An Automatically Improving Method for Generating Descriptions of Financial Data Quality Grading with LLMs

Yang Zhao[1], **Yohei Ikawa**[1], **Bishwaranjan Bhattacharjee**[2]
[1]IBM Research – Tokyo, Japan [2]IBM T. J. Watson Research Center, USA
yangzhao@ibm.com yikawa@jp.ibm.com bhatta@us.ibm.com

Abstract

Generating descriptions for financial data quality grades (e.g., poor, fair, excellent) enhances both data quality assessment and the trustworthiness of AI models. Traditionally, grading criteria have been manually compiled by humans, a process that is time-consuming and requires domain-specific expertise. In this work, we propose an automated, automatically improving framework for describing financial data quality grades at arbitrary levels. Specifically, we first train a financial classifier to categorize data into multiple quality grades, with the theoretical capability to support arbitrary grading levels. Then, a collected list of financial hypernyms is used to optimize the description for each financial grade using two search strategies. The quantitative results show that the financial knowledge–aware editor improves description accuracy and the QWK correlation score by over 10 points respectively on a hold-out test set, while the qualitative results indicate better performance in terms of informativeness and trustworthiness. We release the code and data here[1].

1 Introduction

Grading financial data involves assigning a score to a document to indicate its relevance and quality within the financial domain. For example, a financial text may be graded as poor (score 1), fair (score 2), or excellent (score 3) to reflect its quality and domain relevance. Among other factors, generating descriptions for different grades plays an important role in several aspects: first, descriptions help establish clear criteria for each grade, enabling users to place greater trust in AI models; second, they can serve directly as annotation guidelines, helping users design LLM-based annotation prompts to filter high-quality data from large corpora such as FineWeb (Penedo et al., 2024), which has become increasingly popular in recent years.

Previously, many studies have relied on manually developed, domain-specific data grading criteria. For example, FineWebEdu[2] enlisted human annotators to create five data quality grading criteria, which were then integrated into annotation prompts to guide LLMs in extracting education-specific data. Despite such successes, prompt design still depends heavily on domain-specific human expertise, and in less familiar domains, generating accurate grade descriptions becomes even more challenging.

To address this shortcoming, we propose a automatically improving method for generating descriptions in financial data quality grading. Table 1 shows a initial 3-grade description and optimized 3-grade description. We herein focus on two research questions: (1) **How can we obtain quality grading for financial data at arbitrary levels?** (2) **How can we automatically generate informative descriptions for each grade?**

To answer these questions, in Section 2, we introduce a two-stage approach to obtaining binary annotations (financial/non-financial) for the financial data, and train a financial document classifier to generate probabilities, which are segmented into grading scores. In Section 3, we automatically optimize grade descriptions using a curated set of financial hypernyms and a financial knowledge–aware (Fin-aware) editor. This editor guides LLMs to produce and iteratively refine descriptions using the model's own feedback via PPO. We further explore two search strategies that improve both description accuracy and Quadratic Weighted Kappa (QWK) correlation by more than 10 points over baseline methods. Qualitative evaluation also confirms that the resulting descriptions are more informative and trustworthy.

[1]https://github.com/code4nlp1713/code

[2]https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

Proceedings of The 10th Workshop on Financial Technology and Natural Language. EMNLP-2025, Suzhou, China

Score	3-Grade description	Automatically Optimized 3-Grade description
1	Score 1 if the document is poor.	Score 1 if the document is poor at explaining financial concepts and institutions, or does not discuss the performance or valuation of assets or liabilities at all.
2	Score 2 if the document is fair.	Score 2 if the document presents general financial information in a utilitarian manner but uses vague terms and lacks specific details about the financial status of the entity or individual.
3	Score 3 if the document is excellent.	Score 3 if the document demonstrates clarity and depth in discussing financial topics, including performance metrics, risk management, financial strategies, and potential uncertainties, while providing transparent and accurate data.

