Noise-Robust Financial Numerical Entity Attribute Tagging

Anonymous ACL submission

Abstract

Financial Numerical Entity (FNE) understanding aims to recover the meaning of numerical mentions in financial reports. Existing studies primarily focus on concept name prediction and face two important limitations. First, labels derived from inline XBRL may contain errors because filings are usually prepared manually. Second, other important FNE attributes, such as reporting-time relation, measurement scale, and accounting sign, are less emphasized. We propose **NO**ise-**R**obust Tagging for Rich Financial Numerical Entity **A**ttributes (NORA) to address these gaps. NORA uses task-aware instance-specific weighting to attenuate the influence of noisy labels during training, and we further propose the Neighborhood Prior-adjusted KNN (NPK) filtering method for more reliable evaluation on real-world noisy test sets. In addition, we construct a large-scale benchmark containing 6.6 million instances with multi-attribute labels and filing metadata. Experiments show that NORA performs strongly compared with state-of-the-art noisy-label baselines, including Co-teaching, Mixup, SSR, and SelfMix. Moreover, NORA is robust under both unfiltered and noise-filtered test settings. It achieves the best Accuracy, Macro F1, and Weighted F1 for concept name and time-relation prediction, while remaining competitive on scale and sign prediction. These results demonstrate the value of jointly modeling rich FNE attributes while accounting for label noise in real-world financial filings.

1 Introduction

Financial Numerical Entity (FNE) understanding is the task of identifying the meaning of numerical mentions in financial text by predicting the attributes that make them interpretable in context (Huang et al., 2023; Yang et al., 2020). In financial reports such as 10-Ks or 10-Qs, this means determining not only the accounting concept (i.e., the tag) an FNE is associated with, but also attributes such as whether it is an instant or duration fact, its relation to the current reporting period (e.g., past, current, or future), its measurement scale (e.g., million), and the sign of the value with respect to the accounting concept (see Figure 1 for an illustrative example). The problem is important because downstream systems for risk analysis, compliance monitoring, and business analysis all depend on assigning the correct semantics to numbers in disclosure documents (Fourny, 2023).

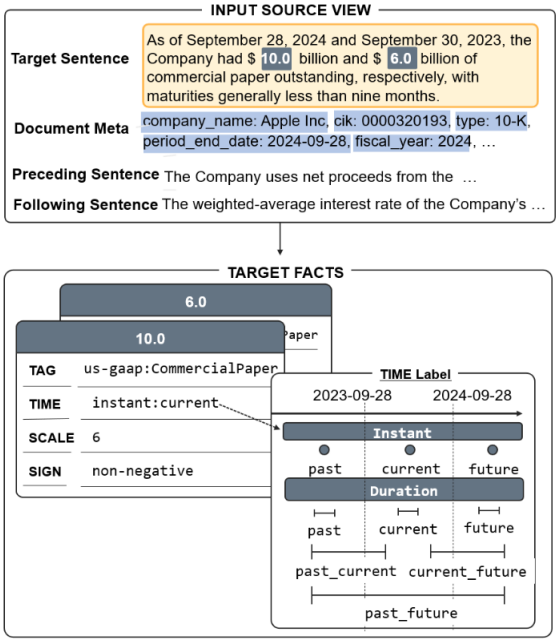

Figure 1: Illustration of input context and corresponding FNE attributes

Existing studies have laid important groundwork for FNE understanding. Benchmarks such as FiNER-139 and FNXL have advanced financial entity tagging by standardizing datasets and evaluations for identifying numerical entities and linking them to accounting concepts (Loukas et al., 2022; Sharma et al., 2023). However, two important gaps remain. First, large-scale supervision in this domain is often obtained from inline XBRL

(iXBRL), yet iXBRL is inherently noisy as a source for ground truth. Several studies document recurring filing mistakes such as mismatches between embedded tags and surrounding narrative, missing elements, incorrect monetary values, and sign errors (Chang et al., 2021; Al-Okaily et al., 2024; Luo et al., 2023). Second, prior benchmarks largely reduce the task to concept-name tagging and do not jointly model other crucial attributes such as time relation, measurement scale, and the accounting sign of numeric entities. As a result, existing formulations understate both the semantic richness of the task and the practical difficulty of learning from noisy labels.

To address these limitations, we propose the **NO**ise-**R**obust Tagging for Rich Financial Numerical Entity **A**ttributes (NORA) framework. NORA treats FNE understanding as a joint prediction problem over related attributes, including concept name (i.e., tag), reporting-time relation, measurement scale, and accounting sign, and is designed to be robust to noisy labels induced by real-world iXBRL filings. The key idea is to incorporate learnable instance-specific weighting so that the model can attenuate the overfitting downside of noisy labels while still benefiting from large-scale automatically derived supervision. Moreover, we develop the SSR-inspired Neighborhood Prior-adjusted KNN (NPK) noisy-label filtering method (Feng et al., 2024), which enables model evaluation on real-world noisy testbeds. As a companion dataset to NORA, we curate a large-scale corpus of 6.6 million narrative instances aligned to iXBRL facts with rich FNE attributes from more than 110,000 10-K and 10-Q filings, directly targeting the gap between idealized benchmark settings and realistic financial reporting environments.

Our main contributions are as follows:

- We formulate FNE understanding as a rich multi-attribute prediction task that jointly models concept name, reporting-time relation, measurement scale, and accounting sign.
- We construct a large-scale benchmark from real-world iXBRL filings, comprising 6.6 million narrative instances with rich FNE attributes aligned to facts from more than 110,000 SEC filings.
- We propose NORA, a noise-robust learning framework that uses learnable instance-specific weighting to mitigate overfitting to noisy supervision and consistently outperforms strong noisy-label learning baselines on most tasks.
- We develop the NPK noisy-label filtering method that enables more reliable model evaluation using real-world noisy filings.

2 Related Works

2.1 Financial Numerical Entity (FNE) Tagging

FNEs are numerical mentions in financial reports whose meanings are determined by the surrounding text (Loukas et al., 2022; Sharma et al., 2023). Unlike general named entities, an FNE cannot be fully interpreted from its surface span alone; approximately 91% of tagged tokens in these reports are numeric, meaning the correct label depends almost entirely on the surrounding textual and structured reporting attributes. FNE tagging thus requires models to connect these mentions with complex accounting semantics.

Benchmark datasets have evolved to reflect the increasing complexity of this task. Initial efforts like FiNER-139 standardized evaluation for the 139 most frequent XBRL tags (Loukas et al., 2022). However, real-world filings utilize an ontology of thousands of potential GAAP metrics. This led to the development of the Financial Numeric Extreme Labelling (FNXL) dataset, which expands the label set to 2,794 distinct tags (Sharma et al., 2023). Together, these resources have established FNE tagging as an extreme classification problem, characterized by a heavy-tail label distribution where many tags appear very infrequently. The FLAN-FinXC (Khatuya et al., 2024) took a generative approach by instruction-tuned a FLAN-T5 to output XBRL tag documentations given a FNE in context. An unsupervised tag matcher then output the final concept name. This approach achieved good performance on the FNXL and FiNER-139 benchmarks.

