PROFACTUM.AI

White Paper

2024 v.1.00 (draft)

Sergiy Chernyshov[1]

Abstract: Profactum.AI, an innovative Australian technology startup, is poised to revolutionise the legal industry through the integration of generative artificial intelligence (GenAI) and retrieval-augmented generation (RAG). Our flagship development, an AI-powered legal assistant chatbot, is designed to navigate an extensive body of legal texts. In this white paper, we introduce the inaugural version of our application, emphasizing its unique features and capabilities. Additionally, we provide an overview of our strategic roadmap for ongoing development and commercialization, ensuring the sustained growth and evolution of our cutting-edge legal solutions.

Table of contents

I. Introduction

PROFACTUM.AI introduces the next-generation technology for legal research and inference by utilizing advanced artificial intelligence (AI) coupled with retrieval-augmented generation (RAG) across a comprehensive legal database.

[1] Software developer and cybersecurity expert, CTO and Co-founder of PROFACTUM.AI.

and user's documents In this paper, we introduce the initial functionalities of our application and outline our vision for future development and commercialization.

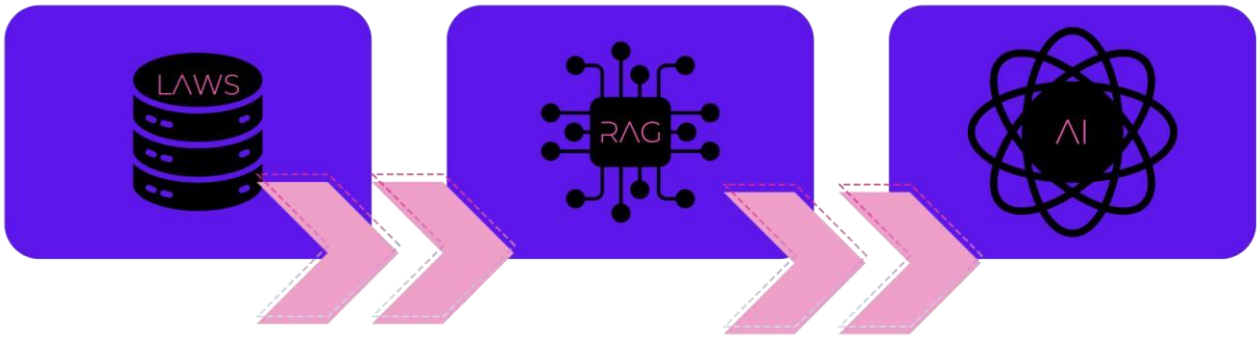
The rise of large language models, spanning both proprietary and open-source frameworks, has unlocked unprecedented opportunities to revolutionize legal research and practice. While some lawyers have successfully employed OpenAI's ChatGPT to enhance their work, others have encountered issues such as "hallucination," [1] where GenAI generates inaccurate or fabricated information. Although legal professionals can typically identify these errors, we recognize the risks associated with AI use by those without legal expertise. We do not yet consider AI sufficiently advanced to resolve legal disputes independently. Moreover, it is evident that simply deploying AI chatbots is insufficient, even for seasoned legal practitioners. Our research and development efforts have focused on creating technology that enables AI to perform analysis and reasoning while rigorously adhering to the legal corpus. This approach is designed to prevent AI hallucinations and ensure that AI-driven conclusions are substantiated by legal references.

The following chapter explores the challenges of applying retrieval-augmented generation (RAG) techniques in enterprise-grade solutions, detailing our design and its performance. The third section delves into PROFACTUM.AI's commercialization strategy, business development, and future roadmap.

II. Our technology

2.1 RAG issues and the solution

The concept of retrieval-augmented generation (RAG) began to take shape globally in 2020 [2]. As defined by Google, RAG is "an AI framework that combines the strengths of conventional information retrieval systems (such as databases) with the generative capabilities of large language models (LLMs)" [3].



In simpler terms, RAG allows AI to efficiently interact with external knowledge that isn't inherently part of the AI model. This capability is crucial because retraining AI models to include new information is not only complex but also constrained by their limited "context" window—essentially the amount of information the model can process at one time, measured in "tokens." This limitation is akin to a computer's RAM or the human brain's capacity for immediate information processing. RAG expands this capacity, enabling AI to handle extensive datasets, such as a collection of laws.