Table 1: Examples of 3-grade and optimized 3-grade financial data quality descriptions. In the 3-grade scale, 3 denotes the highest quality and 1 the lowest.

2 Related Work

Several studies on prompt optimization are relevant to our work. Prompt optimization methods include paraphrasing (Jiang et al., 2020; Yuan et al., 2021; Haviv et al., 2021) and reinforcement learning (RL) approaches (Deng et al., 2022; Zhang et al., 2022), though prior RL methods often yield uninterpretable prompts or have limited action spaces. Kong et al. (2024) automate prompt rewriting via RL but target simpler, single-sentence tasks, while our method addresses longer, multi-criteria prompts (∼300 words). We further incorporate financial knowledge to improve description quality, distinguishing our approach from earlier work.

3 Financial Data Grading Annotation

We use FineWeb dataset (Penedo et al., 2024) and randomly select 600k documents for annotation. Because financial documents[3] are scarce in FineWeb, directly annotating such a large set is inefficient; thus, we adopt a two-stage approach to annotate ground-truth grading.

Stage 1: We first prompt LLMs to generate a list of around 200 financial keywords (see Appendix A for details on keyword generation) and sort the 600k documents in descending order based on their overlap with financial keywords in each document's bag of words. Annotation then begins from the head and tail of the sorted list for financial and non-financial classification, respectively.

Stage 2: We then employ Human-LLM collaborative annotation for binary classification, as it is much easier and more reliable than multi-scale annotation for both LLMs and humans. We use Mixtral-8x7B-Instruct[4] model to annotate documents and then ask a financial expert[5] to review the LLM's annotation to correct them using the same binary annotation instruction in Appendix B. After removing the identified error cases (4% error rate in the LLM's annotations), we obtain a ground-truth financial dataset consisting of 3,840 positive (high-quality financial) and 3,840 negative documents. Please see Appendix E for annotation details and data statistics.

Subsequently, we shuffle the financial and non-financial document sets separately, taking the first 1k financial and the first 1k non-financial documents respectively as the test set, with the remaining documents used to train a RoBERTa-based (Liu, 2019) financial classifier. The financial classification accuracy on the test set is 98.8%.

Dataset	Grade Levels	Annotated Documents
3-grade	Poor: 0.0 (<0.001) Fair: 0.5 (±0.015) Excellent: 1.0 (>0.999)	900 total (300 per level) 450 val., 450 testing
4-grade	Poor: 0.0 (<0.001) Fair: 0.33 (±0.015) Good: 0.66 (±0.015) Excellent: 1.0 (>0.999)	1.2k total (300 per level) 600 val., 600 testing

Table 2: Dataset description for 3-grade and 4-grade financial document classification.

Probability Segmentation as Quality Grading We take 580k unannotated documents from the FineWeb dataset and apply our financial classifier to assign a probability score[6] to each document. As expected, most probabilities are close to either 0 or 1, while thousands fall in the middle range (see Table 7 in Appendix F. In this study, we empirically define 'middle' probability thresholds for different quality levels, as Table 2 shows.

[3]A financial document is herein defined as any finance-related text within large corpora such as FineWeb.

[4]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1. It is licensed under Apache License 2.0.

[5]The annotator holds a Ph.D. degree and works in the financial industry.