However, existing FNE-related benchmarks still simplify the full tagging problem. Most formulations emphasize the span and/or concept name of a numerical mention, while the complete interpretation of FNEs in filings also depends on attributes such as reporting time, measurement scale, and accounting sign.

A second limitation is that filing supervision is naturally noisy. Prior work documents quality issues in XBRL reporting, including reporting inconsistencies, missing elements, incorrect monetary

values, and sign mistakes (Bartley et al., 2011; Du et al., 2013). These concerns remain relevant in iXBRL settings, where embedded facts can still reflect preparation, mapping, or reporting errors (Basoglu and White, 2015; Chang et al., 2021). Recent studies further show that XBRL and iXBRL data quality can affect downstream analysis and market-facing applications (Al-Okaily et al., 2024; Luo et al., 2023). These observations are directly relevant to FNE tagging because large-scale labels are often derived from iXBRL facts embedded in financial reports. Training and evaluation on noisy ground truth may lead to sub-optimal or even misleading results. In contrast to prior work, we formulate FNE tagging as joint attribute prediction over numerical mentions and study it under naturally noisy, document-grounded supervision.

2.2 Robust Learning under Noisy Labels

Learning from noisy supervision has been studied extensively in deep learning and NLP, but its implications for FNE tagging remain underexplored. Representative approaches include *mixup*, which regularizes training through interpolation (Zhang et al., 2018); Co-teaching, which exchanges small-loss examples between two networks (Han et al., 2018); SelfMix, which combines sample partitioning with semi-supervised learning for text classification (Qiao et al., 2022); and SSR, which uses neighbor-based sample selection and subset expansion for efficient noisy-label learning (Feng et al., 2024). A broader survey is provided by Song et al. (2023).

Among these directions, loss reweighting is particularly relevant to FNE tagging because it directly controls how much each training instance contributes to optimization. Prior importance-reweighting work shows that, under symmetric random classification noise and mild assumptions, a properly reweighted loss can recover the classifier that would have been learned from clean data (Wang et al., 2018). However, such approaches often rely on fixed reweighting functions or pre-specified noise rates, which are difficult to justify for iXBRL-derived supervision.

In financial filings, label reliability is unlikely to be uniform across either instances or attributes. Facts in iXBRL are prepared through company-specific reporting processes, and errors may arise from reporting conventions, taxonomy variation, extraction artifacts, or local document context. A numerical mention may therefore be reliable for one attribute, such as time relations, but mislabeled for another, such as concept names. This setting calls for instance-specific and task-aware loss reweighting, yet this direction has received limited attention in noisy label studies in genenral and FNE tagging in specific. Our work addresses this gap by using learnable instance-specific weights to reduce the influence of noisy labels during multi-attribute FNE tagging.

Noisy supervision also creates a practical evaluation challenge. Most noisy-label studies assume access to a clean test set (Song et al., 2023), but this assumption is often unrealistic for large-scale financial filings, where manually constructing a noise-free evaluation set is expensive and difficult to scale. If test labels are derived from the same imperfect iXBRL sources as training labels, benchmark scores may reflect both model performance and annotation noise. To mitigate this issue, we develop the NPK noisy-label filtering method, inspired by neighbor-based sample selection (Feng et al., 2024), to construct more reliable evaluation subsets from real-world noisy filings.

3 Methodology

NORA addresses two challenges in FNE understanding: learning from noisy labels in training data and evaluating models when the test set may also contain noisy labels. For training, NORA introduces instance-specific weighting for each classification head, together with a regularization term that controls how aggressively noisy instances are attenuated. For evaluation, we introduce the NPK mechanism based on the data smoothness assumption, which has been shown to be effective by Feng et al. (2024).

3.1 Noise-Robust Training in NORA

NORA builds on a pretrained transformer encoder and adds a reweighting component to each classification head to reduce the downside of training with noisy labels. The input sequence consists of `[CLS]`, the target sentence, filing metadata in JSON format, the previous sentence, and the following sentence when available. The metadata includes the company name, CIK, document type (e.g., 10-K or 10-Q), report-period end date, and current fiscal year. If the resulting sequence exceeds the context window, we truncate it to fit the encoder limit.

This study focuses on noisy-label learning for FNE tagging and treats the location of the FNE

span as given. Although the framework can be extended to jointly perform FNE identification and tagging, we keep span identification outside the current scope to isolate the effect of noise-robust attribute prediction.

Because label reliability may depend on the full textual and metadata context, NORA estimates an instance-specific noise level from the global [CLS] representation. Let $\mathbf{h}_{\text{CLS}}$ denote the contextual representation of the [CLS] token. We first project this representation through a shared dense layer with $\tanh$ activation and then apply a task-specific sigmoid layer. For task a, the estimated noise gate is:

$$\tilde{\mathbf{h}}_{\text{CLS}} = \tanh(\mathbf{W}\mathbf{h}_{\text{CLS}} + \mathbf{b}),$$
$$g_a = \sigma(\mathbf{W}_a \tilde{\mathbf{h}}_{\text{CLS}} + b_a).$$

Since $g_a \in (0, 1)$, it can be interpreted as the estimated noise level of the instance for task a. When $g_a \approx 0$, the instance is treated as reliable and informative for that task. When $g_a \approx 1$, the instance is treated as noisy or ambiguous and contributes little to the task loss.

For the task prediction head, NORA mean-pools the contextualized token representations over the target FNE span and passes the resulting span representation to task-specific heads. We consider tasks $T = \{\texttt{tag}, \texttt{time}, \texttt{scale}, \texttt{sign}\}$. Let L_a denote the raw loss for task a. NORA applies the following noise-modulated loss:

$$\tilde{L}_a = (1 - g_a) \cdot L_a + \lambda_a \cdot g_a^2.$$

The first term downweights the raw loss when the instance is estimated to be noisy. The second term regularizes the gate and prevents the trivial solution in which the model sets g_a close to 1 for every instance. The hyperparameter λ_a controls this trade-off: larger values force the model to learn more from instances, whereas smaller values allow it to ignore noisy or highly ambiguous instances more aggressively.

The total loss for an instance is the weighted sum over all tasks T:

$$L_{\text{total}} = \sum_{a \in T} w_a \tilde{L}_a,$$

where w_a is a task-level weight that balances the contribution of each prediction objective.

Tag Loss

Due to the severe class imbalance and long-tail distribution of accounting concepts, we use Class-Balanced (CB) Cross-Entropy Loss (Cui et al., 2019). CB loss scales the standard cross-entropy penalty by the inverse effective sample size for each class:

$$L_{\text{tag}} = \frac{1}{B} \sum_{i=1}^{B} \frac{1 - \beta}{1 - \beta^{n_{y_i}}} L_{\text{CE}}(p_i, y_i),$$

where n_{y_i} is the number of samples in the ground-truth class y_i, B is the mini-batch size, and β controls the strength of the class-balance adjustment.

