TVS Sidekick: Challenges and Practical Insights from Deploying Large Language Models in the Enterprise

Paula Reyero Lobo[1,2], **Kevin Johnson**[1], **Bill Buchanan**[1], **Matthew Shardlow**[2], **Ashley Williams**[2], **Samuel Attwood**[2]
[1] TVS Supply Chain Solutions, Chorley, UK
[2] Manchester Metropolitan University, Manchester, UK
{P.ReyeroLobo, M.Shardlow, Ashley.Williams, S.Attwood}@mmu.ac.uk
{kevin.johnson, bill.buchanan}@tvsscs.com

Abstract

Many enterprises are increasingly adopting Artificial Intelligence (AI) to make internal processes more competitive and efficient. In response to public concern and new regulations for the ethical and responsible use of AI, implementing AI governance frameworks could help to integrate AI within organisations and mitigate associated risks. However, the rapid technological advances and lack of shared ethical AI infrastructures creates barriers to their practical adoption in businesses. This paper presents a real-world AI application at TVS Supply Chain Solutions, reporting on the experience developing an AI assistant underpinned by large language models and the ethical, regulatory, and sociotechnical challenges in deployment for enterprise use.

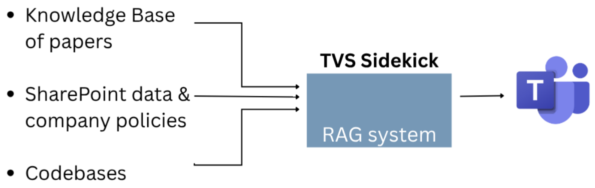

Figure 1: Overview of TVS Sidekick, an AI assistant that leverages LLMs to answer queries with relevant enterprise data using retrieval augmented generation (RAG) via a Microsoft Teams extension.

1 Introduction

Recent developments are driving industry interest in the field of Large Language Models (LLMs). Key developments of note are the abundant availability of commercial language modelling solutions (Devlin et al., 2019; Brown et al., 2020; Thoppilan et al., 2022) and the increased public awareness of the capabilities of LLMs (Mialon et al., 2023; Qu et al., 2025). However, to successfully utilise these models, organisations must navigate important societal challenges related to ethics, sustainability, and compliance (Hagendorff, 2024; Laux et al., 2024).

TVS SCS UK is a top-tier third-party logistics (3PL) provider in Europe and the UK, offering comprehensive supply chain solutions. 3PL customers increasingly adopt intelligent technology-led solutions to optimise their supply chain operations and reduce costs (Pournader et al., 2021; Li et al., 2023). To stay ahead of the competition, TVS SCS UK are leveraging LLMs to create a competitive advantage and enhance their internal operational efficiency. TVS SCS UK has decided not to use third-party software integrators or product vendors for its solutions, which would negatively impact their agility and innovation. Instead, they have started their journey towards an AI transformation through an in-house AI team.

TVS Sidekick is the flagship product of this in-house team. TVS Sidekick is built upon the principles of Retrieval Augmented Generation (RAG) (Lewis et al., 2020). All relevant company documents, as available via their internal cloud-based systems, are vectorised and compared to the input query, with the LLM then performing information extraction for the purposes of question answering with custom prompting (Qu et al., 2025). Users interact with TVS Sidekick via a Microsoft Teams extension (Figure 1).

As TVS SCS UK advances its AI transformation through the development of Sidekick, it must also navigate a complex legal and regulatory landscape. At the centre of this landscape are the European Union Artificial Intelligence Act (EU AIA) (European Commission, 2014) and related standards, such as ISO/IEC 42001 for AI Management Systems (International Organization for Standardization., 2023). Furthermore, TVS SCS UK must overcome a range of sociotechnical challenges that accompany the deployment of LLMs, such as issues of fairness, transparency, and accountability

Proceedings of the First Workshop on Ethical Concerns in Training, Evaluating and Deploying LLM associated with RANLP 2025, pages 17–26, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-103-5-003

(Crockett et al., 2023; Ojewale et al., 2025), which limit their practical adoption in enterprise environments.

In this paper, we report on TVS SCS UK's experience developing Sidekick, navigating the relevant legislation and regulations, and overcoming the challenges they have encountered along the way.