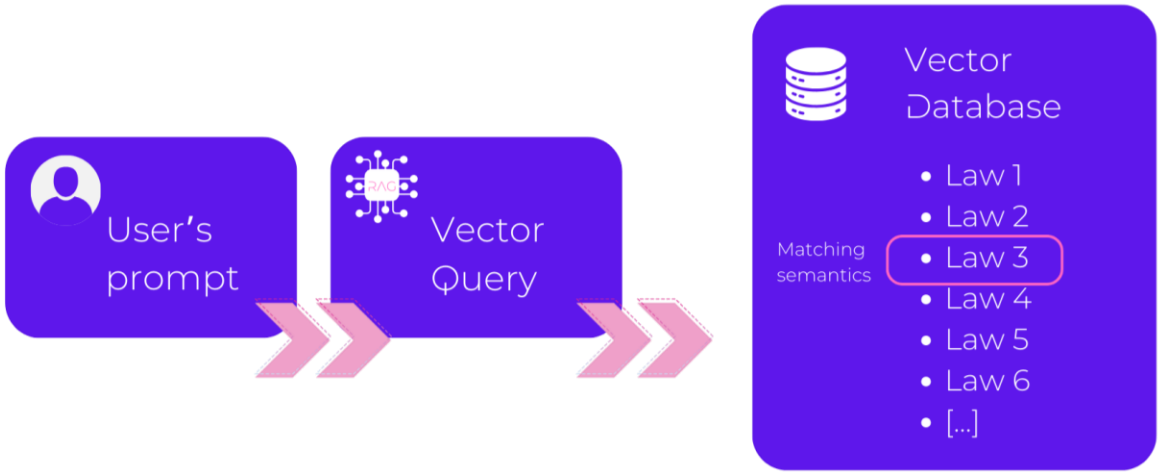
RAG can be deployed on consumer-grade computers using various free tools like 'NVIDIA Chat with RTX' [4] or 'GPT4ALL' [5]. However, to function efficiently, it requires a relatively powerful GPU,

specifically an RTX series with at least 8GB of VRAM. Despite its advantages, there are two significant limitations to its application:

1. **Limited Capabilities of Smaller LLMs:** For instance, while LLAMA 3.1 8B can operate on a consumer-grade laptop, it falls short when it comes to handling more complex tasks, such as legal work.

2. **Time-Consuming Vectorisation Process:** Vectorizing a large body of laws from any jurisdiction is a labor-intensive task. For example, when this process was conducted on a gaming laptop, it took several weeks to complete. Vectorization involves formatting user data in a way that AI can interpret, using an embedding model like BERT [6] to transform the data into digital vectors that represent semantic meanings. This preprocessing step results in the creation of a vector database that encapsulates all the input data.

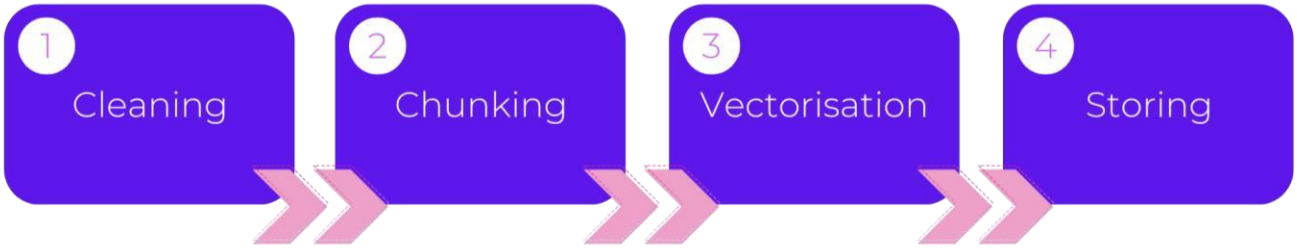
In practice, when a user submits a query, the application converts the query into a vector—a digital representation of the query's semantics—and then searches the vector database for relevant matches. The matched data is subsequently provided to the AI for further processing.



Vectorizing over 220,000 Australian laws poses a significant challenge, both in hardware and software. The vectorization of 10GB of Australian legal texts is a computationally demanding task, even for an enterprise-grade workstation equipped with a powerful GPU. However, a vector database alone does not suffice to deliver a complete user experience in such a chatbot application. It also necessitates integration with a traditional database management system (DBMS) like MySQL or MongoDB, integrated with the vector database. Unfortunately, popular online resources lack detailed guidance on these complexities or viable architectural solutions. Additionally, the libraries and drivers involved often encounter compatibility and stability issues, making the process of installation and adaptation to specific hardware configurations require deep expertise and extensive debugging time.