[6]We extract the probability of label 1 from the softmax

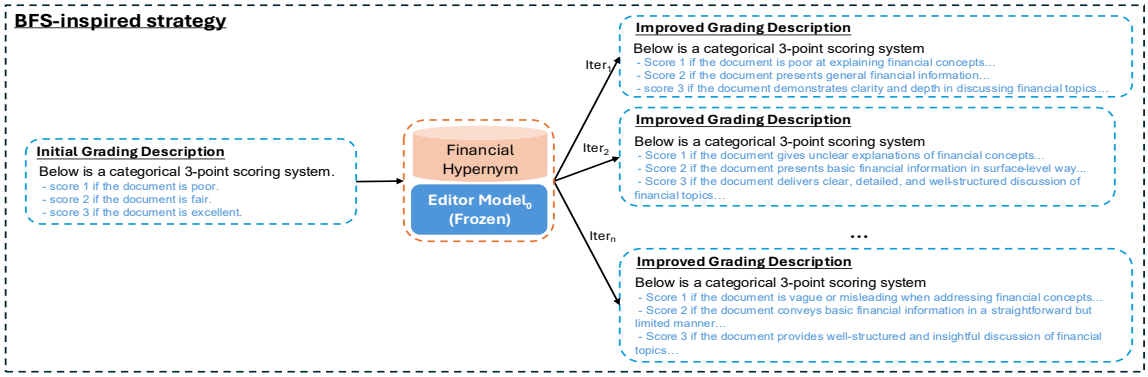

Figure 1: BFS-inspired strategy for automatically improving descriptions in data grading.

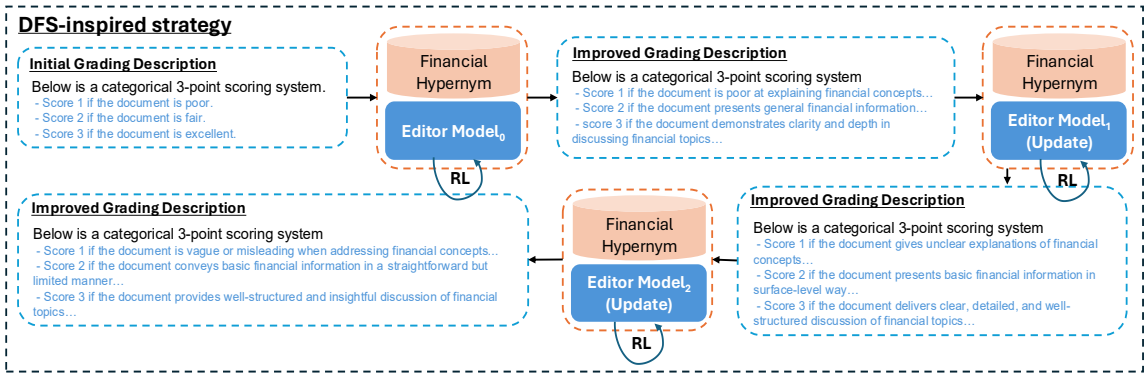

Figure 2: DFS-inspired strategy for automatically improving descriptions in data grading.

4 Proposed Method

To generate description for each grade, we frame the problem in the context of data annotation using LLMs: typically, an annotation prompt containing data grading criteria is manually crafted and provided to LLMs to generate classification results. For example, in the education domain, HuggingFace researchers manually designed a scoring prompt[7]. However, we reverse the problem herein: *given classification results (low/moderate/good/best) of documents, can we generate description for each grade without human effort?* We explore two search strategies for generating descriptions, inspired by Breadth-First Search (BFS) (Moore, 1959) and Depth-First Search (DFS) (Tarjan, 1972).

Formally, we define the description for data grade as X and the data annotation prompt containing this description as $P(X)$. The description X is iteratively refined by an editor model, LLM_{edt} into X'. We provide the prompt $P(X)$ to an evaluation language model, LLM_{eval}, which generates the predicted grade Y_{pred}. The ground-truth grade is denoted as Y_{true}. We define the difference between Y_{pred} and Y_{true} using the Quadratic Weighted Kappa (QWK) Correlation (Cohen, 1968) whose domain ranges from -1 and 1. Our goal is to find the optimal X' that maximizes QWK score.

4.1 BFS-Inspired Strategy

Based on the problem formulation, the BFS-inspired strategy aims to generate as many description as possible iteratively from the initial description X_0 (see the Appendix C). As shown in Figure 1, every time, editor model LLM_{edt} will only edit description according to editing prompt (see the Appendix D). We generate N new descriptions and select the one with the highest QWK score as the output of the BFS-inspired strategy.

