Executive Summary

SEC-Based Prompt Collection for LLMs

A human-in-the-loop data system for generating high-quality financial prompts

Goal: Build a scalable, quality-controlled data pipeline to generate prompt–answer pairs from SEC filings that test LLM reasoning, multi-document synthesis, and factual accuracy.

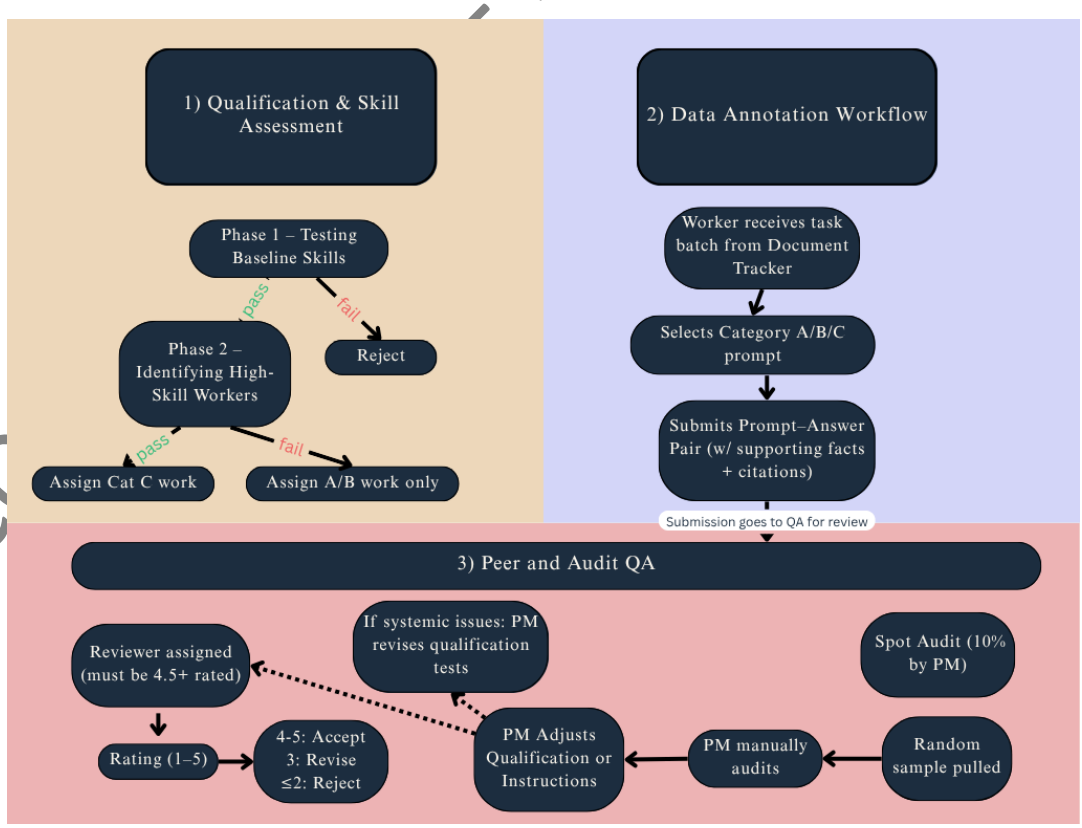
Core Components:

- **Contributor Qualification (2 phases):** Skill stratification to gate complex prompts
- **Prompt Categories (A/B/C):** Tiered by complexity and evaluative value
- **QA Pipeline:** Peer review + PM audit with revision and deactivation logic
- **Compensation Model:** $17–22/hour equivalent with bonuses by complexity

Outputs: 100+ prompts, structured across categories, using a document assignment tracker to ensure coverage across financial sectors and prompt complexity.

Use Cases:

- LLM evaluation in high-stakes, high-fidelity domains
- Training/benchmark data for RLHF
- Human-in-the-loop QA systems for high-precision language tasks



SEC-Based Prompt Collection for LLMs

Table of Contents

This document outlines a scalable, quality-first data collection project for generating financial prompts from SEC filings, suitable for LLM training or evaluation pipelines.

1) Financial Qualification & Skill Assessment

Goal:

Stratify workers by skill level to ensure high-quality financial prompts for Categories A, B, and C. This qualification process identifies contributors with the ability to read and analyze financial filings, with Phase 1 screening for foundational ability and Phase 2 identifying candidates capable of complex multi-document analysis.

