Operationalising LLMs for Compliance-Critical Letter Writing in Financial Services

Devesh Batra **Alexandros Anatolakis** **John Hartley** **Jude King**
Greig A Cowan **Raad Khraishi**

NatWest AI Research

Abstract

Large Language Models (LLMs) are transforming how financial institutions tackle labour-intensive tasks like drafting personalised, regulation-compliant letters. This paper presents a deployed LLM-based system that has generated tens of thousands of complaint-resolution response letters for a major UK bank since November 2024, following its demonstration of a 30% boost in letter quality and a 62% reduction in drafting time during the development phase. It has already doubled the bank's three-day complaint resolution rate, improved customer satisfaction, and reduced manual workloads. In this paper, we detail our end-to-end LLM framework for regulated complaint handling, our continuous "LLM-as-judge" approach for compliance monitoring and prompt optimisation, and our large-scale deployment strategy. In addition, we share lessons learned from overcoming scepticism through iterative development and user engagement. We also discuss the technical architecture and operational insights that have enabled a secure, robust, and future-proof deployment in this high-stakes domain.

1 Introduction

The increasing digitisation of banking has catalysed the automation of tasks using advanced computational tools, notably in customer complaint response management. This automation not only reduces operational latency and human error but also enhances customer experience through hyper-personalisation and allows staff to focus on more complex decision-making tasks [1]. Addressing customer complaints necessitates strict compliance with regulatory standards, such as those imposed by the Financial Conduct Authority (FCA) in the UK. Banks are required to provide timely, final written responses summarising complaint investigations and outcomes, which is critical for maintaining customer trust and avoiding referral to the Financial Ombudsman Service (FOS) that incurs additional costs. Given the high handling costs associated with complaints, improving efficiency in resolution processes is essential. This paper explores the potential of LLMs to automate the drafting of response letters, thereby streamlining customer service operations.

Drafting professional letters consistently poses a significant challenge, being time-consuming and prone to errors, particularly when adhering to complex guidelines. LLMs, with their advanced linguistic capabilities, are well-suited to automate the creation of compliant and empathetic communications. Their deployment can significantly enhance the quality and consistency of customer interactions in banking, while also mitigating risks associated with non-compliance. However, the implementation of LLMs in sensitive sectors like banking requires rigorous evaluation and robust safeguards to ensure compliance and maintain public trust.

Related work: LLMs are increasingly applied in customer communications to generate structured, coherent, and consistent responses in emails, chats, and letters [2, 3]. Their capacity to adhere to brand-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

specific tone and terminology guided by explicit instructions enhances both trust and personalisation, while ensuring formal standards are met [4, 5, 6]. Notable efficiency gains in overall customer support have also been reported [7]. In financial services, LLMs automate routine correspondence such as bank statements and personalised investment summaries, which improves efficiency and compliance [8, 9, 10]. A hybrid approach, where LLM-generated content is subsequently verified by compliance officers, further ensures that legal and governance requirements are maintained [11, 12].

Contributions: This paper provides: (1) A comprehensive framework for designing and implementing an LLM-driven solution for response letter generation in the banking sector, (2) a novel methodology for prompt design that caters to the nuanced demands of financial communications, (3) an innovative evaluation methodology for assessing LLM-generated letters, integrating human-centric and automated approaches to ensure quality and regulatory adherence.

2 Development Methodology

Figure 1 illustrates the components of our response letter generation framework. The framework's entry point is the output from the human complaint handler investigation, which serves as the informative features for letter generation. The process then branches into three parallel components, each generating a distinct section of the letter independently. Next, the prompt manager routes input data to the appropriate prompt templates, with 16 possible routing options. The prompts for each section are then formatted according to the relevant input data, subject-matter expert (SME) feedback, and regulatory requirements. Next, an LLM generates the letter based on the formatted prompts. Finally, the generated letter undergoes a quality assessment using a custom LLM-as-a-judge implementation.