4.2 DFS-Inspired Strategy

The DFS-inspired strategy, different from BFS-inspired one, will update the parameter of LLM_{edt} using converted Quadratic Weighted Kappa (QWK) score as reward via PPO (Schulman et al., 2017) reinforcement learning framework, as shown in Figure 2. Also, each time, LLM_{edt} builds upon

layer of the financial classifier.

[7]https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier/blob/main/utils/prompt.txt

	3-grade hold-out test set			4-grade hold-out test set		
	Accuracy	F1	QWK Corr.	Accuracy	F1	QWK Corr.
Initial description	38.9	31.4	19.0	27.8	22.2	22.0
BFS	72.2	71.0	79.7	48.0	45.4	67.1
DFS	56.4	57.5	58.6	30.3	24.2	39.3
Our Fin-aware BFS	**83.1**	**82.2**	**88.2**	59.2	58.1	78.0
Our Fin-aware DFS	79.1	78.8	85.1	**61.0**	**60.7**	**78.6**

Table 3: Performance on hold-out test sets for 3-grade and 4-grade evaluations is measured using accuracy, macro-F1, and Quadratic Weighted Kappa (QWK). We select the description with the highest QWK score after 50 iterations on a 600-document validation set and evaluate it on a separate 600-document hold-out set.

the current best description to generate a new one, continuously evolving itself over N iterations.

4.3 Integration of Financial Hypernym

Given the vast search space of LLM_{edt}, finding optimal descriptions can be time-consuming. To address this, we incorporate financial hypernyms extracted from the FineWeb corpus without relying on external resources or human annotation. These hypernyms serve as high-level descriptors, enabling the model to rewrite descriptions without delving into overly specific financial terms or events. For example, *Citigroup* is replaced with *bank*, and *property* with *asset*. To obtain financial hypernyms, we first use our financial classifier to select 5.6k documents with a probability above 0.99 from the 580k dataset. We then extract the most frequent financial nouns and adjectives, removing numbers and common Wikipedia words. Following (Peng et al., 2022), we prompt the RoBERTa model (Liu, 2019) with crafted templates such as *In a financial context, word is a type of <mask>.* or *In a financial context, something word is <mask>.*, selecting the most probable <mask> token as a high-level substitute for *word*. This process yields 120 financial hypernyms.

5 Experiments and Result

5.1 Models and Experimental Setup

For LLM_{eval} and LLM_{edt}, we use Mixtral-8x7B-Instruct-v0.1[8] with 4-bit quantization. We refer readers to Appendix G for training details and experimental setup.

5.2 Result

Table 3 presents results on a test set that was not used during the training process. Accuracy and Macro-F1 measure the exact grade match with ground-truth data grades, while QWK quantifies the correlation between predicted and ground-truth grades, considering the ordinal nature of the grading score. We have the following observations:

(1) Both BFS-inspired and DFS-inspired optimizations yield better results than the initial grade description. Incorporating financial knowledge (Fin-aware BFS/DFS) further boosts Accuracy, F1, and QWK scores by over 10 points each, confirming the effectiveness of this simple financial hypernym integration.

(2) DFS underperforms BFS on both the 3-grade and 4-grade test sets. However, when augmented with financial hypernyms, Fin-aware DFS nearly matches Fin-aware BFS on the 3-grade test and slightly surpasses it on the more challenging 4-grade test, suggesting the potential of combining DFS-inspired search with RL to update the editor model (In BFS, editor model is not updated).

To further confirm whether and to what extent the generated descriptions contain financial hypernyms, we compute the word overlap between grade descriptions produced by BFS/DFS and Fin-aware BFS/DFS, respectively. Table 4 shows the percentage of financial hypernyms in each description. It is interesting to note that the methods with the highest percentage of financial hypernyms (8.2% for Fin-aware BFS and 12.3% for Fin-aware DFS) achieved the best performance in Table 3, implying that the proportion of financial hypernyms may impact performance, although a model with 0% financial hypernyms also led to high accuracy (79.1%).

