Investment Strategies Ideation Using Large-Language Models and Structured Multi-Modal Data

Michael Robbins, Taewan Yoon, Martina Paez Berru

MathWorks Finance Conference | September 30, 2025

Agenda

1. Introduction: The Quant's Dilemma
2. Our Process: A five-step pipeline
3. Sourcing & Acquisition
4. Cleaning & Structuring
5. Analyzing Structures
6. Ingestion & Vector Analysis
7. Prompting & Evaluation
8. Conclusion & Takeaways

The Core Problem

The Central Question

"How do you come up with your strategies? Where do you find your ideas?"

Current State

Most professional investors rely on being "plugged in"—a process prone to behavioral biases and often more art than science.

Challenge for New Quants

Finding viable strategies without decades of experience, especially when faced with overwhelming volumes of academic papers.

Our Mission

Build a systematic pipeline that surfaces worthwhile research, creating a repeatable process for strategy ideation.

Team

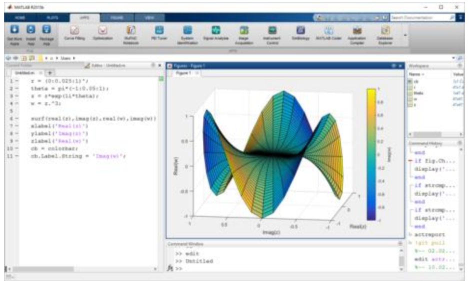
Michael Robbins has taught at Columbia for many years. He has held six Chief Investment Officer roles, including one for a bank with 8.5 million clients.
He wrote Quantitative Asset Management, McGraw-Hill, 2023.

Contact
michael.robbins@QuantitativeAssetManagement.com

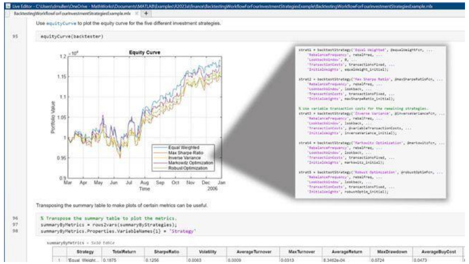
Taewan Yoon is an AI researcher and a dual MBA/MS candidate at Columbia University. He is a tech evangelist for Anthropic, and has worked at Amazon and Coupang. He is also an active member of the CFA Society New York.

Martina Paez Berru is an MS candidate in Business Analytics at Columbia University. She has worked with Ardian and Deep Venture Partners as a Data Scientist.

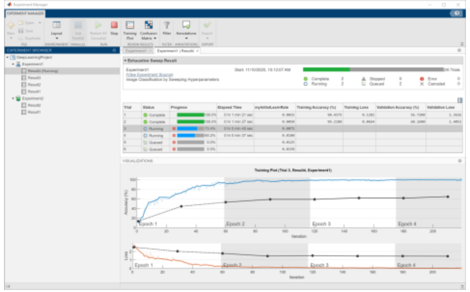
MathWorks Tech



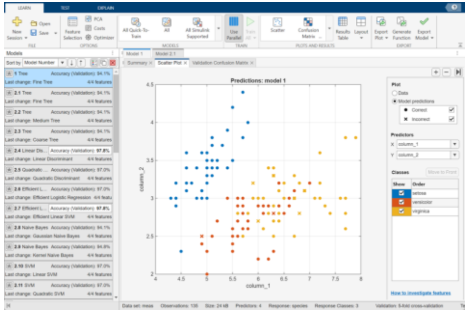
MATLAB, especially
Graph & Network Algorithms



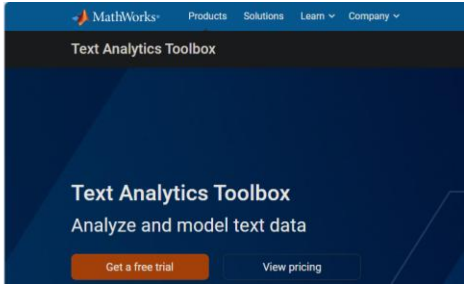
Backtesting Framework



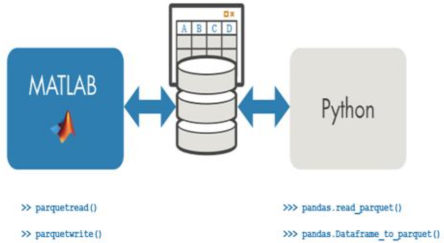
Experiment Manager



Classification Learner



Text Analytics



MATLAB/Python Integration
& MATLAB/Java Integration

Experiment Design

Structured Data Conversion

Transform raw PDFs into structured JSON, preserving the author's original context to reduce hallucinations

