Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

Validating Large Language Model Annotations

Anne Lundgaard Hansen

2026-020

Please cite this paper as:
Hansen, Anne Lundgaard (2026). “Validating Large Language Model Annotations,” Finance and Economics Discussion Series 2026-020. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2026.020.

NOTE: Staff working papers in the Finance and Economics Discussion Series (FEDS) are preliminary materials circulated to stimulate discussion and critical comment. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors. References in publications to the Finance and Economics Discussion Series (other than acknowledgement) should be cleared with the author(s) to protect the tentative character of these papers.

Validating Large Language Model Annotations

Anne Lundgaard Hansen [*]

Federal Reserve Bank of Richmond
Board of Governors of the Federal Reserve System

First draft: September 25, 2025.
This draft: March 11, 2026.

Abstract: This paper proposes a validation framework for LLM-generated measurements when reliable benchmarks are unavailable. Validity is established by testing whether an LLM can reconstruct passages from annotated labels while maintaining semantic consistency with the original text. The framework avoids circular reasoning by establishing testable prerequisite properties that must be met for a validation to be considered successful. Application to news article data demonstrates that the framework serves as a practical alternative to human benchmarking, which offers advantages in objectivity, scalability, and cost-effectiveness while identifying cases where LLMs capture economic meaning that human evaluators miss.

Keywords: Large Language Models, Validation Framework, Text Annotation, Sentiment Analysis.
JEL Codes: C18, C45, C80.

[*] The author is with the Quantitative Supervision and Research group at the Federal Reserve Bank of Richmond and affiliated with the Financial Stability Cyber and AI Sentinel Lab at the Board of Governors of the Federal Reserve System. E-mail address: `Anne.Hansen@rich.frb.org`. The views expressed in this paper are solely those of the authors and do not reflect the opinions of the Federal Reserve Bank of Richmond or the Board of Governors of the Federal Reserve System. The author thanks Jeff Allen, Steve Ge, Sophia Kazinnik, Seung J. Lee, Huiyu Li, Viviana Luccioli, David MacArthur, Nitish Sinha, Christina Wang, and seminar participants at the Federal Reserve Board for valuable discussion and feedback.

1 Introduction

Large language models (LLMs) are proving to be powerful tools for addressing questions within economics and finance. These models are increasingly used to quantify textual data, e.g., by labeling sentiment scores or classifying topics, which are subsequently plugged into downstream estimation problems.[1] One major concern with such approaches is the black-box nature of LLMs. With billions of parameters (which are unknown to users if the model is propriety) and extraordinarily large training data sets, it is difficult (if not impossible) to tease out the inner workings of these models. What is more, LLMs are designed to satisfy end users and therefore tend to provide answers, even in cases where instructions are unclear and carry multiple interpretations. LLMs are also known to occasionally hallucinate, and it is poorly understood under which circumstances hallucinations occur and how they can be avoided. It follows that researchers *should* question the validity of LLM annotations.

To build trust in LLM-generated measurements, researchers typically benchmark results against human-generated data, at least for a subset of their sample.[2] This approach is also recommended by Ludwig et al. (2025), who further suggest to use human-generated measurements to debias LLMs. However, the validity of human-generated annotations is also questionable. Human measurements are likely influenced by subjectivity and inconsistent treatment of data, e.g., due to learning as the task progresses, changing external environments under which the task is completed, and inattention or straight-up burnout.[3] Against such considerations, LLMs and AI in general have been cited for their ability to generate objective and consistent measurements (Sharma et al., 2025; Mirzakhmedova et al., 2024; Du et al., 2025; Törnberg, 2023). In addition to these concerns, human validation is time-consuming and costly, especially if relying

[1] See, e.g., Bertsch et al. (2025); Fan et al. (2024); Jha et al. (2024); Liu and Shi (2025); Shah et al. (2024); Kirtac and Germano (2024).

[2] See, e.g., Bauer et al. (2024); Chen et al. (2022); Cook et al. (2025); Hansen and Kazinnik (2024); Shapiro et al. (2022).

[3] These concerns are acknowledged in the literature, and often dealt with by averaging the responses of multiple human evaluators in the attempt to cancel out errors (Hansen and Kazinnik, 2024; Malo et al., 2014; Cook et al., 2025). This method, however, only works if errors are symmetrically distributed around zero and the number of human evaluators is large.

on genuine domain experts. Manual labeling tasks are therefore often crowdsourced, with the risk that crowd workers use LLMs for the task.[4]

This paper addresses the question: How can researchers validate the LLM measurements, in the absence of reliable external benchmark data, such as human-generated measurements? I propose minimal requirements that a researcher should confirm before deploying LLMs for measuring inputs to a downstream estimation problem. Then, I present methods to test these requirements.

In my framework, a combination of an LLM and a prompt designed to extract some measure from a text are considered valid as an entity if the measure repeatedly helps reconstructs the original text. This idea is similar to the practice of checking goodness-of-fit of an estimated model from traditional econometrics. In the context of LLMs, the data generating process is the combination of an LLM and a prompt, and the coefficient is the measure one wishes to extract from the text. Just as the interpretation of a coefficient relies on the data generating process in traditional econometrics, the validation framework requires the combination of the LLM and a prompt to validate the label.

To rule out concerns of circular dependency problems, I specify two additional requirements. First, the annotation backtranslation property requires that the LLM system can translate between label and text without introducing errors. Second, the separation property requires that two texts generated based on different labels can be clearly separated. Together, these conditions ensure in most cases that an erroneous measure does not pass the validation framework.

This work builds on the econometric framework for LLMs proposed by Ludwig et al. (2025). Their framework assumes that there exists a measurement that can be observed without using LLMs, albeit it may be costly to do so, e.g., using human-generated labels. In this context, they argue that LLMs should only be used for estimation problems establishing economic relationships when (i) the training data of LLMs do not overlap with the estimation data and (ii) after explicitly accounting for LLM errors by comparing human and LLM-generated results. Contrasting Ludwig et al. (2025), I focus on problems where the researcher does not have access to an external benchmark, e.g., due to human biases and errors. This setting is relevant in many

[4] Veselovsky et al. (2023) estimate that 33-46% of crowd workers use LLMs for completing tasks.

applications, as human measurements may not be reliable.

Benchmarking and evaluating the capabilities of LLMs has garnered considerable attention among researchers and practitioners. This trend is exemplified by the proliferation of leaderboards that rank LLMs according to various standardized benchmark tests targeting diverse competencies, including coding proficiency, knowledge retention, human-like reasoning, and scientific reasoning. However, a model's ranking on such leaderboards provides limited insight into its reliability for specific tasks (Ludwig et al., 2025). Furthermore, these ranking systems have been criticized for their susceptibility to gaming and distortion (Singh et al., 2025). Patwardhan et al. (2025) suggests assessing the trustworthiness of LLMs by evaluating the consistency in answers across multiple model runs. They also consider a cross-validation technique, where the response generated by one LLM is compared with the responses of other LLMs. However, Ludwig et al. (2025) and Reiss (2023) argue that the errors of LLMs are unpredictable and do not necessarily center around true values, implying that the average of a large number of LLM predictions does not necessarily constitute a correct one. Wang et al. (2023) suggests to generate multiple a set of reasoning from the same model and pick the most consistent one. They show that this method improves results in *reasoning* tasks, which are inherently different from annotation tasks.[5] While the validation framework I propose is also using the LLM to validate itself, similar to these contributions, the criteria for validity is rooted in the original passage rather than purely focusing on LLM output.

I proceed as follows: Section 2 defines requirements for validity of LLM annotations. Section 3 proposes a test that assesses whether choice of LLM and prompting strategy satisfies these requirements. Section 4 presents illustrative examples of how the test can be used to validate the ability of LLMs to classify sentiment, clarity, temporal focus, and the identification of specific topic mentioning. This section also applies the validation framework across different sizes and generations of LLMs, showing that larger and more recent models pass the prerequisite tests needed for the framework more frequently. In Section 5, I present a full-scale application showcasing how the tests can be replace human benchmarking in LLM annotation workflows.

[5] Indeed, the ability to generate consistent and logical arguments may not coincide with the ability to produce meaningful annotations.

Limitations are discussed in Section 6, and conclusions follow in Section 7.

2 Requirements for Validity

Consider the problem of using an LLM to produce an annotation $\beta \in \mathcal{S}$ from a passage of text y. For example, y could be news article headlines or earnings call transcripts, and β could be sentiment labeled from the set $\mathcal{S} = \{\text{positive}, \text{neutral}, \text{negative}\}$ or other measures of linguistic characteristics. The passage y can be a subset of a full document, e.g., sentences or passages from earnings call documents.[6] The measure is assumed to be discrete, but can be either numeric or categorical.

The application of LLMs involve the choice of a model and a prompting strategy. Both elements impact the results and are difficult to disentangle: a prompt may work in one way with one LLM and another way with another LLM. I shall therefore refer to these choices as one entity, which I call a function.