Qualitative Evaluation To further evaluate description quality, a human expert rated the outputs in Table 8 and Table 9 in Appendix on three criteria: Fluency, Informativeness (i.e., detailed, spe-

[8]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Generation	3-grade	4-grade
description_BFS	5%	0.7%
description_DFS	2.4%	2.9%
description_BFS_fin	8.2%	2.2%
description_DFS_fin	0%	12.3%

Table 4: Percentage of financial hypernyms in each description. A higher percentage indicates a more informative description in the financial domain.

cific, and actionable), and Trustworthiness (i.e., logically consistent without contradicting basic financial principles). As shown in Table 5, fluency scores are similar across all five descriptions, while descriptions enhanced with financial hypernyms perform better in both informativeness and trustworthiness.

Generation	Fluen.	Infor.	Trust.
description_init	4	2	3
description_BFS	4	3	4
description_DFS	4	4	4
description_BFS_fin	4	4	5
description_DFS_fin	4	5	5

Table 5: Human evaluation for Fluency (Fluen.), Informativeness (Infor.), and trustworthiness (Trust.) of different grading descriptions on a 1–5 Likert scale.

6 Conclusion

We propose an automated, automatically improving method for financial data quality grading that supports arbitrary grading levels via a simple binary domain-data classifier. Financial hypernyms are automatically derived and integrated with two search strategies, yielding significant performance gains and more informative descriptions, as confirmed by qualitative. In the future, we plan to extend our approach to other domains, as it requires minimal binary annotation or potentially no human annotation, given the low error rate of LLMs' annotations.

References

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. Bertese: Learning to speak to bert. *arXiv preprint arXiv:2103.05327*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Weize Kong, Spurthi Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. PRewrite: Prompt rewriting with reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–601, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Edward F Moore. 1959. The shortest path through a maze. In *Proc. of the International Symposium on the Theory of Switching*, pages 285–292. Harvard University Press.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2022. Discovering financial hypernyms by prompting masked language models. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 10–16, Marseille, France. European Language Resources Association.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022. Tempera: Test-time prompting via reinforcement learning. *arXiv preprint arXiv:2211.11890*.

A Financial Keywords Collections

We prompt the Mixtral-8x7B-Instruct model to generate a list of words representative of the financial domain. To encourage diversity in the generated financial keywords, we query the model eight times, each with different decoding hyperparameters. Specifically, we select the *temperature* from {0.6, 0.7, 0.8}, *top_p* from {0.9, 0.95, 1.0}, and *top_k* from {50, 75, 100}. Each time, we use a different combination of these three hyperparameters to generate 500 financial keywords.

Please generate a list of 500 single finance-related words in valid JSON format. Each entry should include the following fields:

1. "word": A single finance-related word.

2. "justification": A single sentence explaining why the word is relevant to the financial domain. Ensure the JSON is syntactically valid and formatted as follows:

```
[
  {
    "word": "example_word",
    "justification": "This is an
    example description of the word's
    relevance to the financial domain."
  },
  ...
]
```

Provide the output in a single JSON array containing exactly 500 entries. Note that financial word should be single word instead of phrases.

Our experimental results show that (1) a significant number of generated words are duplicates, and (2) many outputs are financial phrases (rather than single words), including financial institution names. After deduplication, we obtain a final list of 202 unique financial keywords, such as *EPS, slippage, or cashflow*.

