of relying on a single LLM model (GPT-4.1-mini) include the systemic dependency introduced by the single model used. It is difficult to evaluate the extent to which different LLM models may extract different provisions from the same text, assign different levels of confidence to those provisions, or interpret ambiguous legislative language in different ways. Similarly, the accuracy and completeness of the extracted information cannot be quantitatively evaluated without assessing the reliability of multiple raters—either against human annotators or against one or more additional LLM models. While the schema validation checks confirm that the output of the extraction process is structurally correct, the validation checks do not evaluate the accuracy of the substance of the output. An extraction output may pass all five of the validation checks and still contain inaccurate summaries, misidentifications of section references or incomplete keyword annotations.

The token management strategy employed in order to facilitate the practical application of the methodology presents a further limitation. Because the LLM model has a maximum context window size, documents that exceed the size of the context window are truncated. Truncation of the document may result in omissions of provisions in the latter parts of lengthy legislative documents. For the majority of the corpus, less than 5% of the documents were truncated. However, for the longest documents in the corpus (Bulgaria at 236 pages, Poland at 211 pages and Romania at 211 pages), truncation may have resulted in the exclusion of relevant provisions.

The limits of this analysis are very real and should reduce any overly optimistic

interpretation of the extracted data. However, the output of the pipeline is not created in isolation; instead it can be evaluated against findings from established comparative water law studies that have been conducted by experts. By doing so, such comparisons serve two goals: they provide an external check as to whether the computational results are reasonable at some level, and they clearly indicate what advantages or disadvantages the use of computational methods will have over more traditional methods.

6.5 Comparison with Expert-Driven Studies

Comparisons with expert-driven comparative water law studies can validate the findings of this thesis.

Caponera [4] described the variety of water right regimes based upon legal tradition—riparian rights in common law systems, administrative concessions in civil law systems, and community-based allocation in Islamic and customary law systems. The cluster results generally correspond with Caponera's taxonomy; specifically, the high number of Commonwealth countries in Cluster 0 and the high number of post-Soviet/civil law countries in Cluster 1 correspond to the structural differences between common law and civil law water governance that Caponera found using a manual review process.

Burchi [5] reported an overall trend towards basin-based water governance globally and noted Australia, South Africa, and Zimbabwe as examples of basin-based approaches. Similar to Burchi's study, this research shows that basin-based water governance is present in Brazil, China, and Japan—countries not studied in detail by Burchi. Therefore, the computational approach used here supports expert findings and adds geographic breadth to those findings not available before.

Although the computational approach agrees with expert analysis, there are also areas in which the computational approach appears to fall short compared to expert analysis. Burchi differentiated between countries that had basin-based water governance incorporated in their legislation and those that had implemented basin-based water governance in practice—a differentiation that the computational approach used here cannot make since the computational approach analyzes only the legislative text itself and not how the legislation is implemented. Additionally, the detailed analysis by Caponera of customary water law and the interactions between customary water law and formal water law is beyond the capability of the computational approach used here, since the computational approach analyzes only formal water legislation and does not analyze customary water law. Therefore, the computational approach provides breadth for expert legal scholars who provide depth.

Another area of comparison involves the level of analytical granularity. Expert comparative studies can distinguish among very fine-grained differences in legal language—

e.g., the difference between a "shall" (mandatory) and a "may" (permissive) provision in a statute; the significance of the position of a statute in a larger statutory hierarchy; the implications of a particular enforcement mechanism. Although the LLM-based extraction pipeline captures some of these differences through the pipeline's verbatim text extraction and summary generation functions, the embedding-based similarity analysis necessarily reduces these nuances to a single number. As a result, the computational approach is much better suited to detecting broad trends—i.e., which countries are addressing similar issues, which policy aspects receive the most focus, and which countries cluster together—whereas expert analysis is more capable of explaining how specific legal mechanisms are being employed to pursue governance goals.

Given this relationship between the two methods of analysis, the most effective way to utilize the computational approach is to use it as a filter to help determine which country-pair comparisons are most relevant for expert analysis. That is, rather than replace the detailed examination of a statute that is typical of traditional comparative law, the computational approach uses machine learning techniques to identify which of the tens of thousands of possible country-pair comparisons are most likely to be valuable to an expert, thereby allowing experts to focus their time and effort where it is most likely to generate the highest analytic payoff.

