Language, Knowledge and Intelligent Systems

https://lkisjournal.org

# Comparing Knowledge Graph Completion and Question Answering Accuracy in Legal Information Portals

**Dong-Hyun Heo[1,*], Jennifer Chiu[2], Tony Ho[2]**

[1] Department of Computer Science and Engineering, College of Engineering, Seoul National University, Seoul, Republic of Korea

[2] Department of Computer Science, College of Computing, City University of Hong Kong, Hong Kong, Hong Kong SAR, China

Corresponding author: dong.h.heo@snu.ac.kr

### Abstract

The rapid digitization of legal documents has necessitated the development of advanced information retrieval systems, specifically legal information portals, to facilitate efficient access to statutes, case law, and legal precedents. Knowledge graphs have emerged as a foundational technology within these portals, enabling semantic search and complex reasoning capabilities. However, legal knowledge graphs inherently suffer from incompleteness and noise, which directly impacts the accuracy of downstream applications such as question answering. This paper presents a comprehensive benchmark study evaluating the performance of various knowledge graph completion algorithms and their subsequent effect on question answering accuracy within the context of legal information portals. We construct a specialized legal dataset encompassing thousands of entities and relations derived from real-world judicial records and legislative texts. By systematically evaluating translation-based, tensor factorization, and neural network-based completion models, we establish a baseline for link prediction in the legal domain. Furthermore, we investigate how improvements in knowledge graph completeness correlate with the precision and recall of natural language question answering systems deployed over these graphs. The findings offer critical insights into the bottleneck of current legal question answering architectures and provide a standardized evaluation framework for future research in legal informatics.

Keywords: Knowledge Graphs; Legal Informatics; Question Answering; Link Prediction; Benchmark Study

## 1. Introduction

### 1.1 Background and Motivation

The transformation of the legal sector through digital technologies has led to an exponential increase in the volume of accessible legal text. Legal professionals, researchers, and the general public increasingly rely on digital portals to navigate this vast corpus of information. Traditional keyword-based search mechanisms, while historically useful, often fail to capture the nuanced semantic relationships inherent in legal discourse. Consequently, knowledge graphs have been widely adopted to structure legal information, transforming unstructured text into structured, machine-readable formats. A knowledge graph represents information as a network of entities connected by specific relations, which is particularly suitable for the legal domain where statutes, judicial decisions, jurisdictions, and legal concepts are intricately interconnected. Despite their utility, the automated construction of these graphs inevitably results in missing links and incomplete entity representations. Missing relations can severely degrade the user experience in legal information portals, where the omission of a critical legal precedent or a jurisdictional mapping could lead to erroneous legal interpretations. Therefore, addressing the incompleteness of legal knowledge graphs is not merely an academic exercise but a practical necessity for ensuring the reliability of digital legal services. The challenge is further compounded by the complexity of legal language, which is characterized by specialized terminology, highly structured arguments, and temporal dependencies [1].

### 1.2 Research Objectives and Scope

To mitigate the challenges associated with incomplete legal data structures, researchers have developed numerous knowledge graph completion techniques aimed at inferring missing facts based on existing network topology. However, the evaluation of these techniques is predominantly conducted on general domain datasets, which fail to reflect the structural and semantic peculiarities of legal data. The primary objective of this study is to establish a rigorous benchmark for evaluating knowledge graph completion methods specifically tailored to legal information portals. Furthermore, we aim to quantify the direct impact of these completion methods on the accuracy of downstream question answering systems [2]. Question answering systems represent the primary user interface in modern legal portals, allowing users to pose complex natural language queries and receive precise, fact-based responses. By bridging the gap between graph completion and question answering, this study seeks to demonstrate how underlying structural improvements manifest as tangible benefits in end-user applications. We hypothesize that certain classes of completion algorithms, particularly those capable of modeling asymmetric and hierarchical relations, will yield disproportionate improvements in question answering accuracy. The scope of this paper encompasses the creation of a novel legal benchmark dataset, the empirical evaluation of prominent completion algorithms, and an in-depth analysis of query resolution accuracy in a simulated legal portal environment [3].

