LEXFACTUM.AI

White Paper

2024 v.1.00 (draft)

Oleksii Konashevych, PhD[1]
Sergiy Chernyshov[2]

Abstract: Lexfactum.AI, an Australian technology startup, aims to revolutionise legal work by harnessing the power of generative artificial intelligence (GenAI) and retrieval-augmented generation (RAG). We have developed an AI-chat bot legal assistant that operates on a comprehensive collection of laws. In this paper, we present the initial version of our application, highlighting its distinct features and capabilities. We also outline our strategic roadmap for future development and commercialisation, ensuring the continuous enhancement and expansion of our innovative legal tools.

Table of contents

I. Introduction

LEXFACTUM.AI[3] introduces the next-generation technology for legal research and inference by leveraging artificial intelligence (AI) and retrieval-augmented generation (RAG) over a database of laws

[1] PhD in Law, Science and Technology, CEO and Co-founder of LEXFACTUM.AI.
[2] Software developer and cybersecurity expert, CTO and Co-founder of LEXFACTUM.AI.
[3] LEXFACTUM.AI is a registered business name in Australia.

and user's documents. In this paper, we present the initial capabilities of the application and provide an outline of our plans for future development and commercialisation.

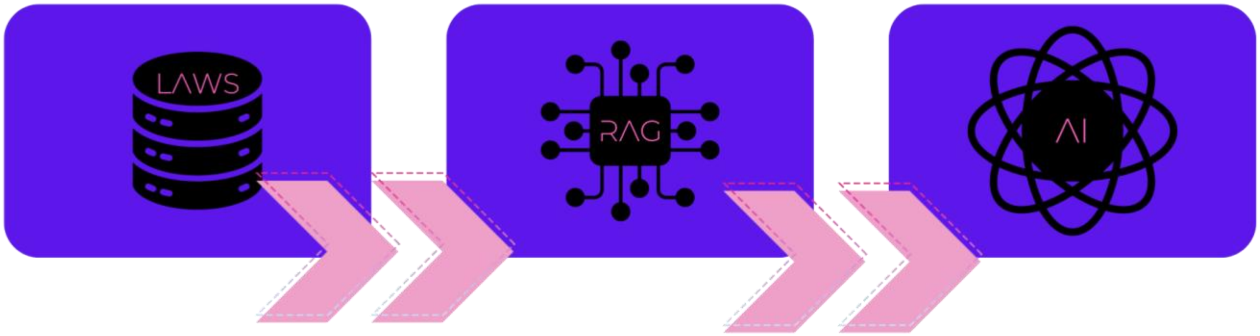
The advancement of large-language models, encompassing both proprietary and open-source solutions, has opened up a realm of possibilities to revolutionise legal work and research. There are reports that some lawyer's use successfully use OpenAI's ChatGPT to aid their work, some of them failed due to "hallucination," [1] which essentially means GenAI can make up facts. Whilst professionals have the ability to discern mistakes of GenAI, it is our understanding that the use of AI by non-legal professionals is a matter of concern. We do not believe that AI is there yet to be able to resolve legal disputes without supervision. At the same time it is clear that mere use of AI chatbots is not enough even for law specialists. Our research and development efforts have been dedicated to developing a technology that empowers AI to perform analysis and reasoning while strictly adhering to the corpus of laws. This approach aims to prevent AI hallucinations and ensure that AI-driven inferences are supported by legal citations.

The following chapter discusses why retrieval-augmented generation (RAG) techniques appeared to be not that easy to apply in enterprise grade solutions, what we designed and how it performs. The third part elaborates on LEXFACTUM.AI's plan of commercialisation, business development and further roadmap.

II. Our technology

2.1 RAG issues and the solution

The development of retrieval-augmented generation (RAG) in the world began in 2020 [2]. According to Google, RAG is "an AI framework that merges the strengths of traditional information retrieval systems (such as databases) with the capabilities of generative large language models (LLMs)" [3].