Time Relation and Scale Losses

An FNE may correspond to an instant fact, which refers to a specific point in time, or to a duration fact, which spans a time period. For instant facts, the time relation to the filing period is categorized as *past*, *current*, or *future*. For duration facts, the relation can be *current* (spanning within current period), *past_current* (from before current period to current), *current_future* (from current to future), or *past_future* (from before current period to the future).

The scale attribute specifies the power of ten by which the displayed number should be multiplied to recover its unscaled value. For example, a value reported as 210 in millions has scale 6, corresponding to 210×10^6. In our data, scale values range from -12 to 12, so we define scale prediction as a 25-class classification task. For both time-relation and scale prediction, we use cross-entropy loss with class weights inversely proportional to class frequencies.

Sign Loss

Because negative accounting signs constitute only about 3% of the data, we use Focal Loss (Lin et al., 2018) to emphasize hard and minority-class examples:

$$L_{\text{sign}} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t),$$

where p_t is the predicted probability of the true class, α_t balances positive and negative classes, and γ is the focusing parameter.

3.2 Robust Evaluation via Neighborhood Prior-Adjusted KNN Filtering (NPK)

Real world financial reports may contain labeling errors. If evaluation is conducted directly on noisy test labels, measured performance may reflect the performance of the modela affected by label quality. To obtain a more reliable estimation, we introduce the NPK filtering procedure that identifies high confidence evaluation subsets from noisy test data.

The NPK procedure is based on the assumption of neighborhood consistency: if an instance is correct, its label should be similar to its neighborhood instances. We implement this idea in three steps.

Semantic feature extraction. We use a pretrained transformer encoder to extract the embedding for each instance. The embeddings are ℓ_2-normalized so that inner product similarity is equivalent to cosine similarity.

KNN retrieval. For scalability, we use FAISS (Douze et al., 2025) to build an index over the normalized embeddings. For each sample x_i, we retrieve its k nearest neighbors in the semantic embedding space.

Prior-adjusted consistency scoring. To reduce majority-class bias in the neighborhood counts, we adjust each class count by its global class prior. For class c, the prior is

$$\pi_c = \frac{N_c}{N},$$

where N_c is the number of instances in class c and N is the total number of instances. Let $m_i(c)$ denote the number of neighbors of x_i assigned to class c. The prior-adjusted neighborhood score is

$$q_i(c) = \frac{m_i(c)}{\pi_c + \epsilon},$$

and the consistency score for the observed label y_i is

$$c_i = \frac{q_i(y_i)}{\max_c q_i(c) + \epsilon}.$$

After computing c_i for all valid samples, we rank examples in descending order and retain the highest-scoring subsets as cleaner evaluation splits. In our experiments, we report results on the top 90% subsets. These filtered splits are not assumed to be perfectly noise-free, but they provide a lower-noise evaluation setting that better separates model errors from residual annotation errors in real-world filings.

3.2.1 Baselines

We compare NORA against four representative baselines for noisy-label learning in deep neural networks: Co-teaching (Han et al., 2018), SSR (Feng et al., 2024), Mixup (Zhang et al., 2018), and SelfMix (Qiao et al., 2022). These baselines cover complementary strategies for handling noisy supervision. Co-teaching represents small-loss sample selection, where two networks teach each other by exchanging examples that are likely to be clean. SSR evaluates whether explicit neighbor-based filtering and subset expansion can identify reliable training instances in the FNE setting. Mixup tests whether interpolation-based regularization improves robustness by smoothing the decision boundary under corrupted labels. SelfMix provides a semi-supervised comparison that partitions samples by confidence and uses unlabeled-style training to reduce dependence on noisy annotations.

4 The NORA Dataset

We construct the NORA Dataset as the companion for our noise-robust approach. Similar to FiNER-139 and FNXL, the dataset is developed based on SEC filings. However, our dataset provides more than 6.6 million sentences with around 12 million annotated FNEs, compared to 1.1 million sentences in FiNER-139 and 79K sentences in FNXL. Moreover, we provide rich attributes for FNEs, such as tag (concept name), time relations, scale, sign (accounting sign). While not investigated in this study, the dataset provide additional attributes such as the exact fact value, the unit of measurement, the decimal precision levels, and reporting categories (i.e., axis and members). We also implement basic noise filter to remove noisy text such as page numbers and TOCs. Table 3 in Appendix provide a comparison of NORA Dataset with FiNER-139 and FNXL.

In addition to the target sentence, the target FNE spans, and annotated attributes, we provide document meta data such as company name, reporting period, and fiscal years so that learning time relations is possible. We also provide the previous and following sentences in addition to the target sentences to expand the context for FNE understanding.

The dataset includes 446,338 unique tags and 20 unique scale. We apply a frequency filter of 1,000 to reduce standard tags below the threshold to `standard_rare`. All custom tags are reduced to the `custom` tag. The final tag set contain 978 unique tags.

5 Experiments

5.1 Experimental Setup

We implement NORA in PyTorch and initialize the text encoder with the `nlpaueb/sec-bert-base` (Loukas et al., 2022). Unless otherwise stated, we set the focal-loss parameters to $\alpha = 0.25$ and $\gamma = 3.0$, and the class-balanced loss parameter to

(a) Unfiltered test set

Task	Method	Acc.	Macro F1	Weighted F1
TAG	Co-teaching	0.6282	0.3629	0.6031
	Mixup	0.6293	0.3848	0.5999
	SSR	0.4196	0.3062	0.4134
	SelfMix	0.4269	0.0528	0.3205
	NORA	**0.6330**	**0.4148**	**0.6107**
TIME	Co-teaching	0.5964	0.3717	0.5671
	Mixup	0.5748	0.3594	0.5445
	SSR	0.5619	0.3543	0.5349
	SelfMix	0.4190	0.1723	0.3558
	NORA	**0.7941**	**0.6317**	**0.7949**
SCALE	Co-teaching	0.9699	0.3743	0.9698
	Mixup	**0.9845**	**0.3851**	**0.9844**
	SSR	0.9573	0.3511	0.9566
	SelfMix	0.8211	0.2367	0.8032
	NORA	0.9829	0.3840	0.9828
SIGN	Co-teaching	**0.9837**	**0.8430**	**0.9832**
	Mixup	0.9820	0.7812	0.9789
	SSR	0.9819	0.8240	0.9813
	SelfMix	0.7746	0.4919	0.8494
	NORA	0.9832	0.8082	0.9810

(b) NPK-filtered test set

Task	Method	Acc.	Macro F1	Weighted F1
TAG	Co-teaching	0.6542	0.2942	0.6211
	Mixup	0.6370	0.3266	0.6062
	SSR	0.5207	0.3666	0.5219
	SelfMix	0.4380	0.0552	0.3300
	NORA	**0.6652**	**0.4314**	**0.6447**
TIME	Co-teaching	0.6382	0.3871	0.6099
	Mixup	0.5765	0.3640	0.5484
	SSR	0.6372	0.4081	0.6153
	SelfMix	0.4193	0.1733	0.3565
	NORA	**0.8001**	**0.6412**	**0.8009**
SCALE	Co-teaching	0.9380	0.2940	0.9244
	Mixup	**0.9844**	**0.3864**	**0.9842**
	SSR	0.9716	0.3751	0.9714
	SelfMix	0.8189	0.2357	0.8007
	NORA	0.9834	0.3850	0.9833
SIGN	Co-teaching	0.9745	0.6052	0.9658
	Mixup	0.9778	0.6729	0.9711
	SSR	0.9836	**0.8493**	**0.9836**
	SelfMix	0.7762	0.4917	0.8508
	NORA	**0.9841**	0.8170	0.9821

Table 1: Model performance on the unfiltered and 90% NPK filtered test sets. Acc. denotes Accuracy.

