Project Proposal: Dataset Construction for Quantitative Finance Using LLM-Assisted Extraction

1. Overview

This project proposes a structured methodology for constructing a high-quality Quantitative Finance dataset using textbook content and LLM-supported extraction. The objective is to systematically identify and collect instruction-tuning samples, reasoning examples, and tool-use/code snippets directly from authoritative academic material. By grounding dataset creation in textbook pedagogy rather than synthetic generation, the project aims to address the domain-specific shortcomings of existing LLMs in quantitative finance.

---

2. Motivation

State-of-the-art LLMs frequently hallucinate financial concepts, misapply mathematical definitions, and fail to maintain precision in derivations and reasoning tasks central to quantitative finance. One major contributor is the lack of academically rigorous training data available for open-source model development. Despite frontier models being trained on textbook-derived corpora, such datasets are not openly accessible.

This project aims to fill that gap by creating a foundational, structured dataset built directly from quant finance textbooks, enabling future LLMs to acquire deeper conceptual understanding and reduce hallucination in quant tasks.

---

3. Method

3.1 Sliding-Window Extraction

The dataset is constructed by scanning textbooks using a **three-page sliding window**:

- Each window contains a contiguous 3-page excerpt from the textbook.
- This excerpt is passed to an LLM in an iterative loop.
- For each window, the LLM identifies and describes the categories of information present:
    - **Instruction-tuning samples** (e.g., problem statements, derivations to complete)
    - **Reasoning samples** (multistep arguments, formula derivations, conceptual explanations)

    - **Tool-use/code samples** (algorithmic procedures, computational expressions)

The pipeline records whichever categories naturally appear. If a given segment contains reasoning but no code, only reasoning data is extracted. This ensures the dataset mirrors the authentic structure and variability of textbook content.

3.2 Category-Based Extraction

For each window, the LLM extracts:

- All instruction-like material suitable for training models on directed tasks.
- All reasoning chains that exemplify structured financial logic.
- All computational or code-relevant content when present.

The goal is comprehensive collection — not forced category balancing — preserving the pedagogical integrity of the original text.

---

4. Purpose and Intended Use

The overarching purpose is to build a **high-quality, academically grounded dataset** tailored to the needs of quantitative finance.

This dataset will:

- Provide structured, domain-specific material that LLMs currently lack.
- Improve the reliability and conceptual depth of models operating in quantitative finance.
- Serve as a fundamental building block for training LLMs on quant tasks such as:
    - Mathematical derivations
    - Financial reasoning
    - Problem-solving
    - Computational and algorithmic implementations

As models gain stability and domain comprehension from this dataset, they can be leveraged to automate higher-level quant workflows — including code generation, strategy development, and analytical tooling — particularly within development environments such as Cursor.

---

5. Significance

This work enables the creation of a dataset that is:

- Authentic to academic sources
- Systematically extracted
- Balanced according to naturally occurring pedagogical distribution
- Directly usable for fine-tuning or foundational model training in quantitative finance

It lays the groundwork for a robust, conceptually grounded training resource that addresses gaps in current LLM performance and supports the development of specialized models for quant reasoning and computation.