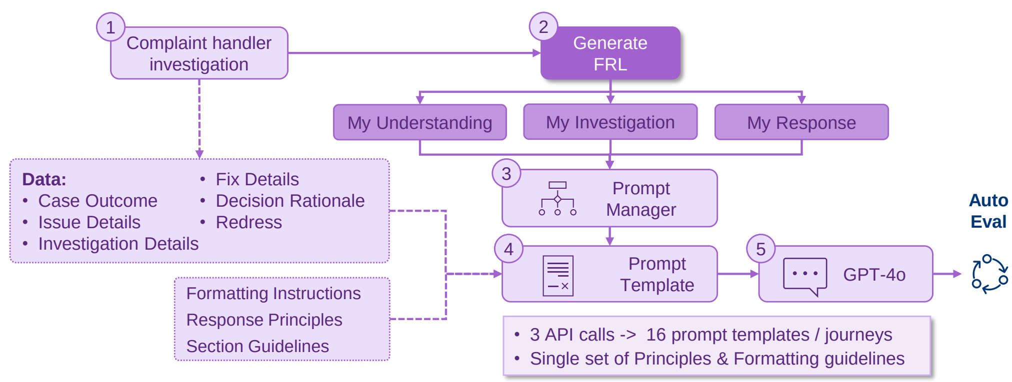

Figure 1: Overview of the response letter generation framework. FRL refers to the Final Response Letter.

Generating response letters. A response letter comprises three key sections about the complaint: summary, investigation, and resolution. Our approach uses a generative model conditioned on complaint details, customer data, handler's notes, bank policies, and regulatory guidelines. Post-GPT-4 LLMs have transformed text understanding and generation [13, 14] by excelling in zero-shot task transfer learning, often eliminating the need for fine-tuning or few-shot examples [15, 16]. These models swiftly parse nuanced financial data [17] and generate lengthy, coherent responses.

Prompt Manager and prompting methods. The content of each section is shaped by key input features, such as the complaint submission method, number of issues, and case outcome, leading to 16 distinct routes for generating letter content. Common elements like formatting and tone-of-voice are maintained across sections. To efficiently manage prompt templates, we developed an internal package for versioning and composable prompts, enabling parallel development. The response letter is created through three section-wise inference calls to an LLM, with each guided by detailed instructions to ensure compliance with regulatory and bank policies. As LLMs are known to exhibit context sensitivity and performance degradation in extended or multi-turn scenarios [18], careful prompt management and template design are critical to maintain high output quality. Prompts were derived from historical exemplars, incorporating structured instructions and formatting standards, allowing for efficient generation of policy-compliant letters without the need for few-shot learning.

Metric	Uplift [%]	95% CI
Correctness	40	[16, 71]
Fluency	39	[20, 61]
Compliance	14	[2, 27]
Structure	34	[14, 59]
Quality	25	[5, 51]
Overall	30	[15, 47]

Table 1: Percentage improvement in quality of AI-generated versus human-written response letters based on human evaluation, with 95% confidence intervals (CI).

Metric	Level	Accuracy [%]	95% CI
Formatting	Letter	85	[77, 90]
Principles	Letter	96	[90, 98]
Guidelines	Section:1	100	[96, 100]
Guidelines	Section:2	92	[85, 96]
Guidelines	Section:3	95	[89, 98]
Hallucination	Section:1	83	[75, 89]
Hallucination	Section:2	83	[75, 89]
Hallucination	Section:3	83	[75, 89]

Table 2: Automated evaluation performance metrics for synthetically generated response letters. CIs are calculated using the Wilson score interval method.

Data selection. Response letters are generated after the complaint handler completes their investigation, with all relevant data captured in the bank's system. Using stratified sampling, 1000 samples were chosen from the top 250 complaint category codes across the period from 2023-01-01 to 2024-04-01 and stratified by factors such as Multiple/Single Issue, Product Type, and Case Outcome. Exploratory analyses of historical complaints, customer feedback, and FOS reports identified key trends and themes.

Feature selection. 14 features, including the categorical Case Outcome, were selected for response letter generation, with data quality criteria applied to maintain a reliable context. This curated dataset underpins our generation framework, ensuring robust performance. Sensitive data (e.g., PII, PAN) is masked to guarantee privacy and regulatory compliance.

