Human Preference Alignment in Financial Advice: A Generative AI Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

Financial advice is a highly regulated domain where unclear communication can cause consumer harm and regulatory breaches. Yet existing recommendation systems often fail to adapt to individual risk preferences and comprehension levels. In this work, we investigate how generative AI can be used to improve both the clarity and personalisation of financial product communications. We first construct a benchmark of clarity by collecting human ratings of real financial product descriptions and construct a novel dataset with 25,000 synthetically generated variations. Using this dataset, we then explored two optimisation strategies for generative models: dynamic generation guided by classifier feedback, and an RLHF-style approach using the classifier as a reward model. Our findings show that clarity is shaped both by textual style and consumer profile, and that integrating preference signals significantly improves comprehensibility. This work contributes a benchmark, models, and methods for aligning generative AI with human preferences in financial communication.

1 Introduction

The provision of financial advice has undergone a profound transformation in recent years, driven by the increasing demand for personalised solutions. Traditional recommendation systems in finance have primarily relied on statistical models, offering generic product suggestions that often fail to reflect the unique risk-reward preferences and financial literacy levels of individual clients. At the same time, the complexity of financial products and the opacity of communication have contributed to persistent challenges in consumer understanding and trust.

A substantial body of prior research has sought to evaluate how well consumers understand financial products by examining the readability of financial texts. For example, the study by Loughran and McDonald (2014a) analyzed financial disclosures using established readability metrics such as the Fog Index, the Flesch Reading Ease Score, and a measure inspired by the U.S. SEC's plain English initiative (Loughran and McDonald, 2014b). Similarly, van Boom et al. (2016) assessed the comprehensibility of insurance contracts by applying multiple indicators, including the Common European Framework of Reference for Languages (CEFR) and the Dutch Flesch–Douma Reading Ease measure. In Burke and Fry (2019), online materials covering payday loans, personal loans, and credit card offers were evaluated through the Fog Index. While these approaches provide valuable insights into the readability of financial documents and online content, relatively little attention has been given to methods for generating texts that are not only measurable but also genuinely easier for consumers to understand.

Generative Artificial Intelligence (AI) offers new opportunities to address these limitations. Beyond assessing the readability of financial documents, generative models have the capacity to adapt financial communications to the linguistic and cognitive needs of diverse users. This capability is particularly

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

significant in a highly regulated domain such as financial advice, where unclear disclosures, or unsuitable recommendations can lead to substantial consumer harm and regulatory penalties.

In this paper, we focus on the clarity and effectiveness of financial promotions and client communications. We investigate how generative AI can be leveraged not only to generate personalised recommendations that respect risk-reward trade-offs, but also to produce explanations and disclosures that align with the preferences and comprehension levels of different client segments. Specifically, we propose and evaluate a model that aligns generative outputs with human preferences for clarity, transparency, and regulatory compliance in financial communication.

By bridging advances in generative modelling with the practical requirements of financial regulation and consumer protection, our work contributes to the growing body of research on trustworthy and human-centred AI in financial services.

2 Financial Clarity Benchmark

To establish a baseline of clarity in financial communications, we conducted a survey in which participants evaluated the comprehensibility of real product descriptions collected from online sources. Respondents were asked to rate clarity on a five-point Likert scale (1 = very unclear, 5 = very clear), the criterion being whether the information provided was sufficient to make an informed decision about the product.

To ensure representativeness, we focus on two of the most common financial products, credit cards and overdrafts, drawing material from 21 UK-based institutions offering these services. All texts were segmented into equal-length paragraphs by category (product description; risk disclosure; cost transparency; feature explanation). To normalize the ratings, we calculated the proportion of responses at each clarity level by dividing the number of ratings for that score by the total number of descriptions evaluated for scpecific financial institution. Our analysis revealed that low-clarity content (scores 1 and 2) made up to 44% of the evaluated text in some cases, suggesting that a substantial proportion of financial communications are not easily understandable to consumers.

