




Trustworthy Generative AI for Financial Services

Marc-André Zöller   

GFT Deutschland GmbH, Eschborn, Germany

Anastasiia Iurshina   

GFT Deutschland GmbH, Stuttgart, Germany

Ines Röder 

GFT Deutschland GmbH, Stuttgart, Germany

Abstract

This work introduces *GFT EnterpriseGPT*, a regulatory-compliant, trustworthy generative AI (GenAI) platform tailored for the financial services sector. We discuss the unique challenges of applying GenAI in highly regulated environments. In the financial sector data privacy, ethical considerations, and regulatory compliance are paramount. Our solution addresses these challenges through multi-level safeguards, including robust guardrails, privacy-preserving techniques, and grounding mechanisms. Robust guardrails prevent unsafe inputs and outputs, and privacy-preserving techniques reduce the need for data transmission to third-party providers. In contrast, grounding mechanisms ensure the accuracy and reliability of artificial intelligence (AI) generated content. By incorporating these measures, we propose a path forward for safely harnessing the transformative potential of GenAI in finance, ensuring reliability, transparency, and adherence to ethical and regulatory standards. We demonstrate the practical application of *GFT EnterpriseGPT* within a large-scale financial institution, where it successfully improves operational efficiency and compliance.

2012 ACM Subject Classification Applied computing → Online banking; Applied computing → Document management and text processing; Human-centered computing → Collaborative and social computing; Computing methodologies → Artificial intelligence

Keywords and phrases Generative AI, GenAI, Trustworthy AI, Finance, Guardrails, Grounding

Digital Object Identifier 10.4230/OASICS.SAIA.2024.2

Category Practitioner Track

1 Introduction

The integration of generative AI (GenAI) in the financial sector holds substantial potential, particularly in applications such as credit assessments [4], personalized financial advice [6], customer support [2], investment research [6], or audit assistance for regulatory compliance [12]. For example, GenAI can be used in the compliance office to optimize processes and to increase efficiency. The regulatory framework mandates stringent adherence to rules and policies to mitigate risks such as a violation of data protection. Yet, the complexity and volume of these internal and legal regulations require high effort from experienced auditors to ensure compliance. As an auditing assistant, GenAI can provide comprehensive overviews of existing rules and policies, highlight changes from previous versions, detect conflicting policies, and ensure alignment with internal guidelines. Furthermore, GenAI can identify gaps in the current regulatory framework, thus supporting more robust compliance strategies.

Despite the promising potential, the deployment of GenAI in finance, characterized by its highly regulated environment and serious risk exposure, faces significant challenges. Ensuring that AI-generated content is accurate, reliable, and free from errors in the form of hallucinations [5] is crucial, as inaccuracies can erode trust among users and stakeholders [14]. Moreover, adhering to a strict ethical framework is essential [10]. GenAI systems must be designed and trained to uphold ethical standards, ensuring that user requests are handled



© Marc-André Zöller, Anastasiia Iurshina, and Ines Röder;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görge, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 2; pp. 2:1–2:5

OpenAccess Series in Informatics

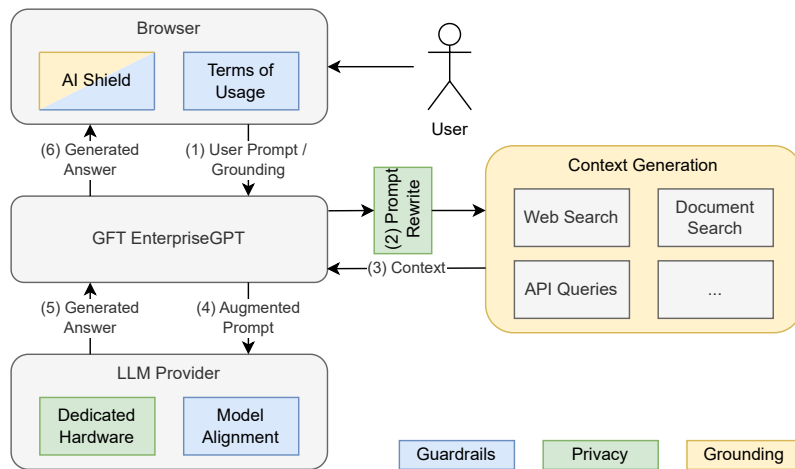


OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

responsibly and that responses are unbiased and transparent. This includes preventing the generation of harmful content or advice that could exploit users. Regulatory compliance is another critical challenge, especially concerning data security. Financial data is highly sensitive, and GenAI systems must comply with stringent regulations like GDPR to maintain trust and avoid legal repercussions, ensuring robust data protection measures are in place [16].