Required Skills:

- Understanding of financial concepts (e.g., revenue, expenses, year-over-year comparisons)
- Ability to navigate 10-K and 10-Q filings
- Clarity and precision in writing factual questions
- Basic numerical reasoning and ability to compare across multiple sections or documents

1.1) Phase 1 – Testing Baseline Skills

Tasks:

Using the 10-K report for NVIDIA Corporation for the fiscal year ending January 26, 2025,

1. *Write a simple factual question with a numerical answer.*
    - *Provide the answer and cite the page number.*

Example:

Prompt: What was NVIDIA's total revenue for FY 2025?
Answer: The revenue was $130.5 billion as stated in the Fiscal Year 2025 Summary (p. 64).

2. *Write a factual question that requires analyzing information from two different sections: The question should require a text-based answer.*
    - *Provide an answer in 2-4+ sentences, show your reasoning, and cite the page numbers used.*

Example:

Prompt: How does NVIDIA describe the geographic distribution of its supply chain, and what risks does it associate with this setup?

Answer: NVIDIA states that the majority of its supply chain is concentrated in the Asia-Pacific region (p. 37, Global Trade section). The company highlights risks related to U.S.–China

geopolitical tensions and notes that export controls from the U.S. government could significantly affect its operations (p. 10, Risk Factors).

*Please ensure your prompt does **not** include any personally identifiable information (e.g. email addresses).*

Phase 1 Pass Criteria:

- **Question formulation:** Clear, factual, and verifiable prompts that reflect proper understanding of section context
- **Question 1:** Clearly formulated, numerical answer directly traceable to a financial table with correct citation.
- **Question 2:** Demonstrates analysis across both sections with correct interpretation; logical and well-reasoned answer; accurate citations.

1.2) Phase 2 – Identifying Category C Workers

Using the Q1 2025 10-Q filings from General Motors, Ford, and Tesla:

- *Write one factual question that compares information across all three companies.*
    - *Your answer in 4-6+ sentences with an explanation showing your reasoning and appropriate citations with page numbers used from each document*

Example:

How did General Motors, Ford, and Tesla approach their EV strategy in Q1 2025?

In Q1 2025, the three automakers pursued different EV strategies. GM anticipates making investments in suppliers or secure a steady supply of critical materials through strategic partnerships (GM 10-Q, p. 34). Ford emphasized hybrid flexibility and regulatory responsiveness, noting delays in full EV transitions due to uncertain U.S. policy environments (Ford 10-Q, p. 32). Tesla, by contrast, doubled down on vertical integration by heavily investing in its battery cell production and development. (Tesla 10-Q, p. 26). These strategies reflect different risk tolerances and interpretations of the evolving U.S. regulatory landscape.

Phase 2 Pass Criteria:

- **Question formulation:** Ask for content across *all three* companies into one coherent prompt (not three disjoint questions)
- Writes a substantial, factually grounded, and well-reasoned response with accurate citations.

Red Flags

- Includes **opinion-based or speculative questions** (e.g., “Why does Tesla think it's better than Ford?”).
- Fails to **cite page numbers** or makes unsupported claims.
- Uses **PII or personal references**
- Repeats information without meaningful analysis or inference in Phase 2.