ML-Driven Enrichment

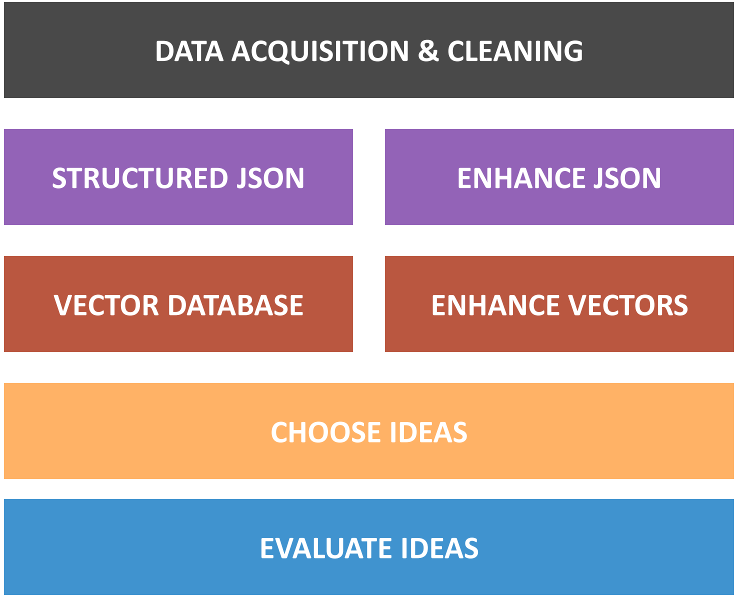
Apply traditional machine learning and computational linguistics to enhance the structured data and its subsequent vector representations

Graph RAG Ingestion

Use a GraphRAG to build our vector database, ensuring the rich, multi-dimensional context is maintained rather than flattened

Structured Prompting

Enriched, context-aware database enabling sophisticated prompting to identify contrarian and niche investment ideas



Industrial Engineering and Operations Research
COLUMBIA | ENGINEERING

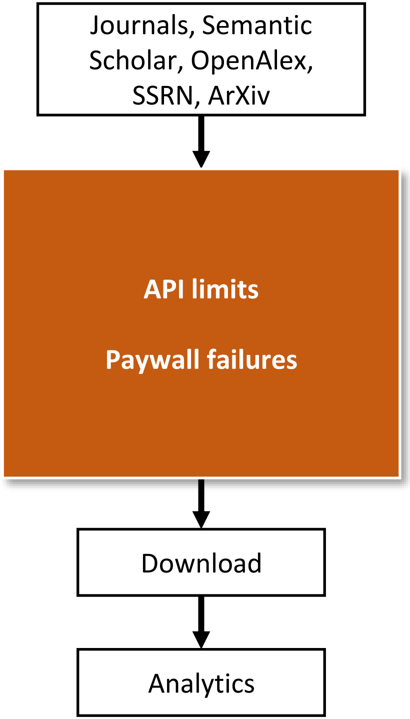
Sourcing & Acquisition of Ideas

The Challenge of Volume

- **Tens of thousands**
  of papers published quarterly
- **API limits**
  (e.g., SSRN 500/hr) make brute-force approaches impossible
- **High paywall failure rates**
  (~20%) even with authorization

Solution

A systematic, three-phase funnel to intelligently filter information

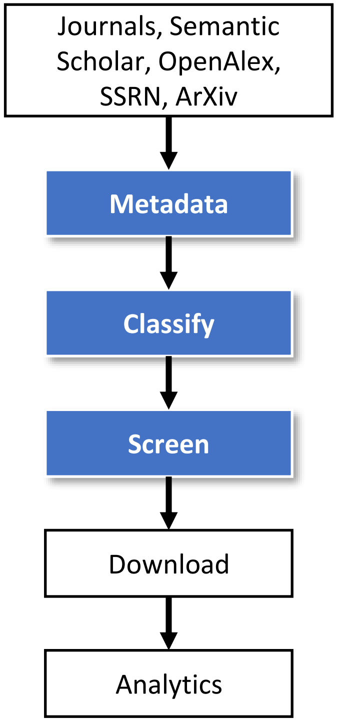


Industrial Engineering and Operations Research
COLUMBIA | ENGINEERING

Sourcing & Acquisition of Ideas

Our Three-Phase Process

- **Harvest**
  Pull metadata from diverse sources without touching full PDFs
- **Classify & Screen**
  Apply analytic learning process that scores and ranks papers based on metadata alone
- **Download & Store**
  Only download full PDFs of highest-potential papers, ensuring efficient use of limited API calls