$\beta = 0.99$. For the multi-task objective, we assign task weights of 10.0 to both Tag and Sign, 0.3 to Time, and 0.2 to Scale. For the noise-gate regularization term, we set $\lambda_{Tag} = 1.0$ and use $\lambda_a = 10.0$ for the remaining tasks.

We optimize all models with AdamW and a weight decay of 0.01. To account for different learning dynamics between the shared encoder and task-specific heads, we use separate learning rates: 1×10^{-5} for the transformer backbone, 5×10^{-4} for the Tag head, 3×10^{-5} for the Scale head, 2×10^{-5} for the Sign head, and 1×10^{-5} for the Time head. Models are trained for 30 epochs with a global batch size of 64 using a cosine annealing learning-rate scheduler with minimum learning rate $\eta_{\text{min}} = 1 \times 10^{-7}$.

For NPK filtering, following SSR, we retrieve $k = 50$ nearest neighbors for each sample in the normalized semantic embedding space and retain the top 90% highest-scoring examples for the filtered evaluation setting.

Evaluation Metrics For all experiments, we report three standard classification metrics: Accuracy, Macro F1, and Weighted F1. Accuracy measures overall prediction correctness, while Weighted F1 accounts for label-frequency differences when aggregating class-level performance. We place particular emphasis on Macro F1 because it gives equal weight to each class and is therefore more informative under the severe class imbalance and long-tail label distributions presented in FNE tagging.

5.2 Results and Analysis

We evaluate NORA under two complementary settings. The first uses the unfiltered test set to measure performance. The second uses the NPK noise-filtered test subsets introduced in Section 3.2, which retain examples with stronger neighborhood consistency and therefore provide a lower-noise view of model behavior. Together, these settings allow us to assess both practical robustness on noisy data and performance when evaluation labels are more reliable.

Unfiltered Test Set Panel (a) of Table 1 reports the results on the unfiltered test set. On Tag, NORA achieves the best performance across all three metrics, with an Accuracy of 0.6330, a Macro F1 of 0.4148, and a Weighted F1 of 0.6107. The gain is most notable in Macro F1, where NORA improves over the strongest baseline by 3 percentage points.

On Time, NORA shows the clearest advantage, reaching 0.7941 Accuracy, 0.6317 Macro F1, and 0.7949 Weighted F1; these scores exceed the next-best baseline by 19.8 percentage points in Accuracy and 26.0 percentage points in Macro F1, indicat-

ing that instance-specific weighting is especially helpful for context-dependent temporal labels.

On Scale, Mixup obtains the best results, but NORA is very close, trailing by only 0.16 percentage points in Accuracy and 0.11 percentage points in Macro F1. On Sign, Co-teaching performs best, while NORA remains competitive with 0.9832 Accuracy and 0.8082 Macro F1 despite the strong class imbalance of this task.

Overall, Panel (a) shows that NORA is strongest on the two most semantically demanding tasks, Tag and Time. Although NORA is not the top performer on Scale and Sign, the gaps are small, indicating limited trade-offs on more important attributes.

Noise-Filtered Test Set We next evaluate on the NPK filtered test subset. This setting allows us to examine whether the conclusions from the full test set persist when evaluation labels are likely to be cleaner. Panel (b) of Table 1 reports the filtered results.

The results mostly preserve the main trends from the unfiltered setting. NORA again obtains the best Accuracy, Macro F1, and Weighted F1 on Tag and Time. These consistent gains suggest that NORA is not simply benefiting from noise in the test labels, but is better capturing context-dependent FNE attributes. The relative performance on Scale remains stable. We note that NORA becomes the best performer on Sign in terms of the accuracy, and is only 3 percentage point behind SSR, which is the best performer in the filtered setting. Overall, the filtered setting supports the same argument as the full setting: NORA provides a robust approach for realistic noisy financial annotation scenarios.

Analysis on Estimated Noise Level One key question is whether the noise gate can identify noisy labels in a real-world setting. To answer this question, we extracted five instances with the largest g_{tag} values and five instances with the largest g_{time} values for manual inspection. A Certified Public Accountant (CPA) independently reviewed these instances and assigned the corresponding labels. Tables 7 to 16 in the Appendix summarize the ten inspected instances. The manual analysis shows that all inspected instances are noisy, in contrast to a randomly sampled comparison set, which contains less than 5% label errors.

As a concrete example, consider the instance in Table 12. The target sentence states: "Our provision for gross-to-net allowances was $3.0 million at March 31, 2024, $[0.6] million of which was recorded as a reduction to accounts receivable and $2.4 million recorded as a component of accrued expenses," where the brackets indicate the target FNE. This sentence comes from a 10-Q filing with a reporting-period end date of March 31, 2024. The estimated g_{time} value is 0.5866, which is substantially higher than the task mean of 0.1193. The original label is `instant; future`, whereas the model predicts `duration; future`. The expert annotation is `instant; current` with the reasoning that the $0.6 million value represents a reduction to accounts receivable measured at a specific point in time, so it is an `instant` fact; because the date is March 31, 2024, matching the filing's reporting-period end date, its relative temporal category is `current`. This example illustrates how a high g_{time} value can flag an instance whose original label is inconsistent with the accounting context.

Table 2 further examines prediction consistency among training instances (i.e., in-sample prediction accuracy) with the largest g_a values. If the noise gate is functioning as intended, instances with higher g_a should be more likely to be noisy or difficult, reflected by greater inconsistency between the observed labels and model predictions. This pattern is evident for Tag and Time. The top-ranked high-g_a instances have relatively low consistency, especially when only the highest-gate instances are considered. Among the top five instances, the percentage of consistent prediction is 0.20 for Tag and 0.40 for Time, meaning that 80% and 60% of these instances, respectively, have labels different from those predicted by the model. The noise gate therefore appears to attenuate the influence of such inconsistent instances by assigning them larger g_a values, causing the model to down-weight them during training. Moreover, prediction consistency generally increases as the top-N set expands, suggesting that instances with smaller g_a values are more likely to correspond to reliable in-sample predictions. In contrast, high-g_a instances for Scale and Sign remain highly accurate across the ranked subsets, suggesting that the noise-attenuation mechanism is less informative for these two tasks.