## 2. Literature Review

### 2.1 Knowledge Graph Completion Methodologies

Knowledge graph completion, often synonymous with link prediction, has been the subject of extensive research over the past decade. The fundamental goal is to predict missing relations between entities based on the observed triples within the graph. Early approaches primarily relied on translation-based models, which map entities and relations into a continuous low-dimensional vector space. These models operate on the principle that the vector representation of a head entity, when added to the vector representation of a relation, should approximate the vector representation of the tail entity. Such methods are computationally efficient and have shown significant promise in general domains [4]. Subsequent advancements introduced tensor factorization techniques, which capture latent semantics by modeling the interactions between entities and relations through multiplicative operations. These approaches have proven highly effective in capturing symmetric and antisymmetric relations, which are prevalent in legal contexts [5]. More recently, graph neural networks have revolutionized the field by enabling the aggregation of neighborhood information, thereby enriching entity embeddings with localized structural context. Graph convolutional networks specifically designed for relational data allow for the integration of multi-hop structural patterns, which is critical for complex reasoning tasks [6]. Despite these advancements, the application of these methodologies to the legal domain remains largely underexplored, necessitating domain-specific benchmarking to determine their true efficacy.

### 2.2 Legal Informatics and Semantic Representation

The intersection of computer science and law has birthed the specialized field of legal informatics, which focuses on the representation, processing, and retrieval of legal information. Historically, legal ontology engineering required massive manual effort by domain experts to define the concepts and rules governing legal systems. With the advent of machine learning, there has been a paradigm shift towards automated knowledge extraction from unstructured legal corpora. Researchers have utilized natural language processing pipelines to identify legal entities such as judges, courts, plaintiffs, defendants, and specific statutory clauses [7]. The relations between these entities are highly specific, including concepts like supersedes, has jurisdiction over, and applies precedent from. The structural integrity of these legal knowledge graphs is paramount because legal reasoning is fundamentally logical and sequential. Errors or omissions in the graph can cascade through reasoning processes, leading to invalid conclusions. Existing literature highlights the difficulty of relation extraction in legal texts due to the prevalence of long-form, multi-clause sentences and implicit cross-references. Consequently, legal knowledge graphs are often sparser than their general-domain counterparts, making the task of graph completion both more challenging and more critical [8].

### 2.3 Question Answering Systems over Knowledge Graphs

Question answering over knowledge graphs involves translating natural language queries into formal logical queries that can be executed against the graph structure. This process typically requires entity linking, relation extraction, and semantic parsing to accurately map the user intent to the underlying data schema. In general domains, template-based and neural machine translation approaches have achieved remarkable success in resolving complex, multi-hop queries. However, legal question answering introduces unique challenges. Legal queries often involve conditional logic, temporal constraints, and comparative analysis, requiring a deep understanding of the legal context. Furthermore, users of legal information portals expect high precision and explainability, as the answers may influence legal strategies or compliance decisions. The literature indicates a significant gap in benchmarking the performance of question answering systems when operating over knowledge graphs completed by various algorithmic approaches. Most studies evaluate question answering systems assuming a static, fully populated graph, which is an unrealistic assumption in real-world legal portals. By integrating the evaluation of completion algorithms with question answering accuracy, this study aims to provide a more holistic understanding of system performance in operational environments.

## 3. Theoretical Framework of Legal Knowledge Graphs

### 3.1 Structural Representation of Legal Knowledge

A legal knowledge graph can be formally defined as a directed multigraph composed of a set of entities and a set of predefined legal relations. Each fact within the graph is represented as a semantic triple consisting of a head entity, a relation, and a tail entity. Entities in the legal domain are highly heterogeneous, spanning abstract concepts such as liability and negligence, to concrete entities like specific court cases, legislative acts, and geographical jurisdictions. The relations defining the interactions between these entities are equally diverse and exhibit distinct mathematical properties. For instance, the relation is_equivalent_to is symmetric, whereas the relation appeals_to is strictly asymmetric [9]. Furthermore, legal hierarchies introduce transitive relations, such as is_a_subsection_of, which are vital for understanding legislative structures. The theoretical framework guiding this study posits that the capacity of a completion algorithm to accurately model these specific relational properties determines its suitability for the legal domain. Traditional translation models struggle with modeling multiple relation types simultaneously, particularly when dealing with one-to-many or many-to-many relationships that are ubiquitous in legal hierarchies.