Importantly, as in Ludwig et al. (2025), I do assume that there exists a "true" measurement. But, whereas Ludwig et al. relies on human annotations to define truth, I assume that the true measurement cannot necessarily be recovered outside of LLMs, e.g., due to human bias and inattention. Let $\mathring{\beta}$ denote this true annotation, and let $\mathring{f}$ be the combination of an LLM and a prompt that generated the observed passage of text y based on $\mathring{\beta} \in \mathcal{S}$. The true data generating process can thus be described by:

$$y = \mathring{f}\left(\mathring{\beta}\right). \tag{1}$$

Other than the explicit dependency on the annotation $\mathring{\beta}$, there are no requirements on the prompt component of $\mathring{f}$. It may incorporate various enhancements such as retrieval-augmented generation (RAG) with external data, detailed instructions, few-shot examples, and chain-of-thought reasoning, or it may simply be a straightforward request to generate a text with characteristic $\mathring{\beta}$.

[6] For many LLMs, chunking of full-document texts is necessitated by restrictions on length of the model's context windows.

The true data generating process given by $\mathring{f}$ and thus $\mathring{\beta}$ are unknown. Instead, the researcher is using a choice of LLM and prompting strategy to annotate the text y. Let f^{-1} denote the chosen method, and let $\hat{\beta} \in\in \mathcal{S}$ be the resulting annotation:

$$\hat{\beta} = f^{-1}\left(y\right). \tag{2}$$

The question is: how can $\hat{\beta}$ be validated as a measurement of $\mathring{\beta}$, when only y is observed?

If the problem was numeric, an econometrician could estimate $\hat{\beta}$ and check that the goodness-of-fit of the fitted values $f(\hat{\beta})$. For example, in a linear regression model, OLS estimation yields $\hat{\beta} = f^{-1}(y) = (X'X)^{-1}X'y$ and the fitted values are given by $\hat{y} = f(\hat{\beta}) = X\hat{\beta}$. To formulate this idea in the context of textual data with categorical labels, suppose that the annotator function f^{-1} is accompanied by a text generator function f that generates a passage of text from a label β. Analogous to the linear regression example, I consider $\hat{\beta}$ a valid annotation for $\mathring{\beta}$ if $\hat{y} = f(\hat{\beta})$ provides a satisfying goodness-of-fit to y. In the context of textual data, I propose to measure goodness-of-fit by semantic similarity.

Continuing the linear regression analogy, measuring goodness-of-fit is not sufficient to build confidence in an estimated quantity. An estimator should at least be consistent, ensuring that estimates converge in probability to $\mathring{\beta}$ in the limit. Similarly, I impose requirements on the functions f and f^{-1}. Specifically, I impose two requirements on the generator function f: The annotation backtranslation[7] and the separation properties. The annotation backtranslation property ensures that f and f^{-1} are mutually consistent, while the separation property ensures that the function f generates texts with different defining characteristics for different labels. These properties are detailed below.

[7] Annotation backtranslation is a variation of the backtranslation property from machine learning, where accuracy of a translated text is assessed by re-translating it back to its original language (Sennrich et al., 2016). Li et al. (2024) also adopts the idea of backtranslation to generate instruction prompts used to simulate training data for fine-tuning language models; they denote their method *instruction backtranslation*.

Annotation Backtranslation Property: The functions f and f^{-1} satisfy the annotation backtranslation property if for any annotation $\beta \in \mathcal{S}$,

$$f^{-1}\left(f(\beta)\right) = \beta. \tag{3}$$

□

Separation Property: The function f satisfies the separation property if for all annotations $\gamma \neq \beta$ with $\gamma \in \mathcal{S}$, $f(\gamma)$ is not semantically similar to $f(\beta)$.

□

The annotation backtranslation property addresses the concern that validation fails for a correct $\hat{\beta}$ due to an erroneous simulation function f, i.e., $\hat{\beta} = \mathring{\beta}$ but $f(\mathring{\beta})$ is not semantically similar to y. In this case, $\hat{\beta}$ would be incorrectly rejected as a valid annotation because of the failure of f. Choosing f and f^{-1} such that (3) is satisfied rules out such cases because if $\hat{\beta} = \mathring{\beta}$ but $f(\hat{\beta}) = \check{y}$, where $\check{y}$ is not semantically similar to y, then

$$f^{-1}\left(\check{y}\right) \neq \mathring{\beta}, \tag{4}$$

which contradicts (3). Similarly, the property also captures cases where a wrong label is validated due to an erroneous simulation function.

Another concern is that $\hat{\beta}$ is incorrect, i.e., $\hat{\beta} \neq \mathring{\beta}$, but $f(\hat{\beta})$ is generating a text that is semantically similar to y. In this case, both f and f^{-1} fails, but in a way such that their combination appears valid. This is problematic because the researcher would accept an incorrect annotation $\hat{\beta}$. The separation property rules out this type of error. Specifically, the separation property ensures that if $\hat{\beta} \neq \mathring{\beta}$, then $f(\hat{\beta})$ is not semantically similar to y.

Given this setup, I propose the following definition for an annotation to be a valid measure of $\mathring{\beta}$:

Valid Annotation: The annotation $\hat{\beta}$ defined in (2) is a valid measure of $\mathring{\beta}$ if for a function f that accompanies the annotation function f^{-1}, such that the annotation backtranslation and separation properties are satisfied, $\hat{y} = f(\hat{\beta})$ is semantically similar to y.

□

This definition establishes truth based on the semantic similarity between text generated with this label and the original text. This standard is intuitively appealing beyond the analogy to linear regression presented above: If the generated text is not similar to the original one, then the annotated label is not capturing the text's essence and should not be considered valid. Note that this definition also invalidates annotation problems that are ill-defined, e.g., scoring aspects of a text that have little relevance for the text's character.

The researcher has considerable flexibility in specifying the function f, including the choice of prompt, model, and hyper-parameters such as temperature. The only requirement is that the function satisfies the prerequisite properties jointly with the annotation function f^{-1}. The researcher may therefore introduce elements from the original text into the prompt, e.g., its characteristics such as style and length, to increase semantic similarity with the original text. This is permissible as long as the inclusion does not prevent the function from simulating semantically distinct texts from different labels (the separation property).

2.1 Addressing Concerns of Circular Dependency

The concept of using LLMs to validate themselves raises important concerns of circular dependency. The annotation backtranslation and separation properties will jointly detect most issues. However, there remains one type of error that could pass the validation framework due to circular dependency. This type of error is highly specialized, requiring multiple aspects to always fail in a specific way, and is therefore unlikely to occur. Nevertheless, examining this scenario in detail is valuable for understanding the boundaries and potential vulnerabilities of the framework.

Consider the simple problem of annotating the sentiment of a text as either positive or negative. If an LLM annotation is valid, the translation between true label $\mathring{\beta}$, original text y, annotated label $\hat{\beta}$, and simulated text $\hat{y}$ clearly separates positive from negative as follows:

$$\begin{tikzcd}
\mathring{\boldsymbol{\beta}} \arrow[r, "\mathring{f}"] & \boldsymbol{y} \arrow[r, "f^{-1}"] & \hat{\boldsymbol{\beta}} \arrow[r, "f"] & \hat{\boldsymbol{y}} \\
+ \arrow[r] & + \arrow[r] & + \arrow[r] & + \\
- \arrow[r] & - \arrow[r] & - \arrow[r] & -
\end{tikzcd}$$

An annotation function f^{-1} that fails the validity requirement either annotates a positive text as negative, a negative text as positive, or both:

$$\begin{tikzcd}
\mathring{\boldsymbol{\beta}} \arrow[r, "\mathring{f}"] & \boldsymbol{y} \arrow[r, "f^{-1}"] & \hat{\boldsymbol{\beta}} \arrow[r, "f"] & \hat{\boldsymbol{y}} \\
+ \arrow[r] & + \arrow[dr] & + \arrow[r] & + \\
- \arrow[r] & - \arrow[ur] & - \arrow[r] & -
\end{tikzcd}$$

This type of error is easily detected by the validation framework, because y will not be semantically similar to $\hat{y}$.[8] But, what happens if not only f^{-1} is invalid but also f generates text with a different sentiment than intended by $\hat{\beta}$? For example, a system where the annotation function is biased towards negative sentiment but the simulation function is biased in the other direction, y and $\hat{y}$ may be semantically similar despite $\hat{\beta} \neq \mathring{\beta}$:

$$\begin{tikzcd}
\mathring{\boldsymbol{\beta}} \arrow[r, "\mathring{f}"] & \boldsymbol{y} \arrow[r, "f^{-1}"] & \hat{\boldsymbol{\beta}} \arrow[r, "f"] & \hat{\boldsymbol{y}} \\
+ \arrow[r] & + \arrow[dr] & + \arrow[r] & + \\
- \arrow[r] & - \arrow[r] & - \arrow[ur] & -
\end{tikzcd}$$

In this case, the LLM system yields a positive text $\hat{y}$ based on a positive origin y but an erroneous negative label $\hat{\beta}$. But, the validation framework captures this case as the prerequisite properties fail. Specifically, the separation property fails because f generates a positive text regardless of the value for $\hat{\beta}$.