6.6 Methodological Limitations

In addition to providing evidence of the validity of the findings presented in this thesis, several limitations of the methodologies utilized should be noted to put the findings in context.

Selection bias. The corpus consists only of countries for which the researchers could obtain digital access to their water legislation. This limits the selection of countries to include only countries with developed digital infrastructures for publishing legislation, and therefore excludes countries that do not have such digital infrastructures. Similarly, the corpus limits countries where water governance occurs primarily through customary rather than formal law, and countries with limited or restricted water legislation. The corpus therefore over-represents countries with well-developed legal publishing infrastructures and under-represents small states, states in conflict, and states with primarily customary water governance.

Homogeneity in LLM Summarization. The similarities were measured on the summaries made by GPT-4.1-mini as opposed to raw legislative documents. As one LLM is generating all of the summaries, there may also exist some artificially similar stylistic and structural patterns across the output due to the model itself. Therefore, countries with legislatively differing water laws can have summaries that are worded similarly or structured in the same way simply because they were generated by the

same model. Therefore, the homogeneity caused by the summarizations may cause an increase in the similarity of each pair of documents, and thus reduce the overall apparent heterogeneity of the document set. Although using extracted text directly from the documents in addition to the summaries helps somewhat alleviate this issue, it is still important to consider this when analyzing the results of the clustering.

Temporal inconsistency. The corpus contains legislation from countries ranging from 1942 (Costa Rica) to 2023 (Vietnam), representing a span of more than 80 years. Over this time-span, the paradigmatic structures of water governance have undergone significant changes. Therefore, comparing legislation from different decades introduces a confounding variable. Differences in countries' legislation may represent differences in the substantive legal choices made by countries, but they may also reflect the era in which their legislation was written, due to the evolution of international water governance standards. A country with a 1950s-era water law will appear differently from a country with a 2020s-era water law not simply because of the differences in the regulatory philosophies represented in their respective water laws, but also because of the differing international standards governing water law in each decade.

English-centric analysis. All similarity analysis was performed on English-language versions of the texts, whether the original language version or a translation. This introduces a bias toward legislative concepts and terminology that are easily translatable into English. Therefore, legal concepts that are culturally bound, have no equivalent in English, or are embodied in the structural characteristics of the original language (for example, the distinction in Arabic draftsmanship between mandatory and permissive provisions) will be flattened or lost in translation, which will reduce the discriminatory ability of the analysis for non-English origin legislation.

Keyword set subjectivity. The similarity scores were calculated based on the researcher-defined keyword sets for each of the policy dimensions. Since different keyword selections would generate different similarity scores and therefore different clustering results, the keyword selections represent a subjective choice of the researcher and therefore introduce a degree of subjectivity in the keyword-based analysis.

No ground truth benchmark. There was no manually annotated ground-truth dataset created to evaluate the accuracy of the LLM extraction. Because of the lack of a benchmark, it is not possible to calculate the precision and recall of the extraction pipeline, nor is it possible to determine potential failure modes. This represents a limitation that is shared by large-scale computational legal analysis, where the cost of creating a manually annotated dataset across 165 countries and three policy dimensions makes it prohibitively expensive.

Single-model dependency. The LLM extraction pipeline was dependent on a single model, namely GPT-4.1-mini for both the complete extraction and the extraction of each of the topic-specific sub-domains. Therefore, the results are dependent on the

specific abilities, biases, and failure modes of a single model. Potential model-specific tendencies, such as a preference for certain formulations, a propensity to over- or under-extract in certain legal traditions, or systemic errors in translating text from languages other than English, could introduce unknown biases in the results.

OCR quality variation. The quality of the Optical Character Recognition (OCR) processing applied to scanned documents varied significantly, particularly for scanned documents that required OCR processing, especially for documents written in Arabic script from Egypt, Iran, Syria, and Tunisia. While Google Cloud Document AI provided high-quality OCR, the quality of the character recognition of degraded scans is inherently low, and therefore errors introduced by poor OCR processing propagate through all subsequent steps in the pipeline.

Single-statute limitation. The corpus included only one legislative instrument per country, typically the most comprehensive water legislation available. However, in reality, many countries govern water resources using multiple complementary statutes—separate statutes for protecting groundwater quality, ensuring environmental quality, supplying and sanitizing water, irrigating crops, and generating hydroelectric energy. Thus, analyzing only the primary water legislation may not accurately reflect the full scope of a country's water governance framework. Countries that distribute their water governance responsibilities across multiple statutes (as do Canada and New Zealand, which appear in Cluster 4 with "narrowly scoped" legislation) may be categorized as having limited water governance even though they have comprehensive water governance frameworks distributed across multiple statutes.