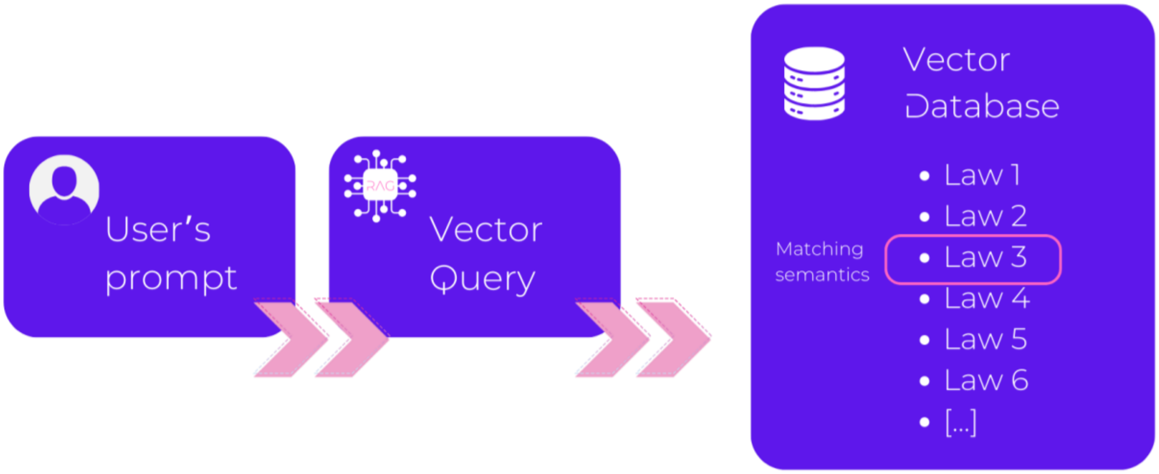


In simpler terms, RAG enables AI to work efficiently with external knowledge that might not be present within the AI model itself. This is necessary because AI models are not only challenging to retrain just to incorporate new information, but they also have a limited "context" window, which refers to the amount of information they can process at once, measured in "tokens." This is similar to a computer's RAM or the human brain in terms of instant information capacity. RAG effectively extends this capacity, allowing AI to process large datasets, such as a collection of laws.

RAG can be implemented on consumer-grade computers using various free applications like 'NVIDIA Chat with RTX' [4] or 'GPT4ALL' [5]. To operate efficiently, it requires a relatively powerful GPU, at least an RTX series with a recommended 8GB of VRAM. However, there are two main limitations to its use:

1. **Limited Capabilities of Smaller LLMs:** For example, LLAMA 3.1 8B can run on a consumer-grade laptop but is insufficient for more complex tasks, such as legal work.
2. **Time-Consuming Vectorisation Process:** Vectorising a large corpus of laws from any jurisdiction is a time-intensive task. For instance, when tested on a gaming laptop, this process took several weeks. Vectorisation involves preparing user data in a format that AI can understand, using an embedding model like BERT [6] to convert the data into digital vectors representing semantic meanings. This preprocessing step creates a vector database that reflects all input data.

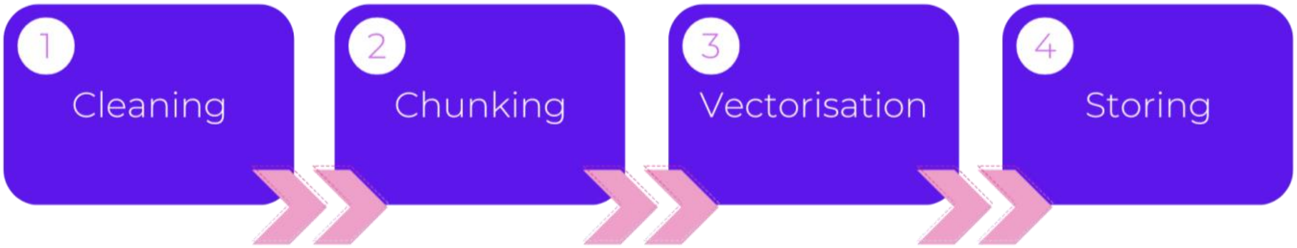
In terms of application, when a user submits a query, the application transforms the query into a vector (a digital representation of the query's semantics) and searches the vector database for matches. The matching data is then provided to the AI for processing.



However, vectorising over 220,000 Australian laws presents a significant challenge both in terms of hardware and software. Vectorisation of 10GB of Australian laws is a computationally intensive task, even for an enterprise grade workstation with powerful GPU. However, the vector database alone is not sufficient for a complete user experience in such a chatbot application. It also requires a traditional database management system (DBMS) like MySQL or MongoDB, integrated with the vector database. Unfortunately, popular online resources do not provide detailed guidance on these complexities or offer viable architectural solutions. Moreover, the libraries and drivers face compatibility and stability challenges. Installing and adapting them to a specific hardware configuration requires deep expertise and time for debugging.

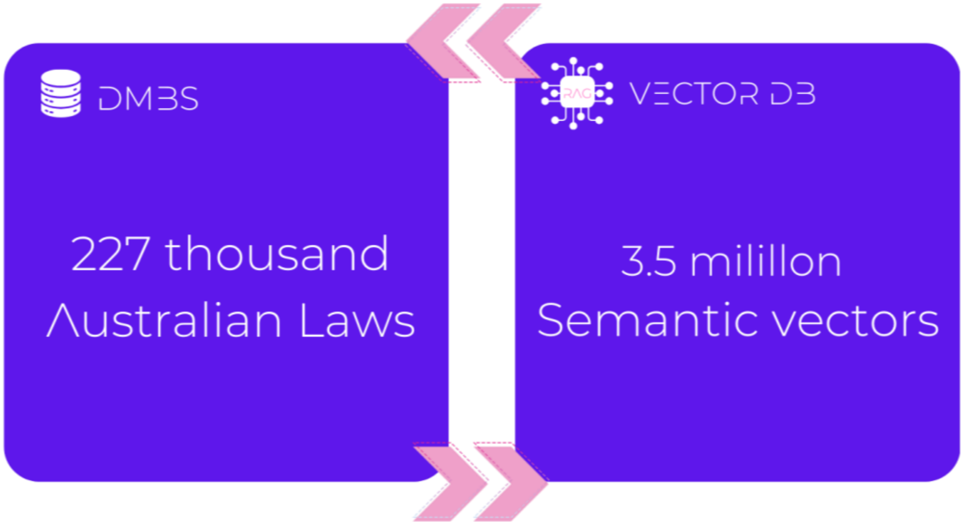
Despite these challenges, our research and development efforts have led to significant improvements in the performance of database pre-processing (vectorisation) and subsequent vector search queries, as well as related queries in our DBMS.

