Article

Not peer-reviewed version

Improving Access to Building Licensing Information in Australia: Design and Development of a Graph-Based Retrieval-Augmented Generation (RAG) Artificial Intelligence (AI) System

Diya Yan * , Jiate Liu , Bocheng Han , Zhengyi Yang , Jun He , Jirong Xu , Riza Yosia Sunindijo , Cynthia Changxin Wang

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0640.v1

Keywords: Artificial Intelligence; Australia; building license; construction industry; Large Language Model (LLM); Retrieval-Augmented Generation (RAG)

Article

Improving Access to Building Licensing Information in Australia: Design and Development of a Graph-Based Retrieval-Augmented Generation (RAG) Artificial Intelligence (AI) System

Diya Yan [1,*], Jiate Liu [2], Bocheng Han [3], Zhengyi Yang [4], Jun He [5], Jirong Xu [6], Riza Yosia Sunindijo [7] and Cynthia Changxin Wang [8]

1 School of Built Environment, UNSW Sydney, Sydney, NSW, 2052, Australia
2 School of Computer Science and Engineering, UNSW Sydney, Sydney, NSW, 2052, Australia
3 School of Computer Science and Engineering, UNSW Sydney, Sydney, NSW, 2052, Australia
4 School of Computer Science and Engineering, UNSW Sydney, Sydney, NSW, 2052, Australia
5 Department of Research, Arcova Pty Ltd, Sydney, NSW, 2052, Australia
6 Department of Research, Jorico Training Pty Ltd, Sydney, NSW, 2052, Australia
7 School of Built Environment, UNSW Sydney, Sydney, NSW, 2052, Australia
8 School of Built Environment, UNSW Sydney, Sydney, NSW, 2052, Australia
* Correspondence: diya.yan@unsw.edu.au

Abstract

Digital technologies have been widely adopted to improve efficiency, transparency, and decision making in the construction industry. However, regulatory processes such as building license and registration applications remain complex, fragmented, and difficult for applicants to navigate, particularly for early career practitioners and small businesses. This study presents the design and development of a graph-based retrieval-augmented generation (RAG) artificial intelligence (AI) system that assists users in applying for building licenses and registrations in New South Wales, Australia. The proposed approach integrates eight complementary frameworks of regulatory burden and service design to identify ten categories of licensing-related burden and translate them into concrete system requirements. The developed prototype provides context aware responses, step-by-step guidance, and tailored information based on user queries, thereby reducing regulatory burden for individuals, companies, and industry bodies. Prototype evaluation against general-purpose AI tools indicates that the system can improve information accessibility and reduce application-related friction in representative licensing scenarios. This study sheds light on AI-enabled regulatory support systems and demonstrates how RAG can be applied to improve accessibility and usability of construction related licensing processes. The findings have implications for policymakers, regulators, and researchers seeking to leverage AI to support digital transformation in the construction industry.

Keywords: Artificial Intelligence; Australia; building license; construction industry; Large Language Model (LLM); Retrieval-Augmented Generation (RAG)

---

1. Introduction

As one of the largest sectors in Australia, the building and construction industry contributes approximately 12 percent of national gross domestic product (GDP) and employs more than 1.3 million people, accounting for around 9 percent of the total workforce [1]. Among all Australian states, the New South Wales (NSW) represents the largest construction market, generating around one third of the total value of construction work done nationally [2] and employing a similar

 doi:10.20944/preprints202602.0640.v1

proportion of the construction workforce [3]. Beyond its economic and workforce significance, the Australian construction industry has been actively engaged in digital transformation, with digital technologies increasingly used to support project delivery, information exchange, and regulatory processes. In NSW, individuals and businesses must hold an appropriate license or registration to carry out any residential building work, including general building work valued at more than AUD 5000 in labor and materials [4]. These licensing requirements are designed to ensure practitioner competency, public safety, and compliance with statutory standards. Under the Home Building Act 1989, it is an offence to carry out, advertise, or contract for regulated work without holding the appropriate license or registration [5]. The scale and strictness of these regulatory requirements highlight both the importance of licensing systems and the challenges faced by applicants in navigating complex and evolving procedures.

To support individuals and businesses applying for building, trade, and specialist licenses and registrations, the NSW Government has invested AUD 166.5 million in a multiyear licensing program to deliver digital end-to-end licensing journeys [6]. In 2025, an additional AUD 62.5 million was committed to the License NSW system to replace legacy technology, enhance system security, and improve the speed and convenience of interactions for license applicants [7]. However, despite these efforts, licensing related information remains fragmented across multiple webpages, policy documents, and external platforms, often requiring extensive reading and interpretation by applicants. Navigating eligibility criteria, documentation requirements, and application pathways continue to be time consuming and cognitively demanding. As listed on the application website, the common reasons for application refusal or resubmission include ineligibility, incomplete documentation, or failure to meet specific regulatory requirements [8]. These challenges contribute to increased administrative burden and delays for individuals and businesses seeking to enter or operate within the construction industry.

In recent years, the rapid advancement of artificial intelligence (AI) and large language models (LLMs) has introduced new opportunities to assist individuals and businesses applying for licenses and registrations in the construction industry. LLM demonstrate strong capabilities in processing large volumes of text, synthesizing complex regulatory information, and generating coherent, context sensitive responses to user queries [9]. Such capabilities are particularly relevant to licensing contexts, where applicants are often required to interpret detailed eligibility criteria, procedural requirements, and supporting documentation across multiple sources. For example, Cheong, et al. [10] investigated the use of LLM-based chatbots to support users in legal advisory contexts through case-based reasoning. The study highlighted the effectiveness of LLMs in helping users identify, organize, and prepare salient information for legal proceedings, rather than providing direct legal recommendations. Similarly, Shittas, et al. [11] introduced a chatbot assistant designed to help users better understand and select open source software licenses by mimicking human-like discussions and recommending appropriate licenses with clear explanatory justifications. Within the built environment domain, Synott and Aksenova [12] examined the adoption of AI in architecture, engineering, and construction consultancy firms, highlighting its benefits in fostering operational efficiency, knowledge management, and decision support. Aboelazm and Dganni [13] further explored the potential of AI applications in the tendering process, demonstrating improvements in efficiency related to document analysis, keyword searching, and information retrieval from tender documentation. These studies suggest that AI-powered tools can reduce information search effort, improve user understanding, and enhance access to regulatory knowledge. However, many existing approaches rely on general purpose language models, raising concerns regarding information accuracy, traceability, and consistency in regulatory contexts, particularly for individuals and businesses applying for building, trade, and specialist licenses and registrations in the Australian construction industry.

The recent adoption of retrieval-augmented generation (RAG) has introduced new opportunities by combining the generative capabilities of LLMs with explicit retrieval from trusted external knowledge sources. RAG enables user queries to be answered based on relevant documents

 doi:10.20944/preprints202602.0640.v1

retrieved from a predefined corpus, ensuring that responses are grounded in authoritative and up to date information rather than relying solely on a model's internal knowledge [14]. However, conventional RAG systems typically retrieve evidence as isolated text chunks based on lexical or semantic similarity. This approach makes it difficult to consistently address complex queries that require linking dispersed constraints across multiple documents and procedural steps [15]. Beyond traditional RAG, graph-based RAG refers to a class of methods that organize knowledge using explicit graph structures, such as entities, relationships, and higher-level communities, and retrieve structurally connected evidence to support generation [16]. Compared with flat retrieval approaches, this method is better suited to multi-step reasoning and improves traceability by exposing explicit reasoning paths and evidence linkages [17]. This capability of graph-based RAG is particularly beneficial for licensing and registration contexts, where eligibility criteria, procedural requirements, and supporting documentation are distributed across different web pages, and even minor interpretation errors can result in delays, resubmissions, or application refusal. Despite these advantages, there are few, if any, studies that apply graph-based RAG to support building, trade, and specialist licensing processes in the Australian construction industry.

To address this gap, this study aims to design and develop a graph-based RAG AI system grounded exclusively in official government documentation to support individuals and businesses in applying for, renewing, and amending building, trade, and specialist licenses and registrations in Australia. Accordingly, this study seeks to address three research questions: 1) How can a theoretical framework be applied to identify regulatory burden associated with building, trade, and specialist licensing and registration processes; 2) How can a system architecture be designed to support licensing-related information seeking and reduce regulatory burden for individuals and businesses; and 3) What future development roadmap can be identified for graph-based RAG AI system in the Australian construction licensing contexts.

By answering these research questions, this study goes beyond prior research in three ways. First, this study integrates eight complementary frameworks, namely the Regulatory Burden Measurement Framework, the Standard Cost Model, the Regulatory Impact Assessment Framework, the Business Regulatory Cost Framework, the Administrative Burden Framework, the Sludge and Administrative Friction Framework, the Innovative Service Design Frameworks, and Information Asymmetry Theory. Together, these frameworks inform the systematic identification of ten types of regulatory burden, including administrative and procedural costs, substantive compliance costs, time and delay related costs, information and learning related costs, opportunity and economic costs, financial and monetary costs, transaction and coordination costs, non-compliance and enforcement related costs, psychological and behavioral costs, and indirect and system level costs. This framework enables regulatory theory to be translated into AI system functionalities, reducing information-related burdens experienced by license applicants. Second, the system is designed using a graph-based RAG approach with official government documentation. This design ensures information accuracy, traceability, and regulatory alignment, while enabling context aware responses and step-by-step guidance tailored to user queries in building, trade, and specialist licensing processes. Third, this study compared the developed prototype with general purpose LLMs and proposed a future development roadmap for RAG based AI systems in construction licensing contexts. The findings demonstrate that the developed prototype can improve information accessibility, reduce application friction, and potentially lower barriers to workforce entry in a highly regulated industry.