6 Conclusions and Future Work

This paper addresses two gaps in financial numerical entity understanding. First, existing work often studies financial tagging under the noise-free assumption, whereas labels derived from real finan-

Top N of g_a	TAG	TIME	SCALE	SIGN
5	0.20	0.40	1.00	1.00
6	0.17	0.50	1.00	1.00
7	0.14	0.43	1.00	1.00
8	0.25	0.50	1.00	1.00
9	0.33	0.44	1.00	1.00
10	0.40	0.40	1.00	1.00
20	0.45	0.35	1.00	0.90
30	0.53	0.40	1.00	0.87
40	0.45	0.43	1.00	0.90
50	0.38	0.46	1.00	0.92

Table 2: Training data prediction consistency at top N of g_a

cial filings can be noisy. Second, prior benchmarks typically focus on a limited tagging formulation and do not fully capture the rich attributes needed to interpret financial numerical entities, including tag, time, scale, and sign.

To address these gaps, we propose NORA, a noise-robust multi-task framework for financial numerical entity understanding. NORA uses task-aware instance-specific gates to reduce the influence of unreliable labels during training while jointly modeling multiple attributes of each numerical mention. We also introduce NPK filtering as a complementary evaluation protocol for analyzing model behavior on subsets with stronger neighborhood consistency. As the empirical foundation for this study, we construct NORA DATASET, a large-scale dataset from SEC filings that preserves realistic document context and provides rich numerical-attribute supervision.

Across both unfiltered and noise-filtered test settings, NORA achieves the strongest results on the most semantically demanding tasks, especially TAG and TIME. The gains on TIME are particularly large, suggesting that instance-level reliability modeling is valuable when the correct interpretation depends on reporting periods, document metadata, and surrounding temporal cues. Our analysis of the learned noise gates further supports this interpretation: high estimated noise level instances for TAG and TIME are more likely to contain label–prediction inconsistencies, and manual inspection confirms that top-ranked high-gate examples often correspond to noisy labels. These findings indicate that NORA not only improves predictive performance, but also provides a useful signal for identifying unreliable supervision in real financial filings.

Future work can extend this study in several directions. First, NORA currently assumes that the target numerical span is given; the next step is to integrate numerical entity detection with attribute prediction in a full end-to-end pipeline. Second, future models could incorporate richer filing structure, including tables, statement sections, dimensions, and taxonomy hierarchies, which may improve rare-tag prediction and reduce ambiguity. Third, the noise-gate signal could support human-in-the-loop annotation by prioritizing high-uncertainty or high-noise instances for expert review. Finally, extending the framework beyond SEC filings to other jurisdictions, languages, and reporting standards would test the generality of noise-robust financial numerical understanding.

7 Limitations

This study has several limitations. First, although NORA DATASET is large and document-grounded, it is constructed from SEC filings and therefore reflects the reporting practices, taxonomy choices, and disclosure conventions of U.S. public companies. The results may not transfer directly to other jurisdictions, private-company filings, or non-English reports.

Second, our formulation treats the FNE span as given and focuses on attribute prediction after the target numerical mention has been identified. This design isolates the effect of noisy-label learning for FNE attributes, but it does not evaluate the full end-to-end pipeline in which numerical entity detection and attribute prediction are performed jointly. Errors in span detection could propagate to the attribute heads in practical deployments.

The main risk of this work is overreliance on automated financial tagging. NORA is designed to improve robustness under noisy supervision, but its predictions can still be incorrect, especially for rare tags, ambiguous temporal references, unusual company-specific reporting conventions, or cases where the source filing itself contains errors. The system should therefore be used as a decision-support tool rather than as a replacement for expert financial review, audit, or regulatory judgment.

A second risk is that noise filtering may create a false sense of certainty. KNN-based filtering can identify examples that are more consistent with their neighbors, but neighborhood consistency is not equivalent to factual correctness.

Finally, financial information systems can affect downstream decisions in investment, compliance, credit assessment, and auditing workflows. Mis-

classified numerical attributes may lead to incorrect aggregation, misleading comparisons, or flawed risk signals. Any deployment should include human oversight, uncertainty monitoring, and validation against authoritative financial records.

References

Manaf Al-Okaily, Hani Alkayed, and Aws Al-Okaily. 2024. Does xbrl adoption increase financial information transparency in digital disclosure environment? insights from emerging markets. *International Journal of Information Management Data Insights*, 4(1):100228.

Jon Bartley, Al Y. S Chen, and Eileen Z Taylor. 2011. A comparison of xbrl filings to corporate 10-ks—evidence from the voluntary filing program. *Accounting Horizons*, 25(2):227–245.

Kamile Asli Basoglu and Jr. White, Clinton E. (Skip). 2015. Inline xbrl versus xbrl for sec reporting. *Journal of Emerging Technologies in Accounting*, 12(1):189–199.

Hyun Woong (Daniel) Chang, Steven Kaszak, Peter Kipp, and Jesse C. Robertson. 2021. The effect of ixbrl formatted financial statements on the effectiveness of managers' decisions when making interfirm comparisons. *Journal of Information Systems*, 35(2):149–177.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-balanced loss based on effective number of samples. *CoRR*, abs/1901.05555.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Hui Du, Miklos A. Vasarhelyi, and Xiaochuan Zheng. 2013. Xbrl mandate: Thousands of filing errors and so what? *Journal of Information Systems*, 27(1):61–78.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. 2024. Noisebox: Toward more efficient and effective learning with noisy labels. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11914–11928.

Ghislain Fourny. 2023. *The XBRL Book: Simple, Precise, Technical*. Amazon Digital Services LLC - KDP Print US.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Allen H. Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Subhendu Khatuya, Rajdeep Mukherjee, Akash Ghosh, Manjunath Hegde, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2024. Parameter-efficient instruction tuning of large language models for extreme financial numeral labelling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7391–7403, Mexico City, Mexico. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Xin Luo, Tawei Wang, Liu Yang, Xinlei Zhao, and Yiyang Zhang. 2023. Initial evidence on the market impact of the ixbrl adoption. *Accounting Horizons*, 37(1):143–171.

Dan Qiao, Chenchen Dai, Yuyang Ding, Juntao Li, Qiang Chen, Wenliang Chen, and Min Zhang. 2022. SelfMix: Robust learning against textual label noise with self-mixup training. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 960–970, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561, Toronto, Canada. Association for Computational Linguistics.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2023. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153.

Ruxin Wang, Tongliang Liu, and Dacheng Tao. 2018. Multiclass learning with partially corrupted labels. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6):2568–2580.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language

model for financial communications. *Preprint*, arXiv:2006.08097.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *Preprint*, arXiv:1710.09412.

A Additional Dataset Details for the NORA Dataset

Comparison with Prior Datasets Table 3 contrasts NORA DATASET with FiNER-139 and FNXL.

A.1 Detailed Schema and Attribute Definitions

Each instance in NORA DATASET is organized into four components. The *document* field stores filing-level metadata such as the reporting entity name, CIK, and reporting period; *context* stores the target sentence containing the numeric mention; *targets* identifies the character-aligned text span to be tagged; and *golds* stores the aligned XBRL attribute values for that target. Table 4 summarizes the attribute definitions used in NORA DATASET.