1.1 Significance of this Study

This study presents practical insights from applied AI research in a real-world business context. To be specific, we contribute to the field in three ways:

- *Technical Contributions.* We describe the design and implementation of Sidekick, an AI assistant underpinned by LLMs that is tailored for enterprise use, including novel approaches to prompt engineering and RAG.

- *Regulatory Contributions.* We present a case study of how a business is aligning its development with emerging legislation and regulations, most notably the EU AIA, by working towards harmonised technical standards (e.g., ISO/IEC 42001).

- *Sociotechnical Contributions.* We explore the sociotechnical challenges that accompany the deployment of LLMs in enterprise environments. We report quantitative statistics relating to the adoption of Sidekick alongside a qualitative analysis of end-user feedback.

1.2 Structure of this Study

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 details the technical implementation of Sidekick. Section 4 presents a case study of how TVS SCS UK is aligning its development with emerging legislation and regulations. Section 5 includes a quantitative and qualitative evaluation of the progress to date. Finally, section 6 concludes this paper and describes directions for future work.

2 Related Work

2.1 LLMs in the Enterprise

Prominent applications, reviewing strategies to augment LLM capabilities. The transformer architecture enhanced language modelling capabilities and has since sparked great attention in industry (Vaswani et al., 2017). This led to many readily available pre-trained models, which proved their superiority in fine-tuning applications (Devlin et al., 2019). With increased data size and model complexity, decoder-only models like the generative pre-trained transformer (GPT) model series have become more attractive for industry due to their few/zero-shot performance (Brown et al., 2020). This paradigm shift led to methods for aligning to user intent (Ouyang et al., 2022) (like reinforcement learning with human feedback) powering popular conversation-focused products like ChatGPT. While these scaled-up models offer business value (e.g. analysing vast data in real-time), issues such as the closed-source nature of existing solutions creates barriers to organisations lacking computational power (Yang et al., 2024).

Focus on approaches including RAG (and pipeline parts showing improvement). Recently, the focus has turned into giving more agency to LLMs to become independent problem solvers. For instance, by consulting with external knowledge sources for factual grounding (Lewis et al., 2020; Thoppilan et al., 2022). More broadly, a significant step forward is the combination of "tools", namely tool-augmented LLMs (Mialon et al., 2023), including retrieval-augmented language models for efficiently handling new data. Such approaches generally consist of four stages: task planning (i.e. break down user query into tasks), tool selection, tool calling, and response generation (Qu et al., 2025). Similarly, critical advances require frameworks for enabling LLMs to recall previous interactions (Zhang et al., 2024), allowing for multimodal data processing (Sun et al., 2025; Song et al., 2025), or to improve responses based on past interactions (Wang et al., 2024).

This paper presents a case study of recent LLM developments in practice, specifically through the technical implementation of an AI assistant that processes heterogeneous enterprise data sources using knowledge augmentation strategies, including novel approaches to prompt engineering and RAG.

2.2 Responsible and Ethical AI

Challenges in training, evaluating, and deploying LLMs and emerging AI regulation. While AI shows great potential and business opportunities, many concerns arise from embedding biases, contributing to climate degradation, threatening human rights and more (UNESCO, 2021). An active research area has emerged for responding hard normative questions related to AI, such as bias and

Principles	Requirements
Human oversight & accountability	AI to support/augment humans, with humans clearly accountable.
Technical robustness and safety	AI tools work as expected, minimising potential harms.
Transparency	Clear notification of AI involvement, clear and traceable outputs.
Privacy & data governance	Follow existing privacy rules with quality, robust data.
Diversity & fairness	Output free of bias and does not discriminate or treat unfairly.
Social & environmental wellbeing	AI is sustainable and beneficial to all.

Table 1: Key emerging principles and requirements from global AI regulations (British Standards Institution, 2025).

fairness, transparency, and accountability (Jobin et al., 2019). Institutions at global, international, and national levels have responded with recommendations for responsible and ethical AI, consisting of principles and practices such as a human rights-centred approach to AI (UNESCO, 2021), or AI assurance methodologies (i.e. to "measure, evaluate, and communicate the *trustworthiness* of AI systems" (Department for Science, Innovation & Technology, 2024)). The advent of LLMs only adds a layer of complexity to the ethical debate (Hagendorff, 2024), raising additional concerns (regarding transparency, copyright, and safety) (European Commission, 2025) that require specific regulation for generative AI technologies.

