Singapore University of Technology and Design

Capstone Term 7

Capstone 9 Project 07

AsiaCloud_AI4PDPA

Author	*Student ID*
Ang Ching Xuen	1006976
Fan Xiangwei	1005533
Gregory Lim Eu Rhen	1007485
Issac Anand Rajaram	1007208
Matthew Andrei Salatin Purba	1007094
Qi Hengchang	1007166
Sherman Kho Jun Hui	1006890

November 30, 2025

Acknowledgements

We are grateful to all who supported the AI4PDPA project throughout Term 7.

From AsiaCloud Solutions, we express our Mr. Ng Lye Kiat, Mr. Yuri See, and Ms. Valerie Ng, for their insights into real-world compliance practices and continual support.

From the Singapore University of Technology and Design, we extend special thanks to Dr. Sun Zhu, Dr. Susan Wong, and Dr. Thomas Schroepfer, for their steady guidance and constructive feedback on project direction and expression.

Executive Summary

Rapid digitalisation in Singapore has increased the volume of personal data handled by organisations and heightened the risks of misuse, unauthorised access, and data breaches. While the Personal Data Protection Act (PDPA) provides the regulatory framework governing responsible data practices, many organisations—especially small and medium enterprises (SMEs)—continue to struggle with interpreting and applying these obligations in day-to-day operations. Existing compliance resources are fragmented, technically complex, and often costly, creating a gap for accessible, practical guidance.

To address this challenge, this project collaborates with AsiaCloud to develop AI4PDPA, a modular, AI-driven platform designed to help SMEs understand and apply PDPA requirements more effectively. The platform incorporates a PDPA chatbot powered by retrieval-augmented generation (RAG), a document template generator, a PDPA training module with automated scoring, and a dashboard that presents aggregated compliance insights. Together, these components aim to provide real-time, contextualised assistance without requiring legal expertise or expensive consultancy services.

In Term 7, the team established the system architecture, conducted literature analysis, identified ethical and PDPA compliance requirements, developed three iterative prototypes of the chatbot. Key technical progress included implementing the chunking and storage pipeline, deploying the system on Amazon Web Services (AWS) using containerised services, developing initial front-end interfaces, benchmarking large language models (LLMs) to compare their accuracy and reliability, and integrating early safety and refusal mechanisms aligned with PDPA obligations.

Looking ahead, Term 8 will complete the full platform build. Priorities enhancing the chatbot experience, integrating server-based chat history, completing the document-generation and training modules, and building an analytics dashboard that provides value for both SMEs and AsiaCloud. To support long-term scalability, the team will transition from a manual deployment process to a GitHub Actions-based Continuous Integration and Continuous Delivery (CI/CD) pipeline. The new workflow will automate testing, container builds, and deployment to AWS, reducing operational overhead and ensuring that only validated builds reach production—aligned with modern best practices for development operations (DevOps). When fully developed, the platform aims to provide SMEs with a low-barrier, affordable, and practical solution for PDPA compliance, while also giving AsiaCloud a strong foundation for a future commercial offering. By combining AI technologies with regulatory grounding and ethical safeguards, the project seeks to enhance organisational readiness, reduce compliance risks, and contribute to Singapore's broader Smart Nation and data protection objectives.

Contents

1 Introduction

1.1 Motivation

Rapid digitalisation has transformed how organisations collect and use personal data to deliver services, optimise operations, and support decision-making (Zul, 2022). In Singapore, this shift is amplified by national drives toward Smart Nation goals and widespread AI adoption (Choudhury, 2024). As data use increases, so do risks associated with misuse, unauthorised access, and large-scale breaches. Recent cases illustrate these risks. For example, Marina Bay Sands was fined SGD 315,000 after a breach affecting more than 665,000 customer records, demonstrating that even mature organisations face difficulties in maintaining strong data protection practices (Chan, 2025). To safeguard individuals while supporting innovation, the PDPA was introduced to set baseline obligations for all organisations handling personal data (Personal Data Protection Commission, 2023). However, compliance remains challenging. PDPA guidelines are extensive, spread across multiple documents, and written for readers with legal or technical expertise. As regulations evolve, many organisations—particularly SMEs—struggle to interpret requirements correctly. This difficulty is evident in common operational errors, such as the Credit Counselling Singapore incident where email addresses were exposed due to incorrect use of the "To" field (Personal Data Protection Commission, 2017). Such examples highlight the need for accessible and scenario-specific compliance guidance.

1.2 Problem Statement

Although resources produced by the Personal Data Protection Commission (PDPC) are publicly available, they are not always accessible to non-specialists. Many SMEs lack legal support and rely on ad-hoc interpretations or outdated practices. Existing commercial compliance tools are often costly or too complex, while general-purpose AI chatbots may hallucinate, mix regulatory frameworks, or offer unsafe advice. This project seeks to address these gaps by developing AI4PDPA, a modular, AI-assisted compliance platform for Singaporean SMEs. Its objectives are threefold. Firstly, AI4PDPA aims to provide accurate, PDPA-grounded, and scenario-specific guidance through a controlled AI chatbot. Secondly, AI4PDPA aims to streamline compliance workflows using structured document templates and PDPA training modules. Thirdly, AI4PDPA aims to support AsiaCloud's long-term commercialisation plans with a scalable, cost-efficient architecture and aggregated usage analytics.

1.3 Scope

The project spans two academic terms. Term 7 focuses on foundational work, which includes defining the overall system architecture, developing an initial chatbot prototype, conducting preliminary testing, and refining project requirements through iterative feedback sessions with the industry partner. This early phase establishes the technical direction, key constraints, and safety considerations necessary to support a scalable PDPA compliance platform. Term 8 prioritises full implementation across the platform's core modules. Planned work includes completing the chatbot's features—such as safe-answer behaviour, PDPA-specific grounding, and more reliable retrieval—together with developing a server-based chat history system to support analytics and longer-term storage. The team will also implement PDPA document template generation, build a training module that allows administrators to upload training content and create quizzes, and develop an administrative analytics dashboard that presents aggregated usage insights for AsiaCloud. The system's architecture will emphasise modularity, cost efficiency, and extensibility so that AsiaCloud can build upon this proof-of-concept and scale it into a commercial product after the project concludes.

1.4 Stakeholders

The platform is designed for users who interact with PDPA obligations in day-to-day SME operations. SMEs form over 99% of Singapore's enterprises (Lim, 2025), and many have limited resources for formal compliance, making accessible PDPA support especially important. Operational staff—including administrative, customer service, sales, and human resources personnel—handle personal data frequently but often rely on informal or low-digitisation workflows such as spreadsheets or email (Navarrete, 2019). These users require quick clarification when performing tasks such as collecting customer details, responding to access requests, or drafting consent notices. The chatbot and document templates are designed to support these immediate operational needs. Managers and internal data stewards typically assume PDPA responsibilities in addition to their primary roles. They require clearer guidance for interpreting PDPA obligations, onboarding staff, and ensuring consistent handling of personal data across the organisation. This group benefits from structured training modules and aggregated analytics that highlight recurring compliance gaps. SME owners and leadership teams use higher-level insights to evaluate organisational compliance readiness. They rely on the dashboard to understand frequently asked PDPA questions, training participation, and potential areas of risk that require process changes or additional staff support. Beyond end users, AsiaCloud Solutions is a key stakeholder as both industry partner and future maintainer. Their interests include evaluating user behaviours, validating market demand, and identifying features suitable for a commercialised product. The platform's modular architecture and server-based chat logging were designed to support this longer-term vision. While not within the immediate scope of the project, the architecture allows for future expansion to industry associations, training providers, and larger organisations that require lightweight PDPA reference tools.

1.5 Success Criteria

The platform's success is evaluated across four dimensions: accuracy, safety, usefulness, and scalability. Accuracy refers to providing PDPA-aligned responses grounded in official guidelines and enforcement cases. Safety focuses on refusal behaviour for sensitive or out-of-scope queries and on reducing hallucinations through retrieval-based grounding. Usefulness concerns the practicality of the chatbot, document templates, training modules, and analytics in supporting common SME compliance tasks. Scalability relates to the system's modular, cost-efficient architecture, ensuring suitability for SMEs and future commercial expansion by AsiaCloud. These criteria will be assessed through controlled testing, benchmarking, and stakeholder feedback in Term 8.

1.6 Constraints

The project operates under several constraints. Regulatory constraints require strict alignment with PDPA and limit guidance to publicly available PDPC materials. Data constraints restrict testing to synthetic or anonymised cases, as real organisational data cannot be used. Ethical and safety constraints include avoiding hallucinations, avoiding inappropriate advice, and discouraging users from entering sensitive personal data. Architectural constraints require a lightweight, cost-efficient, and modular design suitable for SMEs and scalable for AsiaCloud. Finally, evaluation constraints limit testing to controlled environments, as large-scale deployment studies fall outside the scope of the project.