Industrial Engineering and Operations Research
COLUMBIA | ENGINEERING

Cleaning & Structuring Data

The Challenge — Context Loss

Standard PDF extraction flattens text, destroying structure vital for analysis

Our Solution — Structured JSON

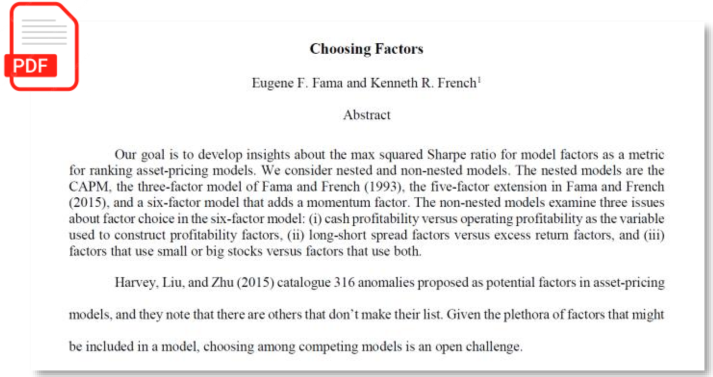
Converting papers into nested JSON format to preserve hierarchy and context

Technical Approach

Custom ingestor using **GROBID**, **PyMuPDF**, and **scipdf-parser**

Data Enrichment

Enhanced with metadata from OpenAlex and Semantic Scholar

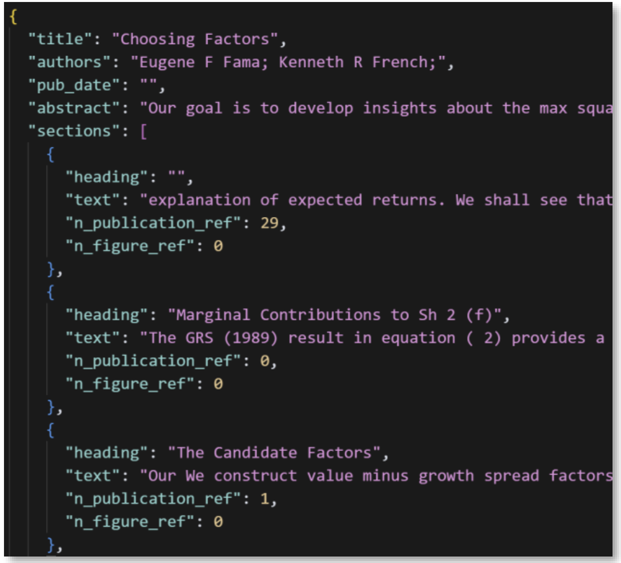

Choosing Factors

Eugene F. Fama and Kenneth R. French[1]

Abstract

Our goal is to develop insights about the max squared Sharpe ratio for model factors as a metric for ranking asset-pricing models. We consider nested and non-nested models. The nested models are the CAPM, the three-factor model of Fama and French (1993), the five-factor extension in Fama and French (2015), and a six-factor model that adds a momentum factor. The non-nested models examine three issues about factor choice in the six-factor model: (i) cash profitability versus operating profitability as the variable used to construct profitability factors, (ii) long-short spread factors versus excess return factors, and (iii) factors that use small or big stocks versus factors that use both.

