RAG-Based Synthetic Scenario Modeling

Arda Tasci (ardatasci@nyu.edu)

New York University

Introduction

Goal: Produce realistic synthetic market data using natural language prompting.

Example:

Input: Natural language prompt
('WWIII Begins')
↓
Output: Realistic synthetic data
(Based on how markets react to war)

Motivation & Purpose:

- **Black-swan events are not well-represented in historical data.**
- Model stress testing may require narrative, event-driven scenarios.
- Current methods do not conveniently account for such cases.

Methodology

1. Collect financial knowledge base into FAISS vector store
2. Match user query with historical event(s) via Cosine Similarity
3. Retrieve related time series via Bloomberg API
4. Train Hidden Markov Model (HMM) on retrieved time series
5. Generate synthetic distribution and path

Prompt: "Death toll of new Covid strain reaches thousands".
Most similar event: Covid-19 Recession.

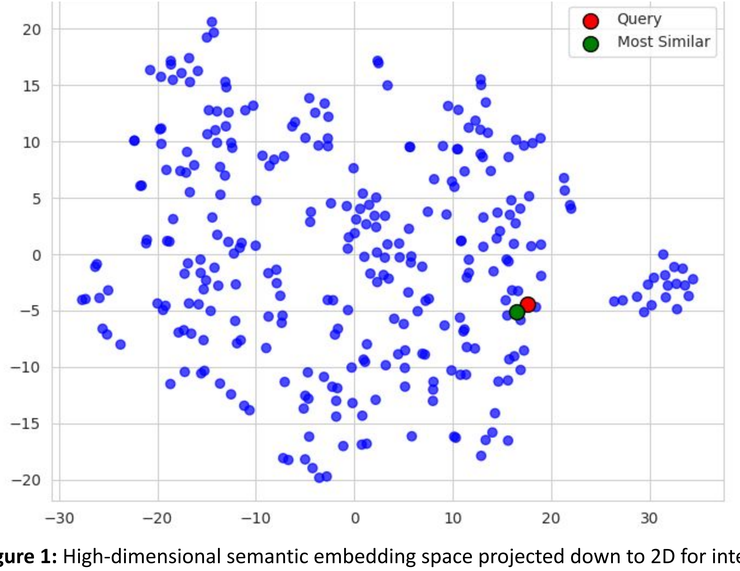

Figure 1: High-dimensional semantic embedding space projected down to 2D for interpretability. Closeness between points is proportional to similarity.

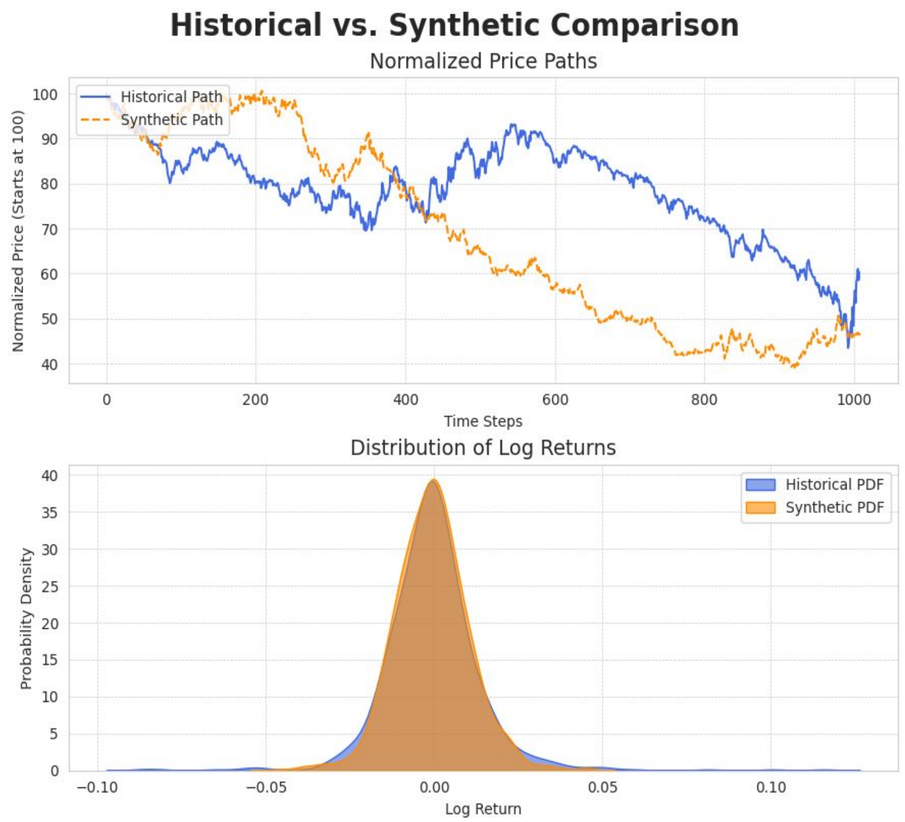

Figure 2: Synthetic scenario for prompt replicates historical distribution of Covid recession over 48 months.

Key Findings

State	Mean (μ)	Variance (σ^2)	Wasserstein Dist.
1	0.0052 / **-0.0005**	0.00319 / **0.00020**	**0.00122**
2	-0.0011 / **-0.0005**	8.35×10^{-5} / **0.00020**	**0.00122**
3	0.0163 / **-0.0005**	**1000.0** / **0.00020**	**0.00122**

Table 1: Comparison of synthetic and real log return distributions for each state in Prompt. Values are shown as *synthetic / real*.

- Despite some parameter mismatches (e.g., variance in S3), Wasserstein distance remains consistently low over 1440 time steps (days).

Limitations:

- Assumes markets are efficient (will react to similar events similarly).
- 'Expected' behaviour may drift over time.
- Model has difficulty replicating PDF tails.

Conclusion

Our framework provides a more efficient, nuanced, and realistic tool for stress testing than traditional time-series backtesting.

Future work will focus on:

- Expansion of knowledge base via Bloomberg News Archives.
- Experimentation with diffusion modeling.
- Agentic capability for chained scenario generation.

Acknowledgements

The author would like to thank: NYU UGSRP, Armstrong, James (California Polytechnic Institute SLO), Krawczyk, Austin (UC Davis), Wang, William (Stanford University) for guidance, peer review, and proofreading.

Works Cited

Wiese, Magnus, et al. "Quant GANs: Deep Generation of Financial Time Series." Quantitative Finance, vol. 20, no. 9, Apr. 2020, pp. 1419–40.

Lin, Zinan, et al. "Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions." Proceedings of the ACM Internet Measurement Conference, ACM, 2020, pp. 464–83.