Database preparation consists of several essential parts (for the purpose of this paper, we skip some technical details):



1. **Cleaning:** We clean the text by removing capital letters, punctuation, and other noise elements to ensure the AI embedding model receives clean input.
2. **Chunking:** We split all laws into chunks, typically one paragraph in length but not exceeding 500 tokens (about 2,000-2,500 symbols). To ensure continuity, we overlap the text from each chunk with the next one. This process resulted in almost 3.5 million chunks.
3. **Vectorisation:** We then use an AI embedding model capable of extracting semantic meaning from these chunks to perform vectorisation. Not all embedding models can effectively handle vectorisation of paragraphs (as a semantic unit). The resulting data is stored in a vector database designed to handle the volume of vectors.
4. **Storing:** In parallel, we create an auxiliary database using a conventional DBMS to support the user application's functionality.

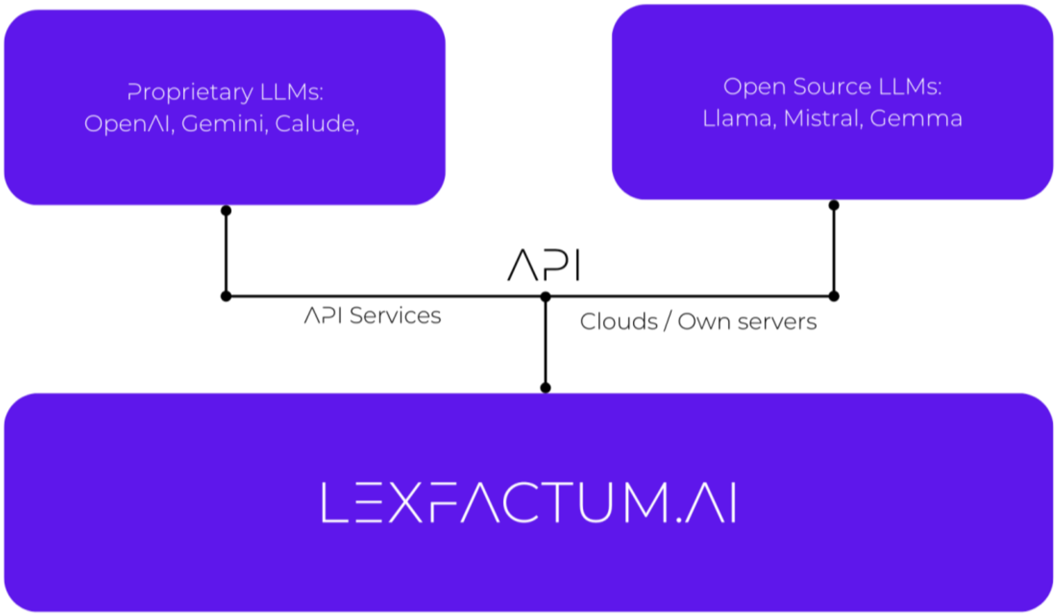
A well-calibrated system with two 56-core CPUs and an NVIDIA A5000 GPU can now prepare the 10GB dataset in around 6 hours, compared to the initial 8 days. Preparing one chunk takes approximately 70 milliseconds.



Once the databases were prepared, we developed the back-end and front-end of the application, implementing the features required for legal work.

2.2 Large Language Models and challenges

The crucial component of the application is a Large Language Model (LLM). The generative AI capabilities must be both robust enough to handle the complexity of legal work and efficient, as our technology is designed to operate *fully locally and offline*. In Section 2.4, we discuss the development of two editions of the application: one is an online web application intended for mass use, while the other is an enterprise solution—a premium technology designed to meet the high standards of privacy and data protection required by our most demanding customers.