2 Literature Review

2.1 Background

Rapid digitalisation has increased the scale at which organisations collect and process personal data, making data protection a key concern, especially in Singapore (Zul, 2022). In response to these challenges, the PDPA was established to safeguard the personal data of individuals while supporting the business interests of organisations (Personal Data Protection Commission, 2023). To aid compliance, the PDPC publishes advisory guidelines to help organisations and individuals understand the PDPA (Personal Data Protection Commission, 2022).

2.2 Challenges

Despite the availability of these resources, many organisations and individuals struggle to understand and comply with PDPA requirements correctly because regulations continually evolve (i-Sprint Innovations Pte Ltd., 2024). According to the PDPC's 2015 industry survey, about 58% of organisations required support to achieve compliance, reflecting knowledge and resource gaps (Personal Data Protection Commission of Singapore, 2015). Additionally, there is growing interest in leveraging AI to automate the retrieval of knowledge and comprehension of regulatory compliance (Gültekin-Várkonyi, 2025). There is, therefore, an opportunity to combine AI and compliance to create tools that can interpret legal frameworks and make regulatory knowledge more accessible.

SMEs form the backbone of Singapore's economy (Lim, 2025). However, SMEs often face greater obstacles in meeting compliance requirements compared to larger organisations due to limited financial resources and the growing complexity of regulatory obligations (Bello et al., 2024). As a result, many do not allocate sufficient time, budget, or staff to interpret and implement compliance regulations (Compliance Consultant, 2025). Digital capability gaps further worsen the compliance challenge. According to a survey conducted by Capterra, 32% of SMEs still rely on spreadsheets to manage customer information, while another 35% use manual methods or email communication, which are insufficient under modern data protection guidelines (Navarrete, 2019). These informal, decentralised data management methods make it difficult to track consent, accurately update records, and securely ensure retention and disposal. In this context, there is a clear need for accessible, low-barrier compliance support tools tailored to SMEs. Solutions such as AI-driven PDPA chatbots can lower the knowledge and resource barrier by providing SMEs with immediate, accurate guidance without requiring legal expertise, formal training, or expensive consultancy services.

2.3 Solutions

A variety of resources are available to support PDPA compliance, primarily provided by the PDPC. These include advisory guidelines, compliance checklists, and the Data Protection Essentials programme for SMEs (Personal Data Protection Commission, 2025). However, users often struggle to identify which specific sections apply to their situation or to interpret PDPA requirements without legal or technical expertise (Lonzetta & Hayajneh, 2020).

In addition to PDPC materials, several commercial and professional solutions are available. Many organisations engage outsourced Data Protection Officers or legal consultants to interpret PDPA requirements on their behalf. Others adopt enterprise-grade privacy management systems such as Varonis, which offer data governance dashboards and risk assessment modules. However, these systems are costly, and they require ongoing subscription fees, which are impractical for SMEs with limited budgets.

2.3.1 Gaps

However, existing solutions lack conversational, real-time interaction. Users must manually search through documentation, navigate dashboards, or rely on external consultants. Taken together, these findings highlight the absence of an accessible, affordable, and PDPA-specific digital tool capable of providing real-time compliance support. This gap underscores the need for an AI-powered PDPA chatbot that allows users to obtain accurate, scenario-specific guidance without requiring legal expertise or substantial financial investment.

The increasing complexity of legal and regulatory frameworks has driven organisations to adopt AI solutions to automate and streamline compliance processes (Bleach, 2024). LLMs allow systems to interpret, analyse, and generate human-like responses to text-based queries (Vaniukov, 2024). In the legal industry, AI tools are being used for document analysis, legal advice support, and contract drafting, reducing mundane work and improving efficiency (SMU Social Media Team, n.d.). A key innovation in this area is AI chatbots, which act as conversational agents to answer user queries about legal and compliance matters. In the legal industry, AI chatbots have already begun transforming how professionals access and interpret information. For instance, Harvey AI, developed on OpenAI's Generative Pre-trained Transformer (GPT) technology, has partnered with law firms and consulting giants such as PwC to assist in legal research, contract review, and compliance analysis (PwC, 2023). This growing adoption highlights the potential for AI to streamline legal workflows and enhance user understanding of complex legal texts. Similarly, there is growing recognition that such tools can extend to the data protection and compliance domain. Developments in this area align with Smart Nation initiatives, which encourage the use of AI to improve governance, productivity, and security (Government of the Republic of Singapore, 2023). Thus, AI-powered PDPA chatbots represent a promising approach to bridging compliance gaps, especially for resource-constrained SMEs.

AI chatbots present several privacy and data protection risks, particularly when deployed in compliance-related contexts. Large language models generate outputs probabilistically and may produce inaccurate, misleading, or hallucinated information. OpenAI publicly acknowledges this behaviour in its model documentation, noting that such models can generate plausible-sounding but incorrect statements (OpenAI, 2023). In regulatory settings, such inaccuracies carry amplified consequences because organisations may unintentionally rely on erroneous interpretations when making compliance decisions. Established research in human-automation interaction also shows that users tend to over-trust automated systems, especially when the system appears competent and fluent, which increases the likelihood that incorrect responses are accepted without critical evaluation (Goddard et al., 2012).

A second major concern relates to how chatbots handle user inputs. If systems are not designed with strict minimisation or protection controls, sensitive information may be captured through server logs, analytics tools, or external telemetry. PDPC guidance emphasises that the handling of identifiers—such as National Registration Identity Card (NRIC) numbers—involves heightened obligations due to the risks of identity theft and misuse if the data is retained or transmitted unnecessarily (Personal Data Protection Commission, 2022). Cisco's 2023 Data Privacy Benchmark Study also reports that users frequently submit sensitive information into digital tools without fully understanding how it will be stored or processed, increasing organisations' responsibility to implement strong safeguards (Cisco Systems, 2023).

Jurisdictional accuracy is another key challenge. General-purpose LLMs often incorporate patterns from global datasets, which can lead to inappropriate blending of regulatory concepts across jurisdictions. The Organisation for Economic Co-operation and Development (OECD) warns that AI systems trained on mixed-jurisdiction data may unintentionally provide legal interpretations that do not align with local laws, creating significant compliance risks (Organisation for Economic Co-operation and Development, 2022).

Finally, AI chatbots can amplify misinformation risks. IBM's research on trustworthy AI highlights that the natural language fluency of modern LLMs can obscure underlying inaccuracies, making hallucinations especially dangerous in high-stakes domains such as law, healthcare, and compliance (Chen & Xiao, 2023). These risks reinforce the need for strict data-minimisation practices, well-defined refusal behaviour, jurisdiction-specific safeguards, and transparent user guidance to ensure that AI chatbots support—rather than undermine—organisational compliance efforts.

2.3.2 Frontiers

Retrieval-augmented generation is an increasingly important technique for improving the factual accuracy, grounding, and explainability of LLM-based systems. Rather than relying solely on internal model parameters, RAG retrieves relevant information from an external knowledge base and incorporates it into the model's context before generating an output (Lewis et al., 2020). This approach helps reduce hallucinations by tying responses to verifiable documents rather than latent model associations. RAG is particularly well suited for compliance and regulatory domains. By grounding responses in authoritative materials—such as PDPC advisory guidelines, enforcement decisions, and published regulatory notes—it ensures that outputs remain aligned with official interpretations and reduces the risk of misleading or inconsistent advice (Personal Data Protection Commission, 2022). Industry research has also shown that retrieval-based grounding improves factual reliability in enterprise AI systems, especially in legal and knowledge-intensive applications where accuracy is critical (Packowski et al., 2024). Jurisdictional control is another key advantage. General-purpose LLMs may generate guidance influenced by foreign regulatory frameworks such as the General Data Protection Regulation (GDPR), which may be inappropriate in the Singapore context. RAG mitigates this risk by restricting the retrieval corpus to Singapore-specific PDPA documents, ensuring that responses remain locally relevant and legally consistent (Organisation for Economic Co-operation and Development, 2022). RAG also supports maintainability. Because the external knowledge base can be updated independently of the model, new PDPC guidelines, decisions, or advisories can be integrated without requiring model retraining (Lewis et al., 2020). This modularity makes RAG particularly suitable for regulatory settings where requirements evolve over time. Overall, RAG provides a robust foundation for AI-assisted PDPA guidance by combining the generative flexibility of LLMs with the factual grounding necessary for compliance-critical tasks.

2.3.3 Ethics

AI systems designed for compliance must align with core PDPA obligations—including Purpose Limitation, Accuracy, Protection, and Retention Limitation—which require organisations to minimise the collection of personal data, implement appropriate safeguards, and ensure that information used to support decisions is accurate and fit for purpose (Personal Data Protection Commission, 2022). Privacy-by-design principles guide the system's architecture. User inputs are processed transiently, with no persistent server-side logging of identifiers. Where optional chat history is enabled, records are stored locally on the user's browser rather than transmitted to the backend. Following PDPC recommendations, the interface also includes guidance discouraging users from submitting NRIC numbers or other sensitive identifiers, reducing the risk of inadvertent data collection (Personal Data Protection Commission, 2022).