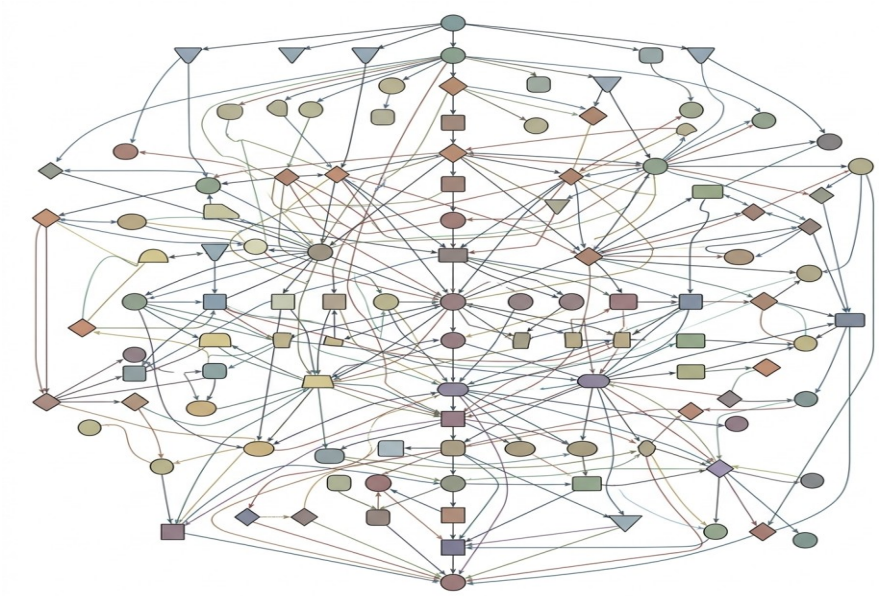

*Figure 1: Comprehensive Semantic Topology of Legal Knowledge Graph Architecture*

### 3.2 Semantic Inference and Multi-hop Reasoning

The utility of a legal knowledge graph is inherently tied to its ability to support semantic inference, which is the process of deriving new, unstated facts from the existing explicitly stated facts. In the context of a legal portal, a user might query whether a specific local ordinance complies with a federal mandate. Answering

this query may require traversing multiple relational hops: from the ordinance to its governing municipality, from the municipality to the state constitution, and finally to the federal statute. This multi-hop reasoning process is heavily reliant on the completeness of the path. If any single edge in this chain is missing, the reasoning process fails, and the question answering system is unable to provide a correct response. Therefore, the objective of knowledge graph completion is not merely to increase the absolute number of triples, but to selectively infer high-value edges that bridge disconnected subgraphs and facilitate continuous reasoning paths. Our theoretical model suggests that algorithms incorporating structural neighborhood aggregation are theoretically better equipped to identify and infer these critical bridging edges, thereby disproportionately enhancing the performance of complex, multi-hop question answering systems [10].

# 4. Benchmark Methodology and Dataset Construction

## 4.1 Data Acquisition and Corpus Selection

To establish a robust benchmark, we constructed a specialized legal dataset, hereafter referred to as the Legal Graph Benchmark. The foundational data for this benchmark was systematically collected from public access legal information portals, encompassing appellate court decisions, supreme court rulings, and corresponding legislative codes from multiple jurisdictions. The raw text corpora amounted to several gigabytes of unstructured legal documentation. The selection criteria focused on ensuring a balanced representation of civil, criminal, and administrative law to guarantee the generalizability of our findings. The raw texts were subjected to a rigorous data cleaning pipeline to remove formatting artifacts, standardize citation formats, and resolve coreference ambiguities. This preprocessing phase was crucial for minimizing noise before the application of automated information extraction tools. The resulting clean corpus served as the bedrock for the subsequent entity and relation extraction phases.

## 4.2 Triplet Extraction and Annotation Protocol

The extraction of semantic triples was executed through a hybrid approach combining automated natural language processing pipelines with human-in-the-loop validation. An initial set of legal entities and relations was identified using a pre-trained named entity recognition model fine-tuned on legal corpora. We defined an ontology comprising distinct relation types specifically tailored to capture legal dynamics, such as amends, repeals, cites, affirms, and reverses [11]. Following the automated extraction, a team of domain experts systematically reviewed a statistically significant sample of the extracted triples to assess precision and recall. Erroneous triples were removed, and missing crucial relations were manually annotated to establish a high-quality ground truth dataset. To simulate real-world incompleteness, we intentionally masked a percentage of the validated triples during the training phase, reserving them exclusively for testing the completion models. Negative sampling techniques were employed to generate plausible but incorrect triples, which is a necessary step for training contrastive learning models.