2. Theoretical Framework

To guide the system architecture design, it is important to understand the regulatory burdens faced by individuals and organizations when applying for building licenses and registrations, enabling the AI tool to be designed with targeted functionalities to reduce these burdens. Regulatory burden refers to the costs imposed on individuals and organizations as a result of government regulation, beyond direct financial charges such as fees or taxes [18]. These burdens typically arise

 doi:10.20944/preprints202602.0640.v1

from the time, effort, and resources required to understand regulatory requirements, demonstrate compliance, and navigate approval processes.

A literature search was conducted using Scopus with the keyword "regulatory burden", followed by a snowballing technique based on the reference lists of relevant studies. This approach enables the systematic identification of both foundational and influential frameworks, while reducing the risk of omitting widely cited or conceptually significant work. As a result, eight frameworks were selected to inform the theoretical framework of this study and to translate regulatory burden into system design objective. Table 1 summarizes the selected frameworks and the categories they encompass. Figure 1 outlines the theoretical framework.

Table 1. Factors mapping.

Category	Examples	1	2	3	4	5	6	7	8
Administrative and procedural costs	Administrative compliance activities including application preparation, form filling, documentation, record keeping, reporting and notifications Interaction and communication with authorities including repetitive submissions	X	X	X	X	X	X	X	
Substantive compliance costs	Costs of meeting regulatory requirements including training, qualifications, equipment acquisition, system implementation, maintenance and ongoing compliance Professional and advisory service fees	X		X	X				
Time and delay related costs	Application, approval and waiting time delays Process inefficiencies and opportunity losses arising from delays	X		X	X	X	X	X	
Information and learning related costs	Information search, interpretation and learning costs Uncertainty, cognitive load and information asymmetry					X	X	X	X
Opportunity and economic costs	Foregone income, delayed market entry and missed business opportunities Reduced productivity and constrained economic activity	X		X	X		X		X
Financial and monetary costs	Application, license and transaction fees Direct monetary compliance costs including consultant and third-party expenses	X	X	X	X				
Transaction and coordination costs	Search and discovery costs across regulatory systems Coordination, monitoring, verification and negotiation across multiple agencies			X	X			X	X
Non-compliance and enforcement related costs	Penalties, fines and enforcement related legal costs Rectification and compliance remediation costs	X		X					
Psychological and behavioral costs	Stress, anxiety, frustration and disengagement associated with compliance processes Loss of confidence in regulatory systems					X	X	X	
Indirect and system level costs	Market distortion and competition impacts Informality risks and equity or distributional effects	X		X	X				

 doi:10.20944/preprints202602.0640.v1

* 1= Regulatory Burden Measurement Framework [19]; 2 = Standard Cost Model [20]; 3= Regulatory Impact Assessment Framework [21]; 4= Business Regulatory Cost Framework [22]; 5= Administrative Burden Framework [23]; 6= Sludge and Administrative Friction Framework [24]; 7= Innovative Service Design Frameworks [25]; 8= Information Asymmetry Theory [26].

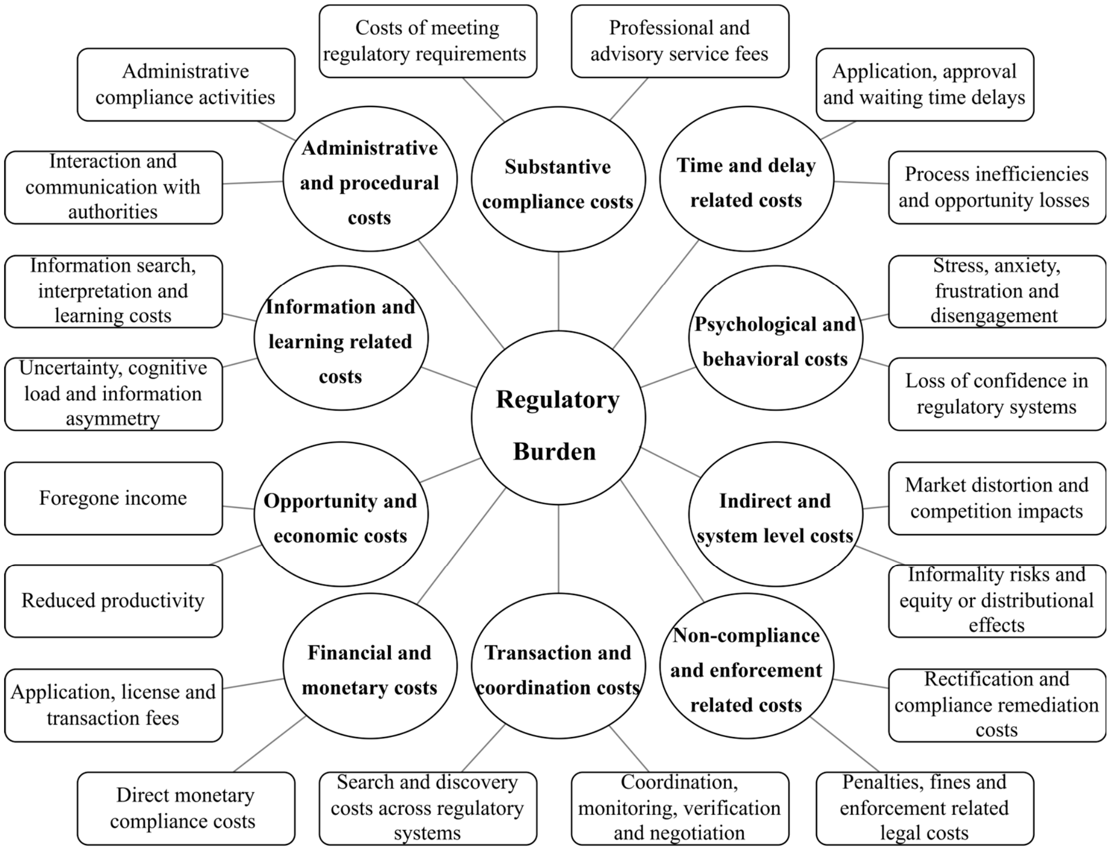

Figure 1. Theoretical Framework.

Administrative and procedural costs encompass the time and effort required to understand regulatory obligations, prepare and submit documentation, and communicate with regulatory authorities [27]. These costs arise from interacting with administrative systems rather than from performing the regulated activity itself. In the context of building, trade, and specialist licenses and registrations, administrative and procedural costs include completing application forms, preparing supporting documentation, providing identity verification, maintaining records, submitting renewals, and responding to requests from regulators. Applicants may also incur costs associated with navigating multiple portals, understanding eligibility criteria, and complying with procedural rules that differ across license categories. This type of cost is often addressed through the development of electronic portals and digital service platforms that streamline administrative procedures and standardize application processes [28].

Substantive compliance costs refer to the costs incurred to meet the substantive requirements imposed by regulation to achieve the intended policy outcomes [29]. These costs relate to what applicants must do or acquire to qualify for a license, rather than how they demonstrate compliance. For building, trade, and specialist licenses, substantive compliance costs include completing mandatory training and qualifications, undertaking required apprenticeships or supervised practice, maintaining continuing professional development, purchasing compliant tools or equipment, obtaining compulsory insurances, and meeting safety or technical standards. As suggested by Nandan Prasad [30], substantive compliance costs can be reduced through clearer regulatory guidance, early identification of mandatory requirement, and minimize unnecessary duplication.

doi:10.20944/preprints202602.0640.v1

Time and delay related costs refer to losses incurred due to time spent waiting for regulatory processes to be completed [31]. These costs may arise even when the application is ultimately approved. In licensing and registration processes, applicants may experience delays during application assessment, background checks, verification of qualifications, or approval stages. Such delays can prevent individuals from legally undertaking work, bidding for contracts, or progressing their careers. In NSW, the current processing time for building licenses is approximately 19 weeks [5]. This delay is largely attributable to government led assessment and verification procedures, including eligibility checks, qualification validation, and regulatory approvals, rather than actions undertaken by applicants themselves.

Information and learning related costs arise from the effort required to locate, interpret, and understand regulatory information [32]. These costs are particularly significant when information is fragmented, complex, or written in technical language. Applicants for building, trade, and specialist licenses often need to consult multiple government webpages, legislation, guidelines, and industry standards to understand eligibility criteria, application steps, and ongoing obligations. Examples include time spent searching for accurate information, interpreting legal terminology, understanding which license category applies to a specific scope of work, and learning how regulatory requirements differ across jurisdictions or license classes. These costs are where AI is most usefully applied, as LLMs can synthesize dispersed regulatory documents, translate technical and legal language into plain explanations, and provide context specific guidance to support applicants in understanding licensing and registration requirements [9].

Opportunity and economic costs represent the value of alternative activities or income that applicants forgo for engaging in licensing and registration processes. In the licensing context, individuals may need to take unpaid leave to attend training, assessments, or examinations, or may delay entering the workforce while completing regulatory requirements [33]. These costs are primarily driven by the regulatory requirements themselves. However, clearer guidance and improved information provision can reduce the risk of mistakes, rework, and unnecessary delays, thereby mitigating the downstream economic impacts experienced by applicants.