Research stresses the importance of technical guardrails—such as retrieval-based grounding, domain-specific prompting, and conservative generation settings—to mitigate misleading or non-jurisdictional outputs (Chen & Xiao, 2023). The platform incorporates controlled refusal behaviour for high-risk queries, including those involving identifiable personal data or legal questions outside Singapore's context. Responsible AI literature also highlights techniques such as validated knowledge sources, rate limiting, and transparency prompts as effective mitigation strategies in safety-critical systems (Google DeepMind, 2024). These principles inform the system's design, which limits data retention, enforces strict query filtering, and provides users with clear explanations of system scope and limitations. If future features—such as server-side chat history, organisational analytics, or multi-user administration—are introduced, additional safeguards will be required. These include encryption at rest, role-based access controls, access logging, and explicit retention and deletion policies to remain compliant with PDPA obligations (Personal Data Protection Commission, 2022).

3 System Design

To support our goal of helping Singaporean SMEs comply with the PDPA, we design a generation pipeline consisting of three steps: moderation, retrieval, and generation.

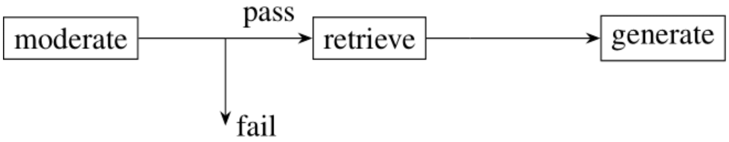


Figure 3.1: Pipeline representation of the system.

Given a user query, the moderation step checks whether the query is safe to handle, the retrieval step finds information relevant to the query, and the generation step crafts a response to the query. On a high level, our system consists of a frontend, backend, and infrastructure. The frontend handles the exposed logic, the backend handles the non-exposed logic, while the infrastructure hosts the backend and frontend.

3.1 Backend

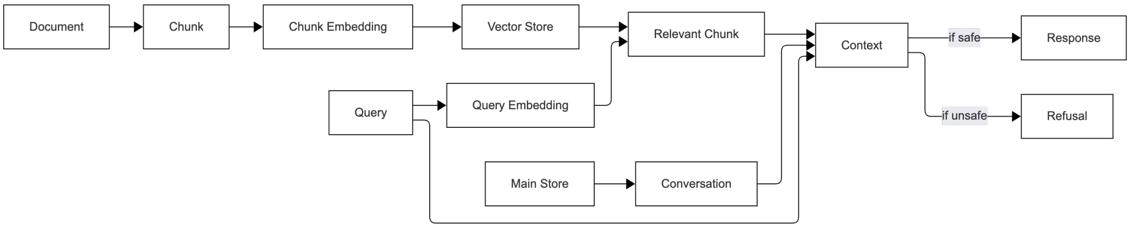


Figure 3.2: Graphical representation of the backend.

To ensure factuality, the retrieval step only retrieves information from documents defined as truthful. To ensure transparency, the generation step names each source from which information is retrieved.

Given a set of documents, the system performs the following steps offline.

1. **Chunking.** The system divides each document into smaller pieces called “chunks”. This is performed to make downstream information consumption more manageable.
2. **Embedding.** The system maps each chunk to a numerical representation capturing the semantics of the chunk, called an “embedding”. This is performed to allow for downstream semantic comparisons with user queries.
3. **Indexing.** The system stores each chunk and its corresponding embedding in a store optimised for embedding retrieval. This is performed to allow for efficient downstream retrieval of relevant chunks for a user query.

Given a user query, the system performs the following steps online.

1. **Moderation.** The system checks if the query seems safe to handle. If it seems unsafe, the pipeline returns a generic response (e.g. “Sorry, I cannot help with that.”) to the user; otherwise, the pipeline

moves to the next step. This is performed to protect the user and our system from any adverse effects that might result from sending an unsafe query to the chat model.

2. **Retrieval.** The system finds the most relevant chunks to the query by computing a similarity score between the query and each chunk, and returns the highest-scoring chunks. This is performed to retrieve relevant information for the query.
3. **Generation.** The system queries the chat model with the conversation, query, and chunks, and returns the response of the model.

We had to make several design decisions for our vector store, inference provider, embedding model, and chat model.

	Chroma	**Pinecone**	**Qdrant**
License class	Open	Closed	Open
GitHub stars	∼25k	∼3k	∼27k

Table 3.1: Comparison between Chroma, Pinecone, and Qdrant.

From our research, we found 3 vector store abstractions: Chroma, Pinecone, and Qdrant. We settled on Qdrant as the open license class gives greater developmental control, and the larger community support—as measured by the number of GitHub stars—facilitates debugging.

	GroqCloud	**OpenRouter**	**Vertex AI**
Relative model count	Smallest	Largest	In-between
Embedding model support?	No	Yes	Yes

Table 3.2: Comparison between GroqCloud, OpenRouter, and Vertex AI.

From our research, we found three inference provider abstractions: GroqCloud, OpenRouter, and Vertex AI. We settled on OpenRouter as its model support provides the greatest developmental flexibility.

	TE3-Small	**BGE-M3**	**QE-8B**
Price	\$0.02/M tokens	\$0.01/M tokens	\$0.01/M tokens
Context size	8,192	8,192	32,000

Table 3.3: Comparison between TE3-Small, BGE-M3, and QE-8B.

For our embedding model, we wanted a model with a context size greater than 512 tokens for input flexibility. Sorting the models officially supported by OpenRouter with that property by price, the top three models were OpenAI's Text Embedding 3 Small (TE3-Small), BAAI's BGE-M3 (BGE-M3), and Qwen's Qwen3 Embedding 8B (QE-8B). We settled on QE-8B, as the higher context size gives input flexibility.

	gpt-5-chat	**gemini-2.5-flash**	**llama-4-maverick**
Price	\$11.25/M tokens	\$2.80/M tokens	\$0.75/M tokens
Average latency	0.59s	0.88s	0.34s
Context size	128,000 tokens	1,024,576 tokens	1,024,576 tokens
Input modalities	Text, image, file	Text, image, file, audio, video	Text, image

Table 3.4: Comparison between gpt-5-chat, gemini-2.5-flash, and llama-4-maverick.

For our chat model, we wanted a model with an average latency within 1 second and throughput no less than 48 tokens per second for responsiveness, a context size no less than 128,000 tokens for in-conversation memory, input modalities including text and image for input flexibility, and a maximum output size no less than 8,000 tokens for output flexibility. Sorting the models officially supported by OpenRouter with those properties by price, the top three models were OpenAI's GPT 5 Chat (gpt-5-chat), Google's Gemini 2.5 Flash (gemini-2.5-flash), and Meta's Llama 4 Maverick (llama-4-maverick). We settled on these three models as candidates for more rigorous downstream evaluative ablations to decide model deployment, following the industry practice of architecting systems that facilitate model switching (Ng, 2025).

3.2 Frontend

The frontend exposes the PDPA chatbot to SME users through a web interface and aims to minimise interaction friction.

We considered two user interface (UI) interaction modes: anonymous access and authenticated access. Anonymous access requires no registration and lets users ask questions immediately, whereas authenticated access requires registration and lets users retain preferences and past conversations through user accounts.

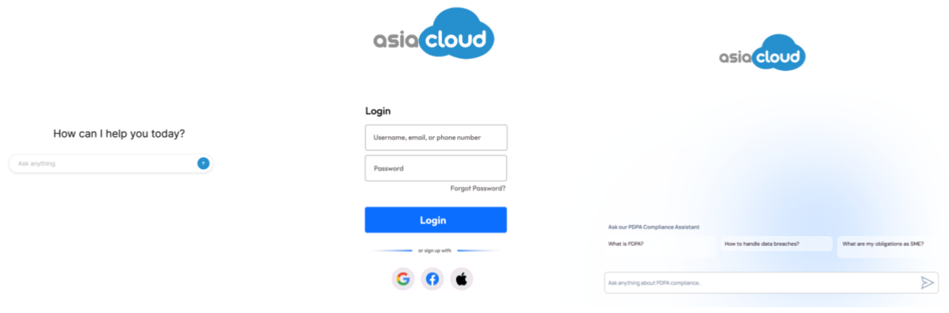

How can I help you today?
Ask anything
asia cloud
Login
Username, email, or phone number
Password
Forgot Password?
Login
or sign up with
asia cloud
Ask our PDPA Compliance Assistant
What is PDPA?
How to handle data breaches?
What are my obligations as SME?
Ask anything about PDPA compliance..

Figure 3.3: Sketches of the frontend UI.

Given SME needs and the scope of our minimum viable product, we adopt anonymous access to reduce barriers and shorten time-to-value, deferring account-based features to a later phase.