Sort Documents with Financial Keywords

For each document in the 600k FineWeb dataset, we first use NLTK[9] for word tokenization, convert all words to lowercase, and convert them into a bag of unique words, BOW_i for document i. We then compute the proportion of overlap between each document's bag of unique words and the financial word list, $\mathbf{F}$, using the following formula:

$$\text{value}_i = \frac{|\mathbf{F} \cap BOW_i|}{|BOW_i|}$$

[9]https://github.com/nltk/nltk

B Financial Data Annotation Instruction

Below is the instruction used by both LLMs and humans for binary financial data annotation. We found that using the text 'The document is financial text.' or 'The document is not financial text.' is more effective than outputting a label of 1 or 0 in the prompt output format. We later convert these textual outputs into labels 1 and 0.

You are an expert in financial data quality with deep expertise in analyzing financial documents. Your task is to evaluate the given document to determine its relevance and quality as financial text.

Document: DOCUMENTS GO HERE

Carefully assess the document and output one of the following responses:

- 'The document is financial text.'

- 'The document is not financial text.'

Provide only the response, without additional description.

C Prompt with Initial description for Financial Data Grade

We start the experiment using a document annotation prompt, similar to this one[10], with the following initial description. We use {poor/fair/excellent} for 3-grade initial description.

Below is an document from a web page. Evaluate it using the categorical {n}-point scoring system described below:

- score 1 if the document is poor.

- score 2 if the document is fair.

- score 3 if the document is good.

- score 4 if the document is excellent.

The document:{DOCUMENT}

After examining the document:

- Briefly justify your total score, up to 100 words.

- You must prepend the score exactly using the following format:

'financial score: <total points>.'

[10]https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier/blob/main/utils/prompt.txt

D Editing Prompt

LLMs use an editing prompt to generate new descriptions for financial data grading criteria. In the BFS-inspired method, descriptions are generated based on the same initial description from Appendix C. In contrast, the DFS-inspired method iteratively builds upon the current best description to produce a new description. To incorporate financial hypernyms, we ask LLMs to selectively use financial word in the word list by appending *based on financial topic words from the following list, using them selectively.* after *Rewrite... requirements.* Also, we add *financial topic words: {CONCATENATED_FINANCIAL_WORDS}* right before *{n}-points*.

Below is a categorical {n}-point scoring system designed to evaluate the financial value of a document. Rewrite the following {n} points via rephrasing and/or adding specific requirements. Use illustrative description if needed.

{n}-points: {REVISED_POINTS}

Each point should begin with '- score X if the document...'
Output the new {n} points only.

E Human-LLM Collaborative Annotation

First, we use the Mixtral-8x7B-Instruct model[11] to assign a label of 1 to high-quality financial documents and 0 otherwise, using the binary annotation instruction in Appendix B. Among 6k documents from the head of the sorted list, 4k are labeled as 1, whereas for 6k documents from the tail, 5.9k are annotated as 0. Next, a human annotator manually reviews the 4k documents labeled as 1 and identifies error cases in 4% (about 160 documents) of them, while finding very few errors among the documents labeled as 0[12]. Finally, we remove the identified error cases, resulting in a ground-truth positive dataset of 3,840 documents and negative dataset of 3,840 documents. Table 6 shows that basic statistics of annotated financial data.

[11]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1. It is licensed under Apache License 2.0.

[12]We do not ask the human annotator to review all 5.9k documents but only the first 4k, as our goal is to create a balanced training dataset for both positive and negative class.

	Training set	Test set
# of document	5,680	2,000
Average words	458.7	466.3
STD	211.8	210.2

Table 6: Statistics of annotated financial dataset. We use NLTK toolkit for word tokenization.

F Financial classifier probability distribution on FineWeb documents

Prob. Range	# of documents
(0.0, 0.05)	542,628
(0.05, 0.1)	3,772
(0.1, 0.15)	2,117
(0.15, 0.2)	1,536
(0.2, 0.25)	1,109
(0.25, 0.3)	899
(0.3, 0.35)	752
(0.35, 0.4)	718
(0.4, 0.45)	674
(0.45, 0.5)	656
(0.5, 0.55)	638
(0.55, 0.6)	649
(0.6, 0.65)	715
(0.65, 0.7)	745
(0.7, 0.75)	802
(0.75, 0.8)	897
(0.8, 0.85)	1,016
(0.85, 0.9)	1,326
(0.9, 0.95)	2,163
(0.95, 1.0)	16,188
In total	580,000

Table 7: Probability Distribution over FineWeb documents.