The validation framework thus detects cases where the annotation and simulation functions are biased in one direction. The only remaining case of concern occurs if there are counteracting biases that occurring in both directions:

$$\begin{tikzcd}
\mathring{\boldsymbol{\beta}} \arrow[r, "\mathring{f}"] & \boldsymbol{y} \arrow[r, "f^{-1}"] & \hat{\boldsymbol{\beta}} \arrow[r, "f"] & \hat{\boldsymbol{y}} \\
+ \arrow[r] & + \arrow[dr] & + \arrow[dr] & + \\
- \arrow[r] & - \arrow[ur] & - \arrow[ur] & -
\end{tikzcd}$$

[8] It is also likely that the annotation backtranslation property will fail.

This type of error will not be detected by the validation framework proposed in this paper. However, I deem these cases to be highly unlikely as they require errors to occur systematically and in a way where biases in f^{-1} and f are perfectly counteracting each other.

3 Testing Validity

This section presents methods for testing the requirements for validity outlined above. An annotation $\hat{\beta}$ as defined by (2) should be rejected as a valid measurement for the passage y if a passage simulated from $\hat{\beta}$, $f(\hat{\beta})$, is sufficiently different from y. Semantic similarity can be measured using cosine similarity between vector representations of y and $f(\hat{\beta})$. As $f(\hat{\beta})$ is non-deterministic, I propose assessing validity based on the average cosine similarity between y and a large number of outcomes for $f(\hat{\beta})$. In other words, $\hat{\beta}$ is not valid if the statistic,

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \text{cos sim}\left(y, f(\hat{\beta})_i\right), \tag{5}$$

where f is chosen such that f and f^{-1} satisfy the annotation backtranslation and separation properties, is sufficiently small. To determine a threshold for this condition, I propose to simulate the test statistic under the null hypothesis that $\hat{\beta}$ is a valid annotation, akin to bootstrap testing from the traditional econometric toolbox. The bootstrap can be performed using the following steps:

1. Generate a text given $\hat{\beta}$: $\tilde{y} = f(\hat{\beta})$.
2. Simulate the test statistic given $\tilde{y}$: $\tilde{\tau}_b = \frac{1}{N}\sum_{i=1}^{N} \text{cos sim}\left(\tilde{y}, f(\hat{\beta})_i\right)$.
3. Repeat steps 1-2 a large number of times (B) to obtain a distribution of test statistics under the null hypothesis: $\{\tilde{\tau}_b\}_{b=1}^{B}$.
4. Reject at significance level α if the α'th percentile of $\{\tilde{\tau}_b\}_{b=1}^{B}$ exceeds τ.

The distribution of $\bar{\tau}_b$ shows the range of values one would expect to observe if $\hat{\beta}$ is the true measurement for y. If the statistic based on the observed passage y falls in the far left tail of this distribution (as defined by the significance level), the annotation $\hat{\beta}$ should not be considered a valid measurement.

The requirement for validity hinges on the annotation backtranslation and separation property. These assumptions are testable within similar frameworks.

Testing the Annotation Backtranslation Property: For a choice of f and f^{-1}, and a given annotation β, e.g., $\beta = \hat{\beta}$, the annotation backtranslation property can be assessed using the following steps:

1. Fix an annotation β.
2. Generate a passage given β: $\tilde{y} = f(\beta)$.[9]
3. Generate an annotation for $\tilde{y}$: $\tilde{\beta} = f^{-1}(\tilde{y})$.
4. Define a binary variable that takes the value one if $\tilde{\beta} = \beta$ and zero otherwise: $I_i = \nVdash_{\tilde{\beta}=\beta}$.
5. Repeat steps 2-3 a large number of times. The average $\frac{1}{N}\sum_{i=1}^{N} I_i$ defines the accuracy and should be close to one.

Testing the Separation Property: For a choice of f and f^{-1}, and a given annotation β, e.g., $\beta = \hat{\beta}$, testing the separation property is a test of the null hypothesis that for any label $\gamma \neq \beta$, $f(\gamma)$ is not semantically similar to $f(\beta)$. Consider the test statistic,

$$\xi = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \text{cos sim}\left(f(\beta)_i, f(\gamma)_j\right). \tag{6}$$

The null hypothesis is rejected if ξ is sufficiently large. The following bootstrapping algorithm can be used to compute a rejection threshold:

1. Generate N passages given an annotation $\gamma \neq \beta$: $\tilde{y}_i = f(\gamma)$ for $i = 1, 2, \ldots, N$.
2. Simulate the test statistic given $\tilde{y}_i$: $\bar{\xi} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \text{cos sim}\left(f(\beta)_i, \tilde{y}_j\right)$.

[9] If β is chosen such that $\beta = \hat{\beta}$, this step can be performed by reusing the passages generated during the test of validity, or vice versa.

3. Repeat steps 1-2 a large number of times (B) to obtain a distribution of test statistics under the null hypothesis: $\{\bar{\xi}_i\}_{i=b}^{B}$.

4. Reject at significance level α if ξ exceeds the $(1-\alpha)$'th percentile of $\{\bar{\xi}_i\}_{b=1}^{B}$.

3.1 Interpretation

If an LLM and associated prompting strategy pass both the validity test and the prerequisite tests, an annotation $\hat{\beta}$ generated by this model-prompt combination can be regarded as valid according to the definition put forth in this paper. Passing the validation test is not only validation of the LLM annotation, but also the function (prompt and model) used to generate that label and the simulation function that translates the label into text.

In contrast, if the validation test is rejected at conventional significance levels (regardless of the test outcomes of the prerequisite tests), $\hat{\beta}$ should not be applied in further analyses. Technically, $\hat{\beta}$ could still be correct. For example, randomly choosing between sentiment labels negative, neutral, positive will on average yield a correct result in one third of the times. However, it will often be wrong and the results are therefore not reliable. It is important to note that the validation test rejects the annotation and simulation functions jointly. The source of rejection could be the model, the specific prompts, how the text is parsed (e.g., passages for which multiple labels apply will likely not be validated by the framework), the labeling task including the set of possible labels $\mathcal{S}$. Adjusting any of these settings, such as deploying an alternative prompting strategy or using a different LLM, may change the test outcomes.

What happens if the validation test passes, but the prerequisite tests fail? The prerequisite properties are necessary conditions for the validation framework to work. Failing these tests will therefore invalidate the framework and the annotation cannot be concluded to be neither valid nor invalid.

4 Illustrative Examples

I illustrate the proposed tests on five different text passages, all related to financial and economic applications. A full-scale application to a larger number of passages is presented in the following section. To test the methods on a wide variety of textual data, I consider passages taken from a bank's 10-K filing, an earnings call transcript, the title and subtitle of a news article, a speech given by a governor of the Federal Reserve System, and a comment from a Reddit conversation thread. These passages vary substantially in linguistic style and length. Together, they offer a varied panel that illustrates the effectiveness of the tests. The passages are defined below:

10-K Filing: *"An adverse change in market conditions in particular segments of the economy, such as a sudden and severe downturn in oil and gas prices or an increase in commodity prices, severe declines in commercial real estate values, or sustained changes in consumer behavior that affect specific economic sectors, could have a material adverse effect on our clients whose operations or financial condition are directly or indirectly dependent on the health or stability of those market segments or economic sectors, as well as clients that are engaged in related businesses."* (JPMorgan Chase & Co., Form 10-K, December 31, 2024; slightly altered to remove firm name).

Earnings Call: *"This client value proposition combined with disciplined pricing helped drive a 9% year over year increase in net interest income. Another strong highlight this quarter was expense management, which enabled us to deliver more than 600 basis points of operating leverage. Continued innovation and the deployment of advanced technology and tools helped us to hold expense growth to just 1% year over year while revenue grew significantly. As a result, our efficiency ratio improved falling below 50% for the quarter. We continue to invest in high-tech, which drove higher digital engagement and we continue to invest in high touch."* (Bank of America, Transcript of 2025 Q3 earnings call, October 15, 2025).

News Article: *"Crypto Investors Celebrated for Most of 2025. Then Came the Hangover. Bitcoin finished the year in the red as investors grappled with the AI trade and macroeconomic risk."* (Wall Street Journal, January 1, 2026).

