Skitag AI

Business Case Assessment: GenAI for GRC

March 2024

GenAI for GRC

Business Case Assessment

Key findings

- LLMs can automate and augment the capabilities to speed up feedback. It is an efficient and scalable alternative to manual and human based feedback and insights generation.
- Use LLM + Retrieval Augmented Generation (RAG) tasks for a domain-specific implementation.
- There are challenges concerning AI regulations, and risk derived from the use of AI-based technology.

What's the problem? Reduce the time to feedback when a GRC issue comes to the team. **What is the value proposition?** Integrate AI-based technology (LLMs) to augment the capabilities of the GRC team in terms of automatically recommended feedback and bring real-time insights when assisting an audit. Reduce cost of operations, and time to feedback. Make it scalable. Be compliant with AI regulations, AI Ethics, data privacy and data security. Be agile, and resilient in case of disruption, or in case of unavailability (e.g. changes in regulations landscape). Make it efficient, and environmentally friendly.

- *Annex A: ChatGPT +RAG Integration*
- *Annex B: Risk Mitigation*
- *Annex C: GRC Suite roadmap [OKRs]*

Driving forces

- **Competing LLMs** (e.g. OpenAI, Google, etc.): There are many competing closed source LLMs implementations. It is an advantage in terms of availability, but it is a vulnerability in terms of rapid adoption of new features. The organisation that implements this technology must move quickly and be very agile. On the other hand, there are a few emerging open source LLMs. Nevertheless, they are still underdeveloped in comparison to closed sources. This alternative could be compelling if agility is not a must have.
- **Technology adoption** (users and competitors): Users are getting used to interacting with automated, rapid, and astonishing AI augmented features. New tech products must have this AI-based interaction to be competitive. The more tech features, the better.
- **Computing costs**: Today's main disadvantage of using LLMs is the cost of training from scratch, or the cost of fine-tuning specific models for domain-specific use cases. Nevertheless, the prospects are suggesting that it will be reduced in the short and medium term (e.g. Neuromorphic computing, Quantum computing). In that sense, there are some prospects pointing to LLMs at the Edge.
- **Regulations**: Technology is moving faster than regulations. That might create some friction between product innovation and government restrictions. Even though there are some AI sandbox initiatives (e.g. ES, EU, UK, or US), there is still some uncertainty in terms of risk management and mitigation that might affect the use and availability of LLMs models.

LLMs (e.g. OpenAI) Integration models

- **Retrieval Augmented Generation (RAG)**: Quick time to market, and time to revenue. Better domain-specific adaptation than just embedding existing GenAI. Cost of operations will be dependent on the sizes of the tokens of the prompts and the data retrieval. Easy to integrate, and low AI skills required. Suitable for a low-code no-code deployment.
- **Fine tuning**: High cost of operations. Very dependent on the technology vendor (i.e. low agility, and low resiliency). Not a scalable option (i.e. custom models for each domain/client). High AI skills required. **Note**: There is the alternative of **Prompt tuning** (Google Research), which is cheaper. Nevertheless, it is still vendor dependent, and it might not pay the extra costs compared to the **RAG** approach.
- **Consume**, or **embed** existing GenAI models is discarded because of generalist outputs (and the threat of risky hallucinations), while **training a GenAI from scratch** is discarded because of the very high cost of operations in terms of computing and data acquisition costs, among other aspects (e.g. less agility, innovation, and resiliency).

Challenges to adoption

- **Regulations**: Take into consideration the GDPR, the EU AI Act, the NIST Risk Management Framework, OECD recommendations for GenAI, the ISO/IEC 42001:2023 AI Management System, and the EU Ethics Guidelines to be compliant with AI regulations. Participate in sandboxes to be sure of being compliant (e.g. EU AI Pact, UK, or US testbeds). Implement a self-assessment for risk discovery and mitigation.
- **Private or privileged data leakage**: Implement Privacy-enhancing technologies or PETs (e.g. Differential Privacy, de-identification, anonymisation, etc) to remove sensitive data from the knowledge base, and from the output of the model.
- **Inaccuracies** (hallucinations): Find the setting of the model to avoid answering prompts that are out of the scope of the domain-specific implementation.
- **Bias**: Implement **Red Teaming** techniques to find possible bias in the outputs of the model.
- **Misuse and adversarial prompting** (e.g. misinformation): Implement **Red Teaming** techniques to find vulnerabilities and suggest action plans in case of adversarial attacks.
- **Feedback loop**: Avoid (or limit) the use of AI-generated data to enrich the knowledge base of the model, or to compensate for any bias that it might have. Have in mind that using AI-generated synthetic input might end up in a feedback loop that increases the bias of the model.