Financial and monetary costs are direct out-of-pocket expenses incurred by applicants due to regulatory requirements. For building, trade, and specialist licenses, these costs include application fees, renewal fees, assessment fees, training costs, insurance premiums, medical checks, and certification expenses. Although those costs are unavoidable, improved understanding of regulatory requirements can reduce the likelihood of rework, repeat applications, and unnecessary fees.

Transaction and coordination costs arise from the need to interact with multiple people, systems, and institutions to complete regulatory processes. Applicants for licenses and registrations often need to coordinate with training providers, employers, referees, insurers, assessors, and government agencies. According to Adepoju, et al. [34], these costs can be reduced through clearer process integration and improved information flow, which minimize the need for repeated interactions with multiple stakeholders and reduce duplication in documentation and communication.

Non-compliance and enforcement related costs refer to costs incurred when applicants fail to meet regulatory requirements or unintentionally breach licensing conditions. In the context of building and trade licenses, these costs may include fines, penalties, legal fees, costs associated with corrective actions, and time spent resolving compliance issues [5]. Non-compliance and enforcement related costs can be avoided if applicants clearly understand licensing obligations and regulatory boundaries before undertaking regulated work.

Psychological and behavioral costs refer to the stress, anxiety, frustration, and cognitive burden associated with regulatory processes [35]. Licensing and registration processes can be perceived as complex, opaque, and high risk, particularly for first time applicants or individuals from non-traditional backgrounds. Examples include anxiety about making errors in applications, uncertainty about eligibility, frustration caused by unclear guidance, reduced confidence in navigating regulatory systems, and decision fatigue resulting from repeated compliance tasks. The ability of AI can reduce this type of cost by lowering cognitive burden, reducing uncertainty, and providing timely and consistent guidance throughout the licensing and registration process [36].

Indirect and system level costs arise from broader impacts of regulatory systems that affect individuals, industries, and markets beyond direct compliance activities. In the construction sector, complex licensing systems may contribute to barriers to workforce entry, skills shortages, reduced competition, and uneven access to opportunities. According to the Bennett and Estrin [37], reducing indirect and system level costs can improve overall system efficiency, lower barriers to entry, and enhance workforce participation.

3. System Architecture Design

A system architecture defines the high level structure of a computer system, specifying its core components and the interactions among them to achieve the intended functionalities [38]. In this study, the proposed system architecture comprises three layers: a knowledge base layer, a reasoning layer, and a user interface layer. Each layer is deliberately designed to address specific categories of regulatory burden identified in the theoretical framework. Figure 2 illustrates the overall system architecture design.

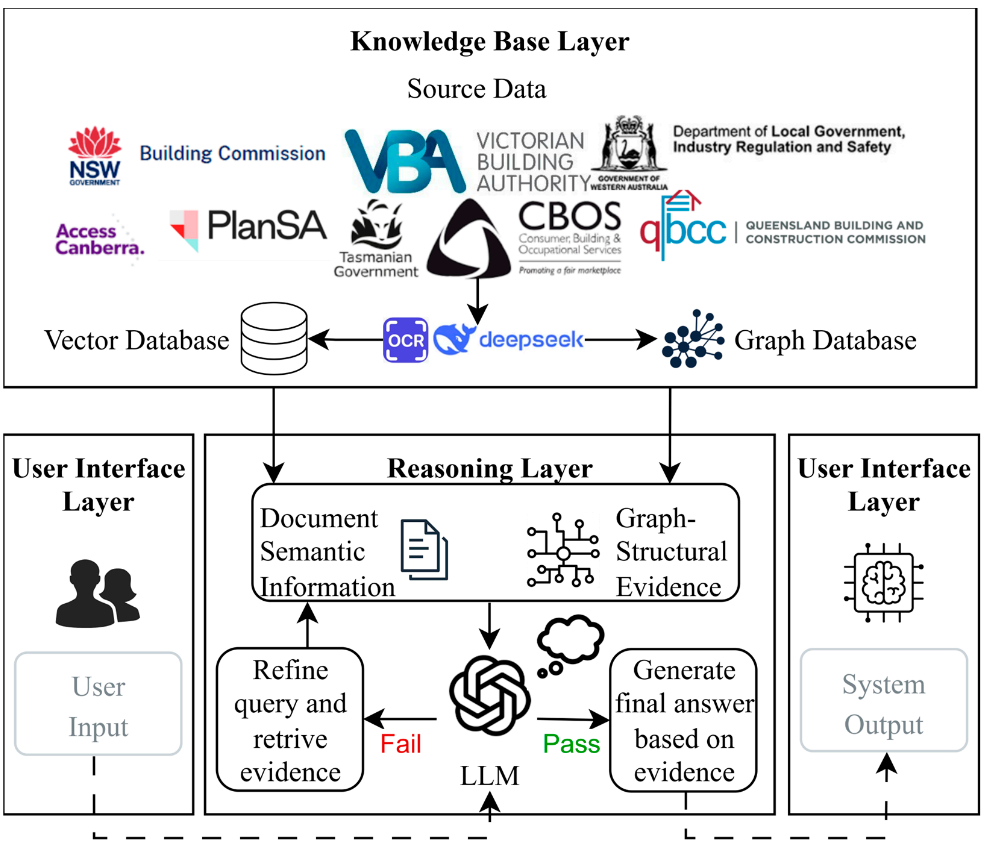

Figure 2. System Architecture Design.

3.1. Knowledge Base Layer

The knowledge base layer serves as the foundational component of the system by supplying structured and domain specific information to support downstream reasoning processes. Rather than relying on general purpose knowledge, this layer ensures that the AI system operates on authoritative and context relevant regulatory content, thereby reducing the risk of irrelevant outputs and hallucinated responses [39]. Its design is motivated by three core requirements in licensing and registration contexts: 1) authority, ensuring that responses are grounded in official government

 doi:10.20944/preprints202602.0640.v1

sources; 2) traceability, allowing every recommendation to be directly linked to verifiable regulatory excerpts; and 3) retrievability, requiring that regulatory information is systematically segmented and indexed to enable efficient and reliable access.

3.1.1. Source Data

Source data is a key element that differentiates this AI system from general purpose LLMs. To ensure the validity, reliability, and authority of the information used, all source data were obtained exclusively from official government websites that provide licensing and registration guidance for the construction industry. Using government websites as the primary data source ensures that the AI system is grounded in current and legally authoritative regulatory information, thereby minimizing risks associated with outdated, incomplete, or unofficial interpretations. Additionally, the Australian government websites offer print and export functions, which allow each webpage to be stored as a stable and verifiable document in portable document format (PDF). These documents were systematically collected and archived to form the raw source dataset for the knowledge base from 17 government agencies, including national bodies such as the Australian Building Codes Board and Safe Work Australia, as well as state and territory regulators responsible for building, trade, and specialist licensing, such as NSW Fair Trading, SafeWork NSW, Victorian Building Authority, Queensland Building and Construction Commission, Building and Energy Western Australia, Plan South Australia, Consumer Building and Occupational Services Tasmania, Building Practitioners Board, and Access Canberra.

The source data in PDF format cannot be used directly by the AI system because the content is unstructured, difficult to search, and does not explicitly show how requirements are connected across documents. Therefore, a dedicated processing pipeline is designed using a dual representation approach to transform these PDFs into structured, machine-readable representations that support accurate reasoning and retrieval. First, a text evidence corpus is indexed with dense embeddings to support semantic matching at the chunk level. Second, a lightweight regulatory graph is constructed to encode requirement-level dependencies and cross-page relations. This design is intended to mitigate two complementary failure modes: purely “flat” chunk retrieval may miss prerequisite constraints that are semantically distant but logically required, whereas graph-only retrieval can be sensitive to incompleteness and noise introduced during graph construction. Combining both representations is therefore expected to improve coverage and robustness while preserving provenance.

3.1.2. Relation to the Theoretical Framework

The knowledge base layer is designed to directly reduce three key types of regulatory burden: information and learning related costs, administrative and procedural costs, and non-compliance and enforcement related costs. First, by consolidating official regulatory materials from multiple government agencies into a single, searchable repository, the system significantly reduces the effort and time required to locate, read, and understand fragmented licensing information. Using authoritative sources also supports more accurate interpretation of regulatory requirements, thereby reducing uncertainty and information asymmetry faced by license applicants. Second, the availability of clearly structured and source grounded information helps lower administrative and procedural costs by explicitly presenting eligibility criteria, evidence requirements, and application steps in an accessible format. This improves users’ ability to prepare applications and supporting documentation efficiently, while reducing the likelihood of incomplete or incorrect submissions. Lastly, because all content in the knowledge base is derived directly from official government websites, the system reduces the risk of inadvertent noncompliance. By ensuring that guidance reflects current and legally valid regulatory requirements and allowing users to verify information against original sources, the knowledge base layer helps minimize rectification efforts and compliance remediation costs.

3.2. Reasoning Layer

 doi:10.20944/preprints202602.0640.v1

The reasoning layer refers to the component responsible for processing user inputs, retrieving relevant information from the knowledge base, and generating context specific recommendations. This layer transforms user queries into grounded responses by retrieving relevant evidence from the knowledge base and synthesizing it into actionable guidance.

3.2.1. Graph-Based RAG

RAG improves answer accuracy by requiring the system to base its responses on external regulatory documents retrieved at the time a question is asked [14]. This approach reduces reliance on the language model's internal memory alone and helps ensure that answers are grounded in verifiable and authoritative sources [40]. Graph-based RAG further enhance this process by representing regulatory requirements and their relationships as a structured network [41]. This allows the system to consider dependencies between rules and to combine related pieces of evidence when generating answers [42]. Such approaches are particularly important in licensing contexts, where missing or misunderstood requirements can lead to application resubmission, processing delays, and additional costs for applicants.