Speech: *"Living and teaching in Michigan during the Great Recession, I saw firsthand how the financial system's fragility contributed directly to job losses. One example is how the default of Lehman Brothers contributed, via a chain of events, to declines in employment in Michigan. Lehman's failure in September 2008 led a money market fund to "break the buck"—the fall in the value of its assets meant it could no longer redeem shares for the $1 that investors expected*

> *to receive—prompting a run on the funds. In turn, the funds pulled back from riskier assets, including asset-backed commercial paper. But the major auto finance companies depended on that commercial paper to finance loans to consumers; hence, they came under stress.3 With less credit available, auto sales plummeted, and Michigan was hit very hard. Many people—including some of my family members, my students' and colleagues' family members, friends, and neighbors—lost their jobs and experienced significant hardship"* (Lisa D. Cook, "A Policymaker's View of Financial Stability" delivered at Georgetown University's McDonough School of Business Psaros Center for Financial Markets and Policy, Washington, D.C., November 20, 2025).
>
> **Reddit Comment**: *"Interest rates aren't really high, these are normal. You got spoiled previously. Mid 5s is where rates we will be if new buyers get lucky."* (Reddit comment by @memorabiliafan, April 20, 2025).[10]

For these passages, I apply the test to various common tasks. Specifically, Sections 4.1-4.5 test the validity of measuring simple sentiment, granular sentiment, clarity, the temporal focus, and the discussion of certain topics. The main set of experiments is presented using the Claude 4.5 Sonnet model (referred to as the LLM), a frontier LLM at the time experiments were conducted, released September 29, 2025.[11] Section 4.7 is devoted to discuss validation of LLMs across generations. The Amazon Titan Text Embedding v2 model is used to generate vector embeddings that are nuanced and context-aware for accurately computing cosine similarity. All applications are facilitated through the AWS Bedrock API.

4.1 Experiment I: Measuring Simple Sentiment

The first set of experiments consider the problem of extracting sentiment scores from each of the passages. Starting simple, I first test the validity of the LLM annotating each passage as positive, neutral, or negative. The validation test and its prerequisites require the specification

[10] The Reddit conversation is available at `https://www.reddit.com/r/Mortgages/comments/1k3tn81/comment/mo5t9d9/`, retrieved January 3, 2026.

[11] The model is implemented with a temperature of 1.0 and unconstrained nucleus sampling (corresponding to a top-p parameter of 1.0), ensuring maximum variation in the simulations. No system prompt is specified, deferring to the model's default behavior.

of an annotation prompt and a simulation prompt. The former requests the model to classify a text provided as an input using one of the three labels. Abstracting from instructions on how to return output in a JSON object, the annotation prompt is given as follows:

```
TASK: Read the following passage and classify the sentiment using one of the
↪ following labels: [Positive, Neutral, Negative].
PASSAGE: {text}
```

Specification of the simulation prompt is more subtle. On the one hand, instructing the model to simulate a passage similar to the original passage may increase the test statistic. On the other hand, providing too detailed instructions may inflate the distribution of cosine similarity under the null hypothesis leading to a rejection of validity. It may also cause the separation property to fail, invalidating the validation test. For all passages, I found that describing in broad terms the content, length, and linguistic style of the original passage in the prompt balances this trade-off. The simulation prompt, given a sentiment label, is thus defined for the 10-K Filing passage as follows (abstracting from instructions on the returned JSON object):[12]

```
TASK: Write {n} arbitrary passages from a bank's annual 10-K filing with sentiment
↪  characterized as {label}. The passages should describe how market conditions
↪ may affect the bank's clients. Write the passages in passive voice. The length
↪ of each passage should be around 90 words.
```

Results are shown in Table 1. The table shows results for both prerequisite tests and the validation test. Panel (a) shows test outcomes for the 10-K filing. The model annotates the passage with negative sentiment. The annotation backtranslation property is therefore tested given the negative label. Specifically, the simulation prompt provided above is used with the LLM to simulate one hundred passages with a negative label. The annotation prompt is then used with the LLM to score the sentiment of all simulated passages. The table shows that all simulated texts are scored with a negative label, resulting in an accuracy of 100%. The backtranslation property is therefore satisfied. The separation property is tested for all but the estimated sentiment of the original passage, i.e., it is tested for the positive and neutral labels. The test statistics are respectively 0.58 and 0.60, which are lower than their associated critical

[12] Simulation prompts for the other passages follow a similar structure. All prompts are available in the online appendix published on my website, https://sites.google.com/view/anneh/.

values of 0.60 and 0.62. The test of this property is therefore passed as well. Given these test outcomes, the validation test is meaningful. The validation test passes as the test statistic of 0.623 exceeds the critical value of 0.621. It follows that the LLM annotation of the 10-K filing passage can be considered valid.

The validation test also passes for the remaining passages, reported in panels (b)-(e). These conclusions therefore support the hypothesis that LLMs are able to correctly classify sentiment on a simple scale.

Notably, the performance of LLMs sometimes depend on the definition of passages. The tests provided in this paper can also help guide how texts should be optimally parsed for obtaining the highest accuracy. For example, in addition to the Reddit comment listed above, I also tested the validity of the LLM scoring the sentiment of a conversation between multiple users. Specifically, I considered the following extended version of the Reddit passage:

"[Firm_Care_7439:] Will we ever get COVID type rates ever again? [FastSunlul:] No because I'm finally old enough and with money therefore it won't happen. [Khandious:] Yeah, Rates will be 2.75% on Monday - June 12, 2028 @ 9:53AM EST. [memorabiliafan:] Interest rates aren't really high, these are normal. You got spoiled previously. Mid 5s is where rates we will be if new buyers get lucky. [Big-Business1921:] Absolutely! If my calculations are correct, I anticipate it happening around 2092."[13]

The LLM rates this passage as negative. However, the validation test fails even when the simulation and annotation functions pass the annotation backtranslation and separation tests. Indeed, even from a manual reading of this passage, the sentiment is unclear as sarcastic ping-pong between users makes it difficult to objectively rate the passage. For the following experiments, I proceed only with the simple Reddit comment presented above.

4.2 Experiment II: Measuring Detailed Sentiment

Applications in finance and economics often involve scoring texts on a detailed scale, e.g., ranging from one to five. In the second experiment, I continue to focus on sentiment, but expand the scale on which sentiment is measured to the following five labels: positive, mostly

[13] The Reddit conversation is available at `https://www.reddit.com/r/Mortgages/comments/1k3tn81/comment/mo5t9d9/`, retrieved January 3, 2026.

positive, neutral, mostly negative, and negative. The simulation and annotation prompts from the first experiment are maintained, only with slight adjustments to accommodate the new scale.

Table 2 reports the results. The LLM continues to rate the 10-K filing passage as negative, and all tests for this case pass. The results, both in terms of assigned label and test outcomes, are also identical to the those from the simple sentiment experiment for the earnings call and Federal Reserve governor speech. In contrast, the passages from the news article and the Reddit conversation are now classified as mostly negative, and these classifications do not pass as valid as the annotation backtranslation property fails with accuracy of just 5-7%. The test results thus indicate that the task of scoring sentiment on a more granular scale is more often invalid. This result is consistent with human benchmark data, which typically involve more disagreement when the granularity of labels increases. For example, there is more disagreement about whether a passage is considered negative or mostly negative versus whether a passage is considered negative or neutral.

4.3 Experiment III: Measuring Clarity

Next, consider the problem of measuring the textual clarity of the passages using the labels clear and vague. While sentiment is a well-defined concept, clarity is more ambiguous and can be interpreted in various ways. For the annotation backtranslation property to be satisfied, it is therefore important to define the concept of clarity in the annotation and simulation prompts.[14] Specifically, I define the annotation prompt as follows:

```
TASK: Read the following passage and classify the clarity using one of the
↪ following labels: [Clear, Vague].
A passage should be classified as ’Clear’ if it is objective, has one clear
↪ interpretation, and explicitly states information rather than implying it.
A passage should be classified as ’Vague’ if it is subjective, is subject to
↪ multiple possible interpretations, or uses hedging words such as ’sort of’, ’
↪ perhaps’, and ’kind of’.
```

[14] For example, the annotation backtranslation test fails for the 10-K filing passage with an accuracy of 65% when clarity is not defined.

```
PASSAGE: {text}
```

The simulation prompt is given analogously. For example, for the 10-K filing passage:[15]

```
TASK: Write {n} arbitrary passages from a bank’s annual 10-K filing. The passages
↪ should describe how the bank’s business model and technological developments
↪ impacted income and performance for the quarter. {clarity_instructions}. The
↪ length of each passage should be around 90 words.
```

where `clarity_instructions` depends on the label as follows:

```
if label == ’clear’:
        clarity_instructions = "The passages should be written in clear language, i
        ↪ .e., they should be objective, have one clear interpretation, and
        ↪ explicitly state information rather than implying it."
elif label == ’vague’:
        clarity_instructions = "The passages should be written in vague language, i
        ↪ .e., they should be subjective, potentially carrying multiple
        ↪ interpretations, and/or use hedging words such as ’sort of’, ’perhaps’,
        ↪  and ’kind of’."
```

The test results for all passages are shown in Table 3. The passages represent both clear and vague texts as classified by the LLM. For all passages, the choice of simulation and annotation functions satisfy the annotation backtranslation and separation properties. The validation test, however, fails for three out of five passages: the 10-K filing passage (vauge), the Federal Reserve governor speech extract (clear), and the Reddit comment (vague). Clarity, even when properly defined, is thus more difficult to classify than sentiment. The validation test passes for both a passage classified as clear (the passage from the earnings call transcript) and vague (the news article passage). The LLM annotation of clarity is thus often invalid and there is no pattern in the distribution of test outcomes across labels. These results underscore the importance of testing validity before using such annotations in downstream applications.

[15] Simulation prompts for the other passages follow a similar structure. All prompts are available in the online appendix.