	Angular	**React**	**Vue.js**
Interruptible reconciliation support?	No	Yes	No
GitHub stars	∼100k	∼240k	∼210k

Table 3.5: Comparison between Angular, React, and Vue.js.

From our research, we found 3 frontend frameworks: Angular, React, and Vue.js. We settled on React as the support for interruptible reconciliation keeps interfaces responsive under load, and the larger community support—as measured by the number of GitHub stars—facilitates debugging.

	Angular	**React**	**Vue.js**
Interruptible reconciliation support?	No	Yes	No
GitHub stars	∼100k	∼240k	∼210k

Table 3.6: Comparison between Angular, React, and Vue.js.

From our research, we found 2 full-stack React frameworks: Next.js and Remix.

	Next.js	**Remix**
Incremental static regeneration support?	Yes	No
Built-in API route support?	Yes	No

Table 3.7: Comparison between Next.js and Remix.

Here, API refers to “application programming interface”.

We settled on Next.js as the support for incremental static regeneration improves load performance, and the support for built-in API routes simplifies server-side integration.

We design a modular component architecture to facilitate development. Routing and layout reside in the main application layer, and reusable interface elements are factored into separate components. Global state, including conversation history and UI configuration, is maintained in a central store built with Redux Toolkit. This separation of concerns improves readability and supports future extensions. Configuration is driven by environment variables that provide the backend API base uniform resource locator (URL) and related parameters. This allows the same codebase to run in development, staging, and production without code changes. To keep latency predictable and token usage bounded, the frontend limits each conversation to a fixed number of recent messages, discarding older turns from the request while retaining them in the on-screen history.

Given a user query, the frontend performs the following steps online.

1. **Input capture.** The user message is added to the local conversation state and rendered immediately in the chat interface.
2. **Request construction.** The frontend selects the most recent messages, then builds a request containing the conversation history, current query, retrieval settings, and model parameters.
3. **Streaming consumption.** The frontend opens a streaming connection to the backend and incrementally reads server-sent events. Each content fragment is cleaned, converted from Markdown to sanitised Hypertext Markup Language (HTML), and merged into the latest assistant message in the state store.
4. **Interface rendering.** As the state store updates, the interface re-renders the assistant message in real time, giving the impression of word-by-word generation. Conversation state, including errors, is persisted in browser storage, which supports multi-conversation views and future export features.
5. **Error handling.** Network and parsing errors are caught and surfaced as concise messages in the chat, so failures degrade gracefully and remain visible to users and developers.

3.3 Infrastructure

For deployment, the system requires a cloud platform that can host both the frontend and backend as containers, expose them securely to the internet, and scale with changes in user traffic.

From our research, we identified three cloud provider abstractions: Amazon Web Services, Microsoft Azure (Azure), and Google Cloud Platform (GCP).

	AWS	**Azure**	**GCP**
Market share	29% (Statista, 2025)	20% (Statista, 2025)	13% (Statista, 2025)
Global region count	36	60	42
Availability zone count	114	126	127
Minimum service count	200	200	100

Table 3.8: Comparison between AWS, Azure, and GCP.

AWS was selected for its large global footprint, mature support for containerised workloads, and close alignment with AsiaCloud's existing infrastructure, making it suitable for rapid prototyping and later production deployment.

Within AWS, we found six compute service abstractions from our research: Elastic Compute Cloud (EC2), Lambda, Fargate, App Runner, Lightsail, and Elastic Beanstalk.

	Billing metric	**Scaling metric**
EC2	Instance-hours	Instances
Lambda	Requests, GB-seconds	Executions
Fargate	vCPU-hours, GB-hours	Tasks
App Runner	vCPU-hours, GB-hours, requests	Instances
Lightsail	Bundled instance-hours	Instances
Elastic Beanstalk	Instance-hours, LCU-hours, GB-months	EC2 instances

Table 3.9: Billing and scaling metrics for AWS services.

Here, vCPU refers to "virtual central processing unit" hours, GB refers to gigabyte, and LCU refers to "load balancer capacity unit".

To handle dynamic traffic, integrate with autoscaling, and support future features, AWS App Runner was chosen as the primary deployment service, allowing the number of running instances to grow or shrink with demand.

To deploy the frontend and backend, the system performs the following steps.

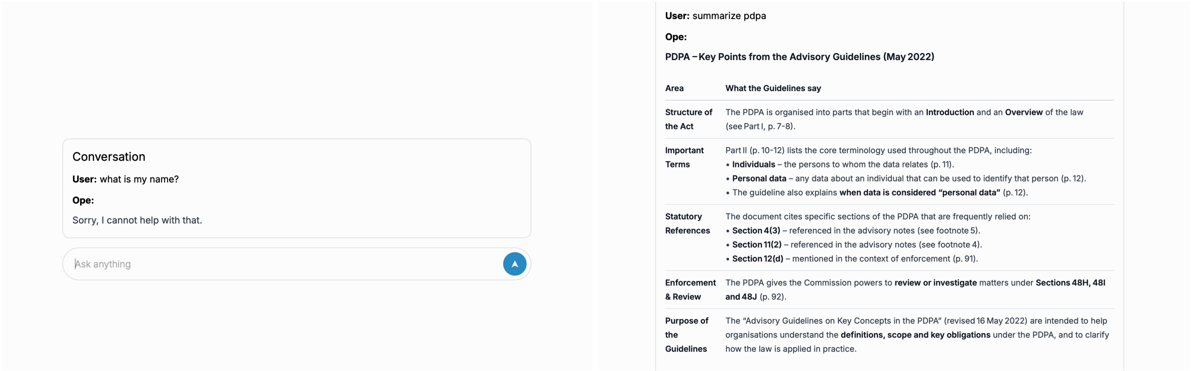
1. **Application containerisation.** Docker builds separate container images for the frontend and backend, capturing their runtime dependencies.
2. **Image storage.** Amazon Elastic Container Registry (ECR) stores the versioned images for deployment.
3. **Service deployment.** App Runner pulls the images from ECR, provisions compute instances, and exposes the services over a secure protocol.
4. **Operation scaling.** App Runner manages health checks, restarts failed instances, and performs basic autoscaling, which minimises operational overhead and lets us focus on application logic.

4 System Implementation

4.1 Term 7

Term 7 saw the creation of the first system implementations in two stages: Iteration 1 and Iteration 2. Iteration 2 builds upon Iteration 1 incrementally, to ensure a manageable growth of system complexity.

In Iteration 1, we developed our first prototype: Prototype 1.



Conversation

User: what is my name?

Ope:

Sorry, I cannot help with that.

User: summarize pdpa

Ope:

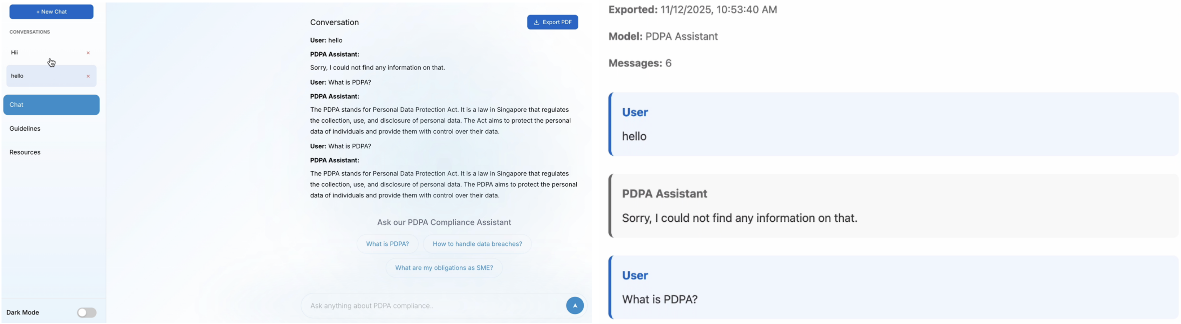
PDPA – Key Points from the Advisory Guidelines (May 2022)

Area	What the Guidelines say
Structure of the Act	The PDPA is organised into parts that begin with an **Introduction** and an **Overview** of the law (see Part I, p. 7-8).
Important Terms	Part II (p. 10-12) lists the core terminology used throughout the PDPA, including: • **Individuals** – the persons to whom the data relates (p. 11). • **Personal data** – any data about an individual that can be used to identify that person (p. 12). • The guideline also explains **when data is considered "personal data"** (p. 12).
Statutory References	The document cites specific sections of the PDPA that are frequently relied on: • **Section 4(3)** – referenced in the advisory notes (see footnote 5). • **Section 11(2)** – referenced in the advisory notes (see footnote 4). • **Section 12(d)** – mentioned in the context of enforcement (p. 91).
Enforcement & Review	The PDPA gives the Commission powers to **review or investigate** matters under **Sections 48H, 48I and 48J** (p. 92).
Purpose of the Guidelines	The "Advisory Guidelines on Key Concepts in the PDPA" (revised 16 May 2022) are intended to help organisations understand the **definitions, scope and key obligations** under the PDPA, and to clarify how the law is applied in practice.