Building on these foundations, this study adopts a Graph-based RAG framework, which integrates adaptive and agent-based decision logic to address two persistent challenges in regulatory question answering. The first challenge is coverage risk, where dispersed prerequisites, exceptions, or scope limitations may be overlooked because they appear across multiple documents or sections. The second challenge is cost risk, which arises when excessive information is retrieved or overly long prompts are generated, increasing processing effort without improving answer quality.

To address these issues, the framework follows a staged strategy. It begins with lightweight text retrieval to efficiently handle straightforward questions. If the retrieved information is judged to be insufficient, the system then selectively expands the search using regulatory relationships captured in a structured rule network. This expansion is guided by the specific question being asked and propagates only along relevant regulatory links, as implemented in the prototype. Through this adaptive process, the framework improves the completeness of answers for dependency rich regulatory queries, while maintaining efficiency for simpler information requests.

3.2.2. Relation to the Theoretical Framework

The reasoning layer is designed to reduce three key categories of regulatory burden: substantive compliance costs, time and delay related costs, and opportunity and economic costs. By retrieving and combining information from multiple authoritative regulatory documents, the system provides clearer and more complete interpretations of licensing requirements. This reduces applicants' reliance on professional advisory services or repeated consultations to understand complex compliance obligations. As a result, users can more accurately identify mandatory training, qualifications, documentation, and other substantive requirements, helping to avoid both under compliance and unnecessary expenditure. In addition, the reasoning process reduces time and delay related costs by identifying prerequisite conditions, dependencies, and correct application sequences at an early stage. By guiding users toward more complete and consistent submissions before formal lodgment, the system lowers the likelihood of approval delays, resubmissions, and prolonged waiting periods caused by missing or conflicting information. Finally, by shortening the regulatory learning curve and reducing procedural rework, the reasoning layer mitigates opportunity and economic costs. Applicants can progress through licensing workflows more efficiently, which supports earlier market entry, reduces foregone income, and improves overall productivity.

3.3. User Interface Layer

The user interface layer represents the component through which users interact with the system. It serves as the communication bridge between users and the reasoning layer by capturing user inputs, transmitting queries for analysis, and presenting generated outputs in clear and readable text.

 doi:10.20944/preprints202602.0640.v1

Through this interaction, users can obtain regulatory guidance and system recommendations in an accessible and intuitive manner. A web based platform was selected as the delivery format due to its broad accessibility, ease of deployment, and capacity to integrate seamlessly with backend services and analytical component [43]. The user interface was guided by principles of clarity, simplicity, and accessibility, ensuring that information is easy to read, navigate, and interpret, even for users with limited regulatory or technical expertise.

3.3.1. User Input and System Output

A web-based conversational interface is adopted to maximize accessibility and to support iterative clarification during user interaction. The interface accepts free text questions written in plain English, allowing users to engage with the system without requiring prior familiarity with licensing terminology, regulatory structures, or procedural language. User interaction is designed to be flexible and intuitive, enabling users to ask follow-up questions and progressively refine their understanding of regulatory requirements. This conversational approach reduces cognitive effort and lowers barriers to engagement, particularly for individuals and small businesses with limited regulatory expertise.

System responses are presented in a usability focused format. Regulatory guidance is translated into clear, plain English explanations and organized as structured checklists and step-by-step instructions to help users identify required actions efficiently. Where appropriate, responses include direct links to official government webpages and application resources, enabling users to verify information and take follow through actions. Through this design, the user interface layer improves information usability, supports incremental learning, and enhances overall accessibility of the licensing process.

3.3.2. Relation to the Theoretical Framework

The user interface layer is designed to reduce three categories of regulatory burden: transaction and coordination costs, psychological and behavioral costs, and indirect and system level costs. By offering a single conversational access point to licensing and registration information, the interface reduces the effort associated with searching across multiple agency websites, navigating different systems, and coordinating information from fragmented sources. Users can obtain relevant guidance, supporting links, and next steps within one interface, thereby lowering the time and effort required to locate, verify, and act on regulatory information. Additionally, the use of plain English explanations and a conversational interaction style help reduce psychological and behavioral costs. By replacing complex legal or procedural language with clear, predictable question and answer exchanges, the interface lowers stress, frustration, and uncertainty commonly experienced during compliance processes. The structured presentation of information, such as dot point summaries and step by step guidance, further improves transparency and user confidence. Lastly, by improving accessibility to regulatory information and lowering interaction barriers, the user interface layer contributes to reducing indirect and system level costs. The system supports more equitable access for individuals and small businesses with limited regulatory literacy, reduces reliance on informal or noncompliant information sources, and helps remove information driven barriers to entry into regulated markets.

4. Prototype Development

After the system architecture and design were finalized, a working prototype was implemented using Python. This programming environment was selected due to its extensive ecosystem of libraries that support document processing, information retrieval, and integration with LLMs.

4.1. Knowledge Base Layer

4.1.1. Source Data Collection

The source data was collected through official government websites. Due to budget and time constraints, the prototype focuses on the construction industry in NSW, Australia. Relevant government websites, including the New South Wales Government, SafeWork NSW, NSW Fair Trading, and the NSW Building Commission, were accessed. The regulatory documents were downloaded and stored in PDF directly from the original webpages. In total, 56 webpages were included, covering eligibility requirements, application procedures, supporting evidence, fee structures, renewal conditions, and ongoing compliance obligations associated with building related licenses and registrations.

The collected regulatory documents were organized using a dual index repository designed to support two complementary forms of retrieval. First, the documents were divided into smaller, logically coherent sections and stored as a searchable text collection that enables information to be retrieved based on meaning rather than exact wording. Second, a lightweight regulatory relationship structure was created to represent dependencies between requirements and links across different documents.

4.1.2. Source Data Processing Pipeline

To enable reliable retrieval by LLMs while preserving clear links to original regulatory sources, a license-oriented processing pipeline was implemented. This pipeline transforms archived regulatory documents stored as PDFs into structured text segments and a lightweight regulatory graph, ensuring that all generated outputs remain traceable to authoritative government materials. The algorithm for license-oriented LLM-aware recursive chunking is presented Appendix A.

Step 1: PDF to Markdown conversion using DeepSeek-OCR 2. Each archived regulatory webpage is stored as a PDF snapshot. These PDFs are first rendered into page level images and then processed using DeepSeek-OCR 2 to produce structured Markdown outputs that preserve layout information such as headings, lists, and tables [44]. Compared with plain text extraction, the Markdown format retains explicit multi-level section markers, such as #, ##, and ###. These markers are critical for identifying the structure of licensing guidance and enable subsequent segmentation to align with regulatory workflows rather than arbitrary text length.

Step 2: LLM aware recursive segmentation using document headings. Each Markdown document is then segmented using its heading structure. Headings define a natural hierarchy of sections and subsections, which is traversed from top to bottom to generate an initial set of text segments. This process is formalized as:

$$C_0 = SPLIT\,(r) \tag{1}$$

where r represents the document level section. In simple terms, sections are kept intact when they are of reasonable length, subdivided when they are too long, and only split internally when no meaningful headings are available. When headings are insufficient, a conservative fallback method is used that respects sentence boundaries and list items, ensuring that individual license requirements are not broken apart.

Step 3: Second pass refinement of text segments. After initial segmentation, a second refinement step ensures that all text segments fall within a suitable length range for reliable retrieval and reasoning:

$$L_{min} \le l(c_i) \le L_{max} \tag{2}$$

Segments that are too short are merged with neighboring segments, while segments that are too long are further divided using the same structure aware logic. Throughout this process, references to the original document's location are preserved.

Step 4: Graph construction and provenance linkage. Each final text segment is treated as a single unit of regulatory evidence. Simple information extraction techniques are applied to identify relationships between regulatory concepts, such as conditions, obligations, or references. These relationships are combined to form a lightweight regulatory structure:

 doi:10.20944/preprints202602.0640.v1

$$G = \bigcup_{c \in C} \mathrm{OPENIE}(c). \tag{3}$$

Crucially, every extracted relationship is linked back to the exact text segment from which it was derived. This linkage ensures full traceability, allowing any system output to be verified against authoritative regulatory sources.

Step 5: Outputs and indexing. The pipeline produces two tightly coupled outputs: a final set of structured text segments and a regulatory relationship structure. Each text segment is stored with provenance information, including the original document, webpage location, and section path. The knowledge base is indexed to support both text-based retrieval and relationship-based expansion. For text retrieval, keyword matching and meaning based similarity are combined into a single score:

$$\mathrm{SCORE}_{\text{text}}(c \mid q) = \lambda \cdot \mathrm{BM25}(q, c) + (1 - \lambda) \cdot \cos\,(e_q, e_c), \tag{4}$$

In plain terms, this allows the system to balance exact terms matching semantic similarity, improving robustness across different question styles while maintaining links to official regulatory text.

4.2. Reasoning Layer

Aligned with the overall system architecture design, the prototype reasoning layer implements a graph-based RAG workflow. In this workflow, relevant regulatory text is first retrieved from the document repository. When a question involves linked or prerequisite requirements, the system then selectively expands its search using the regulatory relationship structure. This design adopts graph-based RAG approaches and their graph enhanced extensions, which are well suited to complex regulatory domains [14,40-42].