1.3) Flowchart: Qualification, Collection & QA Overview

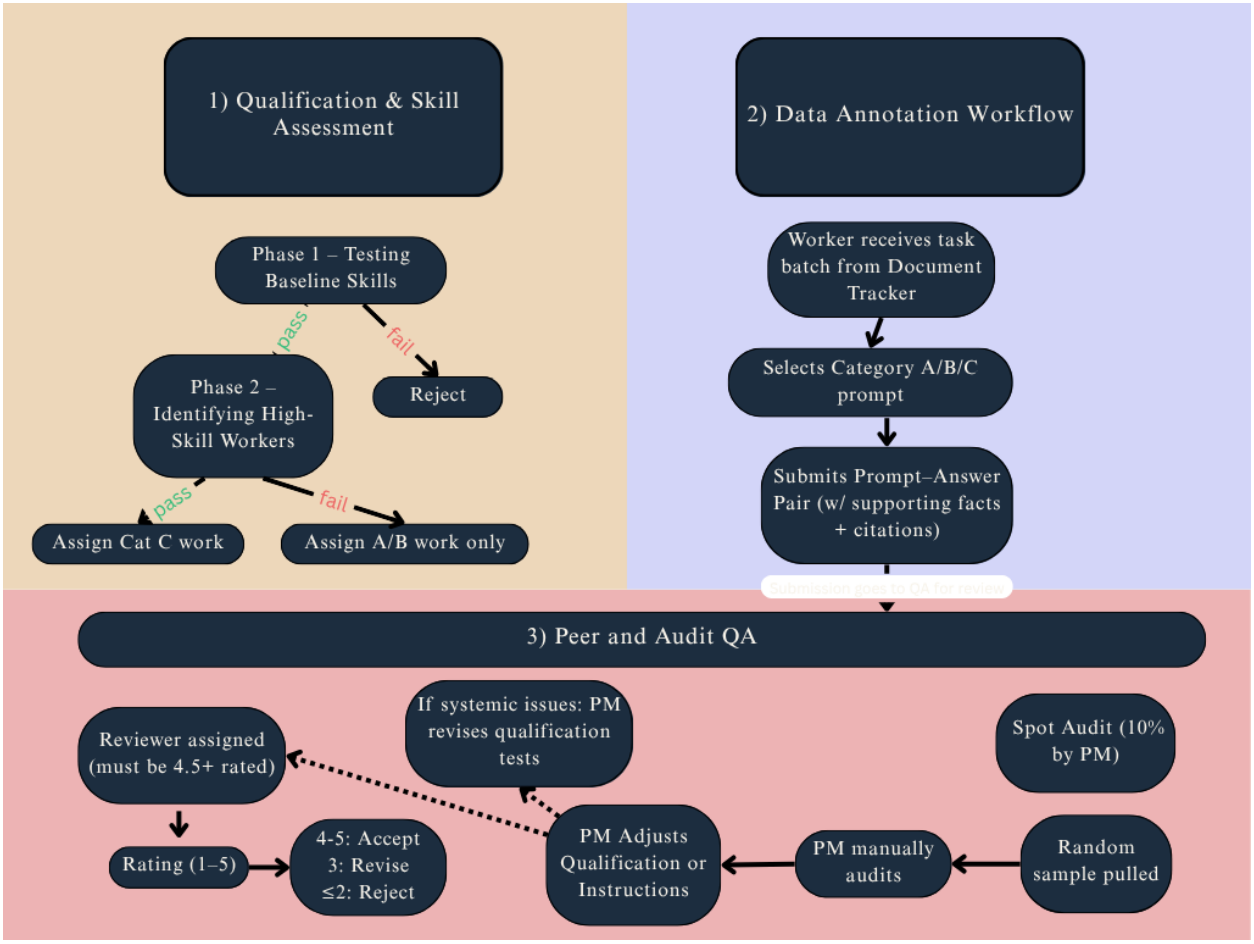


Figure 1. Overview of Qualification, Annotation Workflow, and QA Feedback Loops

2) Financial Data Annotation Workflow

2.1 Overview

Goal: To collect 100 high-quality prompt–answer pairs, this project follows a structured distribution:

- **40% Category A (simple, single document)**
- **40% Category B (complex, single document)**
- **20% Category C (multi-document analysis)**

This 40–40–20 split balances feasibility and evaluative utility. While Category A tasks are faster to complete, Categories B and C provide richer insights into LLM reasoning capabilities. Category C is capped at 20% due to the high complexity and limited pool of qualified contributors.

To reduce variability, ensure topical coverage, and speed up task execution, all workers must select documents from a **pre-vetted Google Sheet**. This central tracker includes:

- A selection of U.S. companies across sectors
- Direct SEC URLs (10-K and 10-Q only)
- Filings from **October 2023 or later**
- Pre-saved PDF links

Providing this document set avoids duplication, streamlines review, and prevents confusion about file types, dates, or industries. The sheet also supports load balancing as workers will be assigned non-overlapping documents based on their task category.

Pricing is based on a $20/hour average wage and estimated completion times:

- Category A: $2.00 (~6–8 minutes)
- Category B: $3.25 (~10–12 minutes)
- Category C: $8.25 (~20–25 minutes)

Converted into estimated hourly pay:

- Category A: $2.00 for ~7 minutes → ~$17.14/hour
- Category B: $3.25 for ~11 minutes → ~$17.73/hour
- Category C: $8.25 for ~22.5 minutes → ~$22.00/hour

This is supplemented by performance-based bonuses to reward accuracy and effort, especially for Category C contributors.

- **Accuracy Bonus**: $5 awarded to any contributor who completes at least 5 prompts (across any category) with an average reviewer score of 4.5 or higher.
- **Complexity Bonus**: $10 bonus for completing five or more Category C prompts.

A recent academic study on data annotation work shows that hourly pay combined with clear instructions increases accuracy.[1] However, a blog post from LLM training data provider Pareto[2] suggests expert annotators prefer per-task pay to match their productivity. Our hybrid model balances these insights, offering per-task pay that scales with complexity, alongside QA-based incentives to maintain high quality.