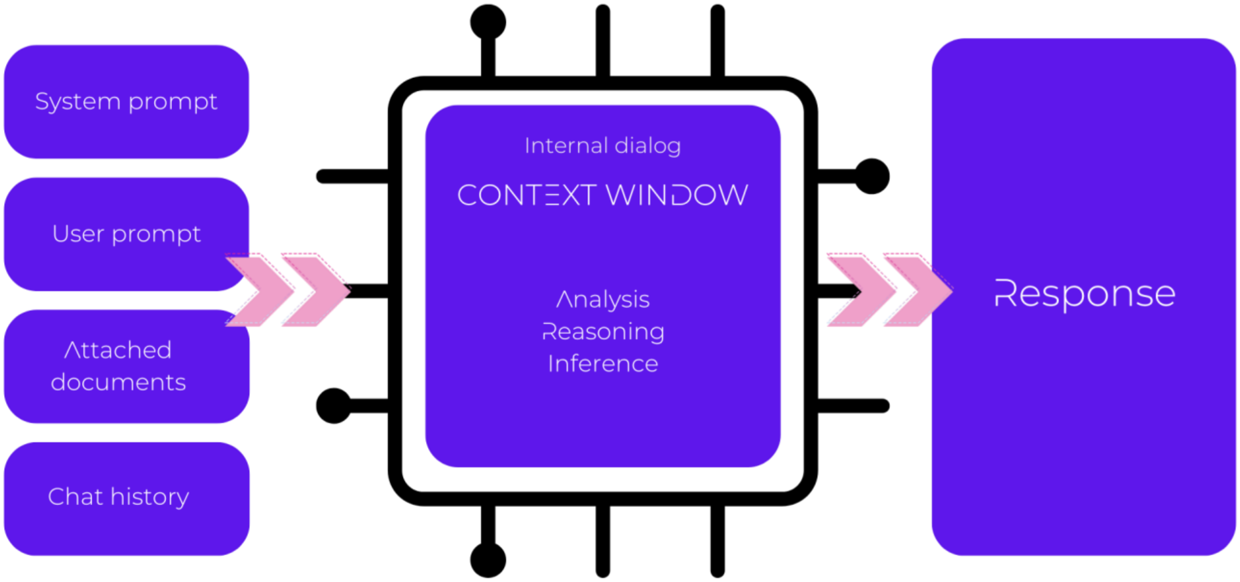


OpenAI currently leads the market, with its proprietary model outperforming both other proprietary models and open-source alternatives in many benchmarks. However, this advantage has recently shifted due to advancements by Claude, Google's Gemini, and others, particularly in various performance aspects. For instance, Gemini excels with its large context window of 1 million tokens, which is crucial for extensive legal analyses. Meanwhile, the highly efficient Groq proves useful in multiple interim processes within our AI reasoning flow. Additionally, recent developments in AI reasoning methods show that even open-source models can surpass their proprietary counterparts. A popular approach involves using a three-step inference process with different models, followed by a final summarisation using one of the models (see Fig. N1).

Each model has several critical parameters, including the size of the dataset and the size of the context window. The effective use of these parameters defines the sophistication of the application. For instance, the Meta's Llama 3.1 8B model (where 8B refers to the size of the training dataset) is fast and intelligent enough to perform internal utility tasks like prompt rephrasing or query routing. The more advanced Llama 3.1 70B model is capable of more complex tasks, such as legal analysis and reasoning.

The context window size, or the maximum number of tokens the model can process at once, is significant for the user experience and application quality. Legal documents can be very extensive; for example, the Corporations Act 2001 (Cth) represents the volume of over 200,000 tokens. The Llama model, with an 8,000-token context window, cannot process such a large document all at once. Therefore, RAG serves both as the application engine for semantic search across all Australian laws and as the tool for browsing individual documents.

The context window is a precious resource because each prompt must include not only the user's question, system instructions, and relevant legal text but also the history of the current chat session. The context window effectively acts as the model's memory. The context window has only 25-30% of its volume practically useful for operation. It means that for better answers, for example, with Llama, the mentioned input should fit within the 2000-2500 token limit. The issue is that the model needs 'space' for reasoning. The model unpacks the data and analyses it, conducting an 'internal dialogue.' If the input fills more than 25% of that space, the quality of the model's answers decreases. Our empirical research revealed that it dramatically drops when the input constitutes over 80% of the context window. Consequently, our preferences gravitate towards models equipped with more extensive context windows, as they guarantee the caliber of reasoning, albeit being more computationally demanding for hardware.