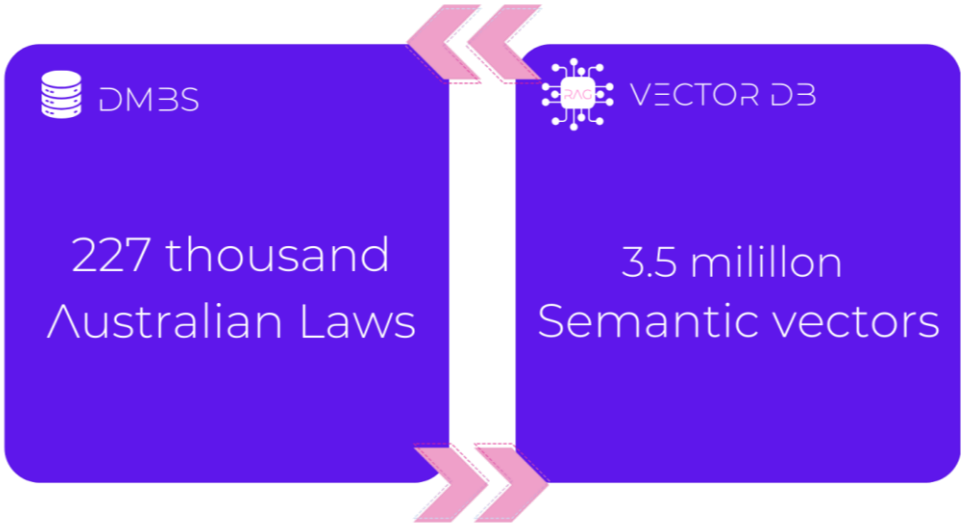
Despite these challenges, our research and development efforts have resulted in substantial improvements in the performance of database preprocessing (vectorization) and the efficiency of subsequent vector search queries, as well as related queries within our DBMS.

Database preparation consists of several essential parts (with some technical details omitted for the purposes of this paper):



1. **Cleaning:** The text is first cleaned by removing capital letters, punctuation, and other noise elements to ensure that the AI embedding model processes clean input.
2. **Chunking:** The legal texts are then divided into chunks, typically one paragraph in length, but not exceeding 500 tokens (approximately 2,000-2,500 characters). To maintain continuity, each chunk overlaps slightly with the next. This process resulted in nearly 3.5 million chunks.
3. **Vectorisation:** Next, an AI embedding model is used to extract semantic meaning from these chunks, performing vectorization. It's important to note that not all embedding models are equipped to handle the vectorization of paragraphs as semantic units. The resulting vectors are stored in a database specifically designed to manage large volumes of vector data.
4. **Storing:** Simultaneously, an auxiliary database is created using a conventional DBMS to support the functionality of the user application.

A finely-tuned system equipped with two 56-core CPUs and an NVIDIA A5000 GPU can now process the 10GB dataset in approximately 6 hours, a significant improvement from the initial 8 days. Each chunk is prepared in roughly 70 milliseconds.

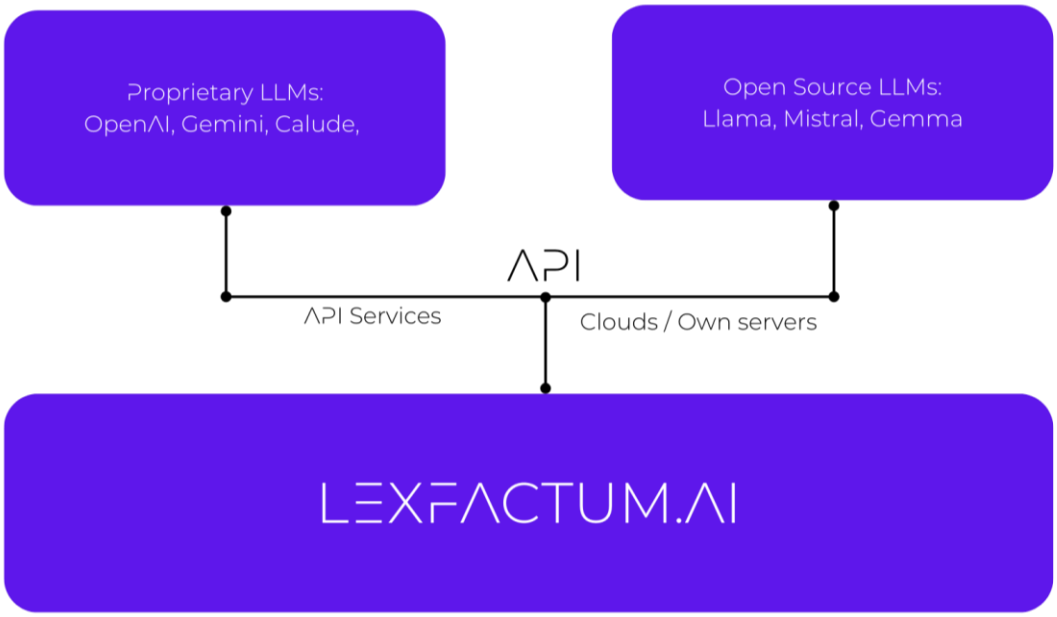


After the databases were prepared, we proceeded to develop both the back-end and front-end of the application, integrating the necessary features for legal work.