Figure 4.1: Preview of Prototype 1.

The prototype showed clear strengths. It implemented the computational graph correctly on the backend and rendered the output of the chat model without token artefacts on the frontend. However, the prototype also had notable shortcomings. It did not maintain memory within conversations and spent unnecessary compute to rebuild the vector store on each run. Overall, Prototype 1 established a functional baseline but required several improvements to support a usable workflow.

In Iteration 2, we expanded Prototype 1 with a set of core functional upgrades to develop our second prototype: Prototype 2.



Conversation

User: hello

PDPA Assistant:

Sorry, I could not find any information on that.

User: What is PDPA?

PDPA Assistant:

The PDPA stands for Personal Data Protection Act. It is a law in Singapore that regulates the collection, use, and disclosure of personal data. The Act aims to protect the personal data of individuals and provide them with control over their data.

User: What is PDPA?

PDPA Assistant:

The PDPA stands for Personal Data Protection Act. It is a law in Singapore that regulates the collection, use, and disclosure of personal data. The PDPA aims to protect the personal data of individuals and provide them with control over their data.

Ask our PDPA Compliance Assistant
What is PDPA?
How to handle data breaches?
What are my obligations as SME?
Ask anything about PDPA compliance..

Exported: 11/12/2025, 10:53:40 AM

Model: PDPA Assistant

Messages: 6

User

hello

PDPA Assistant

Sorry, I could not find any information on that.

User

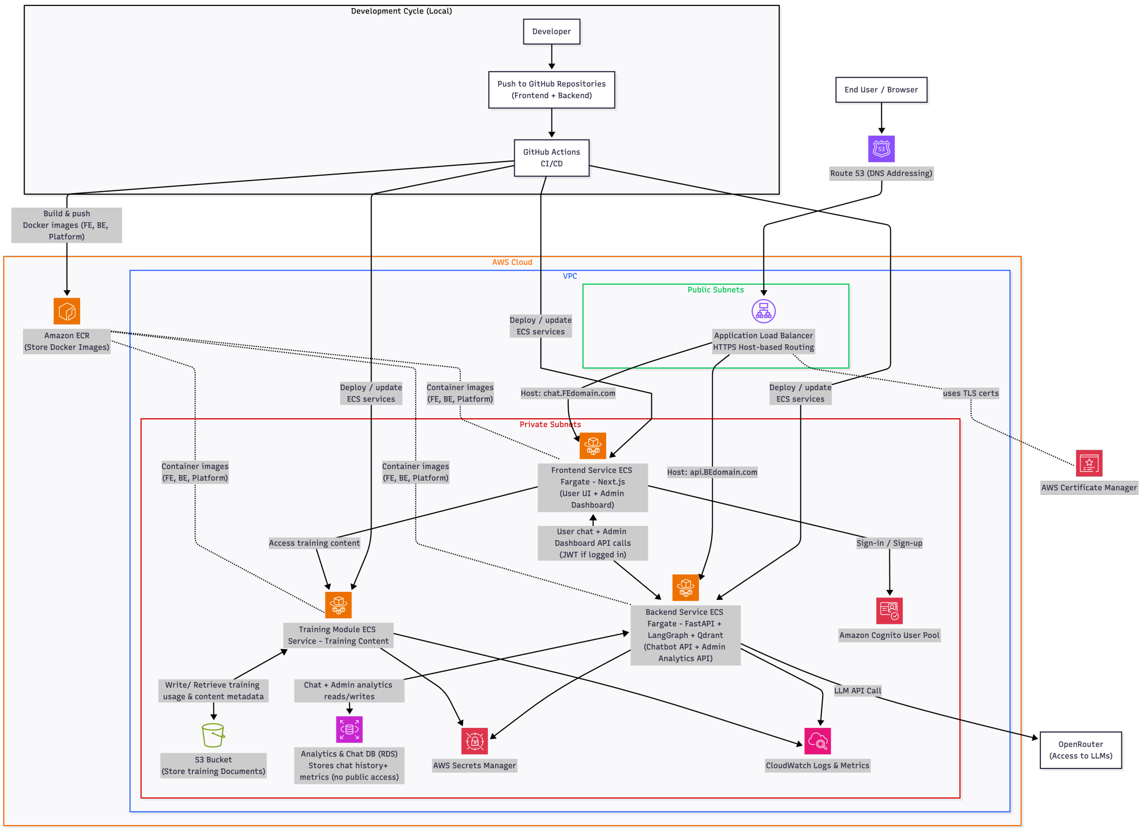
What is PDPA?

Figure 4.2: Preview of Prototype 2.

Prototype 2 introduced several backend improvements. The system ensured turns within a conversation were accumulated, by feeding past messages in the same conversation to the chat model as context on each turn. This made the chatbot more user-friendly, as it meant the chatbot could respond using past messages in the same conversation. The system also used a 256-Bit Secure Hash Algorithm (SHA-256) to generate a fingerprint based on information about the documents, chunking algorithm, and embedding model. This made system startup more efficient, as fingerprints were checked before vector store building. Prototype 2 also introduced several frontend improvements. The chat interface produced progressive, token-by-token responses. A consolidated account page enabled management of preferences, account details, and privacy settings. The interface was redesigned to be responsive across devices, and a dark–light mode toggle was added. Chats could also be exported as Portable Document Format (PDF) files, and past conversations were stored in the browser through LocalStorage. Overall, these additions formed a more complete and usable second prototype.

4.2 Term 8

Term 8 will focus on strengthening deployment reliability and expanding the system into a more complete PDPA compliance platform.



Development Cycle (Local)
Developer
Push to GitHub Repositories (Frontend + Backend)
GitHub Actions CI/CD
End User / Browser
Route 53 (DNS Addressing)
Build & push Docker images (FE, BE, Platform)
AWS Cloud
VPC
Public Subnets
Amazon ECR (Store Docker Images)
Deploy / update ECS services
Application Load Balancer HTTPS Host-based Routing
Deploy / update ECS services
Container images (FE, BE, Platform)
Host: chat.FEdomain.com
Deploy / update ECS services
uses TLS certs
Private Subnets
Container images (FE, BE, Platform)
Container images (FE, BE, Platform)
Frontend Service ECS Fargate - Next.js (User UI + Admin Dashboard)
Host: api.BEdomain.com
AWS Certificate Manager
Access training content
User chat + Admin Dashboard API calls (JWT if logged in)
Sign-in / Sign-up
Training Module ECS Service - Training Content
Backend Service ECS Fargate - FastAPI + LangGraph + Qdrant (Chatbot API + Admin Analytics API)
Amazon Cognito User Pool
Write/ Retrieve training usage & content metadata
Chat + Admin analytics reads/writes
LLM API Call
S3 Bucket (Store training Documents)
Analytics & Chat DB (RDS) Stores chat history+ metrics (no public access)
AWS Secrets Manager
CloudWatch Logs & Metrics
OpenRouter (Access to LLMs)

Figure 4.3: Planned system architecture.

An automated CI/CD pipeline will be introduced using GitHub Actions. Each push to the main branch will trigger automated tests, build container images, publish them to the registry, and update the running services on AWS App Runner. This ensures that only validated builds are deployed and reduces manual operational overhead while maintaining consistency between the codebase and the live environment.

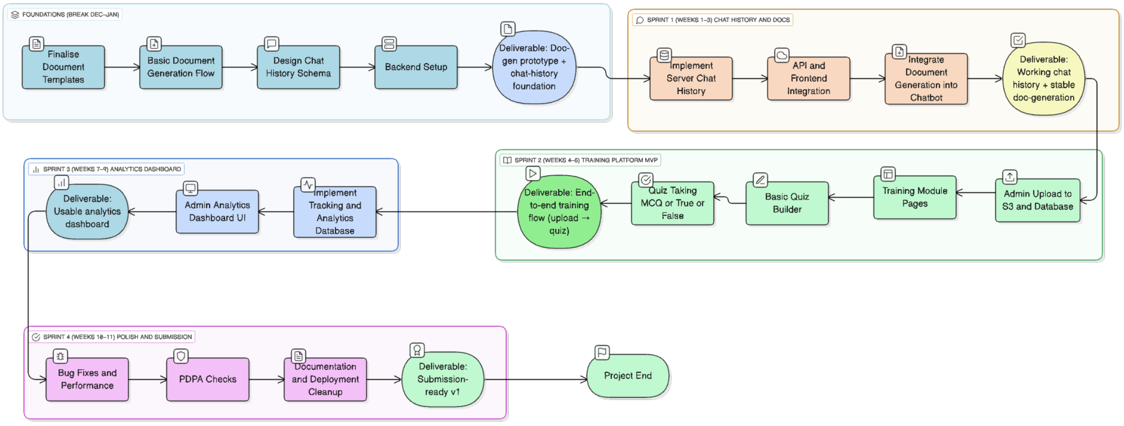

FOUNDATIONS (BREAK DEC–JAN)
Finalise Document Templates
Basic Document Generation Flow
Design Chat History Schema
Backend Setup
Deliverable: Doc-gen prototype + chat-history foundation
SPRINT 1 (WEEKS 1–3) CHAT HISTORY AND DOCS
Implement Server Chat History
API and Frontend Integration
Integrate Document Generation into Chatbot
Deliverable: Working chat history + stable doc-generation
SPRINT 2 (WEEKS 4–6) TRAINING PLATFORM MVP
Admin Upload to S3 and Database
Training Module Pages
Basic Quiz Builder
Quiz Taking MCQ or True or False
Deliverable: End-to-end training flow (upload → quiz)
SPRINT 3 (WEEKS 7–9) ANALYTICS DASHBOARD
Implement Tracking and Analytics Database
Admin Analytics Dashboard UI
Deliverable: Usable analytics dashboard
SPRINT 4 (WEEKS 10–11) POLISH AND SUBMISSION
Bug Fixes and Performance
PDPA Checks
Documentation and Deployment Cleanup
Deliverable: Submission-ready v1
Project End

Figure 4.4: Sprint timeline for Term 8.