Figure 2 illustrates the JSONL instances in NORA DATASET. The key fields in each instance are summarized below.

`document`: Contains essential metadata about the financial filing, such as the company name, document type (e.g., 10-K), fiscal year, and a direct URL to the source iXBRL file.

`context`: Provides the local textual surroundings for the AI model to read. It includes:

- `context_t`: The **target** sentence that contains the numbers being tagged.
- `context_p`: The **preceding** sentence.
- `context_n`: The **following** sentence.

`targets`: Identifies the exact word or number sequence in the target sentence that requires tagging. It uses character-level start and end positions to pinpoint the text (e.g., locating "10.0" and "6.0") and lists the attributes the model is expected to predict.

`golds`: The "ground truth" or correct answers for training and testing. It provides the exact taxonomy tags, the unscaled numerical facts, the time periods, and the measurement units, matched perfectly to the target IDs.

Additional Statistics for NORA DATASET Table 5 lists the ten most frequent tags and their proportions across the train, validation, and test splits. Table 6 summarizes the distribution of time-relation categories across the same splits.

B Analysis of Training Instances with High g_{tag}

To better understand the characteristics of training instances with high g_{tag}, we extracted the instances with the largest g_{tag} values and conducted a manual analysis. A Certified Public Accountant (CPA) reviewed these instances and assigned the correct labels using the same context provided to NORA. The results are summarized in Tables 7 - 11. Overall, the analysis shows that all five inspected instances contain incorrect tag labels. This 100% label-noise rate is substantially higher than that of a randomly sampled comparison set, which contains less than 5% noisy tag labels.

C Analysis of Training Instances with High g_{time}

We conducted a similar manual analysis on instances with the largest g_{time} values. A CPA reviewed these instances and assigned the correct labels using the same context provided to NORA. We summarize the results in Tables 12 - 16. Similar to the high-g_{tag} analysis, all five inspected instances contain noisy labels. These results suggest that the noise gate is effective at identifying noisy labels and attenuating their influence during training.

Feature	NORA Dataset	FiNER-139	FNXL
Dataset Statistics			
Size	6.6M sentences (12M+ tags)	1.1M sentences	79K sentences
Source Filings	10-K & 10-Q (more than 110,000 filings covering 8,000+ companies)	10K financial reports	10-K filings (2.2K companies)
Timeframe	2019–2024	2016–2020	2019–2021
Tag Inventory	446K unique tags (978 final tags)	139 frequent US-GAAP	2,794 US-GAAP
Taxonomy	US-GAAP & Custom concepts	US-GAAP only	US-GAAP only
Task Formulation & Context			
Target Attributes	Multi-attribute (8 total): `tag`, `time`, `scale`, `sign`, `fact`, `measure`, `decimals`, `axis`, `member`	Tag only	Tag only
Provided Context	Target sentence + document metadata (e.g., company, reporting period) + previous sentence + following sentence	Target sentence only	Target sentence only
Data Quality			
Noise Handling	Semantic filtering; removes irrelevant iXBRL content (e.g., page numbers, TOCs)	Noisy (contains embedded TOCs and page numbers)	Noisy (inherits standard filing-level noise)
Validity	Strict (guarantees $\geq$1 valid tag per instance)	Contains missing or inconsistent tags	May contain missing tags

Table 3: Comparison of NORA Dataset with FiNER-139 and FNXL. NORA Dataset offers significantly larger scale, comprehensive multi-attribute targets, richer document-level contexts, and cleaner instances.

Attribute	Description
Attributes used in the task definition	
`tag`	Custom tags are generically labeled as `custom`. Standard tags that occur fewer than 1,000 times within the training set are classified as `standard_rare`. Conversely, standard tags that appear more than 1,000 times retain their exact original names. Standard namespaces include `us-gaap`, as well as others like `dei` and `srt`.
`time`	This attribute is initially categorized into two period types: *instant* and *period*. It is then further classified into simplified timeframes, such as *past*, *current*, *future*, or a combination of these, e.g., *past_current* or *current_future*. The specific accounting period depends on the document type, with 10-K filings corresponding to a one-year period and 10-Q filings corresponding to a three-month period.
`scale`	This represents the scaling factors for the numerical value.
`sign`	A binary indicator of whether the numerical value should be interpreted as negative with respect to the accounting concept.
Additional attributes not used as task targets in this paper	
`fact`	This represents the factual value, which is formatted as a float.
`measure`	The unit of measurement.
`decimals`	The declared decimal precision.
`axis` (list)	A list of dimensions classifying the fact. Each item contains `amseq`, which links the dimension to its member, and `axis`, which stores the axis name.
`member` (list)	A list of members associated with the fact. Each item contains `amseq`, which links the member to its axis, and `member`, which stores the member name.

Table 4: XBRL target attributes in NORA DATASET, divided into attributes used in the paper's task definition and additional attributes retained in the dataset.

```
{
  "job_id": "000032019324000123-00030",
  "document": {
    "company_name": "Apple Inc.",
    "cik": "0000320193",
    "document_type": "10-K",
    "period_end_date": "2024-09-28",
    "fiscal_year": "2024",
    "period_focus": "FY",
    "current_fiscal_year_end": "--09-28",
    "document_link": "https://www.sec.gov/ix?doc=/Archives/edgar/data/0000320193/000032019324000123/
     aapl-20240928.htm"
  },
  "context": {
    "context_p": "The Company uses net proceeds from the commercial paper program for general
     corporate purposes, including dividends and share repurchases.",
    "context_t": "As of September 28, 2024 and September 30, 2023, the Company had $10.0 billion and
     $6.0 billion of commercial paper outstanding, respectively, with maturities generally less than
     nine months.",
    "context_n": "The weighted-average interest rate of the Company's commercial paper was 5.00% and
     5.28% as of September 28, 2024 and September 30, 2023, respectively."
  },
  "targets": [
    {
      "seq_id": 0,
      "start_pos": 66,
      "end_pos": 70,
      "text": "10.0",
      "attribute": ["tag", "fact", "time", "axis", "member", "measure", "decimals", "scale"]
    },
    {
      "seq_id": 1,
      "start_pos": 84,
      "end_pos": 87,
      "text": "6.0",
      "attribute": ["tag", "fact", "time", "axis", "member", "measure", "decimals", "scale"]
    }
  ],
  "golds": [
    {
      "seq_id": 0,
      "value": [
        "us-gaap:CommercialPaper",
        10000000000.0,
        "instant: 2024-09-28",
        [{"amseq": 0, "axis": "us-gaap:ShortTermDebtTypeAxis"}],
        [{"amseq": 0, "member": "us-gaap:CommercialPaperMember"}],
        "iso4217:USD",
        "-8",
        "9"
      ]
    },
    {
      "seq_id": 1,
      "value": [
        "us-gaap:CommercialPaper",
        6000000000.0,
        "instant: 2023-09-30",
        [{"amseq": 0, "axis": "us-gaap:ShortTermDebtTypeAxis"}],
        [{"amseq": 0, "member": "us-gaap:CommercialPaperMember"}],
        "iso4217:USD",
        "-8",
        "9"
      ]
    }
  ]
}
```


Figure 2: Example JSONL data instance from NORA DATASET.