Note on requirement for citations:

While the client does not require citations or page numbers in the final deliverables, we collect them for the internal QA process. As per client instructions, the dataset will be delivered with only supporting facts, without page numbered citations.

2.2) Worker Instructions

Financial Prompt/Response Collection – Worker Instructions

Overview

You will generate high-quality prompt–answer pairs using publicly available SEC filings (10-K and 10-Q). These prompts will be used to evaluate the performance of a large language model (LLM). Your task is to read the filings, formulate factual questions grounded in the documents, provide accurate answers, and supply supporting citations. Adherence to the instructions below is essential to ensure consistency and quality.

2.2.a) General Requirements

Each submission must include the following components:

1. ***Prompt/Question** – A factual, document-based question written clearly and concisely.*
2. ***Answer –** A complete and accurate answer, supported by relevant evidence from the source document(s).*
3. ***Supporting Facts –** Specific data points or reasoning used to formulate the answer.*
4. ***Citations –** Section(s) and page number(s) referenced.*
5. ***Documents –** URLs linking to the source SEC document(s), publication date(s), and saved PDF(s) of each SEC document used in your response.*

Do not include:

- *Personal opinions, subjective language, or speculative claims.*

[1] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4673217
[2] https://pareto.ai/blog/data-labeling-ppt-vs-hourly-wages

- *Any personally identifiable information (PII), such as email addresses.*

2.2.b) Approved Document Selection

To ensure quality and reduce research time, you are required to select from your assigned batch of filings from the official ***Document Tracker Google Sheet*** *provided here. This sheet contains:*

- *Company name*
- *Filing type (10-K or 10-Q)*
- *Filing date (must be October 2023 or later)*
- *Direct SEC filing URL*

Do not reuse documents until all documents in your designated batch have been exhausted.

Do not use documents outside this sheet without prior approval from the project manager.

2.2.c) Prompt Categories

Category A – Simple Single-Document Questions

- *Use a* ***single filing****.*
- *Ask straightforward factual questions that can be answered from one section (e.g., financial statements).*
- *Keep answers concise (1–2 sentences).*
- *Estimated completion time: 6–8 minutes.*

Example

- ***Prompt:*** *What was Company A's R&D expenditure for Q4 2024?*
- ***Answer:*** *Company A spent $7.5 billion on research and development in Q1 2024.*
- ***Supporting Facts:*** *The financial statements show an R&D expense line item of $7.5B.*
- ***Citations:*** *Financial Statements p. 27*
- ***Documents:*** *[SEC URL], publication date and PDF*

Category B – Complex Single-Document Questions

- *Use a* ***single filing****.*
- *Ask questions requiring analysis across sections (e.g., combining risk factors with financial performance).*
- *Provide a substantial answer while showing your reasoning (minimum 2-3 sentences).*
- *Estimated completion time: 10–12 minutes.*

Example

- ***Prompt:*** *How did Company B explain the decline in tablet sales and what risks are associated with this trend?*
- ***Answer:*** *Company B cited decreased education sector demand and increased component costs. It also identified supply chain volatility as a key ongoing risk, which could potentially drive component costs higher.*
- ***Supporting Facts:*** *Decreased education sector demand and increased component costs are referenced in the MD&A section, while supply chain disruptions are mentioned in Risk Factors.*
- ***Citations:*** *MD&A (p. 33), Risk Factors (p. 33)*
- ***Documents:*** *[SEC URL], publication date and PDF*

Category C – Multi-Document Questions

- *Use **2 to 5 filings** (up to 40 maximum), ideally from companies in the same industry.*
- *Ask comparative or summarizing questions across documents.*
- *Provide detailed answers (minimum 4–6 sentences with comparison and summary)*
- *Estimated completion time: 20–25 minutes.*

Example

- ***Prompt:*** *Compare how Companies C, D, and E approached EV production and regulation in Q1 2025.*
- ***Answer:*** *In Q1 2025, the three companies pursued distinct strategies to navigate EV production and changing regulatory environments. Company C emphasized in-house battery manufacturing, whereas Company D focused on increasing flexibility in hybrid production to adapt to changing U.S. government policies. Company E focused on scaling its EV subsidiary through strategic supplier partnerships. Together, these strategies demonstrate differing risk tolerances and timelines for EV expansion across the sector.*
- ***Supporting Facts:*** *The Company C 10-Q MD&A section states that "We expanded battery cell production at our Nevada facility to reduce reliance on external suppliers." Company D's 10-Q states that "Our hybrid models provide a buffer against fluctuating regulation while keeping production flexible." Company E's 10-Q includes the following quote: "Strategic alliances with domestic suppliers and state officials have accelerated EV unit certification under the Inflation Reduction Act."*
- ***Citations:*** *Company C 10-Q (p. 28), Company D 10-Q (p. 15), Company E 10-Q (p. 12)*
- ***Documents:*** *[Three SEC URLs], publication dates and three PDFs*