Harvey, Liu, and Zhu (2015) catalogue 316 anomalies proposed as potential factors in asset-pricing models, and they note that there are others that don't make their list. Given the plethora of factors that might be included in a model, choosing among competing models is an open challenge.



```
{
  "title": "Choosing Factors",
  "authors": "Eugene F Fama; Kenneth R French;",
  "pub_date": "",
  "abstract": "Our goal is to develop insights about the max squa
  "sections": [
    {
      "heading": "",
      "text": "explanation of expected returns. We shall see that
      "n_publication_ref": 29,
      "n_figure_ref": 0
    },
    {
      "heading": "Marginal Contributions to Sh 2 (f)",
      "text": "The GRS (1989) result in equation ( 2) provides a
      "n_publication_ref": 0,
      "n_figure_ref": 0
    },
    {
      "heading": "The Candidate Factors",
      "text": "Our We construct value minus growth spread factors
      "n_publication_ref": 1,
      "n_figure_ref": 0
    },
```

Analyzing Structures

Relevance & Quality – *does the paper matter?*

Used **Computational Linguistics** to analyze sentiment, jargon, and specificity

Similarity & Discovery – *what other papers are like this one?*

Applied **k-Nearest Neighbors (kNN)** on document embeddings to find clusters of similar research

Methodology Vetting – *is the research reproducible and niche?*

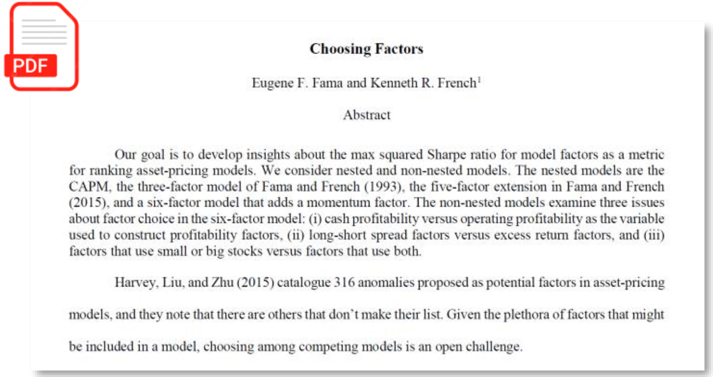
Analyzed **methodology tokens and citation depth** to assess the rigor and uniqueness of the approach

Network & Influence – *who is talking to whom?*

Used **Graph Analysis** (PageRank, centrality) to map the citation network of authors and ideas

Temporal Analysis – *when does an idea emerge and fade?*

Tracked topics and sentiment over time to understand the lifecycle of investment ideas



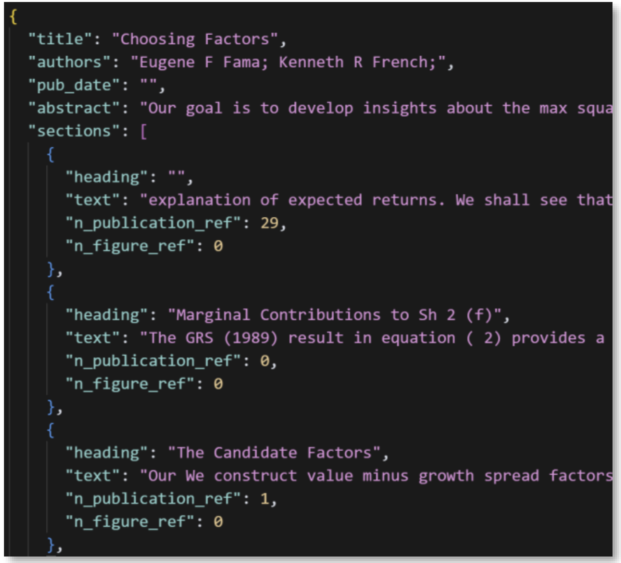
Choosing Factors

Eugene F. Fama and Kenneth R. French[1]

Abstract

Our goal is to develop insights about the max squared Sharpe ratio for model factors as a metric for ranking asset-pricing models. We consider nested and non-nested models. The nested models are the CAPM, the three-factor model of Fama and French (1993), the five-factor extension in Fama and French (2015), and a six-factor model that adds a momentum factor. The non-nested models examine three issues about factor choice in the six-factor model: (i) cash profitability versus operating profitability as the variable used to construct profitability factors, (ii) long-short spread factors versus excess return factors, and (iii) factors that use small or big stocks versus factors that use both.