2.2 Large Language Models and challenges

A key element of our application is the Large Language Model (LLM). The generative AI capabilities must be robust enough to manage the complexity of legal work while remaining efficient, as our technology is engineered to function entirely locally and offline. In Section 2.4, we explore the

development of two versions of the application: one is an online web application intended for widespread use, while the other is an enterprise solution—a premium offering tailored to meet the stringent privacy and data protection standards demanded by our most discerning clients.



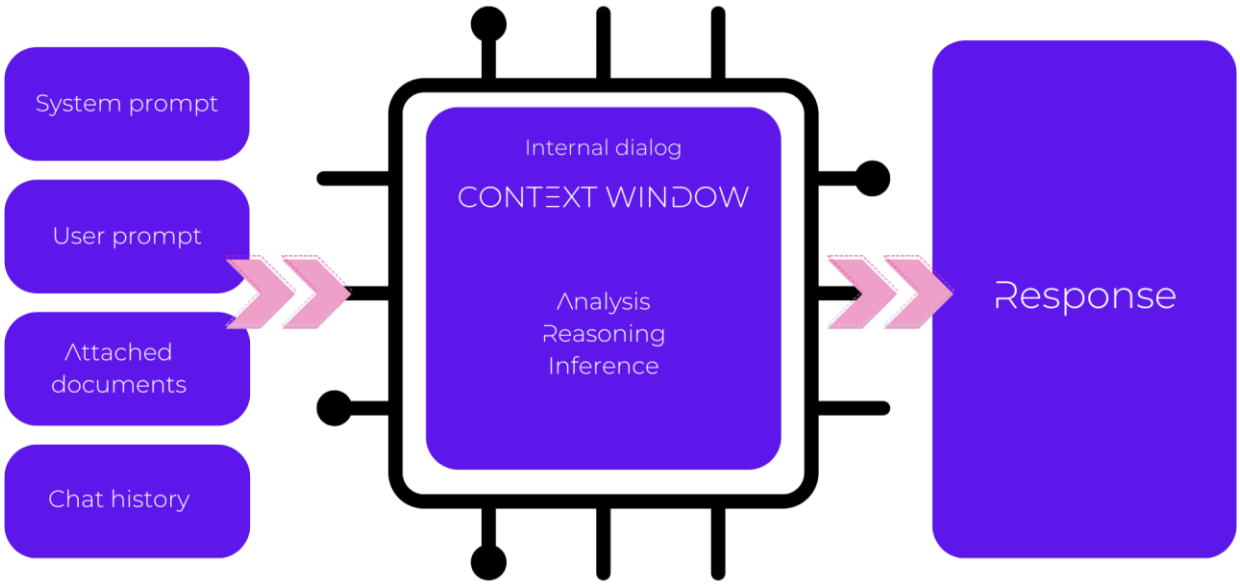
OpenAI currently dominates the market, with its proprietary model outperforming both other proprietary solutions and open-source alternatives in numerous benchmarks. However, this advantage has recently diminished due to advancements made by Claude, Google's Gemini, and others, particularly in various performance metrics. For example, Gemini excels with its expansive context window of 1 million tokens, which is critical for conducting extensive legal analyses. Meanwhile, the highly efficient Groq proves valuable in several interim processes within our AI reasoning workflow. Additionally, recent progress in AI reasoning methods has shown that even open-source models can outperform their proprietary counterparts. A common approach involves using a three-step inference process with different models, followed by a final summarization using one of these models (see Fig. N1).

Each model has several critical parameters, including the size of the dataset and the context window. The effective utilization of these parameters determines the sophistication of the application. For instance, Meta's Llama 3.1 8B model (with 8B indicating the size of the training dataset) is quick and smart enough to handle internal utility tasks like prompt rephrasing or query routing. The more advanced Llama 3.1 70B model is capable of tackling more complex tasks, such as legal analysis and reasoning.

The size of the context window, or the maximum number of tokens the model can process at one time, is crucial for the user experience and overall quality of the application. Legal documents can be extremely lengthy; for instance, the Corporations Act 2001 (Cth) comprises over 200,000 tokens. The Llama model, with its 8,000-token context window, is unable to process such a large document all at once. Thus, RAG functions both as the engine for semantic search across all Australian laws and as a tool for browsing individual documents.