Step 1: Tex-based RAG. RAG operates by retrieving relevant regulatory material at the time a question is asked and using that material to guide answer generation [45,46]. When a user submits a query q, the system identifies the most relevant regulatory text segments from the document collection using a text relevance score. These segments, together with their source information, are provided to the language model so that responses are grounded in verifiable regulatory excerpts rather than generated from memory alone. This initial stage works well for straightforward questions where the required information is contained within a small number of clearly worded sections.

In licensing and registration contexts, requirements are rarely stated in a single location. Eligibility conditions, required evidence, scope limitations, and exemptions are often distributed across multiple documents and linked through logical dependencies. As a result, retrieving only the most textually similar sections can overlook prerequisite or conditional requirements that are essential for a complete and accurate answer.

Step 2: Graph-guided expansion. To address this limitation, the system uses a regulatory relationship structure that represents how requirements are connected. This structure is treated as a network of related regulatory concepts.

To identify which additional requirements may be relevant to a given question, the system applies a Personalized PageRank (PPR) that starts from a small set of query related nodes and gradually propagates relevance to connected nodes [47,48]. The transition process over the network is defined as:

$$P = D^{-1} A \tag{5}$$

where the matrix represents how relevance flows between connected regulatory concepts.

The starting points for this process are determined by a combination of concepts explicitly mentioned in the user query and concepts associated with the initially retrieved text. These two signals are combined as

$$\mathbf{s}_q(v) = \eta \cdot \mathbf{s}_q^{\text{ent}}(v) + (1 - \eta) \cdot \mathbf{s}_q^{\text{ret}}(v) \tag{6}$$

 doi:10.20944/preprints202602.0640.v1

This allows the system to balance what the user explicitly asked about with what was already identified as relevant in the first retrieval stage.

Relevance is then propagated through the regulatory network using

$$\boldsymbol{\pi}_q = (1-\alpha)\mathbf{s}_q + \alpha\mathbf{P}^{\top}\boldsymbol{\pi}_q \tag{7}$$

Conceptually, this process spreads importance from the initial query-related requirements to other requirements that are logically connected, such as evidence items linked to an eligibility condition or restrictions linked to a license category.

Each regulatory concept in the network is linked to one or more supporting text segments. The relevance of a text segment is therefore determined by the relevance of the regulatory concepts it supports:

$$Score_{\text{graph}}(c \mid q) = \sum_{v \in \mathcal{V}} \boldsymbol{\pi}_q(v) \cdot \mathbb{I}\left[c \in \phi(v)\right] \tag{8}$$

This graph-based relevance is then combined with the original text-based relevance to produce a final ranking:

$$Score(c \mid q) = \beta \cdot Score_{\text{text}}(c \mid q) + (1-\beta) \cdot Score_{\text{graph}}(c \mid q) \tag{9}$$

In practical terms, this means that the system considers both how closely a text matches the wording of the question and how important it is within the broader regulatory dependency structure.

Licensing questions vary widely in complexity. Some can be answered using a small number of text segments, while others require consideration of multiple linked requirements. To avoid unnecessary processing, the prototype adopts a staged retrieval strategy.

After each retrieval stage, the system evaluates whether the collected evidence is sufficient to answer the question. Retrieval stops as soon as sufficient evidence is identified:

$$t^* = \min\{t : SUFF(\mathcal{E}_t, q) = 1\} \tag{10}$$

If the initial text-based retrieval is insufficient, the system escalates to graph guided expansion and retrieves additional supporting material only when needed. This adaptive approach reduces unnecessary computation and keeps responses concise, while still ensuring completeness for questions involving complex regulatory dependencies.

4.3. User Interface Layer

The prototype user interface is implemented as a web-based interactive assistant designed to support task oriented regulatory workflows. Users are provided with configurable options to select the underlying LLM, such as GPT 4.1, Claude 3.5, and Gemini 1.5, as well as the preferred response style, such as normal conversational responses, concise bullet point summaries, or detailed step-by-step explanations. This flexibility allows the system to accommodate different user preferences and use cases while maintaining consistent regulatory guidance. Figure 3 illustrates the prototype user interface.

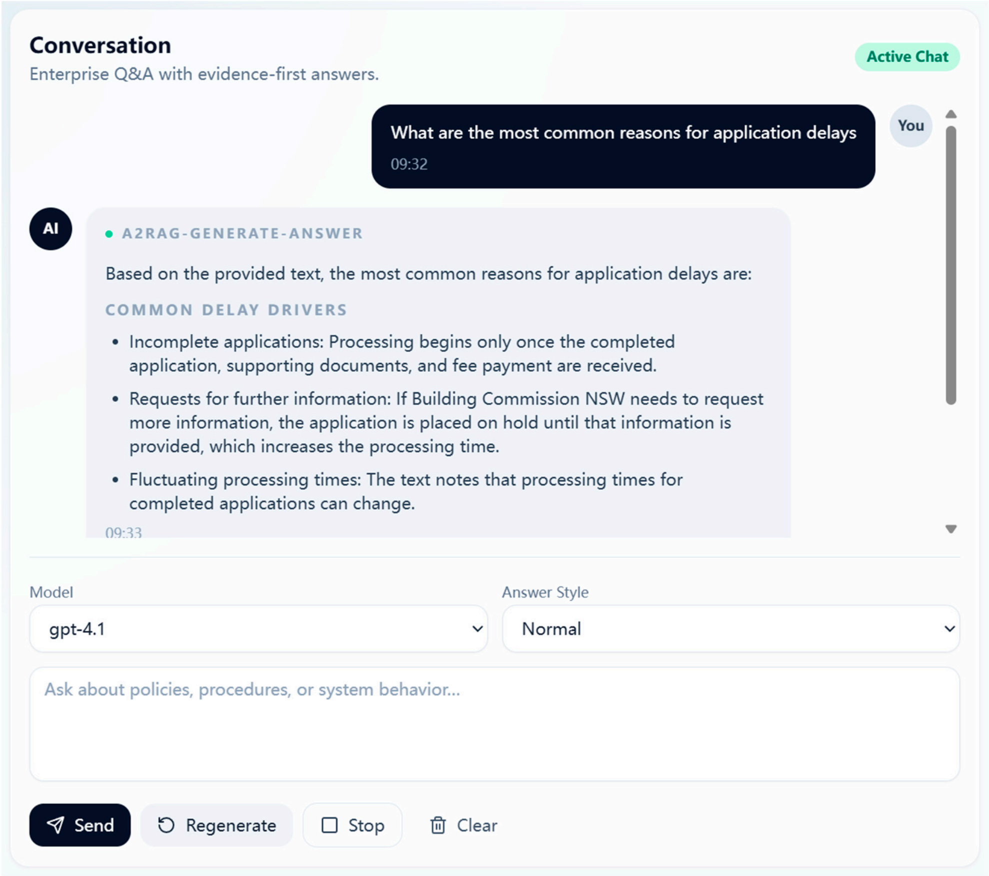

Figure 3. User interface.

5. Prototype Performance

The performance of the prototype was evaluated against two commonly used general purpose AI tools, namely ChatGPT powered by GPT-4o developed by OpenAI, and Google Search AI Overview powered by the Gemini 2.5 developed by Google. A total of 50 mock questions were designed to stimulate an individual applying for a building practitioner license in NSW. These mock questions are presented in Appendix B.

Compared with ChatGPT and Google Search AI Overview, the prototype generated more accurate, context specific, and regulation aligned responses. Table 2 presents three example questions and compares the responses generated by the prototype with those produced by the two general purpose AI tools.

Example Question 1 examined which Certificate IV qualifications are required to apply for registration as a building practitioner. The responses generated by ChatGPT and Google Search AI Overview provided several relevant examples; however, they did not present the complete set of recognized qualifications, resulting in partial and potentially ambiguous guidance. In contrast, the prototype returned the full and formally recognized certification titles aligned with the regulatory requirements. By providing complete and unambiguous qualification information, the prototype reduces information and learning related costs by lowering search effort, cognitive load, and uncertainty associated with interpreting fragmented guidance. It also contributes to reducing substantive compliance costs, as applicants are less likely to pursue incorrect or unnecessary training pathways. Furthermore, clearer upfront guidance helps mitigate time and delay related costs by reducing rework, follow up enquiries, and delays caused by incomplete or incorrect applications.

Example Question 2 addressed the application fee for general building practitioner registration. The responses generated by ChatGPT and Google Search AI Overview were inaccurate, either

 doi:10.20944/preprints202602.0640.v1

reporting outdated figures or conflating fees across different registration categories. In contrast, the prototype provided the correct and current application fee. This outcome highlights the advantage of the prototype in reducing administrative and procedural costs by minimizing corrective interactions with regulatory authorities and avoiding delays associated with fee discrepancies. In addition, improved fee accuracy lowers information and learning related costs by reducing user uncertainty and the need for cross verification across multiple sources.

Example Question 3 asked which building classes a building practitioner may work on when they do not hold a supervisor license. The responses generated by ChatGPT and Google Search AI Overview did not condition the answer on this constraint and instead provided generic scope information, leading to incorrect guidance for the stated scenario. In contrast, the prototype incorporated the license status as a decision relevant attribute and returned scenario specific guidance aligned with the regulatory requirements. This capability reduces substantive compliance costs by clarifying which work is permitted under the user's actual registration status, lowering the risk of undertaking work outside scope. It also reduces non-compliance and enforcement related costs by mitigating exposure to penalties, rectification, or regulatory action arising from incorrect reliance on generic advice. In addition, it reduces information and learning related costs by decreasing uncertainty and the need to cross check multiple sources and can reduce time and delay related costs by avoiding rework and project disruptions caused by scope misunderstandings.