4.4 Experiment IV: Measuring Temporal Focus

LLMs are also often used to assess the temporal orientation of texts, e.g., to classify whether a text is forward-looking, focusing on the present, or backward-looking. Table 4 assesses the validity of LLM annotations of temporal focus in the five passages. The passages represent a mix of forward-looking (10-K filing and Reddit comment) and backward-looking texts (earnings call transcript, news article, and Federal Reserve governor speech), but none of them are classified as focusing on the present. The prerequisite properties are satisfied for all passages, and the validation tests pass for all but the Reddit comment. These results indicate that temporal focus is straightforward for the annotation function to classify, similar to simple sentiment.

Manually reading the Reddit comment, the passage contains elements that would fit all three labels: *"these are normal"* is a statement about the present, *"you got spoiled previously"* is backward-looking, and *"mid 5s is where rates will be"* is a prediction for the future. It is therefore comforting that the validation test fails for this passage.

4.5 Experiment V: Measuring Topics

Finally, I consider the ability of the LLM to identify certain topics within the passages. For each passage, I identify a topic that *is* discussed within the text and one that *is not* discussed in the text. Specifically, topics that are discussed are chosen as "economic risks" for the 10-K filing passage, "technological developments" for the earnings call transcript passage, "crypto investing" for the news article passage, "financial crises" for the speech passage, and "interest rates" for the Reddit comment. The topics not discussed are chosen such that they are plausible topics that could very well have been present in these passages. Specifically, I use "investments" for the 10-K filing passage, "geopolitical risks" for the earnings call transcript passage, "interest rates" for the news article passage, "new technologies" for the speech passage, and "stock market" for the Reddit comment.

I then prompt the LLM to identify whether each of these topics are discussed in the passage using true/false labels. This exercise is different from the previous experiments in the sense that there is a correct and wrong answer. Namely, the labels should be true for the true topics (true positive identification) and false for the false topics (true negative identification).

Table 5 reports results from the true positive identification exercise. The results show that the LLM correctly identifies the topics for all passages, and the validation test passes for all but the earnings call transcript. Turning to the true negative identification reported in Table 6, most of the passages are correctly annotated with the false label and the validation test passes. Two exceptions are observed: For the Reddit comment, the labels is false as expected, but the validation test does not pass. This result likely occurs because the chosen false topic (the stock market) is indirectly related to the actual topic discussed (interest rates). It is therefore up to interpretation whether the passage is discussing the stock market, to some extent. The second exception is the passage from the 10-K filing, for which the topic "investments" is incorrectly identified. Interestingly, the validation test fails with a statistic that is much lower than the critical value (0.55 vs 0.59). The test thus correctly identifies the wrong label.

4.6 Discussion

Table 7 provides an overview of the results for all experiments discussed so far. In the table, a green check mark indicates that the validation test passes in a setting where the annotation backtranslation and separation properties are satisfied, while the cross marks represent cases where validity is rejected. The cross mark is yellow if the prerequisite tests fail and red if the prerequisites are satisfied but the validation test fails. Overall, the test outcomes suggest that sentiment annotation is generally valid as long as the scale is simple and not too granular. Classifying the temporal focus and identifying topics is also often promising. However, scoring clarity even on a simple binary scale often fails. The results also suggest that test outcomes can vary significantly across applications: Annotations of passages from the earnings call transcript and the news article often pass the validation tests, whereas annotations of the Reddit comment are often rejected as valid. Due to this case-dependency, researchers should validate their applications before interpreting LLM annotations or using them in downstream estimation problems.

Comparing the LLM annotations with manual readings of the passages suggests that the test is conservative. There are cases of LLM labels that are sensible, but for which the test fails. For example, the LLM identifies the topic "technological developments" in the passage from the earnings call transcript, but the validation test of this annotation fails. Reading the passage, it is clear that technological developments is a theme in the passage. But, for all cases where the validation test passes, the LLM annotation seems appropriate for the passage.

Finally, the experiments highlight the importance of explicitly defining concepts and context in the simulation and annotation prompts. Specifically for the simulation prompt, the validation test is more likely to pass if the prompt includes details on the content and linguistic characteristics of the original passage. Providing such details is acceptable as long as the prerequisite properties are satisfied.

4.7 Validations Across Model Sizes and Generations

All experiments presented thus far are based on the Claude 4.5 Sonnet, which is a very large, state-of-the-art language model. This section repeats the sentiment experiments (on the simple and detailed scales) with different models varying in size, complexity, and release date. In addition to the baseline model, I consider the Claude 3 Haiku model and two Llama models of different sizes (70B and 8B parameters).[16] These models are considered large, medium, and small language models, respectively. They also represent a different model generation than the baseline model with release dates in March 2024 (Claude 3 Haiku), January 2024 (LLama 70B), and April 2024 (Llama 8B).

Table 8 shows the test results from validating annotations of sentiment for each model, experiment, and passage.[17] In panel (a), sentiment is scored on the simple scale (positive, neutral negative). Labels are identical across all models, but the smaller models do fail to validate the annotations more often, predominantly due to failing prerequisite tests. As such, the largest and most recent LLM (Claude 4.5 Sonnet) is more reliable than previously released and smaller models.

In panel (b) of Table 8, results are shown across models for sentiment scoring on the detailed scale (positive, mostly positive, neutral, mostly negative, and negative). On this scale, there is not full agreement on labels across models. For example, the 10-K filing passage is rated

[16] All models are implemented with a temperature of 1.0.

[17] Detailed test results are available upon request.

negative by the Claude 4.5 Sonnet and Llama 70B models, mostly negative by Claude 3 Haiku, and neutral by the Llama 8B model. Interestingly, the annotations of the outlier models fail the validation either due to failed prerequisites (Llama 8B) or failed validation test (Claude 2 Haiku). Validation of the annotation of the Reddit comment fails across all models, and the models disagree whether the passage is mostly negative (Claude 4.5 Sonnet) or neutral (all other models). These results emphasize the potential issues of interpreting LLM annotations of sentiment scored on a granular scale.

5 Full-Scale Application

While the experiments presented in the previous section are useful for illustrating the approach, relevant applications involve the task of annotating large set of passages rather than a few examples. This section illustrates how the validation test can be implemented in such settings.

5.1 Data

I apply the test to a well-known benchmark data set in language processing, namely the Financial Phrasebank data from Malo et al. (2014). This data set contains around 5000 sentences from financial newspaper articles written in English, for which the sentiment has been manually annotated on a simple positive-neutral-negative scale by the average of 5-8 human evaluators (mostly master's students with majors in finance, accounting, and economics). Choosing this data allows me to compare the validation testing framework with the traditional method of human benchmarking.

5.2 Performance Evaluation Metrics

To evaluate the accuracy of LLM annotations using the validation testing framework, I compute the fraction of sentences that passes the validation test along with the tests of prerequisite properties. This measure of accuracy can then be directly compared with the human benchmark accuracy computed as the fraction of sentences for which the LLM- and human-annotated labels coincide.

It is important to note that there it is no ground truth to the question of which method of evaluating accuracy is more correct. Deviations between conclusions obtained by the validation test and by comparing LLM annotations to human labels can therefore be evidence of the failure of either method, or potentially failing of both methods. Which conclusion to trust lies with the researcher's definition of truth: Is truth defined by the average of human labels, or by the consistency between LLM simulation and annotation?

5.3 Annotation and Simulation Functions

Similar to the experiments presented in Section 4, I use the Claude 4.5 Sonnet model[18] in both the annotation and simulation functions. The annotation and simulation prompts are described below.

Since the study is focused only on financial and economic domains, the human annotators were asked to consider the sentences from the view point of an investor only; i.e. whether the news may have positive, negative or neutral influence on the stock price. As a result, sentences which have a sentiment that is not relevant from an economic or financial perspective are considered neutral. For LLM validation, these details are provided in the simulation and annotation prompts.

The annotation prompt is as follows, excluding instructions on returning output in a JSON object:

```
TASK: Read the following sentence and classify the sentiment using one of the
↪ labels: [Positive, Neutral, Negative].
When classifying the sentence, determine sentiment from the view point of an
↪ investor only; i.e. whether the news may have positive, negative or neutral
↪ influence on the stock price. Sentences which have a sentiment that is not
↪ relevant from an economic or financial perspective are considered neutral.

SENTENCE: {text}
```

The experiments showed the importance of including characteristics of the original texts in the simulation prompt. Since the application involves running the validation test across two

[18] The model is implemented with a temperature of 1.0..

samples each consisting of 100 sentences, instructions has to be automated. I achieve this by including the original sentence as an input to the simulation prompt (`example`) as follows:

```
TASK: Write {n} arbitrary sentences from financial/economic news paper articles
↪ with sentiment characterized as {label}, from an investor’s point of view. {
↪ additional_instructions} The length of each sentence should be around {length}
↪ tokens, but make sure the sentence is complete. The sentences should be similar
↪  in terms of style and topic (not necessarily in terms of sentiment) to the
↪ following sentence: {example}.
```

The parameter `length` is the word count of `example`. The input `additional_instructions` takes a value only if the label is neutral to provide additional guidance on the definition of this label:

```
if label == ’neutral’:
        additional_instructions = "A neutral sentence is neutral in the sense that
        ↪ it not expected to have any impact on stock prices. A sentence not
        ↪ related to finance or economics is therefore considered neutral."
else:
        additional_instructions = None
```

The last sentence of the prompt instructs the model to generate texts are similar in style and topic to `example`. This inclusion increases semantic similarity with the original text, and thus the likelihood of passing the validation framework. Critically, the separation property test prevents the generated text from being excessively dependent on the original content by ensuring that texts generated from different labels are semantically distinct.