Tag Name	Train	Validation	Test
us-gaap:DebtInstrumentInterest RateStatedPercentage	1.84%	1.88%	1.80%
us-gaap:DebtInstrumentBasisSpread OnVariableRate1	1.48%	1.48%	1.44%
us-gaap:DebtInstrument FaceAmount	1.43%	1.46%	1.40%
us-gaap:LineOfCreditFacility MaximumBorrowingCapacity	1.38%	1.39%	1.37%
us-gaap:AllocatedShareBased CompensationExpense	1.12%	1.11%	1.12%
us-gaap:ConcentrationRisk Percentage1	0.93%	0.94%	0.94%
dei:EntityCommonStock SharesOutstanding	0.76%	0.72%	1.67%
us-gaap:AmortizationOf IntangibleAssets	0.71%	0.71%	0.71%
us-gaap:AntidilutiveSecuritiesExcluded FromComputationOfEarningsPerShareAmount	0.70%	0.68%	0.70%
us-gaap:EffectiveIncomeTaxRate ContinuingOperations	0.64%	0.62%	0.65%

Table 5: Distribution of top tags in NORA training set.

Time	Train	Validation	Test	All	
				# of	%
instant; past	18.67%	18.72%	18.43%	2,255,336	18.65%
instant; current	26.81%	26.85%	26.84%	3,243,964	26.82%
instant; future	1.97%	1.93%	2.78%	248,873	2.06%
Total Instant	47.45%	47.51%	48.05%	5,748,173	47.52%
duration; past	24.09%	24.01%	23.68%	2,907,069	24.03%
duration; current	17.22%	17.31%	16.87%	2,079,432	17.19%
duration; future	1.48%	1.51%	1.46%	178,945	1.48%
duration; past_current	9.65%	9.55%	9.81%	1,167,837	9.65%
duration; past_future	0.07%	0.07%	0.08%	8,729	0.07%
duration; current_future	0.05%	0.06%	0.05%	5,869	0.05%
Total Period	52.55%	52.49%	51.95%	6,347,881	52.48%
Total Time Count	9,593,603	1,131,384	1,371,067	12,096,054	100.00%

Table 6: Distribution of Time Category in NORA DATASET

Field	Value
Context	**Target Sentence:** In the first quarter of 2024, payouts were $ [54] (Value: 54.0) million on 399, 372 psus, including dividends reinvested. **Document Meta:** type: 10-Q, period_end_date: 2024-06-30, fiscal_year: 2024, fiscal_period: Q2. **Preceding Sentence:** The resulting payout was 135% of the outstanding units multiplied by the company's average common share price calculated based on the last 30 trading days preceding December 31, 2023. **Following Sentence:** The 11,372 PDSUs that vested on December 31, 2023, with a fair value of $2 million, including dividends reinvested and matching units, will be paid out in future reporting periods pursuant to the DSU plan (as described above).
g_tag	0.0371
Label	`us-gaap:CollateralAlreadyPostedAggregateFairValue`
Pred.	`us-gaap:CollateralAlreadyPostedAggregateFairValue`
Expert Label	`us-gaap:StockGrantedDuringPeriodValueSharebasedCompensationGross`
Reason	The $54 million quantifies the total financial value of the Performance Share Units (PSUs) paid out during the quarter.Consequently, it requires a share-based compensation monetary-type tag like us-gaap: StockGrantedDuringPeriodValueSharebasedCompensationGross

Table 7: High Noise-Level Tag Label Example 1: PSU Payouts

Field	Value
Context	**Target Sentence:** The interest rate paid by the company for each interest period is calculated on the basis of a compounded average daily sofr rate plus [3. 139] (Value: 3.139) %. **Document Meta:** type: 10-Q, period_end_date: 2024-06-30, fiscal_year: 2024, fiscal_period: Q2. **Preceding Sentence:** None **Following Sentence:** The interest rate swap settles quarterly on each of march 1, june 1, september 1, and december 1.
g_tag	0.0370
Label	`us-gaap:RestrictedCashCurrent`
Pred.	`us-gaap:RestrictedCash`
Expert Label	`us-gaap:DebtInstrumentBasisSpreadOnVariableRate1`
Reason	A basis spread tag like `us-gaap:DerivativeBasisSpreadOnVariableRate` is correct because the text explicitly defines the 3.139% value as the fixed margin added to a variable benchmark rate (SOFR) for a financial instrument. This tag accurately matches the economic substance and percentage data type of the text, whereas the restricted cash tags represent entirely unrelated monetary asset balances.

Table 8: High Noise-Level Tag Label Example 2: SOFR Basis Spread

Field	Value
Context	**Target Sentence:** As of october 31, 2023, the amount of interest accrued, reported in other liabilities, was approximately $ [44, 000] (Value: 44000.0) which did not include the federal tax benefit of interest deductions. **Document Meta:** type: 10-K, period_end_date: 2023-10-31, fiscal_year: 2023, fiscal_period: FY. **Preceding Sentence:** We recognize accrued interest and penalties related to unrecognized tax benefits as components of our income tax provision. **Following Sentence:** The statute of limitations with respect to unrecognized tax benefits will expire between august 2024 and august 2025.
g_tag	0.0370
Label	`us-gaap:RestrictedCashAndCashEquivalents`
Pred.	`us-gaap:RestrictedCash`
Expert Label	`us-gaap:UnrecognizedTaxBenefitsInterestOnIncomeTaxesAccrued`
Reason	The text explicitly defines the $44,000 as accrued interest related to federal tax benefits, which ties directly to the unrecognized tax benefits mentioned in the same paragraph. Consequently, `us-gaap:UnrecognizedTaxBenefitsInterestOnIncomeTaxesAccrued` is the proper tag because its US-GAAP definition specifically aligns with capturing the liability for interest generated by these uncertain tax positions.