Harvey, Liu, and Zhu (2015) catalogue 316 anomalies proposed as potential factors in asset-pricing models, and they note that there are others that don't make their list. Given the plethora of factors that might be included in a model, choosing among competing models is an open challenge.




```
{
  "title": "Choosing Factors",
  "authors": "Eugene F Fama; Kenneth R French;",
  "pub_date": "",
  "abstract": "Our goal is to develop insights about the max squa
  "sections": [
    {
      "heading": "",
      "text": "explanation of expected returns. We shall see that
      "n_publication_ref": 29,
      "n_figure_ref": 0
    },
    {
      "heading": "Marginal Contributions to Sh 2 (f)",
      "text": "The GRS (1989) result in equation ( 2) provides a
      "n_publication_ref": 0,
      "n_figure_ref": 0
    },
    {
      "heading": "The Candidate Factors",
      "text": "Our We construct value minus growth spread factors
      "n_publication_ref": 1,
      "n_figure_ref": 0
    },
```

Ingestion & Analysis of Vector Data

The Challenge — Preserving Context

A standard RAG would flatten our enriched JSON files, destroying the critical context and relationships we worked to preserve

Our Solution — LightRAG

We use the LightRAG architecture, building a vector database that maintains the hierarchical structure of our data, enabling true context-aware queries

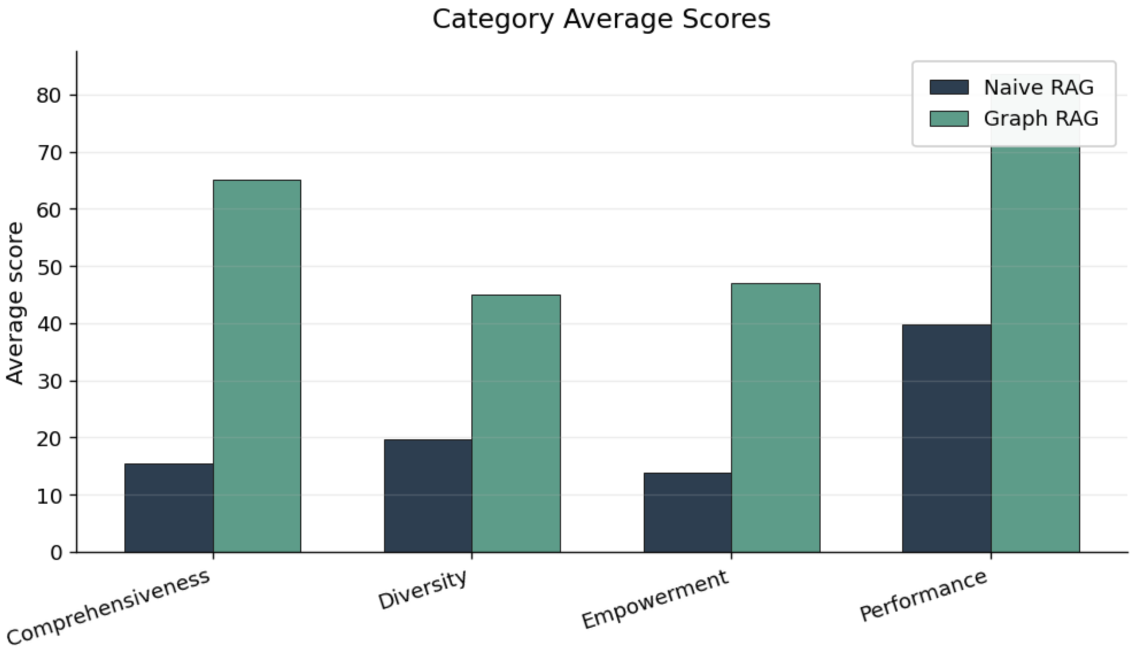
Key Feature — Incremental Ingestion

LightRAG allows us to add new papers and updated metadata over time without rebuilding the entire database

Technical Workflow

Our process is orchestrated in MATLAB to parse the JSON corpus. We then interface with Python via the MATLAB Engine API to leverage specialized NLP libraries

Ingestion & Analysis of Vector Data



Sources: [ ]

LightRAG Architecture & Implementation

Why LightRAG?