3 Model Validation

LLM outputs can be validated through human evaluation, traditional n-gram metrics (e.g., ROUGE [19] and BLEU [20]), and LLM-as-a-judge approaches [21]. We combine human assessment with automated evaluation using LLMs and established metrics. Human evaluation serves as both the gold standard and a baseline for validating automated processes. The auto-evaluation component aims to reduce reliance on domain experts during model development and monitoring.

Human evaluation. 46 colleagues from various business units conducted a blind assessment of both AI-generated and human-written response letters using a 5-point Likert scale, including groundedness (accuracy and relevance), fluency (clarity), compliance (regulatory adherence), structure (sensible organisation), and overall quality (average of the previous metrics). The evaluation included two rounds for prompt refinement.

Automated evaluation. We used an LLM-as-judge to evaluate response letters against four criteria: proper formatting of dates, monetary values, and account numbers; adherence to general principles at the global letter level; compliance with section-level style and context guidelines; and detection of hallucinated or invented facts. We validated our automated evaluation system using a synthetic dataset, addressing challenges related to the quality of test-time letters and the resource-intensive nature of manually sampling true negatives. Starting with SME-recommended exemplars, we created templates and augmented them with false negatives and variations generated by LLMs, enabling the generation of controlled ground-truth data for rigorous validation of our framework.

Bias and fairness. Recognising that LLMs can inherit biases from their training data [22], potentially resulting in unfair outcomes [23], we examined automated evaluation results by protected characteristics, revealing no material difference between groups.

Results. (1) AI-written response letters demonstrate 30% higher quality than human-written ones. 46 SMEs conducted a blind review of 25 human-written and 225 AI-generated response letters using our framework. Table 1 presents the average scores and percentage improvement across our performance metrics. Our results show a statistically significant improvement in AI-generated response letters across all metrics, indicating that SMEs strongly prefer AI-generated response letters over human-written ones.

(2) We assessed the quality of generated response letters using an LLM-as-judge approach (Table 2), with confidence intervals calculated via the Wilson score method. The process exhibits high content and style quality, with low hallucination rates, evidenced by strong performance across Principles, Formatting, and Guidelines metrics. All letters are reviewed by experts before finalisation, ensuring no hallucinations in delivered correspondence. Automated evaluation is used solely for model monitoring and to detect distributional shifts for any required intervention.

(3) The automated evaluation framework facilitates prompt optimisation, reducing the resource demands on SMEs during development and allowing validation of improvements before costly SME validation. Initial evaluations revealed hallucination patterns, such as fabricated compensation payment dates and unclear contextual guidance, which were addressed through optimisation, leading to a 57% reduction in hallucination rates without SME involvement. This method not only enhanced development efficiency but also preserved SME expertise for validating significant improvements.

4 Deployment at a Leading UK Bank

Our response letter generation solution is operational in a hybrid cloud environment at a leading UK bank, assisting complaints handlers daily in drafting customer response letters. The system is accessible through a secure web-based Single Page Application (SPA), with controlled access via enterprise authentication and authorisation. Handlers can review and edit AI-generated drafts before approval for customer dispatch. Data processing is managed by an orchestration layer that oversees batch processing and machine learning pipelines, ensuring continuous model monitoring and real-time application health visibility. Security is upheld through role-based access, token authentication, and audit logging, while data exchanges are protected and sensitive information masked to comply with regulatory standards, ensuring data privacy and operational efficiency.

Since November 2024, the GenAI solution has automated the drafting of tens of thousands of resolution letters, reducing manual effort and enhancing clarity and personalisation of communications. The proportion of complaints resolved within the three-day target window has significantly improved, improving Net Promoter Scores and generating positive feedback from both customers and colleagues.

5 Discussion and Conclusion

The deployment of an LLM-powered complaint response tool in a leading UK bank demonstrates the maturing role of GenAI in regulated industries. Better understanding LLM risk-related behaviours and their alignment with human decision-making is increasingly important for financial services, where robust and predictable AI systems are required [24]. This case highlights how automation can drive efficiency, improve communication quality, and address compliance challenges offering a template for responsible AI integration without undermining regulatory oversight or customer trust.