In addition to deployment enhancements, Term 8 will extend the functional scope of the system. The platform will support persistent server-side chat history, enabling durable and auditable interactions. PDPA documents will be generated directly within the chat interface. A lightweight training module will allow organisations to upload materials, generate quizzes, and provide feedback to users based on their understanding. An analytics dashboard will surface common queries and usage patterns to support organisational insight and operational planning. The final phase of Term 8 will focus on performance improvements, PDPA validation, and deployment refinement to produce a stable first version suitable for evaluation by AsiaCloud and SME users.

5 System Validation

5.1 Verification

To validate system correctness, we created and conducted unit, integration, and end-to-end tests.

1. **Unit.** We conducted tests that checked the outputs of unitary functions on various inputs.
2. **Integration.** We conducted tests that exercised the full computational graph.
3. **End-to-end.** We conducted end-to-end user interaction simulations through the UI.

Across the weeks, we iteratively improved our system to ensure all tests passed.

5.2 Evaluation

To validate system quality, we created and ran evaluative benchmarks to study the capabilities of the models and our latest retrieval method. To do so, we needed to define a measure of performance. We defined an overall utility measure combining objective and subjective components. As a sum of Rational Utility and Subjective Utility, we define Overall Utility by

$$U_{\text{overall}} = U_{\text{rational}} + U_{\text{subjective}}$$

where the utility of a model can be seen as the degree to which it supports user retention and task success. We define Rational Utility by

$$U_{\text{rational}} = 440A + 220\left(\frac{20}{P + 7.43}\right)^{0.7} + 400M^{0.3} + 180\left(\frac{1}{1 + T\exp(T - 3)}\right)$$

where A represents accuracy, P represents price, M represents multimodality, and T represents latency. See Appendix A.1 for more information.

Rational Utility captures objective performance differences. It aggregates several measured categories, each weighted by a contribution constant derived from customer-centric factors such as intention, loyalty, and adoption. These constants were calculated using beta sensitivities scaled by one thousand.

On the other hand, Subjective Utility captures subjective performance differences. Two categories are used, each weighted by a utility constant of five hundred: the first assesses the comprehensibility of responses, while the second assesses their naturalness. For each category, the utility contribution is determined by the proportion of responses meeting the desired quality standard.

Term 7 saw the evaluation of the objective component of the benchmark. First, we crafted two datasets: PDPC22 and APDPC22. PDPC22 contains twenty-two question-answer pairs from an educational quiz-like game by the PDPC, and APDPC22 is an augmented version of PDPC22, which rephrases each question to adopt a more informal tone.

	PDPC22	**APDPC22**
True-false question count	12	12
Multiple-choice question count	10	10
Mean GPT-2 perplexity	62.47	75.84
Mean GPT-2 tokens per row	48.68	50.23

Table 5.1: Properties of PDPC22 and APDPC22.

Then, we examined the properties of PDPC22 and APDPC22, using Generative Pre-trained Transformer 2 (GPT-2) as a reference point because it is open-weights and widely used. Finally, we evaluated the chat models on PDPC22 and APDPC22, with and without our latest retrieval method.

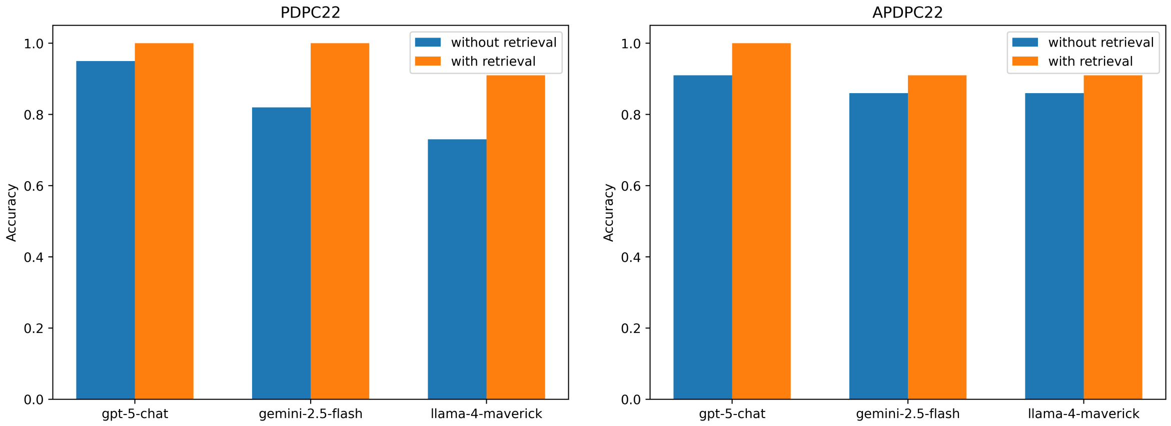


Figure 5.1: Results for PDPC22 and APDPC22.

Across all models, retrieval consistently improved accuracy, though to different degrees. This appears to validate our retrieval method, as it seems to provide a reliable performance boost across different model capabilities.

From these preliminary results, gemini-2.5-flash appears to be the strongest candidate. Term 8 will repeat this experiment with a wider set of forty questions to confirm trends and derive stable accuracy estimates for the Rational Utility calculation, then expand evaluation to subjective metrics through internal testing and, when possible, external trials with SMEs.

6 Conclusion

6.1 Achievements

Term 7 focused on establishing the foundation for a safe, accurate, and scalable PDPA compliance assistant. We validated the problem through literature review, market analysis, and discussions with AsiaCloud, confirming that SMEs face fragmented resources, limited expertise, and a lack of affordable PDPA tools. This validation informed the design of the core system architecture and the selection of the frontend, backend, and cloud technologies.

An initial chatbot prototype was developed with basic retrieval capabilities, allowing PDPA-aligned responses grounded in official PDPC documents. Two iterations improved UI design, retrieval quality, safety behaviour, and early features such as conversation storage, PDF export, and refusal logic. Benchmarking of three LLM candidates was conducted using a preliminary RAG pipeline to reduce hallucination rates.

From an engineering standpoint, Term 7 included setting up the cloud deployment pathway, comparing AWS compute options, and implementing a basic continuous integration workflow. The team also drafted the structure and data flow for the training platform. These efforts provided the technical clarity and feasibility checks required for full implementation in Term 8.

6.2 Insights

First, we found that retrieval-based grounding is essential for accuracy. Early tests showed that even strong models occasionally mixed PDPA with foreign regulations such as the GDPR, underscoring the need for a tightly scoped retrieval corpus and jurisdiction-specific prompting. This guided both our model-selection criteria and the structure of the RAG pipeline.

Second, discussions with AsiaCloud highlighted the importance of modularity for future commercialisation. Each component—chatbot, document templates, training modules, and analytics—must operate independently so AsiaCloud can scale or monetise them selectively.

Risk assessment exercises also reinforced the need for clear safeguards, including refusal behaviour, input sanitisation, and strict data minimisation aligned with PDPA requirements. These considerations influenced prompting strategy, guardrails, and the move toward controlled server-side storage.

Finally, AWS exploration showed that managed compute services such as App Runner offer a good balance of performance, scalability, and operational simplicity. This became the foundation for the Term 8 deployment plan.