PESTLE analysis

- **Political**: US, UK, EU, and Asia are following different approaches and velocities for GenAI adoption and regulation, mainly because of the risk derived from data acquisition, data privacy, and the impact of the outputs of the LLMs. Be aware of incoming regulations and take part in sandboxes.
- **Economic**: Digital Transformation is shaping the global economy. The prospects for technology adoption and business innovation are very favourable for tech business and AI-based products.
- **Sociological**: People are getting used to interacting with AI-based features. However, there are concerns regarding the gap between different populations in terms of access to technology. Business models should have in mind that international organisations are looking for the democratisation of the GenAI.
- **Technological**: The prospects are very good. There are many stakeholders competing to improve LLMs models. Nevertheless, take into consideration the risk of Superhuman capabilities derived from the development of AGI. Have the OpenAI’s Super Alignment task force in the radar to evaluate the impact of this prospect.
- **Legal**: Although there are some regulations in the process of being implemented (e.g. EU, US, or UK), and some international organisation recommendations (e.g. OECD, or WEF) that are being considered as blueprints, have in mind that there may be some uncertainty in terms of risk assessment implementations.
- **Environmental**: Carbon footprint is one of the top issues related to the cost of computing while training a LLM from scratch. This issue should not be misled. Evaluate the environmental impact of similar implementations. In case of similar cost/benefit metrics, decide for the most environmentally friendly implementation.

Scenarios

- **Open Source** LLMs (best-case): It will be better in terms of security, deployment, and maintenance. It won't be dependent on a specific vendor technology. Costs will be reduced. Barriers of entry for other competitors will be lower as well.
- **Owned** LLMs (worst-case): Higher costs of operations and maintenance. Very exposed to cybersecurity issues. Very resilient in terms of technology customisation and availability. Lower rate of adaptation to changes (i.e. lower innovation and agility capabilities). Possible source of alternative business and value creation (i.e. **domain specific LLMs as a Service**).
- **Vendor** (ChatGPT or Gemini) LLMs (selected use case): Easy to implement. Quick technology adoption. User and client technology awareness, and therefore a source of trust. Efficient in terms of cost of development and maintenance. Reduced time to market, and time to revenue.

Product Description

- **GRC knowledge base**. Implement Vector Databases to be used for Data Retrieval (i.e. Retrieval Augmented Generation or RAG). Build a data lake for accessing **domain-specific knowledge base**: ○ **AI** Governance; ○ **AI** Ethics; ○ **ISO**/IEC standards; ○ **Domain** specific and **product** specific regulations.
- **GRC feedback**. Implement a **LLM** together with **Retrieval Augmentation Generation** plugins (RAGs) and **prompt engineering** to automatically drive meaningful actions out of the input and the contextual data. Take into consideration a **Human-in-the-loop** for specific use cases.
- **GRC audit**. Use LLM (+RAG and prompt engineering) to augment the capabilities to bring real-time insights while being audited.

Reasons for choosing this roadmap

The cost of building and integrating existing AI-based new digital products is lowering. Competitive advantages are for those who innovate quickly and with agility. GenAI is enhancing and augmenting human capabilities. ChatGPT and other LLMs are rising as top technology adoption for process and user experience automation, scalability, cost reductions, and shorter time to market and time to revenue, as well as time to feedback. There are validated implementations in customer experience and augmented virtual agent skills. This technology is 24/7 available, in many languages, and it can be adapted to more specific domains of knowledge, as well as other input and output modalities (i.e. multimodal LLMs). There are examples of insightful outputs that can rival with human made analysis.

Companies like OpenAI, and others (e.g. Google, or IBM) are working on the democratisation of GenAI, mainly for economic and financial goals, but also for the potential of reinforcement learning derived from the extensive use and interaction of the society in many different use cases with their technology. The more the interaction, the better the input, and the better their technologies capabilities. It is a scalable business that is moving into a low-code no-code environment, and a reusable block code architecture (plug and play). Business continuity will be measured in terms of technology adoption, mainly in AI-based technology adoption and user interaction. The more technology adopted and integrated in their business processes, the better. It will be a never-ending game of agility and very quickly iterations. It's "The Geek Way" (Speed, Ownership, Science, and Openness).

Finally, a few topics to have in the radar. Energy and resource efficient LLMs will appear. Edge LLMs are a prospect. GPUs, and ASICs will continue progressing. Take into consideration the risks of shortage as well as new hardware developments. Neuromorphic Computing, and Quantum Computing (and Quantum Machine Learning) are promising cost reductions. Synthetic data will explode, which is an advantage as well as a risk. Because of this, its progress must be closely followed. Cyber resiliency will be in the 2024 agenda. Super Alignment might have an impact on AI-based technologies. Human-centric AI-based products must have Humans-in-the-loop. OpenAI GPTs' store could be a source of opportunities as well as a source of competition.

Annex A: ChatGPT +RAG Integration

According to OpenAI ChatGPT Documentation, the functionality of Retrieval Augmented Generation can be integrated through specific plugins. See these links for references:

- OpenAI plugins
- OpenAI GitHub RAG repository.