6.7 Implications for Water Governance

Despite the limitations outlined above, the findings of this thesis have practical implications for water governance at various scales.

For policymakers, the clustering analysis provides a structured basis for identifying peer countries that address similar governance challenges in their water legislation. A policymaker interested in reforming groundwater regulation in his/her country can find countries in the same cluster that have more comprehensive groundwater regulations and therefore can use those countries' legislation as benchmarks for reform. The findings of this thesis that basin-based governance transcends legal families indicate that countries from any legal family can implement basin-based governance if the institutions in place support that governance.

For international organisations such as the FAO, UNDP, and the Global Water Partnership, the corpus and methodology can serve as a tool to monitor the global implementation of water governance principles. The finding that the polluter-pays principle is implemented in practice but not explicitly stated in water legislation, for

instance, has implications for how international organisations assess compliance with environmental governance obligations: reliance on keyword searches for "polluter-pays principle" in national legislation would systematically undercount its adoption.

For legal reform processes, the five distinct legislative archetypes identified in this thesis can provide a conceptual framework for understanding where a country's water legislation falls within the global landscape and what other forms of water governance a country might consider to achieve more balanced water management. Countries with narrowly scoped water laws (Cluster 4) may want to consider the comprehensive water code model (Cluster 1) as a form of reform. Conversely, countries with water laws focused on groundwater (Cluster 0) may want to incorporate elements of basin-based governance to achieve more balanced water management.

For computational legal analysis, this thesis demonstrates that the integration of document processing, machine translation, LLM-based extraction, embedding-based similarity analysis, and clustering can lead to meaningful comparative insights at a scale that was previously unachievable. The framework is likely to be applicable to other comparative law domains—environmental regulation, land tenure, mining codes, labour law—where similar challenges related to linguistic diversity, scale, and heterogeneity exist.

It is essential to stress that the five legal typologies discussed in this paper indicate how countries have codified their law; they do not represent how countries enforce these laws. In recent years, there has been an increasing recognition of a "legislative-implementation" gap (the gap between what has been legislated and what is enforced), especially among low-income countries [5]. The reasons for this include lack of institutional capacity, funding constraints and/or the political will to enforce existing legislation. Therefore, computational comparative law provides tools for studying legislative differences rather than measuring effectiveness of governance. To determine if the legislative differences observed through the cluster analysis are ultimately effective forms of water management requires very different types of data, including but not limited to, measures of governance performance, enforcement/monitoring/compliance information from regulatory agencies, and empirical studies conducted in the field. These additional areas of research represent separate agendas from the focus of this paper.

The findings, limitations, and implications presented in this chapter outline both the utility and limitations of the computational approach to comparative water law. The final chapter will summarize the major contributions of the thesis, discuss the limitations of the interpretations of those contributions, and highlight the most promising avenues for future research.

Chapter 7

Conclusion

7.1 Summary of Contributions

This conclusion brings together the thesis's four major contributions; it discusses the limitations that affect these contributions and outlines possible future research options.

First, the thesis has demonstrated that a complete computational pipeline can be applied to automatically analyze and compare water legislation from 164 countries in more than 35 languages and 10 script systems. The seven-stage pipeline—from building the corpus through text extraction, machine translation, translation quality evaluation, LLM-based legal information extraction, embedding-based similarity analysis, to hierarchical clustering—creates structured, comparable representations of the content of legislative texts allowing for quantitative comparative analysis on a scale that would be impossible to accomplish manually. The pipeline achieved nearly complete coverage, with the LLM-based extraction successfully processing all 165 country-level documents and achieving policy dimension coverage of between 92.3% (for river basin management and polluter-pays principle) and 95.4% (for groundwater regulation).

Second, the multi-metric translation fidelity framework has provided a principled approach to evaluating the reliability of machine-translated legal texts. By integrating COMET reference-free quality estimation (with a corpus mean of 0.83, exceeding the conventional minimum adequacy threshold of 0.80) with embedding-based semantic fidelity analysis and a 30-document manual verification protocol across nine script types, the framework established quantitative confidence levels for the translated corpus. The finding of semantic drift—where translations may achieve adequate COMET scores but exhibit low semantic similarity to the original (e.g., the Georgian cosine similarity is 0.240, the Amharic cosine similarity is 0.117)—constitutes a methodological finding with implications beyond this thesis, indicating a systematic risk when applying machine translation for domain-specific computational analysis.