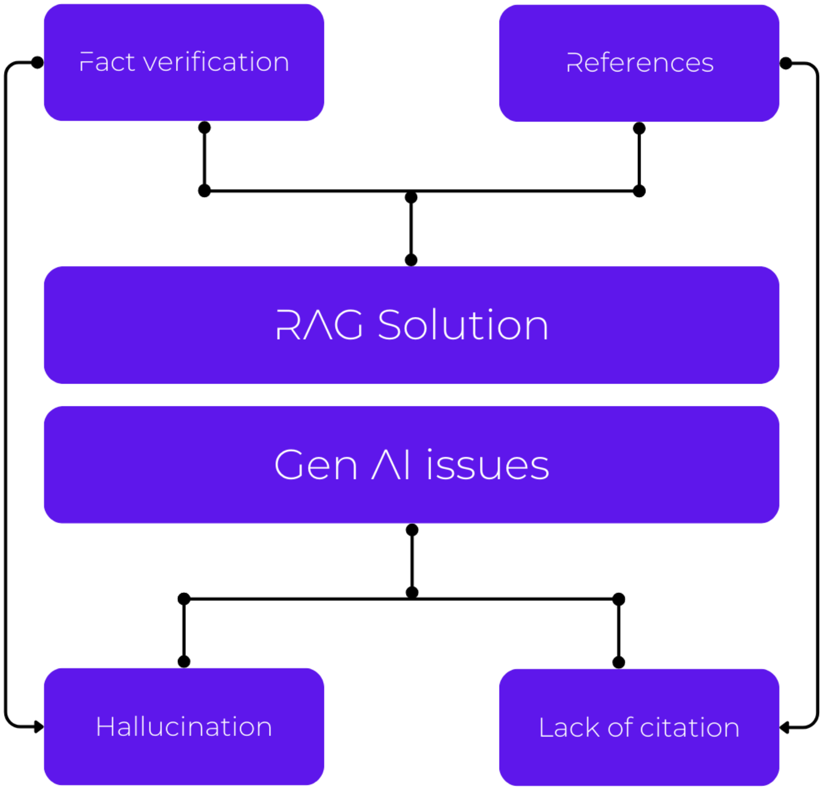
There are various techniques to improve the quality of the results [6]. We have developed a comprehensive back-end solution that implements best practices in RAG and AI design as well as our own findings. Key components of our system perform an extensive work under the hood. After the submission of a user prompt, the system performs internal rephrasing, where the model reformulates user queries into legal language; drafts an answer; identifies relevant laws; analyses them; and finally provides the most comprehensive legal response (for further details see the next section on UIX).

LEXFACTUM.AI's solution is compatible with both proprietary online and open-source offline models, enabling asynchronous queries across any combination of these models. Powered by API, this fully AI-agnostic design allows us to adapt to rapid changes in the AI landscape.

2.3 Main features

Unique features

Why is LEXFACTUM.AI a standout assistant for legal work? While AI chatbots like ChatGPT and Gemini can answer law questions effectively, they often falter by providing 'hallucinations', i.e., inaccurate information and failing to cite their sources, such as specific statutes or court decisions. LEXFACTUM.AI is designed to tackle these issues. Our RAG-empowered application integrates a vectorised database of laws, enabling a semantic, AI-driven search. This ensures that responses are based on accurate, real-world legal texts. Moreover, instead of offering conclusions without explanation, LEXFACTUM.AI provides detailed references to specific statutes and court rulings, down to the paragraph, giving users complete transparency and confidence in the information provided.



Let's explore the user experience in LEXFACTUM.AI. This web application is designed with a familiar layout: a main window for chat and prompts, a left panel for chat history, and an additional right panel for fine-tuning queries.

The right panel is where LEXFACTUM.AI excels, offering specialised features for legal research. We've developed several main types of legal work modes:

1. Problem Solving: Users can describe real-world situations to receive legal advice, either by typing their questions directly or by attaching relevant documents. GenAI offers a range of additional features that are known for example among ChatGPT’s users. These exercises include summarisation, rephrasing, shortening, elaboration and reasoning. All user prompts are cross-referenced against the database of laws and displayed in a separate window, along with relevant text snippets (citations).