6.3 Budget Summary

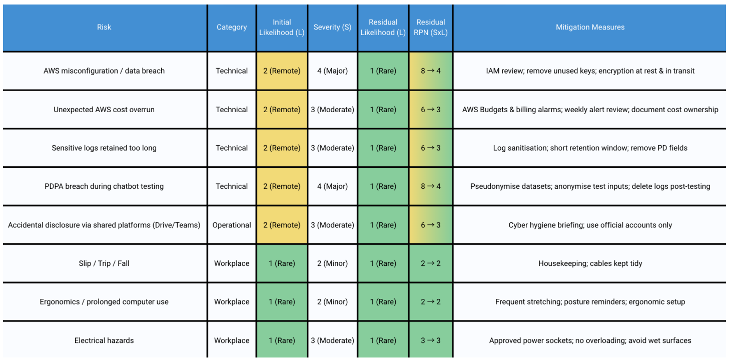
	Unit cost	Units	Total cost
Transportation	SGD 30/cab	16 cabs	SGD 480
Backend	SGD 15/month	7 months	SGD 105
Infrastructure	SGD 108.75/month	7 months	SGD 756

Table 6.1: Project budget summary.

The project's operating costs remain intentionally low to ensure that the platform is viable for SME adoption and long-term sustainability. Based on the current architecture, AWS infrastructure costs amount to approximately SGD 108 per month, covering App Runner compute, storage, load balancing, and operational monitoring. LLM usage through OpenRouter adds an estimated SGD 15 per month, bringing the total estimated monthly cost to roughly SGD 124 per month. One-time expenditures were minimal, with transport and logistics forming the primary non-technical cost. Overall, the cost profile aligns with AsiaCloud's requirement for a lightweight and affordable proof-of-concept that can be scaled or monetised in later phases without significant infrastructure burden.

See Appendix A.2 for more information.

6.4 Risk Assessment



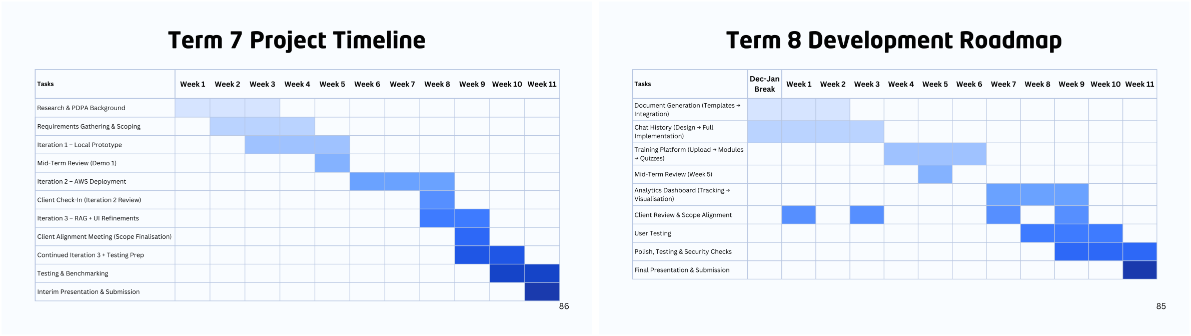
Risk	Category	Initial Likelihood (L)	Severity (S)	Residual Likelihood (L)	Residual RPN (SxL)	Mitigation Measures
AWS misconfiguration / data breach	Technical	2 (Remote)	4 (Major)	1 (Rare)	8 → 4	IAM review; remove unused keys; encryption at rest & in transit
Unexpected AWS cost overrun	Technical	2 (Remote)	3 (Moderate)	1 (Rare)	6 → 3	AWS Budgets & billing alarms; weekly alert review; document cost ownership
Sensitive logs retained too long	Technical	2 (Remote)	3 (Moderate)	1 (Rare)	6 → 3	Log sanitisation; short retention window; remove PD fields
PDPA breach during chatbot testing	Technical	2 (Remote)	4 (Major)	1 (Rare)	8 → 4	Pseudonymise datasets; anonymise test inputs; delete logs post-testing
Accidental disclosure via shared platforms (Drive/Teams)	Operational	2 (Remote)	3 (Moderate)	1 (Rare)	6 → 3	Cyber hygiene briefing; use official accounts only
Slip / Trip / Fall	Workplace	1 (Rare)	2 (Minor)	1 (Rare)	2 → 2	Housekeeping; cables kept tidy
Ergonomics / prolonged computer use	Workplace	1 (Rare)	2 (Minor)	1 (Rare)	2 → 2	Frequent stretching; posture reminders; ergonomic setup
Electrical hazards	Workplace	1 (Rare)	3 (Moderate)	1 (Rare)	3 → 3	Approved power sockets; no overloading; avoid wet surfaces

Figure 6.1: Project risk assessment.

A comprehensive risk assessment was conducted to evaluate technical, operational, and project-execution risks that could affect system performance, data protection, or delivery timelines. Key technical risks included API rate limits, model hallucination, cloud misconfiguration, and retrieval errors, all of which were mitigated through Identity and Access Management (IAM) reviews, billing alerts, retrieval constraints, and controlled refusal behaviour. Operational risks—such as accidental exposure of personal data through user inputs or unsafe document uploads—were addressed using anonymised test cases, input sanitisation rules, and clear user-facing warnings. Workplace and project risks, including dependency on external services or bottlenecks in development capacity, were moderated through modular architecture design, shared code ownership, and an iterative sprint structure. After mitigation, all risks achieved low residual scores, providing a stable foundation for Term 8 development and deployment.

See Appendix **??** for more information.

6.5 Outlook



Term 7 Project Timeline

Tasks	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11
Research & PDPA Background	■	■	■								
Requirements Gathering & Scoping		■	■	■							
Iteration 1 – Local Prototype			■	■	■						
Mid-Term Review (Demo 1)					■						
Iteration 2 – AWS Deployment						■	■	■			
Client Check-In (Iteration 2 Review)								■			
Iteration 3 – RAG + UI Refinements								■	■		
Client Alignment Meeting (Scope Finalisation)									■		
Continued Iteration 3 + Testing Prep									■	■	
Testing & Benchmarking										■	■
Interim Presentation & Submission											■

Term 8 Development Roadmap

Tasks	Dec-Jan Break	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11
Document Generation (Templates → Integration)	■	■	■									
Chat History (Design → Full Implementation)	■	■	■	■								
Training Platform (Upload → Modules → Quizzes)					■	■	■					
Mid-Term Review (Week 5)						■						
Analytics Dashboard (Tracking → Visualisation)								■	■	■		
Client Review & Scope Alignment		■		■				■		■		
User Testing									■	■	■	
Polish, Testing & Security Checks										■	■	■
Final Presentation & Submission												■

Figure 6.2: Outlooks for Term 7 and Term 8.

Through iterative exploration in frontend, backend, cloud infrastructure, testing, and benchmarking, the team clarified the system architecture and identified the most feasible approach for a scalable compliance assistant. Early testing also confirmed the value of retrieval-grounded responses and highlighted key areas—such as chat history design, training workflows, and analytics requirements—that will shape the next phase. Looking ahead, Term 8 will focus on completing the core platform components: server-based chat history, PDPA document generation, the training module, and the analytics dashboard. Development will follow the planned sprint structure, beginning with stabilising the chatbot and document flows, followed by delivering a functional training platform and a multi-metric analytics dashboard for AsiaCloud. The final weeks will be dedicated to system hardening, evaluation, and preparing deployment-ready documentation. Overall, the project remains on track to deliver a modular and cost-efficient proof-of-concept that supports SMEs in navigating PDPA obligations, while providing AsiaCloud with a flexible foundation for future commercial development.

References

Al-Hattami, H. M. (2025). Empowering business research with chatgpt: Academic and student insights through the UTAUT lens [Article 179]. *Discover Computing*, *28*(1). https://doi.org/10.1007/s10791-025-09692-1

Bello, H. O., Idemudia, C., & Iyelolu, T. V. (2024). Navigating financial compliance in small and medium-sized enterprises (smes): Overcoming challenges and implementing effective solutions. *World Journal of Advanced Research and Reviews*. https://doi.org/10.30574/wjarr.2024.23.1.1984

Bleach, T. (2024). Using AI to streamline compliance processes: The future or could too much go wrong? *The Fintech Times*. https://thefintechtimes.com/using-ai-to-streamline-compliance-processes-the-future-or-could-too-much-go-wrong/

Chan, E. (2025). Marina bay sands fined sgd 315000 over data breach that affected more than 665,000 customers. *Channel NewsAsia*. https://www.channelnewsasia.com/singapore/marina-bay-sands-mbs-fined-data-breach-customers-affected-dark-web-leak-pdpa-pdpc-5429346

Chen, P.-Y., & Xiao, C. (2023). Trustworthy ai in the era of foundation models [CVPR 2023 Tutorial]. *CVPR 2023*. https://research.ibm.com/publications/trustworthy-ai-in-the-era-of-foundation-models

Choudhury, A. R. (2024, October). Singapore's smart nation 2.0 policy focuses on ai and building resilience. https://govinsider.asia/intl-en/article/singapores-smart-nation-20-policy-focuses-on-ai-and-building-resilience