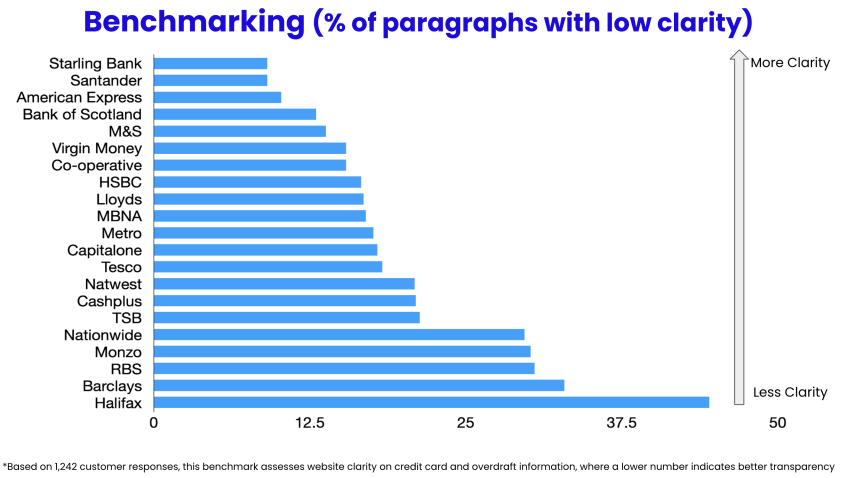

Figure 1: Distribution of clarity scores (1–5) across product descriptions from 21 UK financial institutions.

We further compared the share of low-clarity paragraphs with the number of complaints recorded for each institution in the most recent reporting period, as published by the Financial Conduct Authority. Our findings indicated a positive correlation between unclear communication and higher complaint volumes, supporting the hypothesis that insufficiently clear financial disclosures contribute to consumer dissatisfaction and regulatory risk.

This benchmark enabled us to capture patterns in how consumers perceive the clarity of financial communications across institutions and product types. Our preliminary analysis suggests that clarity

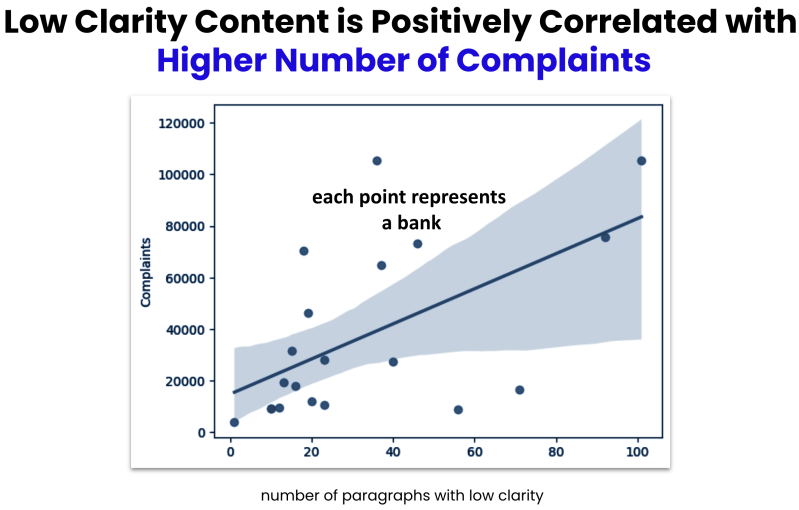

Figure 2: Positive correlation between the proportion of low-clarity content (scores 1–2) and the number of complaints reported to the Financial Ombudsman.

levels vary considerably across providers, highlighting systematic challenges in ensuring that financial promotions are both comprehensible and decision-useful.

3 Model training

For model training we synthetically generated 25,000 variations of financial product descriptions, each written in different styles. Using the same clarity evaluation framework as in our benchmark study, we asked survey participants to rank the clarity of these texts on a five-point scale. Alongside their ratings, we collected demographic information, including age, income, educational qualifications, employment status, marital status, native language, and number of financial dependents.

To ensure data quality, participants were required to provide a short explanation for each rating, justifying why they considered a description clear or unclear. Responses were then filtered and validated to remove inconsistent or low-effort answers.

We trained two types of BERT models (Devlin et al., 2019):

- A multi-class classification model to predict clarity scores based on the text alone.
- A multi-class classification model that incorporated both the text and consumer profile attributes.