2 Methodology

To address these challenges, our platform, called *GFT EnterpriseGPT*, incorporates several measures to enable trustworthy GenAI. In general, these measures fall into three categories: guardrails, privacy and grounding. Figure 1 provides a high-level overview of *GFT EnterpriseGPT*.



■ **Figure 1** High-level system overview of *GFT EnterpriseGPT*. Displayed in blue, green, and yellow are measures to ensure trustworthy GenAI.

First, the user interacts with their browser to generate a prompt. This prompt is locally analyzed. If the prompt is benign, a request, which is processed by the *GFT EnterpriseGPT* backend, is created. *GFT EnterpriseGPT* can use different tools to augment the user prompt with additional context. Finally, the augmented prompt is forwarded to an large language model (LLM) provider, e.g., a local open-source model or a commercial provider. In the following, we will introduce the separate measures for trustworthy GenAI in more detail.

2.1 Guardrails

Guardrails for foundation models and requests are measures to ensure their safe, ethical, and effective use. They help in mitigating risks, preventing harmful outcomes, and ensuring compliance with legal and ethical standards [15]. In general, a distinction can be made between input and output guardrails. Input guardrails verify that requests that are classified as risky will not enter the LLM model, for example, inquiries on a critical topic such as the construction of weapons. In contrast, output guardrails also check for hallucinations [7]. All foundation model providers implement safety measures [15], like alignment fine-tuning. While this provides a solid basis for ethical GenAI, they are not sufficient in practice as they can not be adapted to specific use cases and can be overcome by malicious actors [3].

Therefore, a client-based prompt analysis, called *GFT AI Shield*, is used to detect, for example, unethical requests or violations of internal policies. This analysis is performed purely inside the client and the prompt under inspection is never forwarded to a commercial GenAI provider to prevent data leakage. Only when no violations, for example, a prompt containing credit card information, were detected, the user prompt is forwarded to an LLM for answer generation. The *GFT AI Shield* uses different methods to detect inadequate content:

- Simple pattern matching is used to detect restricted words
- Named-entity recognition [13] using a local neural network checks if restricted categories are used in the prompt
- Using a plugin interface, customers can provide additional classifiers to detect problematic prompts

If a prompt is flagged by one of the filters, the user is informed why the request is not being processed. All violations against the guardrails are logged in a database. As it is always the case with artificial intelligence (AI) applications, the *GFT AI Shield* can not guarantee the perfect correctness of all predictions. To prevent users from simply switching to public GenAI services, which have a lower security standard, the *GFT AI Shield* is tuned to limit false positive errors.

In general, GenAI applications can either be targeted to an internal or public user base of a company. In the case of an internal application, the employees can receive tailored training for responsible GenAI usage. Furthermore, employers can, to some extent, enforce compliant behavior through internal guidelines and terms of usage agreements. However, just relying on compliant behavior is not sufficient as users may make errors or behave maliciously.

2.2 Privacy

Privacy for GenAI using *GFT EnterpriseGPT* is safeguarded through several key measures. Selecting appropriate licensing is crucial, as it defines the terms under which user-generated prompts, model answers, or datasets for fine-tuning, can be used by the foundation model provider. Even though all model providers guarantee that data is not visible to other users, end-users can not verify this guarantee. Using dedicated hardware, such as PTUs [11] or dedicated servers to host open-source models, helps to prevent unauthorized access and data leaks, providing a controlled environment where sensitive information can be processed without exposure risks. Additionally, the *GFT AI Shield* can also be used for enforcing data privacy. As prompts are directly processed on the user's device rather than on remote servers, the need to transmit personally identifiable information (PII) or confidential information across networks to service providers is eliminated, minimizing the risk of data interception or misuse. Finally, as *GFT EnterpriseGPT* also uses tools accessing the internet, for example using a web search, user prompts are rewritten using GenAI to prevent prompt leakage to another third-party service. Together, these measures create a robust framework for protecting user privacy in the deployment and use of GenAI.