2.2.d) Submission Standards

- *All prompts must be grounded in verifiable, factual content from the filing(s).*
- *Use clear, formal English.*
- *Do not repeat the same company or topic multiple times, until all documents have been exhausted.*

2.2.e) Compensation Structure

- *Category A: $2.00 per task*
- *Category B: $3.25 per task*
- *Category C: $8.25 per task*

Bonuses are available based on quality and volume:

- ***Accuracy Bonus**: $5 awarded to any contributor who completes at least 5 prompts (across any category) with an average reviewer score of 4.5 or higher.*
- ***Complexity Bonus**: $10 bonus for completing five or more Category C prompts.*

If there are any questions regarding the instructions, please reach out to the PM at haakon@surge.ai.

2.3) Document Distribution Strategy

Workers will be assigned documents in the following manner:

- Category A and B workers receive a batch of 15 non-overlapping documents, each spanning different companies and industries. This encourages variation across prompts while keeping review load manageable.
- Category C workers receive groups of 5–10 filings from companies within the same sector. This design supports multi-document comparisons.
- To avoid over-reliance on a narrow set of filings, workers are not allowed to reuse the same document until all items in their assigned batch are exhausted.

This system ensures coverage across sectors and filing types while preserving quality control and avoiding bottlenecks during QA and final dataset assembly.

Refresh Policy:

- Document assignments are refreshed weekly or as soon as 80% of a batch has been completed. The PM monitors document usage in real time for balance.
- If all documents in a worker's batch are used, the worker can request reassignment or pause work until a new batch is released.

QA Integration:

This structured distribution supports the QA process by:

- **Balancing load** across reviewers.
- **Ensuring variety** in prompt coverage

3) Peer and Audit QA

3.1) Procedure

To ensure data quality prior to delivery, we implement a two-tiered quality assurance (QA) process involving peer review and spot audits. This structure is designed to maintain standards while preventing bottlenecks.

i) Peer Review by Qualified Raters

Each prompt–answer pair is reviewed by a separate contributor who meets the following criteria:

- For Categories A and B: Reviewers must have completed at least three prompts in the relevant category and maintained an average rating of 4.5 or higher.
- For Category C: All contributors who have passed Phase 2 of the qualification project are eligible to review.

Each submission is rated on a 1–5 scale using a simple rubric provided below.

II) Spot Audits by Project Manager

The Project Manager (PM) will conduct spot audits on approximately 10% of submissions, with a focus on Category C data. Audits ensure consistency across reviewers and catch edge-case errors. Where systemic issues are identified, feedback will be integrated into updated instructions or reviewer training.

3.2) Reviewer Instructions

As a reviewer, your role is to evaluate the quality of each submission based on factual accuracy, clarity, and adherence to project guidelines. Follow the rubric below when rating each submission:

Rating Rubric (1–5):

Score	Description	Action
5	Fully accurate, well-written, meets all guidelines. No edits needed.	Accept
4	Minor issues that do not affect factual integrity or clarity.	Accept
3	Contains issues requiring revision (e.g., vague prompt, missing citation, weak analysis).	Revise
2	Major issues (e.g., incorrect facts, poor structure, misinterpretation of instructions).	Reject
1	Off-topic, clearly low-effort or contains PII.	Reject

Please flag any of the following issues:

- Use of subjective language or opinions not supported by the source
- Inclusion of personally identifiable information (PII)
- Duplicate or unverifiable questions
- Repetitive prompts that fail to introduce new analysis or insight

If a submission receives a rating of 3, leave a short comment explaining the issue and return it to the contributor for revision.

3.3) Handling Low-Performing Workers

To maintain quality, we implement corrective measures:

- **Automated Warnings:** Workers whose average rating falls below 3.5 across five consecutive tasks will receive an automated warning and feedback, including examples of accepted responses.
- **Temporary Suspension:** Workers who continue to underperform will be suspended from contributing further tasks for 3 days.
- **Deactivation:** If no improvement is shown after feedback, workers may be removed from the project.

3.4) Revision Process for Mid-Quality Submissions (Rating 3)

Submissions rated 3 are flagged for revision and returned to the original contributor with reviewer comments. The worker is expected to:

- Address the reviewer's feedback
- Re-submit within 48 hours.
- Retain the original structure, improving only the flagged aspects.

Upon resubmission, the revised task will be re-reviewed.