Our results demonstrated that incorporating consumer characteristics significantly improved predictive performance. Specifically:

- **BERT with demographic attributes** achieved an accuracy of 0.79.
- **BERT without attributes** achieved an accuracy of 0.71.

These findings indicate that clarity is not only a function of the text itself but also depends on the reader's background and financial context, underscoring the importance of personalisation in financial communication.

4 Generative Model Training with Human Feedback

Generative models benefit substantially from **human feedback**, especially in domains such as finance where evaluation criteria like clarity and comprehensibility are inherently subjective. To improve the quality of generated financial communications, we explored two complementary methods.

4.1 Clarity Score with Dynamic Generation and Machine Feedback

In this approach, the generative model was conditioned on a predicted *clarity score*. A classifier trained on our benchmark dataset assessed each generated output and provided a clarity rating. The model iteratively refined its responses until the predicted clarity score exceeded a predefined threshold.

This method allowed for **scalable optimisation**, since it relied on machine-generated feedback loops rather than direct human annotation for every new sample. By continuously filtering generations through the classifier, the model was nudged toward producing texts that conformed to established clarity standards.

4.2 RLHF-style Training with a Reward Model

Our second approach was inspired by **Reinforcement Learning from Human Feedback (RLHF)** (Stiennon et al., 2020), commonly used to align large language models with user preferences. Instead of training a new reward model from pairwise comparisons, we repurposed our *clarity classifier* as a proxy reward function, since it had been trained on human-annotated clarity judgments.

The pipeline followed four steps:

1. **Reward Model:** The clarity classifier provided a reward signal based on human-labeled clarity scores.
2. **Policy Model:** A generative model produced candidate financial texts.
3. **Reward Signal:** The classifier evaluated each output and returned a clarity score.
4. **Optimisation:** The generative model was fine-tuned to maximise the reward, aligning generation with human preferences.

This RLHF-style approach offered **direct alignment with human judgments** while reducing annotation costs, as the reward model was trained once on human survey data and then reused for generative optimisation.

5 Conclusion

In this work, we investigated how generative AI can be leveraged to enhance the clarity and personalisation of financial product communications. By constructing a benchmark of financial clarity, combining real-world product descriptions with 25,000 synthetic variations, and collecting human evaluations alongside demographic attributes, we demonstrated that clarity depends not only on textual features but also on consumer profiles.

Our experiments showed that models incorporating demographic information achieved substantially higher predictive accuracy than text-only baselines, underscoring the importance of tailoring communication to user characteristics. Furthermore, we compared two generative optimisation strategies: dynamic generation guided by classifier-based feedback, and an RLHF-style approach using the classifier as a reward model. Both methods improved clarity, with the RLHF-style method offering stronger alignment with human preferences, while the machine-feedback approach provided scalability.

These findings highlight the potential of generative AI to support clearer, more transparent, and user-aligned financial advice. At the same time, they raise important avenues for future research, including expanding the benchmark to additional product categories, integrating fairness constraints to ensure accessibility across diverse consumer groups, and exploring hybrid frameworks that combine machine and human feedback at scale. Ultimately, aligning financial communication with human preferences has the potential to increase trust, improve decision-making, and strengthen compliance in financial services.

References

Mary Burke and John Fry. How easy is it to understand consumer finance? *Economics Letters*, 177: 1–4, 2019. doi: 10.1016/j.econlet.2019.01.012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *the Journal of Finance*, 69(4):1643–1671, 2014a.

Tim Loughran and Bill McDonald. Regulation and financial disclosure: The impact of plain english. *Journal of Regulatory Economics*, 45(1):94–113, 2014b. doi: 10.1007/s11149-013-9222-4.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Willem H. van Boom, Pieter Desmet, and Mark van Dam. "If It's Easy to Read, It's Easy to Claim" — The Effect of the Readability of Insurance Contracts on Consumer Expectations and Conflict Behaviour. *Journal of Consumer Policy*, 39(2):187–197, 2016. doi: 10.1007/s10603-016-9317-9.