Cisco Systems. (2023). *Privacy's growing importance and impact: Cisco 2023 data privacy benchmark study* (tech. rep.). Cisco Systems. https://www.cisco.com/c/en/us/about/trust-center/data-privacy-benchmark-study.html

Compliance Consultant. (2025, January). Common regulatory compliance challenges for smes. https://complianceconsultant.org/common-regulatory-compliance-challenges-for-smes/

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: Empirical results. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Google DeepMind. (2024). Facts grounding: A new benchmark for evaluating the factuality of large language models [Google DeepMind blog]. https://deepmind.google/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/

Government of the Republic of Singapore. (2023). AI for the public good for singapore and the world. https://file.go.gov.sg/nais2023.pdf

Gültekin-Várkonyi, G. (2025). AI literacy for legal AI systems: A practical approach [Forthcoming; preprint arXiv:2505.18006]. *Iustum Aequum Salutare*, *21*. https://doi.org/10.48550/arXiv.2505.18006

i-Sprint Innovations Pte Ltd. (2024, June). Key challenges in achieving PDPA compliance in 2024. https://ismartcom.com/blog/key-challenges-in-achieving-pdpa-compliance-in-2024/

Kim, Y., Blazquez, V., & Oh, T. (2024). Determinants of generative AI system adoption and usage behavior in Korean companies: Applying the UTAUT model. *Behavioral Sciences*, *14*(11), 1035. https://doi.org/10.3390/bs14111035

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Lim, E. (2025). Sg60: How singapore's smes are shaping a sustainable future for asean. *The Straits Times*. https://www.straitstimes.com/singapore/sg60-how-singapores-smes-are-shaping-a-sustainable-future-for-asean

Lonzetta, A. M., & Hayajneh, T. (2020). Challenges of complying with data protection and privacy regulations. *EAI Endorsed Transactions on Scalable Information Systems*, *8*(30), e4. https://doi.org/10.4108/eai.26-5-2020.166352

Navarrete, S. (2019, August). User study: Crm software adoption in the uk. https://www.capterra.co.uk/blog/854/user-survey-crm-software-adoption-in-the-uk

Ng, A. (2025). Architecting multi-agent systems with andrew ng [YouTube; Sapphire Ventures]. Retrieved November 30, 2025, from https://www.youtube.com/watch?v=yi7doi-QGJI

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. https://arxiv.org/abs/2303.08774

Organisation for Economic Co-operation and Development. (2022). *Oecd framework for the classification of AI systems* (OECD Digital Economy Papers No. 323). Organisation for Economic Co-operation and Development. Paris. https://doi.org/10.1787/cb6d9eca-en

Packowski, S., Halilovic, I., Schlotfeldt, J., & Smith, T. (2024). Optimizing and evaluating enterprise retrieval-augmented generation (rag): A content design perspective [arXiv:2410.12812]. https://arxiv.org/abs/2410.12812

Personal Data Protection Commission. (2017, December). Re credit counselling singapore [2017] SGPDPC 18. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Commissions-Decisions/GroundsofDecisionCreditCounsellingSingapore291217.pdf

Personal Data Protection Commission. (2022). Advisory guidelines on key concepts in the Personal Data Protection Act [Issued 23 September 2013; revised 16 May 2022]. https://www.pdpc.gov.sg/guidelines-and-consultation/2020/03/advisory-guidelines-on-key-concepts-in-the-personal-data-protection-act

Personal Data Protection Commission. (2023). PDPA overview [Last updated 3 November 2023]. https://www.pdpc.gov.sg/overview-of-pdpa/the-legislation/personal-data-protection-act

Personal Data Protection Commission. (2025). Kick-starting your data protection journey [Last updated 6 October 2025]. https://www.pdpc.gov.sg/dp-professional/kick-start-your-dp-journey

Personal Data Protection Commission of Singapore. (2015, September). *Industry survey on the Personal Data Protection Act 2015* (tech. rep.). Personal Data Protection Commission of Singapore. Singapore. https://www.pdpc.gov.sg/help-and-resources/2017/10/industry-survey-on-the-personal-data-protection-act-2015

PwC. (2023, March). Pwc announces strategic alliance with harvey, positioning PwC's legal business solutions at the forefront of legal generative AI. https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-announces-strategic-alliance-with-harvey-positioning-pwcs-legal-business-solutions-at-the-forefront-of-legal-generative-ai.html

SMU Social Media Team. (n.d.). What is legal artificial intelligence (AI) and how will it affect the next generation of legal professionals? https://masters.smu.edu.sg/what_legal_artificial_intelligence_ai_and_how_will_it_affect_next_generation_of_legal_professionals

Statista. (2025). Worldwide market share of leading cloud infrastructure service providers [Accessed 30 November 2025]. https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/

Tanantong, T., & Wongras, P. (2024). A UTAUT-based framework for analyzing users' intention to adopt artificial intelligence in human resource recruitment: A case study of Thailand. *Systems*, *12*(1), 28. https://doi.org/10.3390/systems12010028

Vaniukov, S. (2024, February). NLP vs LLM: A comprehensive guide to understanding key differences. https://medium.com/@vaniukov.s/nlp-vs-llm-a-comprehensive-guide-to-understanding-key-differences-0358f6571910

Zhang, L., Yu, J., Zhang, S., Li, L., Zhong, Y., Liang, G., Yan, Y., Ma, Q., Weng, F., Pan, F., Li, J., Xu, R., & Lan, Z. (2024). Unveiling the impact of multi-modal interactions on user engagement: A comprehensive evaluation in AI-driven conversations. *CoRR*, *abs/2406.15000*. https://doi.org/10.48550/ARXIV.2406.15000

Zul. (2022, November). *Data protection is vital: 85% of Singaporeans concerned about how companies use their data*. TechWire Asia. Retrieved November 28, 2025, from https://techwireasia.com/2022/11/data-protection-is-vital-85-of-singaporeans-concerned-about-how-companies-use-their-data/

A Supplementary Material

A.1 Rational Utility

Rational Utility, as defined by

$$U_{\text{rational}} = 440A + 220\left(\frac{20}{P + 7.43}\right)^{0.7} + 400M^{0.3} + 180\left(\frac{1}{1 + T\exp(T - 3)}\right)$$

decomposes into four objective components: accuracy, price, multimodality, and latency. Each term is a transformed version of a measured quantity, scaled by a contribution constant chosen from prior work on customer intention, loyalty, or adoption. The transformation shape and constant together determine how strongly that quantity influences the final score and where diminishing returns begin.

The first term captures accuracy. Here $A \in \mathbb{R}_{[0,1]}$ is the empirical accuracy of a model on an evaluation dataset. Its contribution is scaled using the largest constant in the framework, reflecting findings that accuracy is the strongest predictor of user intention and retention in service environments (Al-Hattami, 2025). In the context of legal and compliance assistance, factual correctness is especially critical, so accuracy is treated as the dominant factor. Full accuracy yields the maximum contribution allocated to this dimension, while partial accuracy reduces the score proportionally.

The second term captures pricing utility. Here, P is the price per million tokens of combined input and output, as listed on OpenRouter. The functional form is decreasing in P: as a model becomes more expensive to run, its contribution to Rational Utility falls. The exponent induces diminishing sensitivity, so large price differences at the low end matter more than equally large differences at the high end, consistent with behavioural findings on price elasticity. The contribution constant is derived from reported beta sensitivities for price effects on customer concentration, and is capped such that pricing cannot contribute more utility than accuracy (Tanantong & Wongras, 2024). This ensures that an extremely cheap but inaccurate model does not outrank a moderately priced, highly accurate one.

The third term measures the benefit of multimodality. Here, M is the number of input modalities, as listed on OpenRouter. Empirical work suggests that the availability of additional interaction channels modestly improves customer engagement and concentration (Zhang et al., 2024). We encode this with a sublinear exponent, which yields strong gains from adding the first few modalities but quickly introduces diminishing returns. The contribution constant reflects an upper bound comparable to, but slightly below, the weight of accuracy, acknowledging that modality breadth is valuable but should not overshadow factual performance.

The final term reflects latency utility. Here, T is the latency in seconds, as listed on OpenRouter. The chosen functional form behaves similarly to a squashed inverse curve: utility is high and near-saturated for low latencies, then declines sharply as response times exceed a few seconds, and eventually flattens out for very slow models. This shape aligns with user studies showing that delays beyond a small threshold significantly degrade perceived quality and willingness to continue using a system, while shaving off sub-second latency yields smaller marginal gains (Kim et al., 2024). The contribution constant is calibrated from reported beta sensitivities for time-related service attributes, using the mid-range estimate as a compromise between different studies.

Together, these four terms produce a composite Rational Utility score that balances factual quality, cost, interaction richness, and responsiveness. The relative magnitudes of the contribution constants ensure that accuracy remains primary, while still rewarding models that are affordable, multimodal, and fast enough to support practical SME workflows.

A.2 Budget Summary

We used the AWS Pricing Calculator to determine the infrastructure budget allocation.