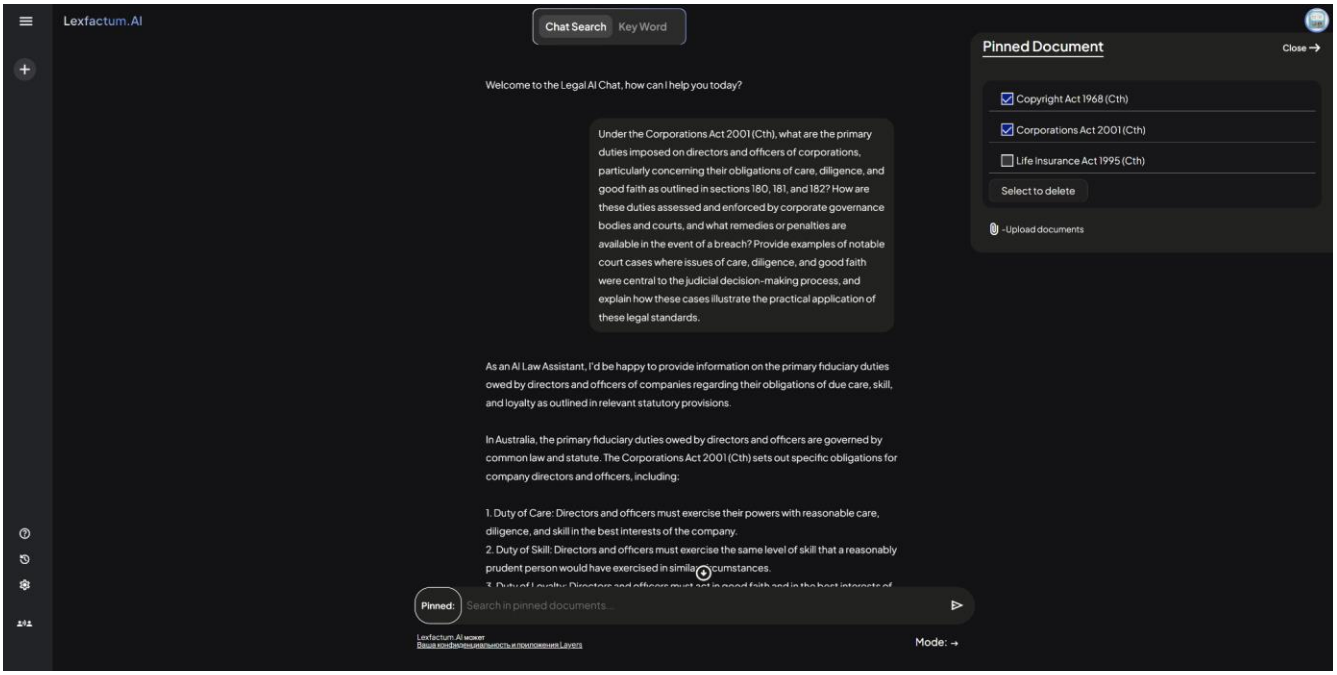
2. Legal Research: This feature allows users to search a database of Australian laws using a vector search, also known as semantic AI search. For example, if you type “a dog crosses the street,” the app creates a semantic vector representation of the phrase and searches for laws with similar meanings. This approach is more effective than keyword searches, which might only find laws mentioning specific words like 'dog' or 'street.' Our AI-assisted search can find relevant traffic rules or animal control laws, providing a more accurate and relevant set of results.

3. Keyword Search: A more traditional keyword search allows for precise legal research when the user seeks specific words (names, dates, facts, etc.).

4. Filters: The right panel allows users to narrow down their legal research to specific fields of law and jurisdictions. The filter options include: Federal level, states and territories, statute law and case law. Statute law can be further refined into primary legislation (legislative acts), secondary laws (regulations, by-laws), and bills (draft laws). Users can combine filters to fine-tune their results.

5. Pinned Collections: The application generates a list of legal acts based on AI and direct keyword searches. Users can pin any act to create custom collections. Once a collection is formed, users can switch to legal research within this collection or even focus on individual acts.

6. User's Documents: Users can upload their documents and pin them in the collection, allowing them to combine legal research over both laws and legal documents, such as agreements, policies, and memoranda.


Lexfactum.AI
Chat Search
Key Word
Pinned Document
Close →
☑ Copyright Act 1968 (Cth)
☑ Corporations Act 2001 (Cth)
☐ Life Insurance Act 1995 (Cth)
Select to delete
-Upload documents
Pinned:
Search in pinned documents...
Mode: →
Lexfactum.AI может
Ваша конфиденциальность и приложения Layers

Welcome to the Legal AI Chat, how can I help you today?

Under the Corporations Act 2001 (Cth), what are the primary duties imposed on directors and officers of corporations, particularly concerning their obligations of care, diligence, and good faith as outlined in sections 180, 181, and 182? How are these duties assessed and enforced by corporate governance bodies and courts, and what remedies or penalties are available in the event of a breach? Provide examples of notable court cases where issues of care, diligence, and good faith were central to the judicial decision-making process, and explain how these cases illustrate the practical application of these legal standards.

As an AI Law Assistant, I'd be happy to provide information on the primary fiduciary duties owed by directors and officers of companies regarding their obligations of due care, skill, and loyalty as outlined in relevant statutory provisions.

In Australia, the primary fiduciary duties owed by directors and officers are governed by common law and statute. The Corporations Act 2001 (Cth) sets out specific obligations for company directors and officers, including:

1. Duty of Care: Directors and officers must exercise their powers with reasonable care, diligence, and skill in the best interests of the company.
2. Duty of Skill: Directors and officers must exercise the same level of skill that a reasonably prudent person would have exercised in similar circumstances.
3. Duty of Loyalty: Directors and officers must act in good faith and in the best interests of

2.4 Online and Enterprise editions

LEXFACTUM.AI developed two editions of the software.