Table 2. Example of responses.

Example Question 1: What Certificate IV qualification do I need to complete to apply for registration as a building practitioner?	
ChatGPT response	"To apply for registration as a building practitioner (General) under the NSW Design and Building Practitioners (DBP) scheme, you typically need to complete one of the following Certificate IV qualifications: CPC40120 Certificate IV in Building and Construction (This is the current nationally endorsed qualification) CPC40110 Certificate IV in Building and Construction (This is the older version and may still be accepted if completed before it was superseded) These qualifications are generally required for roles such as: General Builder (for low-rise or small-scale residential and commercial work) Site Supervisor Construction Manager (depending on scope and experience)"
Google Search AI Overview response	"To apply for registration as a building practitioner (specifically for general building work or a builder's license in NSW), you generally need to complete the CPC40120 Certificate IV in Building and Construction (Building). This qualification covers the essential technical skills, building codes, and business knowledge required for residential and some commercial construction. Key details regarding qualifications: Core Qualification: The CPC40120 Certificate IV in Building and Construction (Building) is the standard requirement. Additional Requirements: In addition to the Certificate IV, you usually need a Diploma of Building and Construction (Building) or specific trade qualifications (like carpentry or bricklaying). Experience: Along with the qualification, you must demonstrate at least two years of relevant, practical, full-time industry experience.

 doi:10.20944/preprints202602.0640.v1

	Alternative Pathway: Older qualifications, such as the Certificate IV in Building (3477) (CPC40110), may be accepted if started before 18 August 2008. It is highly recommended to check with Building Commission NSW for the most up-to-date, specific requirements for your situation.”
Prototype response	“You can meet the requirement through several approved Certificate IV pathways, including: CPC40320 Certificate IV in Building Project Support CPC40110 / CPC40108 / BCG40106 Certificate IV in Building and Construction (Building) CPC40208 / BCG40206 Certificate IV in Building and Construction (Contract Administration) BCG40306 / CPC40308 Certificate IV in Building and Construction (Estimating) BCG40506 / CPC40508 Certificate IV in Building and Construction (Site Management) CPC40120 Certificate IV in Building and Construction”
Example Question 2: What is the application fee for a general building practitioner registration?	
ChatGPT response	“In New South Wales under the Design and Building Practitioners scheme, the cost to apply for a new building practitioner registration (including the general class) varies depending on how long you want the registration to last: 1-year, 3-year or 5-year terms. Here are the application fees for a new building practitioner registration: 1-year registration: $482 3-year registration: $1084 5-year registration: $1808 These fees include a non-refundable processing component.”
Google Search AI Overview response	“Based on current NSW Government records, the application fee for a Building Practitioner registration, specifically for “other classes,” is $180, regardless of whether you hold a previous license or follow Pathway 2. This fee applies to the registration process under the Design and Building Practitioners Act. Key Details: Registration Fee: $180 (non-refundable) for building practitioner - other classes. Additional Costs: If you are adding a category to a building contractor license, the fee is $367, or $109 for a qualified supervisor certificate. Requirements: A MyServiceNSW Account, identification, a National Police Check, and proof of qualifications are required for the application.”
Prototype response	“The fee consists of a general application charge plus a non-refundable processing component. The total payable amount depends on the registration period selected. For a 1-year general building practitioner application, the application fee is $302 plus a $222 non-refundable processing fee, bringing the total to $524. For a 3-year general building practitioner application with a higher category, the application fee is $955 plus a $222 non-refundable processing fee, with a total of $1,177.

 doi:10.20944/preprints202602.0640.v1

	For a 5-year general building practitioner application, the application fee is $1,741 plus a $222 non-refundable processing fee, resulting in a total of $1,963.”
Example question 3: As a building practitioner without a supervisor license, what building classes can I work on?	
ChatGPT response	As a building practitioner without a supervisor license in New South Wales under the Design and Building Practitioners (DBP) scheme, your permitted scope of work depends on your class of registration and whether your role requires supervision of others or overall control of building work. Generally, if you do not hold a supervisor license: You can: Carry out building work under the direction of a registered building practitioner with a supervisor license. Work on class 2 buildings (multi-unit residential) only if registered in a relevant class (e.g. Building Practitioner - General, Specialist, or Design) and only perform the tasks permitted by your class. You cannot: Supervise or certify work. Nominate or act as a principal contractor for regulated building work. Take responsibility for overall coordination or compliance of building work. Examples: You may work as a site foreman or tradesperson involved in execution, but not supervision. You can complete work under a nominated supervisor’s direction, such as: Carpentry or formwork (if separately licensed) Internal fit-out or finishing works General labor on class 2, 3, or 9c buildings, depending on registration”
Google Search AI Overview response	“Without a supervisor license or an endorsed contractor license, a building practitioner in NSW can register to work on Class 2, 3, or 9c buildings (and mixed-use buildings containing them) under the Design and Building Practitioners (DBP) scheme. Registration is required to make compliance declarations, focusing on, but not limited to, remedial or building work. Key Details for Unlicensed/Non-Supervisor Practitioners: Allowed Classes: Class 2 (apartments), Class 3 (residential care/boarding houses), and Class 9c (aged care). Registration Types: You may apply for registration as a building practitioner (general) if you meet the qualification requirements, even without an endorsed supervisor license, to work on these specific classes. Restrictions: Without an endorsed license, you are generally restricted to making compliance declarations for Class 3 and 9c buildings, or specialized work within Class 2. Work Scope: Work includes new buildings, additions, and renovations on these regulated classes. Exemptions: Some minor, non-waterproofing work may be exempt, but structural or critical, regulated work requires registration.

 doi:10.20944/preprints202602.0640.v1

	It is essential to check specific registration requirements for "Building Practitioner - General" (low or medium rise) on the NSW Government website."
Prototype response	"If you do not hold a Contractor License or Qualified Supervisor Certificate (in general building work) under the Home Building Act, your registration will only allow you to work on class 3 and 9c developments and you will not be permitted to do any residential building work."

6. Future Improvement Roadmap

While the proposed prototype demonstrates strong performance in delivering accurate and context aware licensing guidance, several avenues for future improvement have been identified across the knowledge base, reasoning, and user interface layers.

At the knowledge base layer, future development will expand coverage beyond NSW to include all Australian states and territories, enabling nationwide applicability. This expansion will require jurisdiction specific regulatory schemas and cross-state mapping of license categories and terminology. In addition, automated regulatory update mechanisms can be introduced through scheduled crawling of authoritative government sources, version tracking, and change detection, ensuring that licensing rules and procedural requirements remain current without manual intervention. According to Liu, et al. [49], this can be achieved through an automated crawler that systematically identifies combine harvester related information from target webpages, enabling the structured acquisition and organization of web-based knowledge and data. Further improvements will include enhanced multilingual and accessibility support, such as simplified language modes and assistive-technology compatibility, to broaden usability across diverse user groups including non-native English speakers and users with accessibility needs.

For the reasoning layer, future work will focus on improving response accuracy and robustness through tighter integration between structured regulatory knowledge and LLM reasoning. This may be achieved by expanding rule-based validation, improving entity grounding, and incorporating confidence-aware reasoning to reduce hallucinations in edge cases [50]. Response latency can also be reduced through query caching, pre-computed regulatory pathways, and lightweight model routing for common queries. In addition, response naturalness can be improved by refining prompt strategies and output formatting to produce more human-like, conversational, yet regulation-aligned explanations. Reducing LLM token consumption is another important objective and can be achieved through improved chunk selection, hierarchical retrieval, and summarization of regulatory text prior to inference [51].

At the user interface layer, the current web-based interface prioritizes functional clarity but remains intentionally simple. Future development will focus on improving visual design and interaction quality to enhance user engagement and comprehension. For example, Chen and Huang [52] suggested that reducing unnecessary navigation steps, prioritizing visually salient pathways, and streamlining information hierarchy can lower cognitive load and improve user efficiency. Additionally, planned enhancements include support for audio input and file upload, allowing users to submit documents such as qualification certificates or application drafts for contextual guidance. The system can also enable downloadable outputs, such as checklists or step-by-step guidance in document format, to support offline use during application preparation. Furthermore, visual summaries, including automatically generated flowcharts and process diagrams, can be introduced to help users quickly understand complex licensing pathways and decision logic.

7. Conclusions

This study presents the design and development of a graph-based RAG AI system tailored to the Australian construction industry. The research synthesized eight complementary frameworks of regulatory burden and service design to develop a multi-dimensional categorization of ten burden

 doi:10.20944/preprints202602.0640.v1

types in building, trade, and specialist licensing, and demonstrated how these categories can be translated into concrete design requirements for AI-enabled support systems. It further implemented a three-layer graph-based RAG system architecture, grounded exclusively in official government documentation, that was explicitly aligned with these burden categories through its knowledge base, reasoning, and user interface layers. Finally, through prototype evaluation against general-purpose AI tools and expert review, the work illustrated both the potential and current limitations of AI-assisted regulatory guidance and proposed a development roadmap for future RAG-based assistants in construction licensing contexts.