*Table 1: Statistical summary of the constructed Legal Graph Benchmark dataset*

| Dataset Component | Total Count | Average Degree | Density Metric |
|---|---|---|---|
| Legal Entities | 45,210 | 12.4 | 0.0031 |
| Defined Relations | 134 | N/A | N/A |
| Training Triples | 312,850 | N/A | N/A |
| Validation Triples | 42,100 | N/A | N/A |
| Testing Triples | 41,950 | N/A | N/A |

# 5. Experimental Setup and Evaluation Metrics

## 5.1 Model Selection and Implementation Details

Our experimental design evaluates three distinct paradigms of knowledge graph completion: translation-based models, tensor factorization models, and graph neural network models. For translation-based approaches, we implemented standardized versions of models that treat relations as translations in latent space. For tensor factorization, we selected algorithms that compute the similarity between entities and relations using complex mathematical operations, allowing for the representation of symmetric and antisymmetric relation patterns. For the graph neural network paradigm, we utilized models that incorporate multi-relational convolution, aggregating information from an entity's immediate neighbors to generate context-aware embeddings. All models were implemented using widely accepted open-source deep learning frameworks. Hyperparameter optimization was conducted via grid search over the validation set.

Key hyperparameters included learning rate, embedding dimensionality, batch size, and regularization weights. The embedding dimension was standardized across all baseline models to ensure a fair comparison of parameter efficiency. Model training was conducted on high-performance computing clusters utilizing specialized hardware accelerators to manage the large matrix operations inherent in knowledge graph embedding [12].

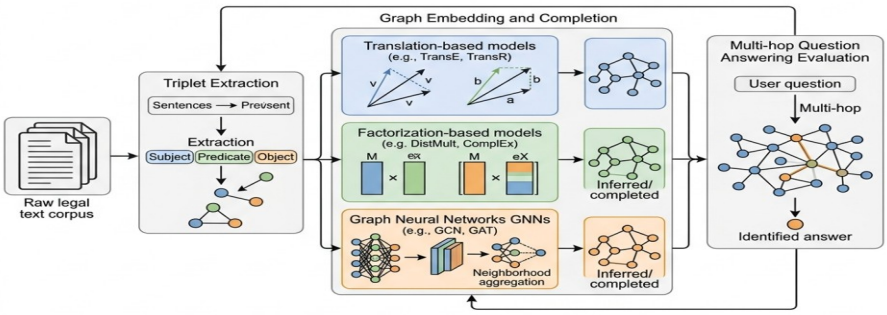


*Figure 2: End*

### 5.2 Evaluation Metrics and Protocols

The evaluation protocol is divided into two distinct phases corresponding to the two primary research objectives: evaluating link prediction accuracy and assessing question answering performance. For knowledge graph completion, we utilize standard link prediction metrics. When testing a missing triple, the model ranks all possible entities in the graph based on the predicted likelihood of forming a valid relation. We report the Mean Reciprocal Rank, which is the average of the reciprocal ranks of the correct entities. Additionally, we report Hits@10, which measures the proportion of correct entities that appear within the top ten predictions. For the question answering evaluation, we developed a specialized query dataset comprising natural language questions paired with their corresponding logical forms and correct answers. These questions range from simple single-hop queries to complex multi-hop reasoning tasks. The question answering system's performance is measured using Exact Match, which requires the system's output to perfectly align with the ground truth answer set, and the macro-averaged F1 score, which provides a harmonic mean of precision and recall over the generated answer sets.

## 6. Quantitative Results of Knowledge Graph Completion

### 6.1 Performance Analysis of Baseline Models

The empirical evaluation of knowledge graph completion models on the Legal Graph Benchmark reveals significant disparities in performance across different algorithmic paradigms. Translation-based models demonstrated a reasonable baseline capability but struggled noticeably with complex legal structures. Specifically, these models exhibited lower Mean Reciprocal Rank scores when processing one-to-many relations, such as a single supreme court decision citing multiple lower court rulings. This limitation stems from their geometric constraints, which force all tail entities of a single relation to cluster too tightly in the embedding space. In contrast, tensor factorization models showed substantial improvements, particularly in handling the symmetric relations prevalent in jurisdictional mappings. By utilizing complex vector spaces or multiplicative interactions, these models could more accurately distinguish between closely related but legally distinct entities. The most robust performance, however, was observed in graph neural network-based approaches. These models consistently outperformed traditional baselines across all metrics. The ability of graph convolutional networks to aggregate structural neighborhood data proved exceptionally

beneficial in the legal domain, where an entity's legal standing is heavily contextualized by its surrounding graph topology [13].