***Global legislation and EU AIA as most far reaching and punitive of regulations*.** The EU AIA is a notable example leading the field of AI regulation, with significant non-compliance penalties to business providing or deploying AI. While legislation approaches and requirements vary across jurisdiction areas (Table 1), AI regulations are developing globally to provide assurances in critical aspects such as human oversight and accountability, technical robustness and safety, or privacy and data governance (British Standards Institution, 2025).

Governments and legislative bodies are working towards practical strategies to implement the principles underlying AI regulations. Harmonised standards are one of the primary mechanisms for helping organisations translate regulatory requirements into technical implementations (AI Standards Hub, 2024). Standardisation should specify minimum technical testing, documentation, and public reporting to limit AI developers and/or users discretion in complying with regulatory requirements (Laux et al., 2024). However, local empirical studies and specific examples of how organisations implement processes that ensure AI regulation principles (Wolf-Brenner et al., 2024) is crucial for a democratic approach to ethical and responsible AI.

From theory to practice. While approaches to ethical AI exist (including bias tests, checklists and risk impact assessments), organisations face barriers that limit their practical adoption (Crockett et al., 2023). Technical approaches alone are not sufficient to establish an ethical AI infrastructure (Ojewale et al., 2025). Instead, participatory approaches involving civil society stakeholders are needed for effective standard setting, implementation, and enforcement (Crockett et al., 2024; Modhvadia et al., 2025). This paper contributes to bridging the gap between theory and practice through the experience of implementing an AI governance strategy in a real-world business context, reporting on the technical, legal and human challenges involved with the adoption of generative AI technologies.

2.3 Positioning this Study

In the logistics sector, real-time data analysis can transform business operations, from internal warehousing and inventory processes to stakeholder management (Pournader et al., 2021). However, empirical research in related areas (Qian et al., 2024; Kapania et al., 2025) shows that benefits and trade-offs in the use of AI technologies manifest differently depending on their application domain.

Despite growing understanding of public attitudes towards AI (Modhvadia et al., 2025; Mhasakar et al., 2025), research on its industrial application remains limited. This study presents insights from the development and use of LLMs at TVS SCS UK, to address the following gaps:

- Examining the implementation and practical application of recent LLM advances within the enterprise context.
- Embedding high-level ethical principles in AI regulatory frameworks into organisational practices.
- Empirical analysis of challenges that emerge with adopting LLMs in a logistics company.

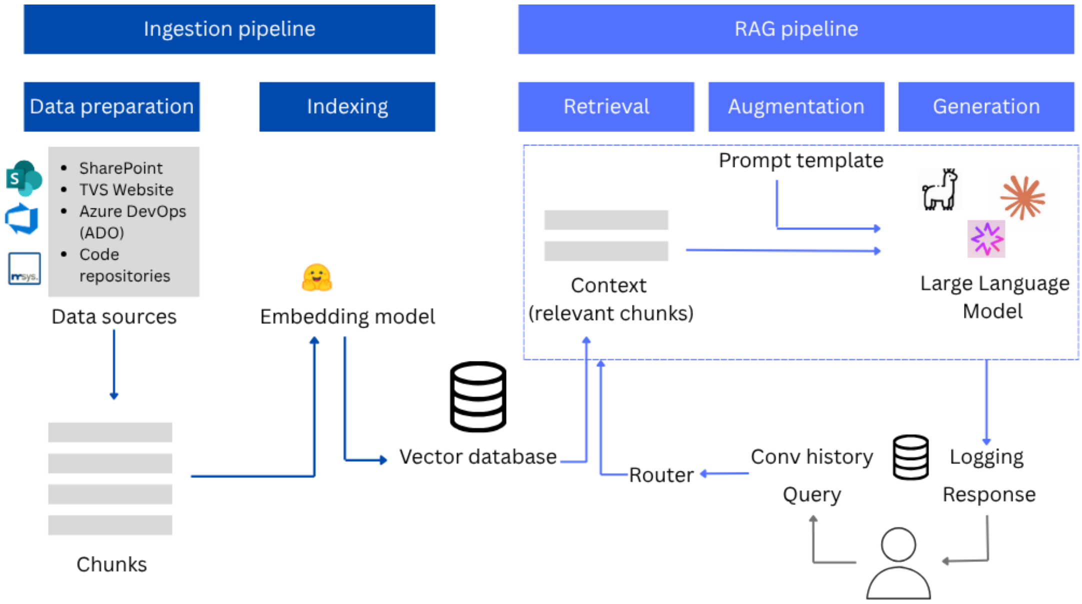

Figure 2: Architecture diagram showing the main components of Sidekick, namely the ingestion and RAG pipelines, with novel approaches to prompt engineering (to handle code queries) and augmentation retrieval (for tool use).