Full GraphRAG

- Expensive (neural nets, traversal)

LightRAG

- Efficient, preserves relationships, scalable
- Incremental ingestion
    - Add/update without rebuilding
    - Ideal for distributed team

Technical Implementation

Embeddings

Sentence transformers (MiniLM-L6-v2)

Graph structure

NetworkX, LightRAG-HKU

Similarity

scikit-learn cosine

Entities/NER

SpaCy

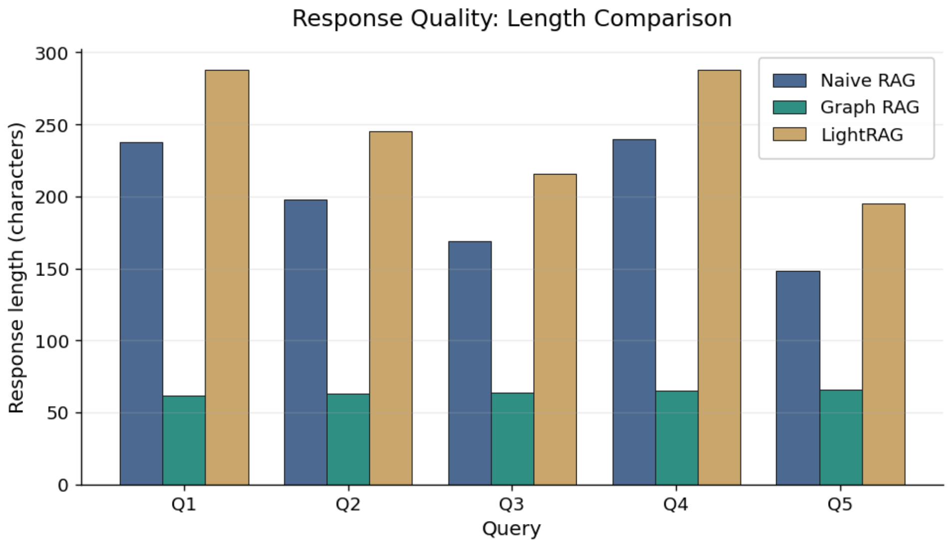
Database

Neo4j

Highlights

- **45,000+ entities with mapped relationships**
- **Machine learning classification**

  Automatic extraction/classification of financial concepts
- **2.4× more comprehensive responses vs. naïve RAG**
- **Successfully preserves document interconnections**



Sources: [ ]

Prompting & Evaluation

The Payoff: Sophisticated Querying

The Culmination of Our Pipeline

The structured data, enriched metadata, and vector database enable questions impossible for out-of-the-box language models.

> *"Show me papers with high linguistic complexity, referenced by authors from top-tier universities, that are not widely cited."*

> *"Find papers with similar vector profiles to this niche Auction paper, but exclude any that rely on end-of-day data."*

Iterative Process

Prompt, evaluate against professional analyst standards, refine, repeat

Pre-calculated Metadata

You can't just drag 4,000 PDFs into an LLM and expect results

Portfolio Manager Intuition

Our system mimics the thought process of experienced professionals

Prompting & Evaluation

Finding Needles in the Haystack

Treasury Auction Paper

An opportunity most investors ignore due to significant barriers to entry:

- Only works four days a year
- Requires difficult-to-obtain data
- Needs high-frequency data processing
- Requires high leverage (operational risk)

Complexity creates a barrier to entry, meaning less competition

Pre-Refunding Announcement Gains in U.S. Treasurys*

Chen Wang† Kevin Zhao‡

July 16, 2025

Abstract

We document substantial and intensifying positive returns in medium- and long-term Treasury bonds on the day before the Treasury Refunding Announcements (TRAs), an important quarterly fiscal event where future issuance plans are unveiled. Pre-TRA gains are distinct from known calendar effects, account for a sizable portion of annual yield and term premium changes, and cannot be attributed to information leakage. We show that reduction in Treasury market uncertainty—particularly fiscal-related uncertainty—prior to TRAs is the key driver. Consistent with this, pre-TRA gains are stronger when immediately following an FOMC meeting, and when national debt approaches the debt ceiling.

*Views and opinions expressed are those of the authors and do not necessarily represent official positions or policies of the OFR or Treasury. We thank Mark Carey, Hoyong Choi, Zhi Da, Corey Garriott, Leyla Han, Byoung-Hyoun Hwang, Zhiguo He, Grace Xing Hu, Francisco Ilabaca, Hanno Lustig, Emanuel Moench, Stacey Schreft, Philipp Schuster, and seminar and conference participants at Quantpedia, Notre Dame, 11th SAFE Asset Pricing Workshop, Quoniam, the CUHK-RAPS-RCFS Conference on Asset Pricing and Corporate Finance, MFA, and FIRS for their helpful comments. All errors are our own. First version: March 18, 2024.