The context window is a valuable resource because each prompt must encompass not only the user's question, system instructions, and relevant legal text but also the history of the current chat session. In this capacity, the context window essentially acts as the model's memory. However, only 25-30% of its

capacity is effectively usable for processing. This means that, to achieve optimal results—particularly with models like Llama—the input should ideally be limited to 2000-2500 tokens. The challenge lies in the fact that the model requires adequate 'space' for reasoning. As it processes the input, the model unpacks and analyzes the data, engaging in an 'internal dialogue.' When the input occupies more than 25% of the context window, the quality of the model's responses begins to diminish. Our empirical research has shown a significant decline in performance when the input exceeds 80% of the context window. As a result, we tend to favor models with larger context windows, as they offer superior reasoning capabilities, although they demand more from the hardware.



There are several techniques available to enhance the quality of results [6]. We have developed a robust back-end solution that incorporates best practices in RAG and AI design, alongside our own innovative insights. The key components of our system perform extensive work behind the scenes. After a user submits a prompt, the system engages in internal rephrasing, where the model reformulates the user's queries into legal terminology; drafts a response; identifies the relevant laws; analyzes them; and ultimately delivers the most comprehensive legal answer (for further details, see the next section on UIX).

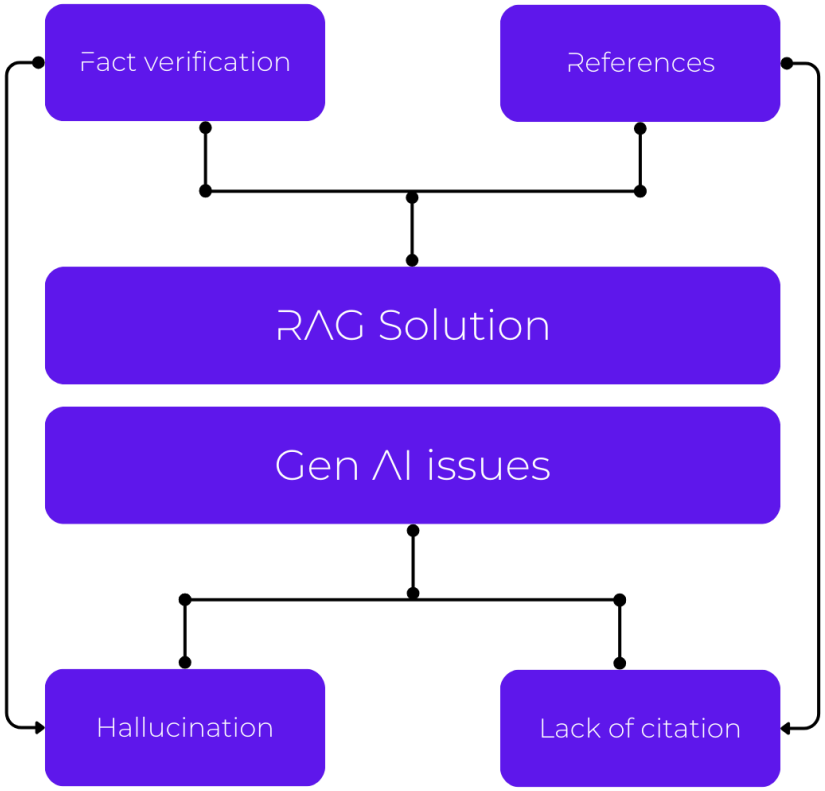
PROFACTUM.AI's solution is compatible with both proprietary online models and open-source offline models, allowing for asynchronous queries across any combination of these models. Powered by APIs, this fully AI-agnostic design enables us to swiftly adapt to the rapid evolution within the AI landscape.

2.3 Main features

Unique features

What makes PROFACTUM.AI a standout tool for legal work? While AI chatbots like ChatGPT and Gemini are capable of answering legal questions effectively, they often fall short by producing 'hallucinations'—inaccurate information that lacks proper citations, such as specific statutes or court decisions. PROFACTUM.AI is specifically designed to address these shortcomings. Our RAG-powered application integrates a vectorized database of laws, facilitating a semantic, AI-driven search. This approach ensures that responses are grounded in accurate, real-world legal texts. Furthermore, instead

of merely offering conclusions without context, PROFACTUM.AI provides detailed references to specific statutes and court rulings, down to the paragraph level, thereby offering users full transparency and confidence in the information provided.