Third, the AquaLex Scrutinizer pipeline has demonstrated that GPT-4.1-mini,

when guided by domain-specific prompts and constrained by five-point JSON schema validation, can extract structured legal provisions from water legislation across various legal traditions with 100% schema validation pass rates. The dual extraction approach—the lightweight topic-specific extraction for focused retrieval and the comprehensive multi-topic extraction for total coverage—provided complementary views of the content of legislative texts.

Fourth, the clustering analysis has identified five distinct water law typologies through K-means clustering and has revealed the hierarchical structure of legislative similarity through Ward's method dendrograms. The typologies were groundwater-centric frameworks (38 countries), comprehensive water codes (9 countries), river-basin-dominant frameworks (9 countries), surface-water-only legislation (3 countries), and narrowly scoped statutes (5 countries). These typologies constitute an empirically grounded classification of global water governance approaches. The substantive findings—including the universal occurrence of "aquifer" as the keyword for groundwater, the impact of Commonwealth legal traditions on river basin terminology ("catchment council" versus "river basin management"), and the absence of verbatim "polluter-pays principle" references despite widespread implicit implementation—represent new insights into the patterns of global water legislation.

7.2 Limitations

The findings of this thesis must be considered in light of several methodological limitations that were extensively discussed in Chapter 6. The primary limitations of this study were: the reliance on a single LLM model (GPT-4.1-mini) without conducting an inter-rater reliability assessment or comparing the model to others; the fact that all the analyses of similarity were conducted in English and therefore computed semantic proximity on translated texts rather than originals; variation in the quality of translation across languages, particularly translation-induced meaning loss affecting Georgian, Amharic, and Lao legislation; temporal inconsistency of the corpus (legislation dated from 1942 to 2023); selection bias due to the availability of digital legislative texts; and subjectivity of the keyword sets used to calculate similarity scores.

While these limitations do not negate the findings of the study, they constrain the interpretations of those findings. Therefore, the framework should be viewed as a means to conduct large-scale exploratory analysis that identifies patterns worthy of further study through expert-driven, language-specific legal scholarship rather than as a final comparative legal assessment.

7.3 Future Work

Several potential directions for future research arise from this thesis, organised here from methodological validation to substantive extension.

First, **model comparison studies** should evaluate the performance of alternative LLMs—including GPT-4o [42], Gemini [43], and open-source models such as Llama and Mistral—on the legal extraction task, in order to assess the robustness of the extraction and reduce dependence on a single model. Studies of this type would require running multiple models on the same corpus and determining inter-model agreement metrics, similar to inter-rater reliability in manual coding. Significant differences among models would indicate areas where the extraction results depend on the model and thus should be treated cautiously.

Second, **precision and recall evaluation** on a manually annotated subset of legislative provisions would provide benchmark values for the extraction pipeline. An applicable approach would be to annotate provisions from a stratified sample of 20–30 countries across different legal traditions and languages and then to evaluate the precision, recall, and F1 scores for each policy dimension of the extracted provisions compared to the annotations.

Third, **robustness analysis** with alternative keyword sets would determine how sensitive the similarity scores and clustering results are to the particular set of policy-relevant terms selected, providing confidence intervals for the typological findings. This could be achieved through a systematic study of keyword perturbations in which keywords are added, deleted or replaced with synonyms, and the resulting changes in cluster assignments are documented.

Fourth, **geographic visualisation** through interactive display of spatial relationships of data through map-based interfaces would increase the accessibility and interpretability of the results, enabling researchers and policymakers to identify spatial patterns in water governance. Interactive choropleth maps of the cluster assignments, policy dimension scores, and translation quality metrics would enable non-technical stakeholders to immediately understand the results and support the identification of regional patterns that are not visible in table or dendrogram formats.

Fifth, **temporal analysis** of multiple versions of water legislation (if available) would allow tracing the evolution of water law over time, and thereby, identify trends in regulatory reforms and policy convergences. Countries that have experienced significant water law reforms—such as South Africa's transition from the Water Act of 1956 to the National Water Act of 1998—would be valuable case studies for exploring how legislative priorities evolve in response to shifts in governance paradigms.

