



## TL; DR

- BERTScore exhibits **low sensitivity to numerical variation**, a significant weakness in finance where numerical precision directly affects meaning.
- We introduce **FinNuE**, a diagnostic dataset constructed with controlled numerical perturbations across earnings calls, regulatory filings, social media, and news articles.
- FinNuE demonstrate that BERTScore, even with domain-specific pretrained checkpoints, **fails to distinguish semantically critical numerical differences**.

## Motivation Case

$s_1$  : [revenue, increased, by, 3, ., 56, %, .]

$s_2$  : [revenue, increased, by, 4, %, .]

$s_3$  : [revenue, increased, by, 40, %, .]

- We would expect  $BERTScore(s_1, s_2) > BERTScore(s_2, s_3)$ , as the numerical difference between  $s_1$  and  $s_2$  (0.44 percentage points) is far smaller than between  $s_2$  and  $s_3$  (36 percentage points).
- Using the widely adopted HuggingFace implementation, we observe the opposite:  $BERTScore(s_1, s_2) = 0.9639 < BERTScore(s_2, s_3) = 0.9764$ .

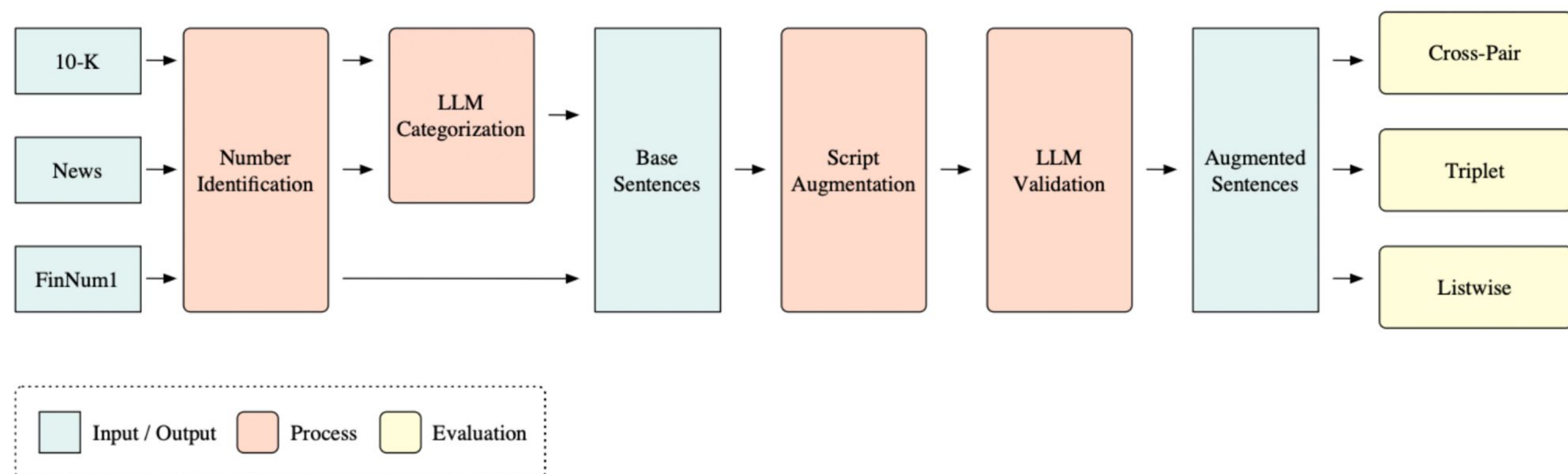
## Data Sources

- **Social media**: From Twitter, adapted from FinNum1[1]
- **Financial news**: From Financial Phrasebank [2]. We retain sentences with at least 50% annotator agreement to ensure quality.
- **Regulatory disclosures**: From 10-K filings (fiscal years 2023–2024) of the five largest company by market capitalization in each of the eleven GICS sectors.

[1] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral Understanding in Financial Tweets for Fine-Grained Crowd-Based Forecasting.

[2] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts.

## FinNuE dataset construction pipeline



## Numerical Augmentation

- **Random**: Generate 9 variants by randomly increasing or decreasing.
- **Rule-based**: Category-specific transformations, see below.

Augmentation	Target Number	Variant	$k$
Date Shift	Sep. 28, 2025	Sep. 28, 2029	9
Duration Convert	1 week	7 days; 1 month	9
Extra Decimal	3.5	3.56	9
Fractional Shift	0.25	0.37; 0.13	9
Scale Change	1,000	10,000	9
Million to Billion	110 million	0.11 billion	1
Last Digit Edit	110	1100; 11	2

## Statistics of FinNuE

Category	Subcategory	Augmentation	
		Random	Rule-based
Monetary	money	1,784	5,179
	quote	1,187	2,761
	change	1,173	3,166
	forecast	585	1,683
	buy price	505	1,175
	support or resistance	290	677
	sell price	112	262
Temporal	stop loss	30	55
	date	3,635	2,033
Percentage	time	591	78
	relative	1,431	1,392
Quantity	absolute	826	1,021
	quantity	1,737	3,139
Product Number	product number	293	458
Indicator	indicator	272	424
Option	exercise price	125	206
	maturity date	75	21
<b>Total</b>		<b>14,651</b>	<b>23,730</b>

## Evaluation Protocol

- **Anchor-based evaluation**
  - **Triplet**: The base sentence should have a higher BERTScore with variants whose numerical values are closer, and a lower score with those further away.
  - **Listwise**: BERTScore should rank sentence variants in correct order — from most to least similar — according to their numerical differences.
- **Cross-pair evaluation**  
 Between two sentence pairs with different contexts, the one with smaller numerical deviation should yield a higher BERTScore.

Checkpoint	Triplet (Accuracy)		Listwise (Kendall's $\tau_b$ )		Cross-Pair (Accuracy)	
	Random	Rule-based	Random	Rule-based	Random	Rule-based
bert-base-uncased	0.9214	0.8309	0.5409	0.3420	0.6727	0.4815
ProsusAI/finbert	0.9186	0.8431	0.5344	0.3580	0.6772	0.4860

## Conclusion and Future Work

We show that BERTScore fails to capture numerical meaning, performing near-randomly when comparing sentences with different numbers. Through **FinNuE**, a controlled diagnostic dataset, we reveal that subword tokenization and greedy alignment cause BERTScore to treat values like 2% and 20% as equivalent. We call for **numerically-aware evaluation metrics** that preserve number boundaries, encode magnitude, and integrate explicit numerical comparison.