Let's delve into the user experience within PROFACTUM.AI. The web application is designed with a user-friendly layout: a central window for chat interactions and prompts, a left-hand panel for chat history, and an additional right-hand panel for fine-tuning queries.

The right panel is where PROFACTUM.AI truly shines, providing specialized features for legal research. We have developed several key modes of operation tailored for legal work:

1. **Problem Solving:** Users can describe real-world scenarios to receive legal advice, either by directly typing their questions or by attaching relevant documents. GenAI provides a range of additional features familiar to users of ChatGPT, such as summarization, rephrasing, shortening, elaboration, and reasoning. All user prompts are cross-referenced with the database of laws and presented in a separate window, accompanied by relevant text snippets and (citations).

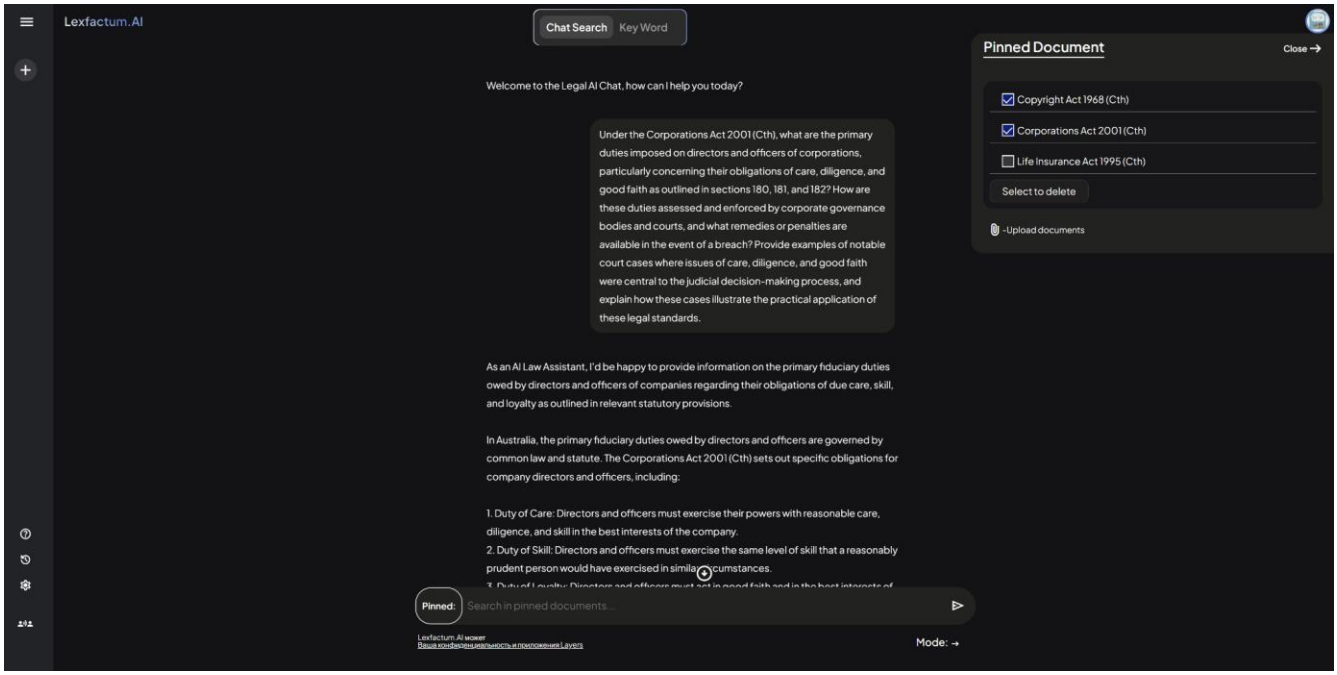
2. **Legal Research:** This feature allows users to search a database of Australian laws using a vector-based, or semantic AI, search. For example, typing “a dog crosses the street” generates a semantic vector representation of the phrase, which the app then uses to search for laws with similar meanings. This approach is more effective than keyword searches, which might only find laws that specifically mention terms like 'dog' or 'street.' Our AI-assisted search can identify relevant traffic rules or animal control laws, providing a more accurate and pertinent set of results.

3. **Keyword Search:** A more traditional keyword search allows for precise legal research when users are looking for specific terms, such as names, dates, or facts.

4. **Filters:** The right panel enables users to refine their legal research by specific fields of law and jurisdictions. Filter options include Federal level, states and territories, statute law, and case law. Statute law can be further categorized into primary legislation (legislative acts), secondary laws (regulations, by-laws), and bills (draft laws). Users can combine filters to fine-tune their search 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 established, users can switch to legal research within that collection or focus on individual acts.