On the other hand, though the costs of fine tuning ChatGPT (e.g. computing, data acquisition, and vendor dependency) are higher than the advantage in terms of performance and quality of outputs compared to the RAG approach, it should be considered as a possible future scenario. It seems to be a straightforward process. On the other hand, the prospects are pointing into a future where the GPTs are more agile, and the computing is even better than today, opening the possibility of LLMs at the edge.

Annex B: Risk Mitigation

- **Data privacy and security**: Be compliant with GDPR in terms of processing Personally Identifiable Information (PII), or Protected Health Information (PHI). Implement Privacy-enhancing technologies or PETs (e.g. Differential Privacy, de-identification, etc.) to avoid private and privileged data leakages. Implement Zero Knowledge Proof (ZKP) methods. Use State-of-the-art encryption methods. Take into consideration the roadmap to Post Quantum Cryptography. Acknowledge the period of prompt content retention. Reduce it or remove it if possible.

- **Intellectual property** (IP **Copyright**): Make sure to be compliant with Intellectual Property or design a disclaimer that protects the GRC suite interests for using third-party AI-based technologies.
- Implement Red Teaming to fight against **inaccuracies**, **bias**, **prompt injection attacks**, **misuse**, **disinformation**, or **misinformation**. Test more, cheaper, and faster.
- **Feedback Loop** and **Synthetic input** (Measure module): Implement synthetic input detection methods (e.g. GenAI generated content). Evaluate the introduction of watermarking while collecting data from users.
- **Reputational**: Be aware of up-to-date information regarding the roadmap of the technology vendor. For instance, OpenAI roadmap to Super Alignment. Evaluate the carbon footprint of implementing AI-based technologies (Environmental impact).
- **Security**: According to the WEF, cyber resiliency is one of the top issues to have in mind for the year 2024. Moreover, this October 2024 enters into force the NIS 2 EU Directive on cyber security. Cybersecurity will be a key subject throughout 2024 for digital products.
- **Operational**: Make the implementation as modular as possible. Be ready for integrating other technologies. Have the open-source initiative in the radar. Evaluate different technology vendors for different geographies. Be ready in case there is a need for a smaller LLMs train from scratch.
- **GenAI Responsibility**: Have in mind the **WEF** recommendations on Responsible GenAI: ○ Build public awareness of AI capabilities and their limitations. ○ Uphold AI accountability. ○ Employ diverse red teams. ○ Adopt transparent release strategies. ○ Enable user feedback (User Experience, and Reinforcement Learning). ○ Embed model and system traceability. ○ Ensure content traceability (or Content Provenance) ○ Disclose non-human interaction (Human-centric, and User experience) ○ Adopt sandbox processes (where possible, e.g. EU AI Pact).

Annex C: GRC Suite roadmap [OKRs]

Q1 GRC Suite with LLM. Play.

- **O1 Build a domain-specific Vector Database: KR1.1** A Vector Database scheme. **KR1.2** Number of existing knowledge base translated to Vector Databases. **KR1.3** An implementation of an infrastructure to manage several domain-specific vector databases.
- **O2 ChatGPT API integration: KR2.1** Definition of the prompts that will be used for feedback. **KR2.2** A list of available plugins for data retrieval. **KR2.3** An implementation/development of plugins for data retrieval.
- **O3 Measure the gap between manual and automated insights: KR3.1** Set the metrics to compare the manual outputs vs automated outputs. **KR3.2** Document the methods to measure the augmented capabilities of **GRC** + ChatGPT.

Q2 GRC Suite with LLM. Work.

- **O1 Implement Red Teaming: KR1.1** One Prompt injections attacks exercise. **KR1.2** One exercise for finding and fighting inaccuracies (or hallucinations). **KR1.3** Document a method to measure bias in augmented insights.
- **O2 Gain Technology Adoption (LLMs) Agility: KR2.1** Prompt engineering documentation for general purpose and specific use cases. **KR2.2** Implement an alternative LLM (e.g. Gemini). **KR2.3** Evaluation report for fine tuning ChatGPT.
- **O3 Gain GRC suite Compliance: KR3.1** Participation in a sandbox. **KR3.2** Self-assessment of AI, Ethics, and GDPR compliance.

Q3 GRC Suite with LLM. Live.

- **O1 Achieve Real-time Feedback: KR1.1** Real-time feedback for GRC requirements. **KR1.2** Real-time augmented capability for audits.
- **O2 Leverage on Augmented Capabilities: KR2.1** Reinforcement learning is becoming marginal. **KR2.2** Human-in-the-loop is becoming marginal.

Q4 GRC Suite with LLM.

- **O1 Continuous innovation: KR1.1** Improve performance by improving prompting engineering documentation. **KR1.2** GRC Suite lifecycle documentation. **KR1.3** GRC Suite lifecycle accountability (audit).
- **O2 Gain technology agility: KR1.1** Implement an alternative LLM (e.g. Gemini): **KR1.2** Low-code no-code implementation.