

## 1 Task Definition



### Expertise

### Expert Panel



### Autonomy

Fin\_Basic, Fin\_Num\_Calc



### Integrity

Role\_Adapt, Call\_API

Fin\_App\_Sec, MT\_Endo

## 2 Data Collection



### Academic Sources

Public Dataset, Professional Testbank



### First-hand Data

Financial Report, Anon. Loan Data, Regulation



### Constructed Sets

Security test cases, Simulated API flows

## 3 QA Generation



### LLM Synthesis

draft candidate Q&A



### Expert Authoring

create realistic cases



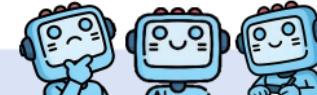
### QA Template

Task Instruction  
Background  
Specific Query  
Format Rules

## 4 QA Refinement

### Iterative Filtering

sample questions → dual independent attempts per LLM → remove easy Q&A



### Difficulty Control

### Expert-AI Verification

multi-round LLM voting → discard ambiguous items → expert review for Q&A quality



### Logical & Semantic Rigor

### Final Expert Review

multi-stage review by 21 anti-fraud experts → validate safety, coverage, and realism



### Professional & Safety Alignment