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



2.4 Online and Enterprise editions

PROFACTUM.AI developed two editions of the software.

Online

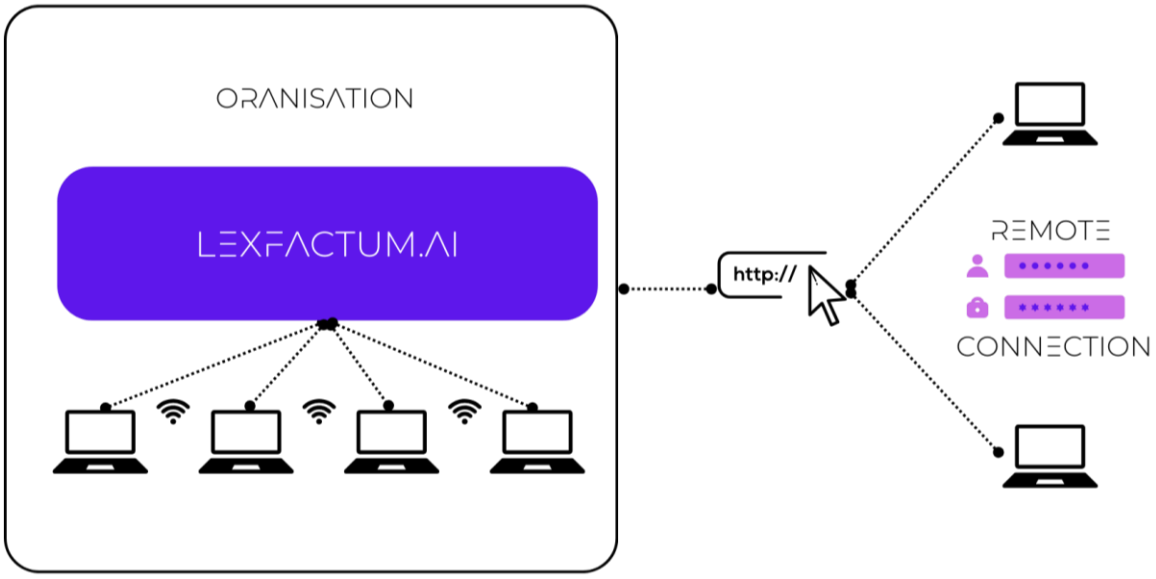
Our online edition functions as a web application hosted in the cloud, offering services to a wide audience through a paid subscription model. It replicates the functionality of the Enterprise edition and is compatible with both open-source and proprietary large language models (LLMs), or a combination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

Currently the online version is available for demonstration in a testing mode and is ready for a full rollout once the necessary infrastructure is established.

Enterprise edition

The enterprise edition is a software solution designed to operate on-premises, within a local network. The database of Australian laws, along with the backend that drives the application—including the AI—functions entirely locally and offline. This setup ensures the highest levels of privacy and data protection, as interactions with the AI are not transmitted to external service providers like OpenAI or Gemini. This level of privacy is achieved by utilizing advancements in open-source LLMs that rival, and in some cases surpass, the performance of commercial alternatives.

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



The Enterprise app is designed for multi-user environments, offering finely-tuned access rights and authorizations. It isolates different databases and ensures that staff operate strictly within their designated access levels.

Deployment Options

Own cloud	Fully local
This service package minimizes upfront costs by leveraging third-party GPU-cloud services. Once the client selects a cloud provider, we will deploy Profactum.AI on the chosen servers. Our application remains isolated from other AI services and providers that might compromise the client's privacy, maintaining security as defined by the chosen cloud provider. We recommend using SOC2-certified cloud providers that meet the highest standards of cybersecurity and privacy. Typically, GPU-clouds	This option grants users complete autonomy from third-party infrastructure, ensuring the highest levels of privacy and data protection. Users can implement their own security protocols and maintain physical control over the hardware. While upfront costs may be higher—depending on the number of users and projected system load Profactum.AI will configure the appropriate hardware infrastructure upon request. Owning the infrastructure also necessitates regular maintenance and support, which Profactum.AI

charge on a monthly or yearly basis, independent of traffic and load.	can provide on an ongoing basis according to the client's needs
	request.

Pricing policy

1. Planning and Resource Allocation

During this phase, our manager will collaborate with the organization to tailor Profactum.AI services and devise a rollout plan. The organization will have the option to choose between two deployment models: Own Cloud or Fully Local. We will offer guidance on hardware configuration requirements, including GPU models, motherboards, CPUs, RAM, and storage, based on the anticipated load (number of users and usage intensity). Additionally, we will develop a growth plan to ensure that the customer's hardware can be upgraded as demand grows.