5.4 Results

Table 9 shows accuracy measures computed using the validation test as well as using the human benchmark. Note that these accuracy measures involve the same set of LLM annotations; they only differ by the way in which accuracy is evaluated. The validation testing framework suggests that the LLM annotations are accurate in 68% (low agreement sentences) to 82% of the times. This range is narrower than the accuracy suggested by human benchmarking which ranges between 65% and 92%. The fact that the validation testing framework is less accurate for the low-agreement data suggests that these sentences are more difficult cases. However, according to the validation testing framework, the difference between the two data sets is not as stark as suggested by human benchmarking.

The table also shows accuracy across the LLM-annotated labels. For both accuracy evaluation frameworks, the sentences labeled as positive by the LLM are the least accurate. For the validity test, the sentences labeled as negative and neutral are associated with similar accuracy. However, according to the human benchmarking method, accuracy is highest (and equal 100% regardless of the level of human agreement in the data sets) for the neutral sentences. This result likely reflect the tendency of humans to classify uncertain cases as neutral, which implies that there are no cases where the LLM classifues a sentence as neutral and human annotators do not.

Table 10 shows in detail how validation test accuracy distributes across all combinations of LLM and human labels. The table reports the number of sentences in each combination along with the associated accuracy in parentheses. While the LLM- and human-generated labels are identical for a majority of the sample, there are respectively seven and 29 sentences in the full-agreement and low-agreement data sets in which the LLM disagrees with the human annotation. These sentences are all labeled with neutral sentiment by humans, and a majority is labeled with positive sentiment by the LLM. These labels would be considered incorrect if relying on human benchmarking. However, the validity test passes for around half of the positive $\sim$neutral (LLM label$\sim$human label) sentences and for all of the negative$\sim$neutral sentences. To understand such cases better, Table 11 shows selected sentences where the LLM and human labels disagrees, but the validation test and its prerequisites pass. These examples emphasize the lack of a ground truth. Even though all or a majority of the human annotators labeled these sentences as neutral, most if not all of these sentences may impact the future stock prices of the firms involved, implying that they may not be neutral. The validation testing framework validates the LLM annotations that capture such nuances.

6 Limitations

Despite the promising results demonstrated by the proposed validation framework, several limitations warrant discussion.

A fundamental limitation of the approach is that validating an annotation function requires the specification of an associated simulation function. When insufficient details are provided in the simulation prompt, the test may falsely reject validity, which may explain the conservative nature of the test as observed in the experimental results.

The simulation prompt is central to how validity is defined within the framework: without a well-formulated function that specifies the data generating process in terms of a label, it becomes conceptually challenging to assess whether that label is correct. If AI is projected onto human intelligence, this requirement seems unfairly strict. For humans, recognition ability used for annotation and production ability used to generate texts are distinct skills, with recognition typically being much easier.[19] For example, literary critics can identify excellent prose without being novelists and most people can identify and enjoy comedy without being comedians. However, LLMs use the same neural network, parameters, and learned representations for both classification and generation. Unlike humans, there are not separate cognitive systems for recognition and production. If the model has learned representations that distinguish negative from positive sentiment, those representations should be available during generation. It is therefore reasonable to require that an LLM can simulate text based on a label to be considered a valid annotator. Drawing an analogy to traditional econometrics, this corresponds to an attempt to evaluate the properties of an ordinary least squares estimator of a coefficient β without assuming a model specification that relates β to the data.

It is, however, important to emphasize that due to the reliance on a simulation function, the framework cannot be used to validate all types of annotations because LLMs are constrained in certain areas in terms of what they will generate. For example, an LLM might correctly identify hate speech, as suggested by Huang et al. (2023) and Zhu et al. (2023), but refuse to generate it. In addition, a model heavily fine-tuned for classification might be worse at controlled generation,

[19] Although Richard Feynman famously remarked, *"what I cannot create, I do not understand."*

creating a gap between valid simulation and annotation that may lead to false rejection of validity.

Another limitation is the reliance on quantifying semantic similarity, implemented here through cosine similarity between vector representations. Consequently, the quality of the validation test is inherently bound by the underlying word embedding model. Specifically, the employed embedding model should be context-aware to correctly define terms that have different meanings in different contexts. Using the same embedding model for all steps of the validation framework mitigates the risk of making wrong conclusions based on wrong vector representations as an erroneous embedding model is unlikely to pass both the validation test and the test of the separation property.

For example, consider two passages of texts y_1 and y_2 given as follows:

> *y1: "Their aggressive depreciation strategy reduced taxable income substantially."*
>
> *y2: "The stock price declined after disappointing sales."*

The first passage has a positive sentiment (a successful business strategy), while the second passage is clearly negative. However, a simple embedding model, e.g., based on a word2vec algorithm using everyday English language, may not capture these contextual nuances. Instead, it might assign excessive weight to individual words that appear negative when taken out of context (such as "aggressive," "depreciation," "reduced," "taxable"). Such a model may therefore estimate a falsely high semantic similarity between $y1$ and $y2$, and thus pass the validation test. At the same time, the model would repeat similar mistakes in the test of the separation property, assigning falsely high similarity between distinct passages which would fail this prerequisite.

Finally, the framework incurs substantial computational costs, as it requires multiple LLM calls to generate and annotate multiple pasages of text. While this concern will likely diminish as models become more efficient and less expensive to run, it represents a current practical limitation. However, it is worth noting that compared to the human benchmarking approach, the computational cost remains negligible, offering a significant advantage in terms of objective validation and scalability.

7 Conclusion

This paper proposes a framework to assess the validity of LLM-generated measurements when reliable benchmarks are unavailable. The framework establishes validity based on whether an LLM can simulate texts from annotated labels that are semantically similar to the original passages, requiring that the annotation and simulation functions satisfy two key properties: annotation backtranslation and separation.

Through systematic experiments on diverse financial and economic texts, I demonstrate that the framework effectively distinguishes between reliable and unreliable LLM annotations. Simple sentiment classification, temporal focus identification, and topic detection generally pass validation, while more nuanced tasks like granular sentiment scoring and clarity assessment often fail. These results align with intuition about task difficulty and provide empirical guidance for practitioners deploying LLMs in research applications.

The application to the Financial Phrasebank dataset shows that the validation framework can serve as a practical alternative to human benchmarking. Importantly, the framework identifies cases where LLM annotations capture economic meaning that human evaluators may miss, particularly for neutral-labeled sentences that contain information relevant to stock prices.

While the approach requires careful specification of simulation prompts and incurs computational costs, it offers significant advantages in objectivity, scalability, and cost-effectiveness compared to human validation. As LLMs become increasingly central to empirical research in economics and finance, rigorous validation methods like the one proposed here are essential for ensuring the credibility and reproducibility of research findings. The framework provides researchers with a systematic tool to assess whether their specific combination of model and prompting strategy produces reliable measurements for downstream analysis.

References

M. Bauer, D. Huber, E. Offner, M. Renkel, and O. Wilms. Corporate Green Pledges. *SSRN Working Paper*, 2024.

C. Bertsch, I. Hull, R. L. Lumsdaine, and X. Zhang. Central bank mandates and monetary policy stances: Through the lens of Federal Reserve speeches. *Journal of Econometrics*, 249:105948, 2025.

Y. Chen, B. T. Kelly, and D. Xiu. Expected Returns and Large Language Models. *SSRN Working Paper*, 2022.

T. R. Cook, A. L. Hansen, S. Kazinnik, and P. McAdam. Under Pressure: Strategic Signaling in Bank Earnings Calls. *Available at SSRN 5382397*, 2025.

H. Du, R. Li, and E. Gehringer. Objective Metrics for Evaluating Large Language Models Using External Data Sources. 2025.

J. Fan, Q. Liu, Y. Song, and Z. Wang. Measuring Misinformation in Financial Markets. *Available at SSRN 4922648*, 2024.

A. L. Hansen and S. Kazinnik. Can ChatGPT Decipher Fedspeak? *SSRN Working Paper*, 2024.

F. Huang, H. Kwak, and J. An. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, page 294–297, April 2023.

M. Jha, J. Qian, M. Weber, and B. Yang. ChatGPT and Corporate Policies. *NBER Working paper*, 2024.

K. Kirtac and G. Germano. Sentiment Trading with Large Language Models. *Finance Research Letters*, 62:105227, April 2024. URL `http://dx.doi.org/10.1016/j.frl.2024.105227`.