This study acknowledges several limitations. First, the theoretical framework adopted in this study is not exhaustive. While it integrates multiple perspectives on regulatory burden, additional frameworks related to legal interpretation, institutional dynamics, or behavioral economics were not included and may capture further dimensions of regulatory cost. Second, the proposed prototype has not yet been evaluated through large scale user testing due to time and budget constraints. As a result, conclusions regarding real world usability, user adoption, and behavioral impacts on regulatory burden reduction remain preliminary. Nevertheless, the study provides a promising and transferable methodology for the design of theory informed AI systems, which can be extended and empirically validated in future research. Moreover, the prototype relies on a single LLM, DeepSeek, without comparative evaluation against other widely used models, such as GPT or Claude. Future studies could address this limitation by systematically comparing multiple LLMs within the same system architecture to assess their relative strengths, limitations, and suitability for regulatory and licensing support tasks. Last but not least, technical constraints associated with AI systems should be recognized. Despite the use of retrieval and graph augmented reasoning to reduce hallucination risks, AI generated outputs may still contain inaccuracies. Accordingly, the system is intended to support, rather than replace, human judgment, and users should rely on official sources and professional advice when making regulatory decisions.

Author Contributions: Conceptualization, D.Y.; methodology, D.Y. and Z.Y.; software, J.L. and B.H.; validation, D.Y. and Z.Y.; formal analysis, D.Y., J.L. and B.H.; data curation, D.Y.; writing—original draft preparation, D.Y. and J.L.; writing—review and editing, R.Y.S. and C.C.W.; visualization, D.Y.; supervision, Z.Y., R.Y.S. and C.C.W.; project administration, J.H. and J. X.. All authors have read and agreed to the published version of the manuscript.

Institutional Review: Not applicable.

Informed Consent: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy considerations.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUD	Australian Dollars
GDP	Gross Domestic Product
LLM	Large Language Model
NSW	New South Wales
PDF	Portable Document Format
RAG	Retrieval-Augmented Generation

doi:10.20944/preprints202602.0640.v1

Appendix A

Table A1. License-oriented LLM-aware recursive chunking algorithm.

Require: PDF documents P; length bounds (L_{min}, L_{max})	
Ensure: Chunk set C with provenance metadata	
M ← ∅	▷Markdown documents
for all p ∈ P do	
m ← DEEPSEEOCR2 TO MARKDOWN(p)	▷layout preserving OCR [44]
M ← M ∪ {m}	
end for	
C ← ∅	
for all m ∈ M do	
T ← BUILDHEADINGTREE(m)	
C0 ← SPLIT(ROOT(T))	
C ← C ∪ REFINE(C0, L_{min}, L_{max})	
end for	
return C	

Appendix B

Table A2. Mock questions.

1. Under the NSW Design and Building Practitioners scheme, am I required to register as a building practitioner if I only prepare regulated designs but do not carry out construction work?
2. I am working in a Class 2 building in NSW but the project started before July 2021. Do I still need to register under the DBP scheme and which transitional provisions apply?
3. Does the requirement to register as a building practitioner apply differently to individuals and body corporate nominees under NSW Fair Trading rules?
4. What Certificate IV qualification do I need to complete to apply for registration as a building practitioner?
5. I hold a Certificate IV in Building and Construction but do not have two years of recent experience. Can I apply for building practitioner general registration and what alternative evidence is accepted?
6. If I am undischarged bankrupt, under what conditions can NSW Fair Trading still consider my application for building practitioner registration?
7. I previously held a building license in another Australian state. What additional requirements apply for mutual recognition in NSW under the current framework?
8. Can a building practitioner general take responsibility for regulated designs for Class 3 buildings, or is a separate design practitioner registration required?
9. If my registration is limited to certain building classes, how does this affect my ability to sign compliance declarations?
10. What are the exact steps to apply for building practitioner registration in NSW, including required documents, fees, identity checks, and expected processing time?
11. At what stage of the application process is the National Police Certificate assessed, and what validity period does NSW Fair Trading require?
12. What is the application fee for a general building practitioner registration?
13. If my application is refused, what review or appeal options are available and within what timeframes?
14. Based on NSW Fair Trading guidance, the DBP Act, and supporting regulations, what are all the mandatory obligations a registered building practitioner must comply with after approval?

 doi:10.20944/preprints202602.0640.v1

15. Which obligations apply only to regulated buildings, and which apply to all building work regardless of class?
16. How do the professional indemnity insurance requirements interact with registration obligations for building practitioners?
17. If I subcontract with regulated building work but do not directly supervise site activities, do I still need to be registered as a building practitioner?
18. As a building practitioner without a supervisor license, what building classes can I work on?
19. Does temporary suspension of registration affect my ability to continue work already commenced on a regulated building?
20. Are there exceptions for practitioners working exclusively on government owned projects or public infrastructure?
21. What documentary evidence must be uploaded to demonstrate competence for building practitioner registration, and which documents are considered insufficient by NSW Fair Trading?
22. Where in the official NSW guidance is the requirement for continuing professional development stated, and what are the minimum expectations?
23. Which sections of the DBP Act explicitly define the responsibilities of a building practitioner general?
24. My building practitioner application was submitted before a recent regulatory update but is still under assessment. Which version of the eligibility criteria applies to my application?
25. If I completed my mandatory training modules before they were updated, do I need to retake them to remain eligible for registration?
26. How do transitional arrangements apply if a regulated building project spans both before and after a change to the DBP Regulation?
27. Which circumstances automatically disqualify an applicant from building practitioner registration, even if all qualifications are met?
28. Can NSW Fair Trading refuse an application solely based on disciplinary history in another profession or jurisdiction?
29. If false or misleading information is provided unintentionally, how does this affect eligibility and future applications?
30. If guidance on the NSW Fair Trading website conflicts with the DBP Regulation, which document takes precedence and why?
31. How should a practitioner interpret differences between fact sheets and the wording of the DBP Act when determining compliance obligations?
32. Are advisory notes legally binding, and how should they be treated compared with legislation?
33. If I am registered as both a building practitioner and a design practitioner, which obligations apply when I perform both roles on the same project?
34. Can one individual act as a building practitioner and body corporate nominee at the same time, and what additional responsibilities arise?
35. How are compliance declarations managed when multiple registered practitioners contribute to the same regulated building?
36. After registration, what ongoing record keeping obligations apply to building practitioners, and how long must records be retained?
37. What events trigger mandatory notification to NSW Fair Trading after registration, and what are the consequences of failing to notify?
38. Under what conditions can NSW Fair Trading suspend or cancel registration, and what procedural safeguards apply?
39. I changed employers midway through a regulated building project. What registration and notification steps are required to remain compliant?

 doi:10.20944/preprints202602.0640.v1

40. If a regulated design is revised after construction has started, who is responsible for issuing updated compliance declarations?
41. What happens if a building practitioner relies on a regulated design prepared by an unregistered design practitioner?
42. From an assessor perspective, what are the most common reasons building practitioner applications are delayed or refused?
43. Which types of experience evidence are frequently considered insufficient, even when applicants believe they meet the criteria?
44. How can applicants proactively reduce assessment delays based on NSW Fair Trading guidance?
45. Which parts of the building practitioner registration process contribute most to administrative delay, and how can they be avoided?
46. What steps in the application process most commonly require re-submission, and why?
47. How can clearer understanding of scope of work reduce the risk of non-compliance penalties?
48. Under the NSW Design and Building Practitioners framework, can a building practitioner rely on experience gained before the scheme commenced, and how must this experience be evidenced in the application?
49. If a building practitioner registration lapses due to non-renewal, what restrictions apply to work undertaken during the lapse period, and what steps are required for reinstatement?
50. When multiple compliance declarations are required for the same regulated building, how does NSW Fair Trading determine accountability if a defect is later identified?

References

1. Jobs and Skills Australia. Construction. Available online: https://www.jobsandskills.gov.au/data/occupation-and-industry-profiles/industries/construction (accessed on 28 June 2025).
2. Australian Bureau of Statistics. Construction Work Done. Available online: https://www.abs.gov.au/statistics/industry/building-and-construction/construction-work-done-australia-preliminary/latest-release (accessed on 30 January 2026).
3. Australian Bureau of Statistics. Labour Force. Available online: https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia-detailed/latest-release (accessed on 30 January 2026).
4. NSW Government. General Building Work. Available online: https://www.nsw.gov.au/business-and-economy/licences-and-credentials/building-and-trade-licences-and-registrations/general-building-work (accessed on 30 January 2026).
5. Building Commission NSW. Assessment Process. Available online: https://www.nsw.gov.au/business-and-economy/licences-and-credentials/building-and-trade-licences-and-registrations/assessment-process-and-time-frames#processing-times (accessed on 30 January 2026).
6. NSW Government. Licensing Program. Available online: https://www.digital.nsw.gov.au/case-studies/licensing-program (accessed on 30 January 2026).
7. NSW Government. Nsw State Budget 2024-25. Available online: https://www.smallbusiness.nsw.gov.au/news-podcasts/news/nsw-state-budget-2024-25 (accessed on 30 January 2026).
8. Building Commission NSW. Reasons Why an Application May Be Refused. Available online: https://www.nsw.gov.au/business-and-economy/licences-and-credentials/building-and-trade-licences-and-registrations/assessment-process-and-time-frames/reasons-why-an-application-may-be-refused (accessed on 30 January 2026).
9. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; Hu, X. Harnessing the Power of Llms in Practice: A Survey on Chatgpt and Beyond. *ACM Transactions on Knowledge Discovery from Data* **2024**, *18*, 1-32, doi:10.48550/arXiv.2304.13712.