Sixth, the framework could be **extended to other legal domains**—environmental law, land tenure, mining regulation, labour law—and thereby, demonstrate the applica-

bility of the framework as a tool for computational comparative law beyond the water domain. At the technical level, the architecture of the pipeline is domain-independent; at the domain-specific level, the pipeline architecture relies on the keyword sets and the LLM extraction prompts that can be easily adapted to new legal domains.

Seventh, **addressing the contextual flattening phenomenon** through improved translation strategies represents an important methodological priority. Possible approaches include training domain-specific neural machine translation models on legal corpora, employing multiple translation systems with ensemble aggregation, and developing post-translation verification processes that mark cases where domain-specific terminology has been replaced with generic terminology.

7.4 Closing Remarks

The global water crisis requires responses that are informed by an inclusive understanding of how different countries govern their water resources. For decades, this inclusive understanding has been limited by the practical difficulty of comparing water legislation across the full range of the world's legal systems, languages and traditions. The thesis demonstrates that methods of computation—integrating document processing, machine translation, large language model-based extraction of legal information, embedding-based similarity analysis, and hierarchical clustering—can bridge this gap, converting a multilingual, multi-script corpus of 164 national water laws into a structured, comparable, and analytically tractable dataset.

Computational methods do not replace the nuanced contextual understanding that expert legal scholars bring to comparative analysis. However, computational methods complement the nuanced contextual understanding of expert legal scholars by providing a broad panorama of global water law that no single legal scholar could produce, revealing patterns and inconsistencies that deserve further study from experts, and creating a scalable methodology that can be updated as legislation evolves and as computational tools improve. Examples of this are the discovery of five water law typologies, the observation that the polluter-pays principle is widely implemented implicitly rather than explicitly invoked, and the characterisation of semantic drift as a systematic risk in cross-lingual legal analysis—examples illustrating the unique contribution of computational methods, which can identify global patterns across scales that manual scholarship—regardless of its rigor—cannot practically accomplish.

Therefore, the thesis contributes not only to the study of water governance but also to the developing field of computational comparative law, demonstrating that the integration of NLP, LLMs, and machine learning enables new avenues for the systematic study of how societies regulate their most basic common resource. As the capabilities of the technologies underlying the framework continue to advance—as the domain

awareness of translation models improves, as the reliability of LLMs as extractors of structured information increases, and as embedding models increasingly accurately capture fine-grained semantic distinctions—the difference between computational and expert analysis will decrease and the potential for computational comparative law to support evidence-based legal reforms will increase.

Bibliography

[1] United Nations, "The united nations world water development report 2023: Partnerships and cooperation for water," UNESCO, Paris, Tech. Rep., 2023.

[2] M. A. Caretta, A. Mukherji, M. Arfanuzzaman, R. A. Betts, A. Gelfan, Y. Hirabayashi, T. K. Lissner, J. Liu, E. Lopez Gunn, R. Morgan, S. Mwanga *et al.*, "Water," in *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2022, pp. 551–712.

[3] OECD, "OECD environmental outlook to 2050: The consequences of inaction," OECD Publishing, Paris, Tech. Rep., 2012.

[4] D. A. Caponera, *Principles of Water Law and Administration: National and International*, 2nd ed. London: Taylor & Francis, 2007, revised and updated by Marcella Nanni.

[5] S. Burchi, "A comparative review of contemporary water resources legislation: Trends, developments and an outlook for the future," *Water International*, vol. 37, no. 6, pp. 613–627, 2012.

[6] S. Burchi and K. Mechlem, "Groundwater in international law: Compilation of treaties and other legal instruments," Food and Agriculture Organization of the United Nations, Rome, FAO Legislative Study 86, 2005.

[7] S. Burchi, "National water law and policy trends and common elements based on fao legislative studies," Food and Agriculture Organization of the United Nations, Rome, Water Reports, 2003.

[8] ——, "Current developments and trends in water resources legislation and administration," in *The Evolution of the Law and Politics of Water*, J. W. Dellapenna and J. Gupta, Eds. Dordrecht: Springer, 2009, pp. 55–72.

[9] Food and Agriculture Organization of the United Nations, "FAOLEX database," https://www.fao.org/faolex/en/, 2023, accessed: 2025-09-15.

[10] International Conference on Water and the Environment, "The dublin statement on water and sustainable development," World Meteorological Organization, Dublin, Tech. Rep., 1992.