The key success factors for the project included user involvement and change management, where engaging complaint handlers in the development and testing phases fostered trust and advocacy. Additionally, the use of agile, cross-functional development allowed for iterative feedback from frontline staff, ensuring that the tool met both technical and compliance requirements. Furthermore, the quality and scalability of the GenAI-generated letters proved to match or exceed human standards, facilitating a successful rollout. Lastly, strict adherence to business processes and regulatory needs was crucial for the tool's adoption. These points underscore the value of user-centred design, agile practice, and business alignment for successful AI deployment in regulated domains.

Challenges persist around managing hallucinations and ensuring accuracy, necessitating continuous monitoring and human oversight. Fully capturing the emotional nuances of sensitive complaints remains difficult, particularly in complex cases. Future work may explore improved prompt engineering, such as chain-of-thought prompting [25], and the use of retrieval-augmented generation [26] to better incorporate policy context. Agentic frameworks may further streamline operations by enabling autonomous information retrieval.

References

[1] Devesh Batra, Conor Hamill, John Hartley, Ramin Okhrati, Dale Seddon, Harvey Miller, Raad Khraishi, and Greig Cowan. A review of llm agent applications in finance and banking. *Available at SSRN 5381584*, 2025.

[2] Vishvesh Soni. Large language models for enhancing customer lifecycle management. *Journal of Empirical Social Science Studies*, 7(1):67–89, 2023.

[3] David GW Birch and Kirsty Rutter. Where are the customers' bots? the ai paradigm shift in retail banking. *Journal of Digital Banking*, 8(2):132–140, 2023.

[4] Kausik Lakkaraju, Sara E Jones, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath C Muppasani, and Biplav Srivastava. Llms for financial advisement: A fairness and efficacy study in personal decision making. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 100–107, 2023.

[5] Hao Zhou, Chengming Hu, Ye Yuan, Yufei Cui, Yili Jin, Can Chen, Haolun Wu, Dun Yuan, Li Jiang, Di Wu, et al. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*, 2024.

[6] Cehao Yang, Chengjin Xu, and Yiyan Qi. Financial knowledge large language model. *arXiv preprint arXiv:2407.00365*, 2024.

[7] Sudhakar Reddy Peddinti, Subba Rao Katragadda, Brij Kishore Pandey, and Ajay Tanikonda. Utilizing large language models for advanced service management: Potential applications and operational challenges. *Journal of Science & Technology*, 4(2), 2023.

[8] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

[9] Edwin Jose and Prasad Prabhakaran. Harnessing large language models (llms) optimizing performance, monitoring, and compliance. *Authorea Preprints*, 2024.

[10] John J Nay. Large language models as fiduciaries: a case study toward robustly communicating with artificial intelligence through legal standards. *arXiv preprint arXiv:2301.10095*, 2023.

[11] Marina Jovic and Salaheddine Mnasri. Evaluating ai-generated emails: A comparative efficiency analysis. *World Journal of English Language*, 14(2), 2024.

[12] Weijiang Li, Yinmeng Lai, Sandeep Soni, and Koustuv Saha. Emails by llms: A comparison of language in ai-generated and human-written emails. 2025.

[13] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

[14] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[17] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.

[18] Robert Hankache, Kingsley Nketia Acheampong, Liang Song, Marek Brynda, Raad Khraishi, and Greig A Cowan. Evaluating the sensitivity of llms to prior context. *arXiv preprint arXiv:2506.00069*, 2025.

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[21] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[22] Giulio Pelosio, Devesh Batra, Noémie Bovey, Robert Hankache, Cristovao Iglesias, Greig Cowan, and Raad Khraishi. Obscured but not erased: Evaluating nationality bias in llms via name-based bias benchmarks. *arXiv preprint arXiv:2507.16989*, 2025.

[23] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

[24] John Hartley, Conor Hamill, Devesh Batra, Dale Seddon, Ramin Okhrati, and Raad Khraishi. How personality traits shape llm risk-taking behaviour. *arXiv preprint arXiv:2503.04735*, 2025.

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[26] 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, 2021.