X. Li, P. Yu, C. Zhou, T. Schick, O. Levy, L. Zettlemoyer, J. Weston, and M. Lewis. Self-Alignment with Instruction Backtranslation, 2024. URL `https://arxiv.org/abs/2308.06259`.

T. Liu and Y. Shi. News Sentiment and Investment Risk Management: Innovative Evidence From the Large Language Models. *Economics Letters*, 247:112124, 2025.

J. Ludwig, S. Mullainathan, and A. Rambachan. Large Language Models: An Applied Econometric Framework. *NBER Working Paper*, 33344, 2025.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

N. Mirzakhmedova, M. Gohsen, C. H. Chang, and B. Stein. Are Large Language Models Reliable Argument Quality Annotators?, 2024. URL `https://arxiv.org/abs/2404.09696`.

A. Patwardhan, V. Vaidya, and A. Kundu. Automated Consistency Analysis of LLMs, 2025. URL `https://arxiv.org/abs/2502.07036`.

M. V. Reiss. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark, 2023. URL `https://arxiv.org/abs/2304.11085`.

R. Sennrich, B. Haddow, and A. Birch. Improving Neural Machine Translation Models with Monolingual Data. In K. Erk and N. A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, Aug 2016. Association for Computational Linguistics.

A. Shah, A. Hiray, P. Shah, A. Banerjee, A. Singh, D. Eidnani, S. Chava, B. Chaudhury, and S. Chava. Numerical Claim Detection in Finance: A New Financial Dataset, Weak-Supervision Model, and Market Analysis, 2024. URL `https://arxiv.org/abs/2402.11728`.

A. H. Shapiro, M. Sudhof, and D. J. Wilson. Measuring News Sentiment. *Journal of Econometrics*, 228(2):221–243, 2022.

N. Sharma, N. Agarwal, and K. Sirts. Towards Consistent Detection of Cognitive Distortions: LLM-Based Annotation and Dataset-Agnostic Evaluation, 2025. URL `https://arxiv.org/abs/2511.01482`.

S. Singh, Y. Nan, A. Wang, D. D'Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. A. Smith, B. Ermis, M. Fadaee, and S. Hooker. The Leaderboard Illusion, 2025. URL `https://arxiv.org/abs/2504.20879`.

P. Törnberg. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning, 2023. URL `https://arxiv.org/abs/2304.06588`.

V. Veselovsky, M. H. Ribeiro, and R. West. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks, 2023. URL `https://arxiv.org/abs/2306.07899`.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models, 2023. URL `https://arxiv.org/abs/2203.11171`.

Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks, 2023. URL `https://arxiv.org/abs/2304.10145`.

Table 1: Validating Classification of Simple Sentiment

The table shows results from tests of the prerequisite properties (annotation backtranslation and separation) and the validation test of the ability of the Claude 4.5 Sonnet model to classify sentiment among the labels {Positive, Neutral, Negative}. The annotation backtranslation property rejection threshold is set as 90% accuracy. All other critical values are based on a 5% significance level. Test statistics and bootstrap procedures are implemented using 100 simulated trajectories.

(a) 10-K Filing

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Negative	100%	90%	Pass
Separation	Positive	0.577	0.600	Pass
	Neutral	0.596	0.621	Pass
Validation	Negative	0.623	0.620	Pass

(b) Earnings Call Transcript

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Positive	100%	90%	Pass
Separation	Neutral	0.626	0.647	Pass
	Negative	0.600	0.614	Pass
Validation	Positive	0.626	0.609	Pass

(c) News Article

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Negative	100%	90%	Pass
Separation	Positive	0.569	0.583	Pass
	Neutral	0.560	0.578	Pass
Validation	Negative	0.605	0.591	Pass

(d) Federal Reserve Governor Speech

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Negative	100%	90%	Pass
Separation	Positive	0.572	0.594	Pass
	Neutral	0.579	0.615	Pass
Validation	Negative	0.611	0.587	Pass

(e) Reddit Comment

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Negative	99%	90%	Pass
Separation	Positive	0.611	0.639	Pass
	Negative	0.612	0.647	Pass
Validation	Negative	0.580	0.575	Pass

Table 2: Validating Classification of Detailed Sentiment

The table shows results from tests of the prerequisite properties (annotation backtranslation and separation) and the validation test of the ability of the Claude 4.5 Sonnet model to classify sentiment among the labels {Positive, Mostly Positive, Neutral, Mostly Negative, Negative}. The annotation backtranslation property rejection threshold is set as 90% accuracy. All other critical values are based on a 5% significance level. Test statistics and bootstrap procedures are implemented using 100 simulated trajectories.

(a) 10-K Filing

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Negative	99%	90%	Pass
Separation	Positive	0.566	0.590	Pass
	Mostly Positive	0.568	0.596	Pass
	Neutral	0.604	0.640	Pass
	Mostly Negative	0.639	0.675	Pass
Validation	Negative	0.620	0.606	Pass

(b) Earnings Call Transcript

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Positive	100%	90%	Pass
Separation	Mostly Positive	0.639	0.662	Pass
	Neutral	0.622	0.643	Pass
	Mostly Negative	0.592	0.619	Pass
	Negative	0.587	0.618	Pass
Validation	Positive	0.630	0.613	Pass

(c) News Article

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Mostly Negative	7%	90%	Fail
Separation	Positive	0.558	0.576	Pass
	Mostly Positive	0.565	0.578	Pass
	Neutral	0.567	0.582	Pass
	Negative	0.621	0.639	Pass
Validation	Mostly Negative	0.605	0.582	Pass

Table continues on next page.

Table 2: Validating Classification of Detailed Sentiment *(continued)*

(d) Federal Reserve Governor Speech

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Negative	97%	90%	Pass
Separation	Positive	0.599	0.632	Pass
	Mostly Positive	0.602	0.624	Pass
	Neutral	0.599	0.633	Pass
	Mostly Negative	0.627	0.668	Pass
Validation	Negative	0.625	0.611	Pass

(e) Reddit Comment

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Mostly Negative	5%	90%	Fail
Separation	Positive	0.622	0.630	Pass
	Mostly Positive	0.605	0.631	Pass
	Neutral	0.591	0.623	Pass
	Negative	0.619	0.640	Pass
Validation	Mostly Negative	0.584	0.584	Fail

Table 3: Validating Classification of Clarity

The table shows results from tests of the prerequisite properties (annotation backtranslation and separation) and the validation test of the ability of the Claude 4.5 Sonnet model to classify textual clarity among the labels {Clear, Vague}. The annotation backtranslation property rejection threshold is set as 90% accuracy. All other critical values are based on a 5% significance level. Test statistics and bootstrap procedures are implemented using 100 simulated trajectories.

(a) 10-K Filing

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Vague	100%	90%	Pass
Separation	Clear	0.601	0.611	Pass
Validation	Vague	0.550	0.599	Fail

(b) Earnings Call Transcript

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Clear	100%	90%	Pass
Separation	Vague	0.597	0.612	Pass
Validation	Clear	0.599	0.598	Pass

(c) News Article

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Vague	100%	90%	Pass
Separation	Clear	0.543	0.569	Pass
Validation	Vague	0.582	0.574	Pass

(d) Federal Reserve Governor Speech

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Clear	100%	90%	Pass
Separation	Vague	0.602	0.623	Pass
Validation	Clear	0.584	0.589	Fail

(e) Reddit Comment

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Vague	100%	90%	Pass
Separation	Clear	0.584	0.601	Pass
Validation	Vague	0.566	0.579	Fail

Table 4: Validating Classification of Temporal Focus

The table shows results from tests of the prerequisite properties (annotation backtranslation and separation) and the validation test of the ability of the Claude 4.5 Sonnet model to classify the temporal focus among the labels {Backward-Looking, Focusing On The Present, Forward-Looking}. The annotation backtranslation property rejection threshold is set as 90% accuracy. All other critical values are based on a 5% significance level. Test statistics and bootstrap procedures are implemented using 100 simulated trajectories.

(a) 10-K Filing

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Forward-Looking	100%	90%	Pass
Separation	Focusing On The Present	0.605	0.632	Pass
	Backward-Looking	0.592	0.625	Pass
Validation	Forward-Looking	0.611	0.607	Pass

(b) Earnings Call Transcript

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Backward-Looking	100%	90%	Pass
Separation	Forward-Looking	0.605	0.640	Pass
	Focusing On The Present	0.618	0.640	Pass
Validation	Backward-Looking	0.610	0.609	Pass

(c) News Article

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Backward-Looking	100%	90%	Pass
Separation	Forward-Looking	0.555	0.572	Pass
	Focusing On The Present	0.557	0.582	Pass
Validation	Backward-Looking	0.622	0.587	Pass

(d) Federal Reserve Governor Speech

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Backward-Looking	100%	90%	Pass
Separation	Forward-Looking	0.587	0.617	Pass
	Focusing On The Present	0.592	0.616	Pass
Validation	Backward-Looking	0.592	0.581	Pass

(e) Reddit Comment

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	Forward-Looking	99%	90%	Pass
Separation	Focusing On The Present	0.590	0.601	Pass
	Backward-Looking	0.577	0.614	Pass
Validation	Forward-Looking	0.565	0.594	Fail

Table 5: Validating True Positive Identification of Topic

The table shows results from tests of the prerequisite properties (annotation backtranslation and separation) and the validation test of the ability of the Claude 4.5 Sonnet model to identify a specific topic within a passage. Topics are chosen such that they are present in the passages. Specifically, topics are "economic risks" for the 10-K filing passage, "technological developments" for the earnings call transcript passage, "crypto investing" for the news article passage, "financial crises" for the speech passage, and "interest rates" for the Reddit comment. Possible labels are {True,False}. The annotation backtranslation property rejection threshold is set as 90% accuracy. All other critical values are based on a 5% significance level. Test statistics and bootstrap procedures are implemented using 100 simulated trajectories.