3 Technical Implementation

This section presents the design and implementation of Sidekick (Figure 2), describing: (i) the integration of relevant company data into a vector database (Ingestion pipeline), and (ii) how this vectorised data is used to process user queries with enhanced LLM capabilities (RAG pipeline).

3.1 Ingestion Pipeline

Vector databases are increasingly used to enhance LLM-generated outputs by providing relevant text fragments ("chunks") that have a similar meaning to the user query (i.e. "context"). To do so, company data needs to be transformed and embedded into a common database that handles semantic similarity searches. The *vector database* acts as a bridge between the two system components, accelerating the retrieval of content that is relevant to the user query.

The first system component integrates information from different company data sources into the vector database, in two main steps:

- Data preparation. First, TVS data is fetched from different *data sources*, i.e. SharePoint, Azure DevOps (ADO), code repositories, and TVS website, with a scheduled hour refresh. Data is then processed to extract *chunks* using a document loader: i.e. parsing (extract or transform to text - for code) and chunking (splitting by semantic or logical boundaries).
- Indexing. Extracting semantic vectors from each chunk with an *embedding model*, and creating an index in the vector database for each data source (to define specific fields).

Sidekick is developed to handle both text and code-related queries. Crucially, using prompt engineering for code integration. First, files are split to objects by logical meaning (i.e. functions, methods, or procedures). An LLM is prompted to generate descriptions to each code file, using its object list to report on the overall purpose, structure, key procedures, functions, and external interactions. Both code and transformed text fragments are stored in the vector database, to expose relevant source code lines as sources when responding to the user query.

3.2 RAG Pipeline

The second system component processes user queries by leveraging company data and conversation history to enhance LLM outputs.

The user query and conversation history (i.e. queries and responses of the last 60-minute session) are sent to a *router*. The router splits the user query into sub-sentences (i.e. specific tasks) and calls an LLM to decide which route to take for the augmentation retrieval. Each route uses a type of "chatterbot", a tool-based LLM optimised to answer questions related to different data sources.

Each task identified from the user query triggers an instance of RAG:

Requests	Standard(s)
Accuracy	23282*
Robustness	24027, 12791
Transparency	12792*
Human oversight	8200, 42105*
Data and Data Management	25012, 5259
Cybersecurity	27001
Record keeping and logging	24970*
Quality management systems	9001, 25059*
Risk management systems	31000, 23894
Conformity assessment	42006

Table 2: Horizontal standardisation request for the EU AIA (AI Standards Hub, 2024), mapped to available ISO/IEC standards. Highlighted standards (*) are yet to be published (20th August 2025).

42001	Requirement	Focus
4.[1/2/3]	Purpose & Requirements	
6.[2/3]	Objectives & Change	
5.[1/2]	Leadership & Policy	
5.3	Roles & responsibilities	
6.1.[1/2/3], 8.[1/2/3/4]	AI Risks	Y
9.1. 9.2.[1/2]	Monitoring & Measuring	Y
10.[1/2], 9.3	Continuous improvement	Y
7.[1/2/3/4], 7.5.[1/2/3]	Awareness & Training	

Table 3: Mapping analysis between ISO/IEC 42001 and existing management systems at TVS SCS UK, highlighting focus areas for implementation ("Y").

- Retrieval: information retrieval from vector database using the same embedding model to extract *context* (top-10 similar chunks) and reformat chunks (its text and metadata as XML or JSON list for code route).
- Augmentation: calls an LLM to extract the required parameters to generate the answer (including prompt template).
- Generation: calls an LLM using the instructions and context from previous steps.

The output generated for each task are combined into a single *response* using the LLM only with generated texts. The user query and response are saved for logging and leveraging conversation history.

4 Navigating Regulatory Challenges of TVS Sidekick: Case Study

This section presents the regulatory challenges that emerge with the development of LLMs, and how they may be overcome in a real-world business context. Specifically, we present a case study on navigating a complex and changing AI regulatory landscape in the enterprise, leading to the implementation of the first harmonised technical standard for responsible AI development and use.