[11] Global Water Partnership, *Integrated Water Resources Management*, ser. TAC Background Papers. Stockholm: Global Water Partnership, 2000, no. 4.

[12] M. M. Rahaman and O. Varis, "Integrated water resources management: Evolution, prospects and future challenges," *Sustainability: Science, Practice and Policy*, vol. 1, no. 1, pp. 15–21, 2005.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020, pp. 2898–2904.

[14] R. Dale, "Law and word order: NLP in legal tech," *Natural Language Engineering*, vol. 25, no. 1, pp. 211–217, 2019.

[15] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset," in *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. ACM, 2021, pp. 159–168.

[16] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An expert-annotated NLP dataset for legal contract review," in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 2021.

[17] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, pp. 4310–4330.

[18] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 passes the bar exam," *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2270, p. 20230254, 2024.

[19] D. Trautmann, A. Petrova, and F. Schiber, "Legal prompt engineering for multilingual legal judgement prediction," in *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, 2022, pp. 39–48.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[21] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics, 2006, pp. 249–256.

[22] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 2685–2702.

[23] R. Rei, J. G. C. d. S. Teixeira, D. Fernandes, N. M. Guerreiro, M. Fonseca, G. Neubig, C. Stewart, A. C. Farinha, and A. Lavie, "COMET-22: Unbabel-IST 2022 submission for the metrics shared task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, 2022, pp. 578–585.

[24] R. Rei, A. C. Farinha, J. G. C. de Souza, P. G. Ramos, A. F. T. Martins, L. Coheur, and A. Lavie, "CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, 2022, pp. 634–645.

[25] D. Cao, *Translating Law*. Clevedon: Multilingual Matters, 2007.

[26] I. Marchetti, "Machine translation of legal texts: Challenges and opportunities," *Comparative Legilinguistics*, vol. 53, pp. 83–110, 2023.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning Representations (ICLR 2013) Workshop Track*, 2013.

[28] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1532–1543.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186, 2019.

[30] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 3982–3992.

[31] OpenAI, "New embedding models and API updates," https://openai.com/index/new-embedding-models-and-api-updates/, 2024, accessed: 2025-10-01.

[32] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021, first published 2019.

[33] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[34] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, 1967, pp. 281–297.

[35] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[36] W. Arts and J. Gelissen, "Three worlds of welfare capitalism or more? A state-of-the-art report," *Journal of European Social Policy*, vol. 12, no. 2, pp. 137–158, 2002.

[37] P. A. Hall and D. Soskice, *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage*. Oxford: Oxford University Press, 2001.

[38] K. Holzinger, C. Knill, and T. Sommerer, "Environmental policy convergence: The impact of international harmonization, transnational communication, and regulatory competition," *International Organization*, vol. 62, no. 4, pp. 553–587, 2008.

[39] M. A. Livermore and D. N. Rockmore, "Distant reading the law," *Virginia Law Review*, vol. 104, p. 1471, 2017.

[40] Google Cloud, "Cloud translation API overview," https://cloud.google.com/translate/docs/overview, 2025, accessed: 2025-09-15.

[41] ——, "Document AI overview," https://cloud.google.com/document-ai/docs/overview, 2025, accessed: 2025-09-15.

[42] OpenAI, "Introducing GPT-4.1 in the API," https://openai.com/index/gpt-4-1/, 2025, released: 2025-04-14. Accessed: 2025-10-01.

[43] Gemini Team, R. Anil, S. Borgeaud *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2024.