Online

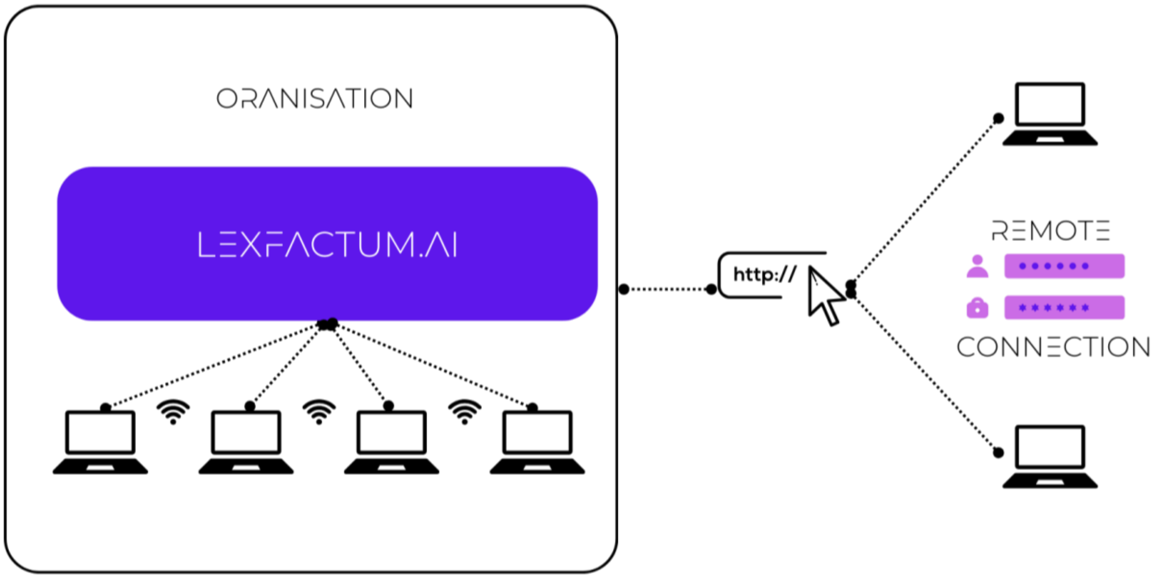
Our online edition operates as a web application hosted on the cloud, providing services to a broad audience through a paid subscription model. It mirrors the functionality of the Enterprise edition and is compatible with both open-source and proprietary large language models (LLMs), or a combination of both. The AI-agnostic backend allows it to work with any LLMs via their APIs, integrating them at various stages of reasoning and inference to enhance the final output. This includes models like OpenAI's Chat, Google's Gemini, Anthropic's Claude and other proprietary models and AI-as-a-service systems such as xAI's Groq.

Currently, the online version is available for demonstration in a testing mode and is prepared for a full rollout once the necessary infrastructure is in place.

Enterprise edition

The enterprise edition is a software product designed to work on premises, within a local network. The database of Australian laws and the backend that powers the application, including the AI, function locally and offline. This ensures the highest levels of privacy and data protection, as conversations with the AI do not get transmitted to external service providers such as OpenAI or Gemini. This level of privacy is made possible by leveraging advancements in open-source LLMs that match and even surpass the performance of commercial ones.

Users can access the application through their browser as a web application, but it is hosted on their enterprise's local network (intranet) rather than the Internet. Optionally, we can configure secure remote access for users working from home.



The Enterprise app is multi-user, allowing for fine-tuned access rights and authorisations. It isolates different databases and manages them so staff operate only within their designated levels of access.

Deployment Options

Own cloud	Fully local
This service package avoids large upfront costs by using third-party GPU-cloud services. Upon client's choice of a cloud provider, we will roll out Lexfactum.AI on such servers. Our app remains isolated from other AI services and service providers that may jeopardise client's privacy, ensuring security to the extent provided by the chosen cloud. We recommend using SOC2 certified cloud providers that adhere to the highest standards of cybersecurity and privacy. GPU-clouds typically charge on a monthly/yearly basis regardless of traffic and load.	This option allows users complete independence from third-party infrastructure, ensuring the desired level of privacy and data protection. Users can introduce their own security protocols and physically control the hardware. While upfront costs are higher, depending on the number of users and projected system load, Lexfactum.AI will configure the appropriate hardware infrastructure upon request. Having its own infrastructure will also require regular maintenance and support, which Lexfactum.AI can provide on a regular basis upon the client's

	request.

Pricing policy

1. Planning and Resource Allocation

In this phase, our manager will work with an organisation to customise Lexfactum.AI services and develop a rollout plan. The organisation will choose between two options: Own Cloud or Fully Local. We will provide guidance on hardware configuration requirements, including GPU models, motherboards, CPUs, RAM, and hard drives, based on the projected load (number of users and intensity of use). Additionally, we will develop a growth plan to ensure that our customer's hardware can be upgraded as demand increases.

2. Installation

This one-time expense covers the costs of system administrators to roll out Lexfactum.AI on the customer’s servers (either on-premises or cloud). This cost is applicable regardless of whether the customer chooses Own Cloud or Fully Local. This phase assumes the customer's infrastructure (be it a cloud or their local data centre) is present and ready to be used. The customer may hire LEXFACTUM.AI for an optional step of developing such infrastructure.