4.1 EU AIA & Harmonised Standards

TVS SCS UK is achieving compliance working towards AI standardisation, which is key to the development and adoption of AI. One key regulation shaping the field of standardisation is the EU AIA, which is leading the global landscape of AI regulation.

Different harmonised standards are being developed to support the implementation of the EU AIA, such as the ISO/IEC 12792 and 24970 standards for addressing the transparency and logging of AI systems, respectively (see Table 2). Building upon relevant standards, including AI Concepts and Terminology (22989) and AI Risk Management (23894), ISO/IEC 42001 is the first international standard for AI Management Systems, aiming to guide organisations in the responsible development and use of AI systems.

Recognising the value of standards to operationalise AI regulation principles for ethical and responsible AI, TVS SCS UK has decided to adopt an AI Management System (AIMS) framework to develop trustworthy AI solutions.

4.2 ISO/IEC 42001 Implementation

TVS SCS UK have developed and deployed formal management systems in important areas such as information security, quality, health and safety, business continuity, and environmental management. To effectively implement an AI management system, TVS SCS UK began with mapping the key requirements of ISO/IEC 42001 to existing standards, focusing on management systems already adopted by the organisation.

The results from this mapping analysis are shown in Table 3. Notably, TVS SCS UK maintains an Information Security management system following ISO/IEC 27001 (International Organization for Standardization., 2022). Processes supporting this standard, especially related to data management and cybersecurity, were aligned with ISO/IEC 42001 requirements. This comparison helped to identify focus areas for developing an AIMS:

Category	Topics
Performance	alignment, reliability, robustness, prompt engineering, usefulness, helpfulness, truthfulness
Safety	privacy, security, safety interpretability, transparency, explainability, fairness, trustworthiness, adversarial attacks
Regulation	regulation, best practice*, governance, compliance, accountability

Table 4: Topics of AI/LLM performance, safety, and regulation feeding into the *Knowledge Base* of papers.

AI Risks. TVS SCS UK maintains a risk management strategy as an integral part of their information security. This ongoing process sets out responsibilities and a methodology to periodically assess risks based on likelihood and impact levels. One of the main challenges introducing AI is the need of staying relevant with current risks. To this end, TVS SCS UK is working towards establishing a *Knowledge Base* that informs AI development and use within the company. The AI team started maintaining an academic database of research reviews including meta-analyses and relevant case studies that is accessible throughout the company; both in full-text and via their in-house AI assistant for the purpose of question answering. Furthermore, a systematic search (Brereton et al., 2007) of AI research papers in relevant topics (Table 4) allows to explore topic distribution and relevant metadata, such as indexed keywords or keywords from the authors, and supports the maintenance and updating of the academic database.

Monitoring & Measuring. Another component of the AIMS framework is to capture monitors and measures on the use of AI, including an internal audit programme. TVS SCS UK is developing a *monitoring system* supporting Sidekick, which includes usage indicators (Table 5) and descriptive metrics of interactions (volume breakdown by department, job title, individual user, and question type). Ultimately, these metrics aim to pragmatically measure the effectiveness of the AI assistant, setting a starting point for other AI performance and safety measures. For instance, obtained through the provision of feedback channels (Torkamaan et al., 2024) to report quality or safety incidents, or the inclusion of LLM observability evaluations (Kenthapadi et al., 2024).

Usage indicators	
Interaction volume	Number of messages (i.e. prompts) and unique users.
Response time	Average response time (s).
User engagement	Average of messages per session (on daily basis).

Table 5: Description of metrics in the *monitoring system* supporting the AI assistant at TVS SCS UK.

Continuous improvement. The effective management of vulnerabilities to the AIMS is crucial for demonstrating continual improvement in the use of AI, with documented validation and verification. TVS SCS UK is establishing processes for maintaining and deploying AI, primarily focused on the evaluation and technical documentation of Sidekick. To this end, a primary evaluation objective has been set to understand the needs and ways in which the AI assistant may best support different company roles and responsibilities. Specifically, through the organisation of periodic *feedback interviews* as part of a continuous evaluation of Sidekick, with target populations whose adoption of AI could bring most benefit to the company. A participatory approach to AI development aims to support a culture of ethical and responsible AI.