Appendix A

LLM Extraction System Prompt

The following system prompt was used to configure GPT-4.1-mini for the legal information extraction pipeline described in Section 4.6. The prompt defines the extraction agent's role, keyword groups, reasoning strategy, output schema, and alternative terminology guidelines.

```
# System Prompt for Water Law Text Extraction

## Role and Purpose
You are WaterLawExpert, a specialized legal analysis agent
designed to extract and organize water management laws from
legal documents. You will analyze text from water law documents
and identify all sections relevant to specific keyword groups.

## Response Rules
- Persist until the extraction task is fully complete.
  Do not terminate your analysis early.
- Follow a methodical approach to identify ALL relevant
  sections, even those that only partially match keywords.
- Be comprehensive in your extraction - include implied
  references where the concept is addressed, even if exact
  keywords aren't used.
- Format your output precisely as a JSON object following
  the specified structure.
- Do not guess or fabricate content - only extract text that
  actually appears in the document.

## Extraction Keyword Groups

### GROUP 1 - GROUNDWATER
- groundwater ownership
```

```
- groundwater permitting
- groundwater monitoring
- groundwater abstraction
- transboundary aquifers

### GROUP 2 - RIVER BASIN
- river basin management
- river basin council
- river basin plans

### GROUP 3 - POLLUTER-PAYS
- polluter-pays principle

## Reasoning Strategy
1. Document Analysis: First, scan the entire document to gain
   a high-level understanding of its structure, chapter
   organization, and key terminology used.
2. Keyword Identification: Identify both direct keyword matches
   AND conceptual matches where the legal provision addresses
   the concept but uses different terminology.
3. Context Evaluation: For each potential match, evaluate the
   surrounding text to ensure you capture the complete legal
   provision, including any relevant qualifiers or conditions.
4. Section Extraction: Extract the full text of relevant
   sections along with their identifiers (Article numbers,
   Chapter references, etc.)

## Output Format
Your response must be a clean, properly formatted JSON object
with this exact structure:
{
  "groundwater": [
    {
      "section_id": "Article X",
      "text": "Full text of the relevant section...",
      "summary": "Brief summary of what this section covers",
      "keywords": ["groundwater ownership",
                   "groundwater permitting"]
    }
  ],
  "river_basin": [
```

```
    {
      "section_id": "Chapter Y, Section Z",
      "text": "Full text of the relevant section...",
      "summary": "Brief summary of what this section covers",
      "keywords": ["river basin management",
                   "river basin plans"]
    }
  ],
  "polluter_pays": [
    {
      "section_id": "Article W",
      "text": "Full text of the relevant section...",
      "summary": "Brief summary of what this section covers",
      "keywords": ["polluter-pays principle"]
    }
  ]
}

## Important: Alternative Terminology

When analyzing water law documents, be aware that legal texts
often use specialized or region-specific terminology. Consider
these alternative terms when searching for relevant sections:

### For GROUNDWATER:
- Subsurface water, underground water, aquifer, well water,
  subterranean water
- Sections addressing wells, boreholes, water tables, or
  aquifer recharge
- Regulations about underground water extraction, drilling
  permits
- Any references to subsurface water quality or quantity

### For RIVER BASIN:
- Watershed management, catchment area, drainage basin,
  water district
- Water resource planning on a regional or geographical basis
- Integrated water management across connected water bodies
- River authorities, basin commissions, or watershed councils

### For POLLUTER-PAYS:
```

```
- Environmental liability provisions
- Discharge permits with fee structures
- Pollution fines and penalties
- Cost recovery for environmental damage
- Requirements for polluters to fund cleanup or monitoring
```

Permitted Keyword Lists for Comprehensive Extraction

The following keyword lists were used as the permitted keyword sets for the comprehensive "AquaLex Scrutinizer" extraction pipeline (Section 4.6). Each extracted provision was annotated with one or more keywords from the relevant group.

Groundwater (11 keywords):

- groundwater ownership
- groundwater permitting
- groundwater monitoring
- aquifer management
- well drilling regulations
- transboundary aquifers
- groundwater recharge
- groundwater rights
- groundwater governance
- groundwater abstraction
- groundwater extraction

River Basin (9 keywords):

- river basin management
- river basin council
- watershed management
- integrated water resources management (IWRM) at basin scale
- transboundary river agreements
- river basin plan

- basin authority
- catchment management
- river basin organization

Polluter-Pays (10 keywords):

- polluter-pays principle
- pollution charges
- discharge fees
- environmental liability for pollution
- remediation costs recovery
- water quality standards
- pollution penalties
- pollution prevention
- cost recovery from polluter
- environmental liability

Appendix B

Country-Cluster Assignments

Table B.1 lists all 65 countries in the keyword and clustering analysis subset, their assigned cluster (K-means, $k = 5$), and the number of extracted provisions per policy dimension. North Korea (KP) was subsequently identified and removed as an anomaly (see Section 5.8).

Table B.1: Country-cluster assignments and provision counts for the 65-country subset.