*Table 2: Performance metrics for link prediction on the Legal Graph Benchmark*

| Model Paradigm | Mean Reciprocal Rank | Hits@1 | Hits@10 |
|---|---|---|---|
| Translation Base | 0.284 | 0.192 | 0.451 |
| Tensor Factorization | 0.357 | 0.265 | 0.538 |
| Graph Neural Network | 0.412 | 0.318 | 0.604 |

### 6.2 Relational Dissection and Error Categorization

A granular analysis of the link prediction results based on relation types provides deeper insights into model behavior. We categorized the relations into hierarchical (e.g., subsection_of), historical (e.g., overruled_by), and associative (e.g., mentions_concept). All models performed exceptionally well on hierarchical relations, likely due to their strict tree-like structures which are easily embedded. However, historical relations posed a significant challenge. The legal timeline often involves multiple appeals, reversals, and amendments, creating long, complex chains of events. Models lacking explicit temporal reasoning mechanisms struggled to correctly order these historical dependencies. Furthermore, associative relations exhibited high variance in prediction accuracy. Associative links are often sparse and heavily dependent on specific terminological nuances within the text. Errors in this category frequently involved the model suggesting plausible but legally inaccurate precedents. This categorization underscores the necessity for future completion algorithms to incorporate domain-specific constraints, such as chronological consistency and jurisdictional boundaries, to further reduce the error rates in legal link prediction.

## 7. Question Answering Accuracy and Error Analysis

### 7.1 Impact of Graph Completeness on Query Resolution

The ultimate validation of our benchmark study lies in observing how the improvements in knowledge graph completion translate to question answering accuracy within legal information portals. We evaluated a standardized semantic parsing question answering system over different iterations of our legal knowledge graph: the raw incomplete graph, and graphs augmented by the predictions from our evaluated completion models. The results conclusively demonstrate a strong positive correlation between graph completeness and question answering accuracy. When operating on the raw, incomplete graph, the question answering system frequently failed on multi-hop queries, returning empty sets due to broken reasoning paths. As the graph was augmented with highly confident predictions from the completion models, the F1 scores improved significantly. Notably, the graph completed by the neural network paradigm facilitated the highest accuracy, confirming our hypothesis that structurally aware embeddings infer the bridging edges most critical for complex reasoning. The question answering system's ability to resolve questions involving statutory exceptions and jurisdictional overrides was markedly enhanced when utilizing the comprehensively enriched graph [14].

*Table 3: Impact of graph completion methods on question answering accuracy*

| Graph Status | Exact Match (Single-hop) | F1 Score (Multi-hop) | Overall Accuracy |
|---|---|---|---|
| Incomplete Baseline | 0.652 | 0.384 | 0.518 |
| Factorization Augmented | 0.710 | 0.521 | 0.615 |
| Neural Network Augmented | 0.745 | 0.638 | 0.691 |

### 7.2 Semantic Parsing Failures and Limitations

Despite the clear benefits of knowledge graph completion, the question answering system still exhibited specific failure modes that highlight ongoing challenges in legal informatics. The primary source of error was not always missing links, but rather the ambiguity inherent in translating complex legal queries into formal logical expressions. Natural language queries in the legal domain often contain implicit constraints that are intuitively understood by human professionals but difficult for semantic parsers to capture. For example, a query asking for "recent relevant precedents" requires the system to operationally define "recent" and "relevant" within the specific context of the user's overarching legal problem. Furthermore, we observed instances of error propagation where a false positive link generated by the completion model led the question answering system to retrieve an entirely incorrect, albeit logically sound, answer. This highlights a critical trade-off in legal information portals: while completing the graph improves recall by connecting

fragmented information, it introduces a risk to precision if the completion model hallucinates incorrect legal relationships.