5 Monitoring & Evaluation

This section presents insights gathered from the deployment of LLMs at TVS SCS UK, highlighting sociotechnical challenges in their enterprise use. Following the on-going implementation of an AI governance model, we specifically report on empirical findings from the monitoring system and initial evaluation of the Sidekick product.

5.1 Adoption & Usage

The implementation of an AIMS framework following ISO/IEC 42001, in particular related to *Monitoring & Measuring* requirements, provides practical insights on the levels of AI adoption and usage in the organisation. Consequently, we report findings from the monitoring system described in Section 4.2.

Figure 3 shows quantitative statistics related to the initial adoption of Sidekick at TVS SCS UK. The monitoring system shows usage indicators and descriptive metrics of interaction volume within a 4-month period (March-June 2025).

Overall, continued use of the AI assistant is

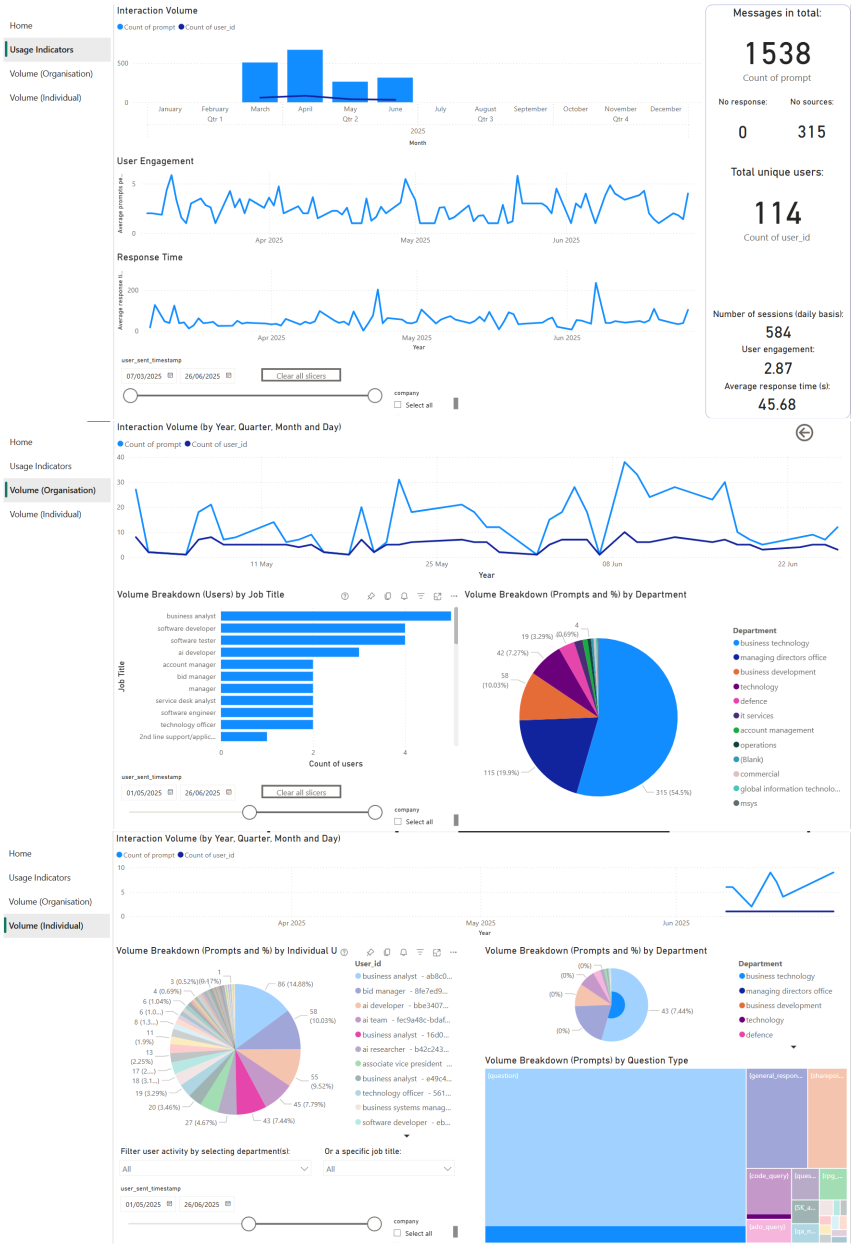

Figure 3: Monitoring system measuring real-time usage data of TVS Sidekick.

shown within the observed time period. This is seen in continued measures both in terms of user engagement and interaction volume (exceeding 500 prompts in the first two months and 250 in the following two months). Despite fluctuations, the conversations do not seem to be long, rarely exceeding an average of five questions per conversation. The response time has peaks on specific dates that increment the time to 46 seconds on average.