Table 9: High Noise-Level Tag Label Example 3: Accrued Interest on Unrecognized Tax Benefits

Field	Value
Context	**Target Sentence:** As of september 30, 2023 and december 31, 2022, $ 0. 5 million and $ [0. 4] (Value: 0.4) million of the grants repayable was included in accrued expenses and other current liabilities and the remaining balance was included in the grants repayable, net of current portion in the condensed consolidated balance sheet. **Document Meta:** type: 10-Q, period_end_date: 2023-09-30, fiscal_year: 2023, fiscal_period: Q3. **Preceding Sentence:** None **Following Sentence:** None
g_tag	0.0368
Label	`us-gaap:UnrecognizedTaxBenefitsIncomeTaxPenaltiesAndInterestAccrued`
Pred.	`us-gaap:UnrecognizedTaxBenefitsIncomeTaxPenaltiesAndInterestAccrued`
Expert Label	`custom`
Reason	The $0.4 million value must be tagged as a current liability because the text explicitly states it is included in "accrued expenses and other current liabilities" for the December 31, 2022 period. Consequently, a custom extension tag can be used for specificity, or fall back to us-gaap:OtherAccruedLiabilitiesCurrent for less specific tagging

Table 10: High Noise-Level Tag Label Example 4: Current Grants Repayable

Field	Value
Context	**Target Sentence:** On november 7, 2023, granite ridge entered into the first amendment to the credit agreement (the " first amendment ") which, among other things, decreased the borrowing base from $ 325. 0 million to $ [275. 0] (Value: 275.0) million and increased the aggregate elected commitments from $ 150. 0 million to $ 240. 0 million. **Document Meta:** type: 10-Q, period_end_date: 2024-09-30, fiscal_year: 2024, fiscal_period: Q3. **Preceding Sentence:** The borrowing base is redetermined semiannually on or about april 1 and october 1 of each calendar year, and is subject to additional adjustments from time to time, including for asset sales, elimination or reduction of hedge positions and incurrence of other debt. **Following Sentence:** This borrowing base decrease was a result of the disposition of certain assets in the permian basin to vital energy.
g_tag	0.0367
Label	`us-gaap:DebtSecurities`
Pred.	`us-gaap:DebtSecuritiesAvailableForSaleRestricted`
Expert Label	`us-gaap:LineOfCreditFacilityMaximumBorrowingCapacity`
Reason	The value 275.0 million represents the company's newly amended borrowing base, which dictates the maximum amount it is currently allowed to draw from its credit facility. Therefore, it aligns perfectly with the us-gaap:LineOfCreditFacilityMaximumBorrowingCapacity tag, as this element accurately captures the authorized limit of a company's available credit line.

Table 11: High Noise-Level Tag Label Example 5: Borrowing Base Amendment

Field	Value
Context	**Target Sentence:** Our provision for gross - to - net allowances was $ 3. 0 million at march 31, 2024, $ [0. 6] (Value: 0.6) million of which was recorded as a reduction to accounts receivable and $ 2. 4 million recorded as a component of accrued expenses. **Document Meta:** type: 10-Q, duration_end_date: 2024-03-31, fiscal_year: 2024, fiscal_duration: Q1. **Preceding Sentence:** None **Following Sentence:** None
g_time	0.5866
Label	instant; future
Pred.	duration; future
Expert Label	instant; current
Reason	The $0.6 million value represents a reduction to accounts receivable at a specific point in time, classifying its structural duration type as an instant metric. Because this exact snapshot date (March 31, 2024) aligns perfectly with the document's reporting duration end date, its relative temporal context is mapped as current.

Table 12: High Noise-Level Time Label Example 1: Gross-to-Net Allowance Reduction

Field	Value
Context	**Target Sentence:** In december 2019, as a result of the initial stipulation and settlement agreement, aep texas (a) recorded an impairment of $ [33] (Value: 33.0) million related to capital investments, which included $ 10 million of 2019 investments, in asset impairments and other related charges on the statements of income, (b) recorded a $ 30 million provision for refund on the statements of income for revenues previously collected through rates and (c) wrote - off $ 4 million of rate case expenses to other operation on the statements of income. **Document Meta:** type: 10-Q, duration_end_date: 2020-06-30, fiscal_year: 2020, fiscal_duration: Q2. **Preceding Sentence:** None **Following Sentence:** None
g_time	0.5714
Label	duration; current_future
Pred.	duration; future
Expert Label	duration:past
Reason	The $33 million is classified as a duration because it is an impairment charge recognized on the income statement, representing an activity over a duration rather than a frozen balance sheet snapshot. It receives the past label because the charge was recorded in December 2019, which historically predates the document's primary Q2 2020 reporting duration.

Table 13: High Noise-Level Time Label Example 2: AEP Texas Impairment Charge

Field	Value
Context	**Target Sentence:** Includes current portion of other liabilities of $ [5] (Value: 5.0) million at march 31, 2022 and $ 4 million at march 31, 2021. **Document Meta:** type: 10-Q, duration_end_date: 2022-03-31, fiscal_year: 2022, fiscal_duration: Q1. **Preceding Sentence:** None **Following Sentence:** None
g_time	0.5714
Label	duration; current_future
Pred.	duration; future
Expert Label	instant; current
Reason	The $5 million value is a balance sheet liability measured at the specific, single date of March 31, 2022, which defines its duration type as an **instant**. Because this exact measurement date matches the end date of the filing's current 10-Q accounting duration, its temporal classification is **current**.

Table 14: High Noise-Level Time Label Example 3: Current Portion of Other Liabilities

Field	Value
Context	**Target Sentence:** At june 30, 2023, the company had an outstanding balance of $ [25. 0] (Value: 25.0) million against its 2023 revolving facility. **Document Meta:** type: 10-K, duration_end_date: 2023-06-30, fiscal_year: 2023, fiscal_duration: FY. **Preceding Sentence:** During the fiscal year ended june 30, 2021, the company repaid $ 55. 0 million against its 2019 revolving facility that was outstanding as of june 30, 2020 and had no outstanding balances as of june 30, 2021 and 2022. The company has $ 110. 2 million availability under the 2023 revolving facility as of june 30, 2023. **Following Sentence:** None
g_time	0.5714
Label	duration; future
Pred.	duration; future
Expert Label	instant:current
Reason	The $25.0 million outstanding facility balance represents a balance sheet snapshot taken at a single, specific point in time on June 30, 2023. As this specific date aligns exactly with the end date of the current 10-K fiscal year filing, it directly corresponds to the current accounting duration.

Table 15: High Noise-Level Time Label Example 4: Revolving Facility Outstanding Balance

Field	Value
Context	**Target Sentence:** The company had proceeds from the sale of trade receivables under the amended arrangement of $ 70. 0 and $ 197. 7, respectively, for the three and nine months ended june 30, 2024 ($ [0. 0] (Value: 0.0) for the three and nine months ended june 30, 2023). **Document Meta:** type: 10-Q, duration_end_date: 2024-06-30, fiscal_year: 2024, fiscal_duration: Q3. **Preceding Sentence:** In accordance with asc 869, transfers and servicing, this amended arrangement is deemed a true sale, as the company retains no rights or interest and has no obligations with respect to the trade receivables. **Following Sentence:** As of june 30, 2024 and september 30, 2023, trade receivables in the amount of $ 32. 1and $ 0. 00. 0, respectively, were sold to the financial institution and are not reflected in the trade receivables in the consolidated balance sheets.
g_time	0.5553
Label	duration; future
Pred.	duration; past_current
Expert Label	duration:past
Reason	The target value of $0.0 measures proceeds explicitly over a continuous time span ("the three and nine months ended June 30, 2023"), clearly defining the duration type as a duration. Since this specific historical duration concluded a full year prior to the current 10-Q accounting duration ending June 30, 2024, it is classified temporally as past.

Table 16: High Noise-Level Time Label Example 5: Trade Receivables Sale Proceeds