†University of Notre Dame. E-mail: chen.wang@nd.edu
‡Office of Financial Research. E-mail: kevin.zhao@ofr.treasury.gov

Prompting & Evaluation

Finding Needles in the Haystack

Gamma Imbalance Paper

An edge exists for those willing to do the difficult calculation work:

- Most research uses over-simplified signal
- True signal is extremely complex to calculate
- Requires massive, expensive options data
- Needs high-frequency, often messy data

Anyone willing to do the hard work has a significant edge

Hedging demand and market intraday momentum

Guido Baltussen[1,3,*], Zhi Da[2], Sten Lammer

[1]Erasmus School of Economics, Erasmus University R
50, Rotterdam 3000 DR, Ne
[2]University of Notre Dame, 239 Mendoza College of
USA.
[3]Robeco Quantitative Investing, Weena 85

26th August 2020

Abstract

Hedging short gamma exposure requires trading in
thereby creating price momentum. Using intraday retu
bonds, commodities, and currencies between 1974 and 2
intraday momentum" everywhere. The return during the
close is positively predicted by the return during the rest
close to the last 30 minutes). The predictive power is ec
significant, and reverts over the next days. We provide
intraday momentum to the gamma hedging demand from
makers of options and leveraged ETFs.

JEL Classification: G12, G15, G40, Q02.

Keywords: Return momentum; Futures trading; Hedgi
Indexing.

Corresponding author. Email: baltussen@ese.eur.nl. E
University Rotterdam, Burgemeester Oudlaan 50, Rotter

We thank SqueezeMetrics for providing data, Kester Bro
research assistance, and Wouter Tilgenkamp, Xiao Xiao, a

Journal of Economic Dynamics & Control 164 (2024) 104880

Contents lists available at ScienceDirect

Journal of Economic Dynamics and Control

journal homepage: www.elsevier.com/locate/jedc

Gamma positioning and market quality ☆

Boyd Buis[a], Mary Pieterse-Bloem[b], Willem F.C. Verschoor[c], Remco C.J. Zwinkels[c,*]

[a] NatWest Markets and Vrije Universiteit (VU) Amsterdam, the Netherlands
[b] Rabobank, Erasmus School of Economics, and ERIM, the Netherlands
[c] Vrije Universiteit (VU) Amsterdam and Tinbergen Institute, the Netherlands

A R T I C L E I N F O

JEL classification:
C15
C63
D40
G23

Keywords:
Dynamic hedging
Feedback effect
Market liquidity
Market quality
Simulation

A B S T R A C T

In this paper, we study the effect of the gamma positioning of dynamic hedgers on market quality through simulations. In our zero-intelligence model, the presence of dynamic hedgers enhances market liquidity under normal conditions. However, positive gamma helps sustain liquidity in stressed scenarios, while negative gamma depletes it. We find that an increase in the net gamma positioning of dynamic hedgers reduces volatility and increases market stability, whereas a negative gamma positioning increases volatility and makes the market more prone to failure. Price discovery typically worsens when dynamic hedgers become more prevalent, regardless of the sign of their positioning. Our findings imply that steering the net gamma position of dynamic hedgers can be considered a policy instrument to improve market quality, especially for instruments with low liquidity or low traded volume.

1. Introduction

In November 2014 an unexpectedly large number of sell orders caused U.S. Treasuries to drop 1.6% before rebounding fully by an equally unprecedented number of buy orders. The intraday largest Treasury move since 2009 is attributed to a large short option position amongst delta neutral traders (Levine et al., 2017). This phenomenon is called a gamma trap or a gamma squeeze: excessive price volatility induced by dynamic hedgers who involuntarily act as momentum traders due to their short gamma position and preference for an overall delta-neutral position.

Risk sensitivities of option contracts to certain parameters are denoted by 'Greeks,' which are the partial first derivatives of the option price in the Black and Scholes (1973) formula to option characteristics. The most prominent of these Greeks is delta, which measures the price sensitivity of the option with respect to price changes in the underlying security. Gamma is a second-order Greek and measures the degree by which this delta changes when the underlying security's price moves.[1] All option contracts exhibit some non-zero gamma; option contracts that are close to maturing and 'at the money' exhibit the largest gamma. Dynamic hedgers are market participants whose objective is to maintain a constant delta. Dynamic hedging trading desks are prevalent at banks, insurers,

☆ The views expressed in this paper represent those of the authors and not those of NatWest Markets or Rabobank.