Country	**Cluster**	**Groundwater**	**River Basin**	**Pollution**
Cluster 0: Groundwater-Centric Frameworks (39 countries)				
Bahamas	0	11	1	1
Bangladesh	0	33	17	4
Belize	0	63	2	30
Bhutan	0	41	30	23
Botswana	0	60	1	4
Brunei	0	30	7	2
Cambodia	0	21	4	1
Dominica	0	21	9	15
Eritrea	0	30	5	15
Eswatini	0	55	33	10
Fiji	0	5	1	3
Ghana	0	14	1	2
Guyana	0	35	13	7
Ireland	0	4	4	5
Jamaica	0	69	4	15
Kenya	0	114	32	13
Lesotho	0	40	13	10
Malawi	0	26	17	4
Malaysia	0	6	8	1

Continued on next page

Table B.1: (continued)

Country	**Cluster**	**Groundwater**	**River Basin**	**Pollution**
Malta	0	12	1	0
Mauritius	0	77	8	1
Namibia	0	88	51	14
Nigeria	0	12	2	0
North Korea*	0	14	11	2
Norway	0	71	51	6
Pakistan	0	51	35	14
Papua New Guinea	0	87	8	3
Rwanda	0	18	13	7
Saint Kitts and Nevis	0	19	4	1
Saint Lucia	0	18	7	3
Saint Vincent	0	41	21	7
Samoa	0	14	2	1
Sierra Leone	0	27	23	4
Tanzania	0	73	22	5
Thailand	0	37	41	6
Tonga	0	14	4	3
Uganda	0	66	22	6
United Kingdom	0	99	11	12
Vanuatu	0	23	16	2
Cluster 1: Comprehensive Water Codes (9 countries)				
Albania	1	97	95	47
Armenia	1	148	29	37
India	1	4	3	22
Kyrgyzstan	1	119	44	41
Mongolia	1	78	36	36
North Macedonia	1	205	83	96
Philippines	1	11	1	84
South Sudan	1	63	39	34
Sweden	1	87	11	109
Cluster 2: River-Basin-Dominant Frameworks (9 countries)				
Australia	2	14	165	21
Brazil	2	16	36	2
China	2	39	60	5
Gambia	2	5	6	0
Grenada	2	12	30	2

Continued on next page

Table B.1: (continued)

Country	**Cluster**	**Groundwater**	**River Basin**	**Pollution**
Japan	2	21	139	0
South Africa	2	102	106	27
Zambia	2	107	133	12
Zimbabwe	2	72	185	4
Cluster 3: Surface-Water-Only Legislation (3 countries)				
Myanmar	3	0	10	3
Sri Lanka	3	0	2	0
United Kingdom (Water Industry Act)	3	0	25	17
Cluster 4: Narrowly Scoped Statutes (5 countries)				
Barbados	4	10	0	2
Canada	4	1	0	29
Nepal	4	6	0	3
New Zealand	4	2	0	31
Singapore	4	1	0	1

*North Korea was subsequently removed from the analysis as an anomaly (see Section 5.8).

Appendix C

Example JSON Extraction Output

The following is a representative excerpt from the JSON extraction output for Kenya's Water Act, illustrating the structure of the LLM-extracted provisions. Each country's output contains arrays of findings for the three policy dimensions (groundwater, river basin, polluter-pays), with original text excerpts, English summaries, keywords, policy themes, and confidence scores.


```
{
  "country_name": "KE_Water Act_eng",
  "source_file": "1_extracted_text/english/KE_Water Act_eng.txt",
  "groundwater_findings": [
    {
      "groundwater_related": true,
      "key_excerpts_original": [
        "\"aquifer\" means an underground geological formation
         able to store and yield water;",
        "\"ground water\" means the water of underground
         streams, channels, artesian basins, reservoirs,
         lakes and other bodies of water in the ground, and
         includes water in interstices below the water
         table;",
        "23 Conservation of ground water.",
        "FOURTH SCHEDULE - ABSTRACTION OF GROUND WATER"
      ],
      "summary_english": "The Act defines 'aquifer' and
        'ground water' explicitly, recognizing groundwater
        as water from underground sources including streams,
        artesian basins, and water below the water table.
        It includes provisions for the conservation of
        groundwater and establishes a specific schedule
        dealing with abstraction of groundwater.",
```

```
      "keywords_english": [
        "groundwater", "aquifer", "underground water",
        "water table", "abstraction", "conservation",
        "artesian basins"
      ],
      "policy_themes": [
        "Permitting & Licensing",
        "Sustainability & Limits"
      ],
      "confidence_score_percent": 98
    }
  ]
}
```