10. Cheong, I.; Xia, K.; Feng, K.K.; Chen, Q.Z.; Zhang, A.X. (a) I Am Not a Lawyer, But...: Engaging Legal Experts Towards Responsible Llm Policies for Legal Advice. In Proceedings of the Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024; pp. 2454-2469.
11. Shittas, G.; Kapitsaki, G.M.; Papoutsoglou, M. Licy: A Chatbot Assistant to Better Understand and Select Open Source Software Licenses. **2025**, doi:10.5220/0013336400003928.
12. Synott, R.; Aksenova, G. Business Model Innovation of Architecture, Engineering and Construction Consultancies in the Ai Era. In Proceedings of the Proceedings of the Institution of Civil Engineers-Civil Engineering, 2025; pp. 1-9.
13. Aboelazm, K.S.; Dganni, K.M. Public Procurement Contracts Futurity: Using of Artificial Intelligence in a Tender Process. *Corporate Law & Governance Review* **2025**, *7*, 60-72, doi:10.22495/clgrv7i1p6.
14. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T. Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks. *Advances in neural information processing systems* **2020**, *33*, 9459-9474.
15. Lai, L.; Luo, C.; Lou, Y.; Ju, M.; Yang, Z. Graphy'our Data: Towards End-to-End Modeling, Exploring and Generating Report from Raw Data. In Proceedings of the Companion of the 2025 International Conference on Management of Data, 2025; pp. 147-150.
16. Tang, X.; Chen, L.; Yang, W.; Yang, Z.; Ju, M.; Shu, X.; Yang, Z.; Tang, Y. Tabular-Textual Question Answering: From Parallel Program Generation to Large Language Models. *World Wide Web* **2025**, *28*, 42.
17. Liu, J.; Chen, Z.; Qiao, S.; Ju, M.; Zhang, D.; Han, B.; Yu, S.; Shu, X.; Wu, J.; Wen, D. A2rag: Adaptive Agentic Graph Retrieval for Cost-Aware and Reliable Reasoning. *arXiv preprint arXiv:2601.21162* **2026**.
18. Levie, J.; Autio, E. Regulatory Burden, Rule of Law, and Entry of Strategic Entrepreneurs: An International Panel Study. *Journal of Management Studies* **2011**, *48*, 1392-1419, doi:10.1111/j.1467-6486.2010.01006.x.
19. Austarlian Government. Regulatory Burden Measurement Framework. Available online: https://oia.pmc.gov.au/resources/guidance-assessing-impacts/regulatory-burden-measurement-framework (accessed on 30 Janurary 2026).
20. Torriti, J. Standard Cost Model: Three Different Paths and Their Common Problems. *Journal of Contemporary European Research* **2012**, *8*, 90-108, doi:10.30950/jcer.v8i1.322.
21. Kirkpatrick, C.; Parker, D. Regulatory Impact Assessment and Regulatory Governance in Developing Countries. *Public Administration and Development: The International Journal of Management Research and Practice* **2004**, *24*, 333-344, doi:10.1002/pad.310.
22. World Bank. Doing Business 2015: Going Beyond Efficiency: Comparing Business Regulations for Domestic Firms in 189 Economies: A World Bank Group Flagship Report; World Bank Publications: 2014.
23. Tarshish, N. The Administrative Burden Framework: A New Horizon for Research and Practice in Social Work? *Social Work* **2024**, *69*, 86-94, doi:10.1093/sw/swad037.
24. Sunstein, C.R. Sludge Audits. *Behavioural Public Policy* **2022**, *6*, 654-673, doi:10.1017/bpp.2019.32.
25. Furrer, O.; Sudharshan, D.; Tsiotsou, R.H.; Liu, B.S. A Framework for Innovative Service Design. *The Service Industries Journal* **2016**, *36*, 452-471, doi:10.1080/02642069.2016.1248420.
26. Marcel, B.; Ortan, T.; Otgon, C. Information Asymmetry Theory in Corporate Governance Systems. *Annals of the University of Oradea, Economic Science Series* **2010**, *19*, 516-522.
27. Baekgaard, M.; Tankink, T. Administrative Burden: Untangling a Bowl of Conceptual Spaghetti. *Perspectives on Public Management and Governance* **2022**, *5*, 16-21, doi:10.1093/ppmgov/gvab027.
28. Chiang, L.; Liao, C.n. The Influence of Digital Standardization on Administrative Efficiency in E-Government: A View of Standards Development Organizations. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research* **2009**, *26*, 455-468, doi:10.1002/sres.946.
29. MohammadRezaei, F.; Faraji, O.; Rezaee, Z.; Gholami-Jamkarani, R.; Yari, M. Substantive or Symbolic Compliance with Regulation, Audit Fees and Audit Quality. *International Journal of Disclosure and Governance* **2024**, *21*, 32-51, doi:10.1057/s41310-023-00178-4.
30. Nandan Prasad, A. Regulatory Compliance and Risk Management. In Introduction to Data Governance for Machine Learning Systems: Fundamental Principles, Critical Practices, and Future Trends; Springer: 2024; pp. 485-624.

 doi:10.20944/preprints202602.0640.v1

31. Marneffe, W.; Vereeck, L. The Meaning of Regulatory Costs. *European Journal of Law and Economics* **2011**, *32*, 341-356.
32. Joshi, S.; Krishnan, R.; Lave, L. Estimating the Hidden Costs of Environmental Regulation. *The Accounting Review* **2001**, *76*, 171-198.
33. Love, P.E.; Davis, P.R.; Worrall, D. Occupational Licensing of Building Trades: Case of Western Australia. *Journal of Professional Issues in Engineering Education and Practice* **2010**, *136*, 215-223.
34. Adepoju, A.H.; Austin-Gabriel, B.; Eweje, A.; Collins, A. Framework for Automating Multi-Team Workflows to Maximize Operational Efficiency and Minimize Redundant Data Handling. *IRE Journals* **2022**, *5*, 663-664.
35. Gray, G.C.; Silbey, S.S. Governing inside the Organization: Interpreting Regulation and Compliance. *American Journal of Sociology* **2014**, *120*, 96-145.
36. Gandhi, T.K.; Classen, D.; Sinsky, C.A.; Rhew, D.C.; Vande Garde, N.; Roberts, A.; Federico, F. How Can Artificial Intelligence Decrease Cognitive and Work Burden for Front Line Practitioners? *JAMIA open* **2023**, *6*, ooad079.
37. Bennett, J.; Estrin, S. Regulatory Barriers and Entry into a New Competitive Industry. *Review of Development Economics* **2013**, *17*, 685-698.
38. Dickerson, C.E.; Wilkinson, M.; Hunsicker, E.; Ji, S.; Li, M.; Bernard, Y.; Bleakley, G.; Denno, P. Architecture Definition in Complex System Design Using Model Theory. *IEEE Systems Journal* **2020**, *15*, 1847-1860, doi:10.48550/arXiv.1909.06809.
39. Roozbahani, Z. A Review of Methods for Reducing Hallucinations in Generative Artificial Intelligence to Enhance Knowledge Economy. *Knowledge Economy Studies* **2025**.
40. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* **2023**, *2*.
41. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From Local to Global: A Graph Rag Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130* **2024**.
42. Zhu, X.; Xie, Y.; Liu, Y.; Li, Y.; Hu, W. Knowledge Graph-Guided Retrieval Augmented Generation. *arXiv preprint arXiv:2502.06864* **2025**.
43. Sukhoroslov, O.; Volkov, S.; Afanasiev, A. A Web-Based Platform for Publication and Distributed Execution of Computing Applications. In Proceedings of the 2015 14th International Symposium on Parallel and Distributed Computing, 2015; pp. 175-184.
44. Wei, H.; Sun, Y.; Li, Y. Deepseek-Ocr 2: Visual Causal Flow. *arXiv preprint arXiv:2601.20552* **2026**.
45. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.S.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.-t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the EMNLP (1), 2020; pp. 6769-6781.
46. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume, 2021; pp. 874-880.
47. Haveliwala, T.H. Topic-Sensitive Pagerank. In Proceedings of the Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA, 2002; pp. 517–526.
48. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The Pagerank Citation Ranking: Bringing Order to the Web*; Stanford infolab: 1999.
49. Liu, H.; Zhang, Y.; Xie, Y.; Zhao, Y. Automatic Updating of a Combine Harvester Knowledge-Based System by Webpages and User-Uploaded Files. *International Journal of Agricultural and Biological Engineering* **2025**, *18*, 165-175.
50. Xia, S.-J.; Zhang, H.; Jiang, Y.; Chen, X.; Chen, Y.; Li, Z.; Wan, Z.; Sun, A. Rethinking Hallucinations: A Cognitive-Inspired Taxonomy and Comprehensive Survey in Large Language Models, Large Vision-Language Models, and Multimodal Large Language Models. *Large Vision-Language Models, and Multimodal Large Language Models (November 09, 2025)* **2025**.
51. Song, D.; Wang, W.; Chen, S.; Wang, X.; Guan, M.X.; Wang, B. Less Is More: A Simple yet Effective Token Reduction Method for Efficient Multi-Modal Llms. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025; pp. 7614-7623.

 doi:10.20944/preprints202602.0640.v1

52. Chen, C.; Huang, K. Fewer Clicks, Lower Emissions: Eye-Tracking Analysis of Eco-Friendly Navigation in Tourism Websites. *Sustainability* **2025**, *17*, 5462.