* Corresponding author.
E-mail addresses: boyd.buis@gmail.com (B. Buis), pietersebloem@ese.eur.nl (M. Pieterse-Bloem), w.f.c.verschoor@vu.nl (W.F.C. Verschoor), r.zwinkels@vu.nl (R.C.J. Zwinkels).

[1] Gamma is the sensitivity of delta with respect to changes in the underlying price. Since delta itself is the sensitivity of the option value with respect to changes in the underlying, gamma is considered a second-order Greek; it is the second derivative of the Black and Scholes (1973) option-pricing formula to the price of the underlying. Other frequently used Greeks are vega (price sensitivity with regards to changes in volatility), theta (price changes due to the evolution of time), and rho (price changes due to changes in interest rates).

https://doi.org/10.1016/j.jedc.2024.104880
Received 23 December 2022; Received in revised form 2 February 2024; Accepted 14 May 2024
Available online 17 May 2024

Bloomberg's Experiment

BloombergGPT Release

Specialized finance model outperforms GPT-3.5 on domain tasks

Short Time Later

General models advance rapidly

GPT-4 Release

New general model overtakes specialized model

The Sobering Lesson

In today's environment, simply specializing in a domain is not a durable competitive advantage.

Strong general models tend to dominate over time.

Bloomberg Professional Services —— Share in

Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance

March 30, 2023

BloombergGPT outperforms similarly-sized open models on financial NLP tasks by significant margins – without sacrificing performance on general LLM benchmarks

NEW YORK - Bloomberg today released a research paper detailing the development of BloombergGPT^{TM}, a new large-scale generative artificial intelligence (AI) model. This large language model (LLM) has been specifically trained on a wide range of financial data to support a diverse set of natural language processing (NLP) tasks within the financial industry.

Takeaway — Building a Resilient Edge

1. **Structured Data First**

   Preserve the author's original context to **reduce hallucinations** and create a reliable foundation

2. **Enrich with Traditional ML**

   Enhance data with computational linguistics and graph analysis, adding insights LLMs don't have

3. **Use a GraphRAG / LightRAG**

   Maintain the rich, hierarchical context of our data, enabling deeper analysis

4. **Target a Contrarian Objective**

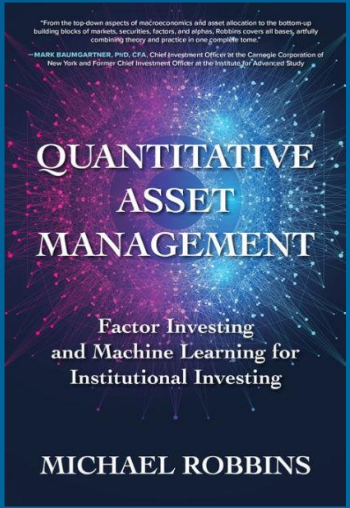
   Hunt for **niche and contrarian** ideas that consensus-driven general models are designed to ignore

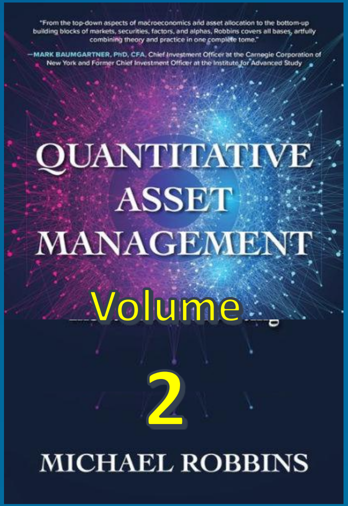
5. **Enable Sophisticated Testing**

   Use **structured prompting** to turn strategy ideation into a defensible, repeatable process

We Need More Research Projects!

- Each semester, we receive research requests from hedge funds, banks, advisors, and other investment management firms.
- We chose **100 examples** from well over **1,500 research projects** and presented them in **Volume 2 of Quantitative Asset Management *(expected)***.





Thank You!

michael.robbins@QuantitativeAssetManagement.com