3. Infrastructure Rollout (Optional)

Creating a data centre of any scale is a specific task requiring special expertise and expenses related to installing hardware on-premises. Our customers can decide to undertake this themselves or involve us as the contractor. If we are hired, we will engage respective contractors and be responsible for rolling out the infrastructure on-premises.

4. Licence Fee

This fee grants access to the Lexfactum.AI application and serves as a recurring support fee. The support includes software maintenance, hardware maintenance (*optional), updates (e.g., regular updates to the law database), and upgrades (e.g., new features), as well as online chat and phone support during business hours.

III. Our plan

3.1 Market overview

The Australian legal tech market is dominated by the same players as in the American market. It is split between the two main competing products: Westlaw by Thomson Reuters and LexisNexis by RELX Group. These products offer not only comprehensive tools for legal research across Common Law countries but also various tools for task and business management, along with customer relationship management software.

Thomson Reuters acquired startup CoCounsel for $650 million US dollar in 2023, and started rolling out their AI capabilities across Thomson Reuters software in the U.S. with plans to add Australia in late 2024-2025. LexisNexis recently introduced their AI capabilities which are also available in Australia such, summarising case law and attached user's documents, and drafting emails to clients. Both programs rely on either OpenAI's ChatGPT or Microsoft's Copilot (LexisNexis) and none of the latter offers solutions on premises.

An Australian startup, Courtaid.ai, received a $150K grant from Microsoft Azure and launched its chatbot application for subscription in 2024. Relying on OpenAI's API, Courtaid.ai is somewhat similar to LEXFACTUM.AI. They do not offer an on-premises edition, as this type of software would require local AI embedding and vectorisation, which is hardware-intensive and demands a higher level of engineering expertise.

3.2 Commercialisation

Both Online and Enterprise editions are commercially viable products. The advantage of the Online as a product is that it can work with AI-services through APIs benefiting from cutting edge technology and their advancements. Our audience will be general users, smaller and mid-size law firms, corporations with in-house legal work, and academia. The subscription model makes it easy to get the services temporarily and cancel it anytime without overheads.

The enterprise solution is a premium product for organisations that have specific demands on data protection and privacy. This edition will work fully offline if needed, or will spread across organisations through the closed perimeter of the local network, leaving an opportunity to extend the service for remote users via the Internet and closed channel of communication. Besides software licence payments, it will require infrastructure maintenance, therefore, less affordable to smaller organisations. Our focus is large law firms, accounting and auditing companies, government agencies and corporations.

Revenue projections are explained in respective annexes.

3.4 Roadmap

We plan to further develop the capabilities of our application. Based on user feedback and our vision for future development, we will enhance the user experience.

Over the next few months, the current MVP will receive several updates, including:

- A more accurate database of laws.
- UIX improvements, such as dedicated buttons for summarisation, shortening, elaboration, and rephrasing (these functions are available in any LLM but currently require manual prompting).
- Custom user prompts creation.

Our reasoning pipeline will undergo enhancements to ensure it adapts to the latest developments in LLMs and incorporates best practices, while leveraging our accumulated knowledge and expertise.

Milestone Upgrades:

1. **Improved Search and Reasoning Capabilities:** Enhanced algorithms for more accurate and relevant results.
2. **Case Folders:** Users will be able to create cases from attached documents (e.g., agreements, policies, emails), collections of laws, and separate chat sessions (threads) within cases.
3. **Collaboration:** Additional tools to facilitate teamwork and collaboration.
4. **LLM Fine-Tuning:** Enhancing the internal model to provide more accurate legal answers and reduce hallucinations.
5. **Agreement Drafting:** Further specialisation of LLMs to improve legal document drafting, unlocking new possibilities in legal work.
6. **Akoma Ntoso Standard:** Expansion into other jurisdictions will begin with the adoption of the widely used standard for legal documents, Akoma Ntoso, recognised in the EU, several African countries, and UN organisations.
7. **General Knowledge Database Integration:** Prioritising tools for integrating LEXFACTUM.AI internal knowledge databases (e.g., wikis and the like) commonly used by corporations
8. **Microsoft Word Add-in:** Integrate LEXFACTUM.AI with MS Word as an add-in to offer a seamless AI experience while editing documents.

III. References

1. AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries, https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries, last accessed 2024/07/30.
2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 9459–9474. Curran Associates Inc., Red Hook, NY, USA (2020).
3. What Is Retrieval Augmented Generation (RAG)?, https://cloud.google.com/use-cases/retrieval-augmented-generation, last accessed 2024/07/24.
4. NVIDIA Chat With RTX, https://www.nvidia.com/en-au/ai-on-rtx/chat-with-rtx-generative-ai/, last accessed 2024/07/25.
5. GPT4All, https://www.nomic.ai/gpt4all, last accessed 2024/07/25.
6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., and Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423.