2.3 Grounding

Grounding for GenAI refers to the process of linking their outputs to real-world contexts, facts, or external data sources. This ensures the models' responses are accurate, relevant, and contextually appropriate, enhancing their reliability and applicability in practical scenarios. Retrieval-augmented generation (RAG) [9] is a standard method for grounding [8]. It augments a user prompt with information retrieved from external sources before sending

it to a foundation model. By implementing our own RAG approach instead of using a commercial provider, a fine-tuned solution per use-case can be developed without exposing all data sources to a third party. While this approach can be used to encode domain-specific knowledge into the foundation model, hallucinations can still occur [5]. To ensure the correctness of results, our platform offers the option to search for references for parts of the generated answer. By reusing the retrieval part of retrieval-augmented generation (RAG), we are searching for a context similar to a specific part of the answer selected for grounding. If such a similar context exists, the correctness of the partial answer can be ensured.

3 Use Case

GFT EnterpriseGPT has been successfully implemented at Landesbank Baden-Württemberg (LBBW) [1]. LBBW is a full-service commercial and central bank in Germany. Currently supporting approximately 9,000 employees at LBBW in their daily work, *GFT EnterpriseGPT* has proven its effectiveness in enhancing operational efficiency and compliance in a large-scale financial institution. This success story underscores the practical viability and benefits of trustworthy GenAI in the finance sector following regulatory requirements.

4 Conclusion

In summary, the integration of GenAI in finance offers transformative potential, enabling more efficient and accurate processes in highly regulated environments. By implementing multi-level safeguards addressing key challenges related to trustworthiness, ethical behavior, and regulatory compliance, the full potential of GenAI can be harnessed. The success of our deployment at LBBW illustrates the practical impact and value of combining measures for building trust with GenAI, paving the way for broader adoption of GenAI in finance.

References

- 1 Landesbank Baden-Württemberg. Lbbw startet mit eigener generativer ki-lösung durch, April 2024. [Online; accessed 26-July-2024]. URL: https://www.lbbw.de/artikelseite/pressemitteilung/lbbw-startet-mit-eigener-generativer-ki-loesung-durch_ah8wecz4u8_d.html.
- 2 Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Working Paper 31161, National Bureau of Economic Research, April 2023. doi:10.3386/w31161.
- 3 Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*, 2024. doi:10.48550/arXiv.2402.09283.
- 4 J. Galindo and P. Tamayo. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1):107–143, April 2000. doi:10.1023/A:1008699112516.
- 5 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. doi:10.48550/arXiv.2311.05232.
- 6 Zengyi Huang, Chang Che, Haotian Zheng, and Chen Li. Research on generative artificial intelligence for virtual financial robo-advisor. *Academic Journal of Science and Technology*, 10(1):74–80, March 2024. doi:10.54097/30r2kk80.
- 7 Colin Jarvis. How to implement llm guardrails?, December 2023. [Online; accessed 03-September-2024]. URL: https://cookbook.openai.com/examples/how_to_use_guardrails.

- 8 Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. 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*, pages 6523–6533. Association for Computing Machinery, 2024. doi:10.1145/3637528.3671467.
- 9 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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.
- 10 Bahar Memarian and Tenzin Doleck. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5:100152, 2023. doi:10.1016/j.caeai.2023.100152.
- 11 Microsoft. What is provisioned throughput?, May 2024. [Online; accessed 30-July-2024]. URL: <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/provisioned-throughput>.
- 12 Rafet Sifa, Anna Ladi, Maren Pielka, Rajkumar Ramamurthy, Lars Hillebrand, Birgit Kirsch, David Biesner, Robin Stenzel, Thiago Bell, Max Lübbering, Ulrich Nütten, Christian Bauckhage, Ulrich Warning, Benedikt Fürst, Tim Dilmaghani Khameneh, Daniel Thom, Ilgar Huseynov, Roland Kahlert, Jennifer Schlums, Hisham Ismail, Bernd Kliem, and Rüdiger Loitz. Towards automated auditing with machine learning. In *Proceedings of the ACM Symposium on Document Engineering 2019, DocEng '19*, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3342558.3345421.
- 13 Peng Sun, Xuezhen Yang, Xiaobing Zhao, and Zhijuan Wang. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278, 2018. doi:10.1109/IALP.2018.8629225.
- 14 Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, pages 272–283, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3351095.3372834.
- 15 Neeraj Varshney, Pavel Dolin, Agastya Seth, and Chitta Baral. The art of defending: A systematic evaluation and analysis of llm defense strategies on safety and over-defensiveness. *arXiv preprint arXiv:2401.00287*, 2023.
- 16 Sreedhar Yalamati. Data privacy, compliance, and security in cloud computing for finance. In *Practical Applications of Data Processing, Algorithms, and Modeling*, pages 127–144. IGI Global, 2024. doi:10.4018/979-8-3693-2909-2.ch010.