(a) 10-K Filing

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	True	96%	90%	Pass
Separation	False	0.554	0.566	Pass
Validation	True	0.606	0.585	Pass

(b) Earnings Call Transcript

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	True	100%	90%	Pass
Separation	False	0.561	0.577	Pass
Validation	True	0.568	0.586	Fail

(c) News Article

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	True	94%	90%	Pass
Separation	False	0.530	0.545	Pass
Validation	True	0.575	0.558	Pass

(d) Federal Reserve Governor Speech

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	True	100%	90%	Pass
Separation	False	0.585	0.600	Pass
Validation	True	0.603	0.600	Pass

(e) Reddit Comment

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	True	100%	90%	Pass
Separation	False	0.539	0.557	Pass
Validation	True	0.583	0.572	Pass

Table 6: Validating True Negative Identification of Topic

The table shows results from tests of the prerequisite properties (annotation backtranslation and separation) and the validation test of the ability of the Claude 4.5 Sonnet model to identify a specific topic within a passage. Topics are chosen such that they are not present in the passages. Specifically, topics are "investments" for the 10-K filing passage, "geopolitical risks" for the earnings call transcript passage, "interest rates" for the news article passage, "new technologies" for the speech passage, and "stock market" for the Reddit comment. Possible labels are {True,False}. The annotation backtranslation property rejection threshold is set as 90% accuracy. All other critical values are based on a 5% significance level. Test statistics and bootstrap procedures are implemented using 100 simulated trajectories.

(a) 10-K Filing

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	True	100%	90%	Pass
Separation	False	0.554	0.589	Pass
Validation	True	0.545	0.590	Fail

(b) Earnings Call Transcript

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	False	100%	90%	Pass
Separation	True	0.544	0.578	Pass
Validation	False	0.619	0.586	Pass

(c) News Article

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	False	100%	90%	Pass
Separation	True	0.524	0.530	Pass
Validation	False	0.526	0.521	Pass

(d) Federal Reserve Governor Speech

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	False	98%	90%	Pass
Separation	True	0.573	0.585	Pass
Validation	False	0.591	0.585	Pass

(e) Reddit Comment

Test	Label	Statistic	Critical Value	Conclusion
Annotation Backtranslation	False	100%	90%	Pass
Separation	True	0.545	0.557	Pass
Validation	False	0.533	0.537	Fail

Table 7: Overview of Results

The table shows an overview of all validation results reported in detail in Tables 1-6. A green check mark represents cases where both the prerequisite tests (annotation backtranslation and separation) and the validation test pass. A red cross mark represents cases where the prerequisite tests pass, but the validation test fails. A yellow cross mark represents cases where validation fails because one or both prerequisite test fails. All tests conducted using the Claude 4.5 Sonnet model.

	Sentiment	Granular Sentiment	Clarity	Temporal Focus	True Positive Topic	True Negative Topic
10-K Filing	✓	✓	✗	✓	✓	✗
Earnings Call Transcript	✓	✓	✓	✓	✗	✓
News Article	✓	✗	✓	✓	✓	✓
Federal Reserve Governor Speech	✓	✓	✗	✓	✓	✓
Reddit Comment	✓	✗	✗	✗	✓	✗

Table 8: Results Overview Across Models

The table shows an overview of validation results for annotations of (a) simple sentiment ({Positive, Neutral, Negative}) and (b) detailed sentiment ({Positive, Mostly Positive, Neutral, Mostly Negative, Negative}). Results are reported across different large language models. A green check mark represents cases where both the prerequisite tests (annotation backtranslation and separation) and the validation test pass. A red cross mark represents cases where the prerequisite tests pass, but the validation test fails. A yellow cross mark represents cases where validation fails because one or both prerequisite test fails.

(a) Simple Sentiment

	Claude 4.5 Sonnet	Claude 3 Haiku	Llama 70B	Llama 8B
Model Size	Very Large	Large	Medium	Small
10-K Filing	Negative (✓)	Negative (✗)	Negative (✓)	Negative (✓)
Earnings Call Transcript	Positive (✓)	Positive (✓)	Positive (✓)	Positive (✓)
News Article	Negative (✓)	Negative (✓)	Negative (✓)	Negative (✓)
Federal Reserve Governor Speech	Negative (✓)	Negative (✓)	Negative (✗)	Negative (✗)
Reddit Comment	Negative (✓)	Neutral (✗)	Neutral (✗)	Neutral (✗)

(b) Detailed Sentiment

	Claude 4.5 Sonnet	Claude 3 Haiku	Llama 70B	Llama 8B
Model Size	Very Large	Large	Medium	Small
10-K Filing	Negative (✓)	Mostly Negative (✗)	Negative (✓)	Neutral (✗)
Earnings Call Transcript	Positive (✓)	Positive (✗)	Positive (✓)	Mostly Positive (✗)
News Article	Mostly Negative (✗)	Mostly Negative (✓)	Neutral (✗)	Mostly Negative (✓)
Federal Reserve Governor Speech	Negative (✓)	Mostly Negative (✗)	Negative (✗)	Mostly Negative (✗)
Reddit Comment	Mostly Negative (✗)	Neutral (✗)	Neutral (✗)	Neutral (✗)

Table 9: Accuracy of Sentiment Scoring of Financial Phrasebank Data

The table shows the accuracy of sentiment ({Positive, Neutral, Negative}) scores generated by the Claude 4.5 Sonnet model. Accuracy is determined using the LLM validation test and by comparing to the human benchmark provided in the Financial Phrasebank data set. Accuracy is computed for 100 sentences drawn randomly from the set of sentences with (a) 100% agreement among human annotations and (b) 50-66% agreement. Sentences for which the pre-requisite tests fail are excluded. The table shows both overall accuracy and the accuracy computed separately for each LLM-generated label.

(a) Sentences with 100% Human Agreement

	Accuracy Validity Test	Accuracy Human Benchmarking
Overall	82.14%	91.67%
By LLM annotation:		
Positive	65.52%	82.76%
Neutral	89.74%	100.00%
Negative	93.75%	87.50%

(b) Sentences with 50-66% Human Agreement

	Accuracy Validity Test	Accuracy Human Benchmarking
Overall	68.29%	64.63%
By LLM annotation:		
Positive	53.85%	50.00%
Neutral	100.00%	100.00%
Negative	87.50%	81.25%

Table 10: Confusion Matrix for Sentiment Scoring of Financial Phrasebank Data

The table shows the distribution of sentences across labels generated by humans (provided by the Financial Phrasebank data set) and an LLM (Claude 4.5 Sonnet model). Numbers in parentheses are the fraction of sentences that passes the validation test. The data set consists of 100 sentences drawn randomly from the set of sentences with (a) 100% agreement among human annotations and (b) 50-66% agreement. Sentences for which the pre-requisite tests fail are excluded.

(a) Sentences with 100% Human Agreement

Human annotation:	Positive	Neutral	Negative	Total
LLM annotation:				
Positive	24 (66.67%)	5 (60.00%)	0	29
Neutral	0	39 (89.74%)	0	39
Negative	0	2 (100.00%)	14 (92.86%)	16
Total	24	46	14	84

(b) Sentences with 50-66% Human Agreement

Human annotation:	Positive	Neutral	Negative	Total
LLM annotation:				
Positive	26 (57.69%)	26 (50.00%)	0	52
Neutral	0	14 (100.00%)	0	14
Negative	0	3 (100.00%)	13 (84.62%)	16
Total	26	43	13	82

Table 11: Selected Sentences from Financial Phrasebank Data

The table shows selected sentences from the Financial Phrasebank data. The selected sentences represent cases where the LLM and human label are different, but the validation test and its prerequisites pass.

LLM Label	Human Label	Source Data	Sentence
Positive	Neutral	100%	*"The machinery now ordered will be placed in a new mill with an annual production capacity of 40 000 m3 of overlaid birch plywood."*
Positive	Neutral	50-66%	*"In addition , YIT has reserved EPI Russia the right to expand the logistics center by about 100,000 m2 ."*
Negative	Neutral	100%	*"The total restructuring costs are expected to be about EUR 30mn , of which EUR 13.5 mn was booked in December 2008 ."*
Negative	Neutral	50-66%	*"Compared with the FTSE 100 index , which rose 51.5 points (or 0.9%) on the day, this was a relative price change of -0.6%."*