2. **Installation**

This one-time expense covers the cost of system administrators to deploy Profactum.AI on the customer's servers, whether on-premises or in the cloud. This cost applies regardless of whether the customer opts for the Own Cloud or Fully Local model. This phase assumes that the customer's infrastructure—be it cloud-based or a local data center is already in place and ready for deployment. The customer may also choose to hire PROFACTUM.AI for an optional service to develop this infrastructure.

3. Infrastructure Rollout (Optional)

Setting up a data center of any scale is a specialized task that requires specific expertise and associated costs for installing hardware on-site. 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 Profactum.AI application and serves as a recurring support fee. Support services include software maintenance, optional hardware maintenance, updates (e.g., regular updates to the law database), and upgrades (e.g., new features). Additionally, online chat and phone support during business hours are provided.

III. Our plan

3.1 Market overview

The Australian legal tech market mirrors the dynamics of the American market, with the same dominant players. The market is primarily divided between two leading products: Westlaw by Thomson Reuters and LexisNexis by RELX Group. These platforms not only provide comprehensive tools for legal research across Common Law jurisdictions but also offer a range of task and business management tools, as well as customer relationship management software.

In 2023, Thomson Reuters acquired the startup CoCounsel for $650 million USD and began integrating AI capabilities across their software in the U.S., with plans to extend these features to Australia in late 2024-2025. LexisNexis has also recently introduced AI capabilities, which are now available in Australia, offering features such as case law summarization, document analysis, and email drafting for clients. Both programs depend on either OpenAI's ChatGPT or Microsoft's Copilot (LexisNexis), and neither offers on-premises solutions.

An Australian startup, Courtaid.ai, received a $150K grant from Microsoft Azure and launched its subscription-based chatbot application in 2024. Similar to PROFACTUM.AI, Courtaid.ai relies on OpenAI's API. However, they do not provide an on-premises edition, as this type of software would require local AI embedding and vectorization—processes that are hardware-intensive and require a higher level of engineering expertise.

3.2 Commercialisation

Both Online and Enterprise editions are commercially viable products. The key advantage of the Online edition lies in its ability to integrate with AI services through APIs, allowing it to leverage cutting-edge technology and advancements. This edition is geared toward general users, small to mid-sized law firms, corporations with in-house legal teams, and academic institutions. The subscription model offers flexibility, enabling users to access the services temporarily and cancel at any time without incurring additional overhead costs.

The Enterprise edition, on the other hand, is a premium product designed for organizations with specific requirements for data protection and privacy. This edition can operate fully offline if necessary or can be deployed across organizations within a secure local network, with the option to extend the service to remote users via the Internet through a secure communication channel. In addition to software license fees, the Enterprise edition will require infrastructure maintenance, making it less accessible to smaller organizations. Our primary target market includes large law firms, accounting and auditing firms, government agencies, and corporations.

Revenue projections are detailed in respective annexes.

3.4 Roadmap

We are committed to further enhancing the capabilities of our application. Guided by user feedback and our vision for future development, we will continue to improve the user experience.

In the coming months, the current MVP will receive several updates, including:

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

Our reasoning pipeline will also be upgraded to ensure it keeps pace with the latest advancements in LLMs and incorporates best practices, all while drawing on our accumulated knowledge and expertise.

Milestone Upgrades:

1. **Improved Search and Reasoning Capabilities:** Implementation of enhanced algorithms for more accurate and relevant search results.
2. **Case Folders:** Introduction of a feature allowing users to create cases from attached documents (e.g., agreements, policies, emails), collections of laws, and separate chat sessions (threads) within those cases.
3. **Collaboration:** Development of additional tools to support teamwork and collaboration.
4. **LLM Fine-Tuning:** Refining the internal model to deliver more accurate legal answers and minimize hallucinations.
5. **Agreement Drafting:** Further specializing LLMs to enhance the drafting of legal documents, thereby unlocking new possibilities in legal work.
6. **Akoma Ntoso Standard:** Expanding into additional jurisdictions by adopting the widely recognized Akoma Ntoso standard for legal documents, which is used in the EU, several African countries, and UN organizations.
7. **General Knowledge Database Integration:** Prioritizing the integration of PROFACTUM.AI's internal knowledge databases (e.g., wikis and similar resources) that are commonly utilized by corporations.
8. **Microsoft Word Add-in:** Integrating PROFACTUM.AI with Microsoft Word as an add-in to provide a seamless AI-driven experience during document editing.

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-benchmarkingqueries, 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/retrievalaugmented-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.