G Experimental Setup Details

The parameters of LLM_{eval} remain frozen in all BFS- and DFS-inspired methods. In BFS-inspired methods, LLM_{edt} is also frozen, whereas in DFS-inspired methods it is updated via LoRA (Hu et al., 2021) (rank 8, alpha 32) within the PPO-based RL framework. We use a learning rate of 2.82×10^{-6} and sampling-based decoding (top_p = 1.0, top_k = 0) to encourage creative writing. For both BFS and DFS, N is set to 50. All experiments are conducted on four A100 GPUs.

Baseline 3-grade description produced by vanilla BFS with **5%** financial hypernyms	**Our best 3-grade description** by Fin-aware BFS with **8.2%** financial hypernyms
• Score 1 if the document is poor and lacks proper financial **analysis**. • Score 2 if the document is adequate, yet misses critical **information** on the financial specifics of the **investment**. • Score 3 if the document is excellent and exhibits deep comprehension of financial intricacies while being presented in a clear, easy-to-understand manner.	• Score 1 if the document is poor at explaining financial concepts and **institutions**, or does not discuss the **performance** or **valuation** of assets or liabilities at all. • Score 2 if the document presents general financial **information** in a utilitarian **manner** but uses vague terms and lacks specific details about the financial status of the **entity** or **individual**. • Score 3 if the document demonstrates clarity and depth in discussing financial topics, including **performance** metrics, **risk management**, financial strategies, and potential uncertainties, while providing transparent and accurate data.

Table 8: Case study: comparison of baseline 3-grade description (vanilla BFS) and best 3-grade description (Fin-aware BFS).

Baseline 4-grade description produced by vanilla DFS with **2.9%** financial hypernyms	**Our best 4-grade description** by Fin-aware DFS with **12.3%** financial hypernyms
• Score 1 if the document lacks crucial details, including the author's identity or copyright **information**. • Score 2 if the document offers basic primary data but misses essential contact details, such as an email address or phone number. • Score 3 if the document encompasses detailed key and supportive **information**, complemented by clear screenshots, relevant links, and informative appendices. • Score 4 if the document offers extensive advantages, comprising functional code samples, valuable learning sources, and a thorough project roadmap, all while exhibiting exceptional writing and **organization**.	• 4-point scoring system for financial documents: • Score 1 if the document is of poor quality and lacks essential financial topic words like **finance**, **investment**, **transaction**, **income**, **exchange**, and **property**. This may indicate an insufficient understanding of the legal context or the boundaries of a financial document. • Score 2 if the document is fair and contains adequate financial topic words. However, the depth of financial **analysis** or detail on topics like **stock**, **payment**, **company**, **liability**, **performance**, or **credit management** may be lacking, making it difficult to understand the impact on the target audience. • Score 3 if the document is good and has all necessary financial topic words, including variables such as **volatility**, **asset**, **debt**, **management**, **regulation**, and **risk**. The document may also address **performance** indicators, specialized financial instruments, and concepts surrounding **wealth** creation, **technology**, and **governance**, with clear **communication**. • Score 4 if the document is excellent and is abundant with financial topic words and concepts, addressing interrelated financial factors and broader economic context. The document may consider risks, constraints, **negative** events, and apply concepts such as **insurance**, **punishment**, and **bankruptcy** to protect against uncertainties. Finally, it is clear, concise, referencing proper **accounting** protocols, and incorporates proper **communication** protocols for the intended audience.

Table 9: Case Study: comparison of baseline 4-grade description (vanilla DFS) and best 4-grade description (Fin-aware DFS).