The descriptive analysis of interaction volume at organisational level reveals that the most active users were primarily in technology (e.g. developers) and business roles (e.g. bid management, business analysts). At the departmental level, these roles correspond to IT (business technology, business development), management, and operational areas such as defence, technology, commercial, and operations.

In terms of individual usage, the breakdown of activity per individual makes a clear distinction between lead and early adopters (i.e. 46 - 253 queries) and occasional users (less than 20 on average). Queries answered with SharePoint data (i.e. *question*) were the most common, followed by responses without retrieval augmentation (*general_response*), codebase file queries (*rpg_query*) and queries related the development environment, i.e. Azure DevOps (*ado_query*).

Understanding current use of AI/Sidekick
Have you used Sidekick/other AI tools?
What have you used it for?
Where was AI/Sidekick most helpful/ unhelpful?
Outlook
In what aspects of your job would AI be most useful?
Do you have any concerns about integrating AI into your workflow?

Table 6: Topic guide of *feedback interviews* supporting the continuous evaluation of TVS Sidekick.

5.2 Qualitative Feedback

The initial round of feedback interviews that feed into the *Continuous improvement* requirement under ISO/IEC 42001 highlights significant challenges when introducing AI in the business context. Primarily, with respect to the perceived benefits and risks of deploying LLMs in the enterprise, due to Sidekick being the flagship product.

In total, 24 interviews with members of the IT department at TVS SCS UK were conducted between March and April 2025. Participants were invited to 30-minute online meetings for a semi-structured interview. The topic guide (Table 6). included questions i) to gather experiences so far in using AI/Sidekick at work and ii) understand how TVS staff want to use Sidekick in the future. Finally, interview minutes were thematically analysed (Byrne, 2022) by two independent coders.

The analysis of qualitative feedback led to better understanding of baseline attitudes towards AI. The following themes were identified:

Enhanced retrieval (Mentioned by: 16). A key advantage of Sidekick over other tools is its specificity to TVS data. Users valued its assistance with SharePoint-related tasks, finding it faster than a manual search and with a "readable and visible" format, especially for the source list.

Good extracting business logic (Mentioned by: 10). Sidekick was particularly helpful in providing business knowledge, with clear use cases for business analysts. Specifically, for understanding the context of TVS data and key definitions of components within business processes.

Not enough technical detail (Mentioned by: 13). Developers emphasized the need for more domain knowledge to explain internal programmes. Particularly, those relying on a legacy programming language with limited technical documentation. The current version of the AI assistant offers a good starting point for understanding key parameters and functions, but remains limited in addressing more specific queries from technical users.

Keen to engage with AI (Mentioned by: 11). Overall, staff were enthusiastic about using Sidekick to standardise code, reduce duplication, refer new starters to source documentation, or avoid ownership issues when using external AI tools. Furthermore, new features were proposed, including learning from user prompts or returning questions to users to resolve ambiguous queries.

Privacy/commercially sensitive questions/Other concerns (Mentioned by: 8). There were no major concerns with the use of Sidekick, provided it was fed with the right information and access levels. Concerns were raised around job security and distrust in AI tools, along with the emphasis on using Sidekick internally due to potential disclosure of information from the client side.

The first round of feedback has been worked into a plan for continual improvement and addressing concerns, informing further developments of TVS Sidekick. TVS SCS UK will continue developing processes to adhere to ethical principles in regulatory standards, sharing practical insights in critical areas such as managing AI risks, providing relevant monitors and measures on AI use, and increasing AI adoption through training and consultation.

6 Conclusion

This paper presented the experience and challenges encountered in a real-world business scenario with the development and deployment of TVS Sidekick, an AI assistant leveraging LLMs for enterprise use. This empirical study provides practical knowledge, including key lessons learned from the implementation and governance of the in-house AI assistant.

Limitations & Ethical Considerations

The findings and insights presented are drawn from a specific organisational context and reflect experiences within a particular time frame and initial phase of evaluation. While the technical specifics and detailed implementation of each component of the governance framework are outside the scope of this work, this paper aims to contribute to the wider community by sharing reflections on navigating technical, ethical, regulatory, and sociotechnical challenges of deploying LLMs in practice.

