Which generative AI can give correct and reliable financial explanations?

Kaan Gögcay

May 28, 2025

Abstract

This study evaluates which generative AI model can deliver accurate and reliable financial explanations. Using the FinBen benchmark, GPT-4o was selected and tested with a diverse set of prompts covering financial concepts and compliance scenarios. Responses were validated using GPT-4o itself. Results show that GPT-4o consistently provides clear and factual financial explanations.

Contents

1 Introduction

Given the growing number of generative AI solutions available today, this project aims to evaluate which model is most suitable for our specific needs. The selected AI must demonstrate a strong understanding of financial products and possess the capability to clearly explain complex financial concepts in an accurate and accessible manner.

2 Research Approach

The study was guided by the sub-question: *"Which generative AI can give correct and reliable financial explanations?"* To answer this, I followed a structured research process involving multiple methods.

First, I identified the most suitable large language models for financial topics, using the FinBen benchmark to select GPT-4o. Then, I tested GPT-4o using a series of financial prompts developed with input from a stakeholder. These prompts covered basic knowledge, scenarios, dilemmas, edge cases, and compliance topics. The responses were validated using GPT-4o to assess their accuracy and reliability. Based on this process, GPT-4o was found to be suitable for explaining financial products clearly and factually.

2.1 Best LLM for Finance

The FinBen [1] is the most popular benchmark for benchmarking financial large language models. FinBen has published an open leaderboard with all the large language models tested. The winner in this benchmark is GPT4-Turbo (modern day GPT-4o). To validate that the model can indeed generate correct and reliable financial explanations, I will test the model with a set of prompts.

2.2 Testing GPT-4o

To test the model, I consulted with my stakeholder to brainstorm potential questions for the model. During this session, my stakeholder explained six financial products and services. Based on this input, I developed the following prompts:

2.2.1 Prompts

2.2.1.1 Basic Knowledge Prompts:

- Explain how a Buy Now Pay Later product works exactly, including the interest-free period and what happens after this period.
- What is the difference between a Fixed Term, Fixed Sum loan and a Running Account / Revolving Credit?
- Explain what direct debit is and on which days of the month it can be set up.
- You want to know if you can change the date of your direct debit, what days of the month are allowed for direct debit.

2.2.1.2 Scenario-based Prompts:

- You have taken out a BNPL product for a new TV of €1200. The interest-free period is 6 months, after which 19.9% interest applies over 36 months. What must you pay monthly if you start repaying immediately after the interest-free period?
- You already have three BNPL products running at different stores. You ask if you can take out a fourth for a new purchase. What are the risks of this?
- You are considering taking out a Fixed Term, Fixed Sum loan but you want to know if you can make extra repayments. What is the correct information you should receive?

2.2.1.3 Dilemma Prompts:

- You have two BNPL plans running. One expires next week with an interest rate of 2%, the other in three weeks with an interest rate of 30%. What are the facts you should know without giving specific advice?
- You ask if it's smart to increase their Running Account limit from €3000 to €5000, while you already have used €2500. What factual information should be shared?

2.2.1.4 Edge Case Prompts:

- If a customer wants to fully repay their Fixed Term, Fixed Sum loan in month 2 of a 24-month contract, do they still have to pay all the interest? Explain how this works.

2.2.1.5 Misleading Prompts (to test robustness):

- You have to pay interest if you fully repay your BNPL product within the interest-free period. Is this correct?
- Is it true that a credit decrease always requires a new credit bureau check?
- A consumer claims they have heard that with a Running Account they never need to pay more than the minimum amount. What is the correct information about this?

2.2.1.6 Compliance and Risk Prompts:

- What warnings should be given to a customer who wants to take out multiple BNPL products?
- What are the legal obligations when offering a credit increase to an existing customer?
- A customer is having difficulty paying the minimum monthly repayment on their Running Account. What factual information should be provided without giving advice?

2.2.1.7 Prompt Structure:

Each prompt was extended with an additional instruction to guide the model towards generating content suitable for video narration. For example, the original prompt:

> Explain how a Buy Now Pay Later product works exactly, including the interest-free period and what happens after this period.

was extended to:

> Explain how a Buy Now Pay Later product works exactly, including the interest-free period and what happens after this period.
>
> Explain this using factual information only, without giving any sort of advice. Write a video script using only spoken narration — no stage directions, introductions, begin/end annotations, any other additional annotations, or visual cues. Address your query directly without unnecessary introductory phrases.

This method ensures that responses resemble spoken video scripts and remain objective and factual.

2.2.2 Prompt Testing

Each prompt and its corresponding response can be found in the document `Prompt_Testing_v1.0`.

2.3 Validating Responses

To validate these responses, an expert will manually assess the responses. Sadly, there are no experts available to assess these results. Therefore, I assessed them myself using GPT-4o. Each response given by GPT-4o is validated by GPT-4o.

3 Conclusion

The generative AI model GPT-4o can explain financial topics in our domain. Making GPT-4o a suitable generative AI for this project.

References

[1] Qianqian Xie, W. H. (2024). *FinBen: A Holistic Financial Benchmark for Large Language Models.* New York: Cornell University.