## 8. Conclusion and Future Directions

### 8.1 Summary of Findings

This benchmark study provides a comprehensive evaluation of knowledge graph completion methodologies and their direct impact on the accuracy of question answering systems within legal information portals. By constructing the Legal Graph Benchmark from real-world legal corpora, we established a rigorous testing environment that accurately reflects the complexities of legal semantic structures. Our quantitative analysis revealed that while traditional translation and factorization models offer baseline improvements, graph neural networks provide superior performance by effectively capturing multi-relational structural dependencies. Crucially, the downstream evaluation demonstrated that augmenting legal knowledge graphs with algorithmic predictions significantly enhances the resolution of complex, multi-hop user queries. The empirical evidence confirms that investing in sophisticated graph completion techniques is a highly effective strategy for upgrading the information retrieval capabilities of modern legal platforms.

### 8.2 Limitations and Future Research

While this study establishes a foundational benchmark, several limitations point toward avenues for future research. The current evaluation framework primarily focuses on structural completion, largely ignoring the rich textual metadata associated with legal entities. Integrating unstructured text representations with structural embeddings could drastically improve the prediction of sparsely connected nodes. Additionally, the temporal dynamics of law—where statutes are enacted, amended, and repealed over time—are not fully modeled in static graph completion approaches. Future research must prioritize the development of temporal knowledge graph models that can accurately capture the chronological validity of legal facts. Finally, addressing the explainability of automated predictions remains a critical hurdle. For algorithmic enhancements to be fully trusted by legal professionals, the systems must not only provide correct answers but also offer transparent, logically sound justifications for both the inferred links and the final query resolutions. Continued interdisciplinary collaboration between computer scientists and legal experts will be essential to realizing the full potential of these semantic technologies.

## References

1. Cheong, H.; Kim, J.H.; Münkel, F.; Spilker, H.D. Do Social Networks Facilitate Informed Option Trading? Evidence from Alumni Reunion Networks. J. Financ. Quant. Anal. 2022, 57, 2095–2139.

2. Acemoglu, D.; Carvalho, V.M.; Ozdaglar, A.; Tahbaz-Salehi, A. The network origins of aggregate fluctuations. Econometrica 2012, 80, 1977–2016.

3. Acemoglu, D. Robots and Jobs: Evidence from US Labor Markets. J. Political Econ. 2020, 128, 2188–2244.

4. Bandiera, O.; Rasul, I. Social Networks and Technology Adoption in Northern Mozambique. Econ. J. 2006, 116, 869–902.

5. Burlig, F.; Stevens, A.W. Social networks and technology adoption: Evidence from church mergers in the U.S. Midwest. Am. J. Agric. Econ. 2023, 106, 1141–1166.

6. Xu, J.; Cui, Z.; Wang, T.; Wang, J.; Yu, Z.; Li, C. Influence of Agricultural Technology Extension and Social Networks on Chinese Farmers' Adoption of Conservation Tillage Technology. Land 2023, 12, 1215.

7. Shaikh, I.A.; O'Brien, J.P.; Peters, L. Inside directors and the underinvestment of financial slack towards R & D-intensity in high-technology firms. J. Bus. Res. 2018, 82, 192–201.

8. Abu Qa'dan, M.B.; Suwaidan, M.S. Board composition, ownership structure and corporate social responsibility disclosure: The case of Jordan. Soc. Responsib. J. 2019, 15, 28–46.

9. Bezemer, P.-J.; Peij, S.C.; Maassen, G.F.; van Halder, H. The changing role of the supervisory board chairman: The case of the Netherlands (1997–2007). J. Manag. Gov. 2012, 16, 37–55.

10. Basioudis, I.G. Auditor's Engagement Risk and Audit Fees: The Role of Audit Firm Alumni. J. Bus. Financ. Account. 2007, 34, 1393–1422.

11. Bhalerao, K.; Kumar, A.; Kumar, A.; Pujari, P. A study of barriers and benefits of artificial intelligence adoption in small and medium enterprise. Acad. Mark. Stud. J. 2022, 26, 1–6.

12. Aghion, P.; Howitt, P.W. The Economics of Growth; MIT Press: Cambridge, MA, USA, 2008.

13. Hu, Y.; Fang, J. Executive alumni and corporate social responsibility in China. China Account. Financ. Rev. 2022, 24, 76–

105.

14. Miyamoto, M. Measuring AI Governance, AI Adoption and AI Strategy of Japanese Companies. Int. J. Membr. Sci. Technol. 2023, 10, 649–657.