References

AI Standards Hub. 2024. Demystifying the EU AI Act – Implications for UK businesses. Available at: https://aistandardshub.org/events/innovate-uk-bridgeai-demystifying-the-eu-ai-act-implications-for-uk-businesses/ [Accessed: 20th August 2025].

Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. 2007. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80(4):571–583. Software Performance.

British Standards Institution. 2025. Understanding ISO/IEC 42001 – a framework for managing AI. Available at: https://iuk-business-connect.org.uk/events/understanding-iso-iec-42001-a-framework-for-managing-ai/ [Accessed: 20th August 2025].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3):1391–1412.

Keeley Crockett, Edwin Colyer, Lauren Coulman, Caitlin Nunn, and Sarah Linn. 2024. Peas in pods: Co-production of community based public engagement for data and ai research. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Keeley Crockett, Edwin Colyer, Luciano Gerber, and Annabel Latham. 2023. Building trustworthy ai solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*, 4(4):778–791.

Department for Science, Innovation & Technology. 2024. Introduction to AI assurance. Available at: https://www.gov.uk/government/publications/introduction-to-ai-assurance [Accessed: 20th August 2025].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

European Commission. 2014. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [Accessed: 20th August 2025].

European Commission. 2025. The General-Purpose AI Code of Practice. Available at: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai [Accessed: 20 August 2025].

Thilo Hagendorff. 2024. Mapping the ethics of generative ai: A comprehensive scoping review. *Minds and Machines*, 34(4):39.

International Organization for Standardization. 2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO Standard No. 27001:2022). Available at: https://www.iso.org/standard/27001 [Accessed: 20th August 2025].

International Organization for Standardization. 2023. Information technology —- Artificial Intelligence – Management Systems (ISO Standard No. 42001:2023). Available at: https://www.iso.org/standard/42001 [Accessed: 20th August 2025].

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2025. 'i'm categorizing llm as a productivity tool': Examining ethics of llm use in hci research practices. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. 2024. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6523–6533, New York, NY, USA. Association for Computing Machinery.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. *Computer Law Security Review*, 53:105957.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. 2023. Large Language Models for Supply Chain Optimization. *arXiv e-prints*, page arXiv:2307.03875.

Manas Mhasakar, Rachel Baker-Ramos, Benjamin Carter, Evyn-Bree Helekahi-Kaiwi, and Josiah Hester. 2025. "i would never trust anything western": Kumu (educator) perspectives on use of llms for culturally revitalizing cs education in hawaiian schools. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Roshni Modhvadia, Tvesha Sippy, Octavia Field Reid, and Helen Margetts. 2025. How do people feel about ai? (Ada Lovelace Institute and The Alan Turing Institute) https://attitudestoai.uk/.

Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mehrdokht Pournader, Hadi Ghaderi, Amir Hassanzadegan, and Behnam Fahimnia. 2021. Artificial intelligence applications in supply chain management. *International Journal of Production Economics*, 241:108250.

Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the dataset practitioners behind large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Yuan Wu, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Comput. Surv.*, 57(11).

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Helma Torkamaan, Steffen Steinert, Maria Soledad Pera, Olya Kudina, Samuel Kernan Freire, Himanshu Verma, Sage Kelly, Marie-Therese Sekwenz, Jie Yang, Karolien van Nunen, Martijn Warnier, Frances Brazier, and Oscar Oviedo-Trespalacios. 2024. Challenges and future directions for integration of large language models into socio-technical systems. *Behaviour & Information Technology*, 0(0):1–20.

UNESCO. 2021. Ethics of Artificial Intelligence | UNESCO. Available at: `https://www.unesco.org/en/artificial-intelligence/recommendation-ethics` [Accessed: 20th August 2025].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Symbolic working memory enhances language models for complex rule application. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17583–17604, Miami, Florida, USA. Association for Computational Linguistics.

Christof Wolf-Brenner, Viktoria Pammer-Schindler, and Gert Breitfuss. 2024. How do professionals in smes engage with ai and regulation? an interview study in austria. In *Proceedings of Mensch Und Computer 2024*, MuC '24, page 646–650, New York, NY, USA. Association for Computing Machinery.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.