Financial Economics Research; Vol.3, No.4; 2026
E-ISSN: 3006-3663 P-ISSN: 3006-3655
DOI: https://doi.org/10.70267/fer.v3n4.7483
Published by: Zeus Press

Can Annual Report Tone Signal Audit Risk? Evidence from Large Language Model Textual Analysis of Chinese Listed Firms

Wen Guan and Yuxuan Zhou*

Nanjing Audit University, Nanjing 211815, China

**Corresponding author: Yuxuan Zhou.*

Abstract

Using a sample of Chinese A-share listed firms from 2021 to 2024, we employ a large language model (LLM) to measure the tone of approximately 20,000 Management Discussion and Analysis (MD&A) sections and examine whether narrative tone signals future audit risk. We document four findings. First, MD&A tone significantly and negatively predicts the probability of receiving a modified audit opinion in the following year; the effect survives controls for the current-year opinion, firm fundamentals, and industry and year fixed effects, and holds within the subsample of firms with clean current-year opinions. Second, the predictive power operates through forward-looking information embedded in tone: firms with more negative tone exhibit significantly lower profitability and a higher likelihood of losses in the subsequent year. Third, in a horse race between the LLM-based measure and a conventional dictionary-based measure built on an established Chinese financial sentiment lexicon, only the LLM tone retains predictive power, indicating that the audit-relevant component of tone resides in contextual meaning that word counts cannot capture. Fourth, an anonymization test addresses look-ahead bias: after stripping all firm identifiers from the text, re-scored tone remains highly correlated with the original measure (Pearson correlation 0.79) and preserves its predictive power, ruling out the concern that the model "remembers" firm outcomes from its training data. In contrast, tone explains neither audit fees nor audit report lag, suggesting that soft information shapes auditors' risk judgments rather than audit pricing. Our study provides a validation framework for LLM-based textual measures in accounting research and offers practical implications for using narrative disclosure in risk assessment.

Keywords

annual report tone, MD&A, audit opinion, large language models, textual analysis

1. Introduction

Beyond the financial statements, corporate annual reports contain extensive narrative disclosure. The Management Discussion and Analysis (MD&A) section, in which managers describe operations, risks, and prospects in their own words, is widely viewed as carrying incremental information that quantitative disclosures cannot fully convey [1-4]. How managers say things may matter as much as what they say: whether difficulties are downplayed or candidly acknowledged, and whether the outlook is stated with confidence or hedged with qualifications, may reveal managers' private expectations about the firm [5,6].

Auditors are among the most important professional users of annual report information. The formation of an audit opinion is fundamentally a risk-assessment process: the auditor must evaluate the risk of material misstatement and the entity's ability to continue as a going concern. If narrative tone indeed contains forward-looking information about a firm's condition, firms with more negative tone should experience more adverse developments in subsequent periods and should therefore be more likely to receive modified audit opinions. Testing this conjecture, however, has long been constrained by measurement: tone is soft information, and the standard measurement technology, counting words against a sentiment dictionary, cannot process context, negation, qualification, or euphemism [8]. The resulting measurement error may obscure true relations, particularly in Chinese, where disclosure language is formulaic and indirect.

Large language models offer a new measurement technology. Unlike bag-of-words approaches, an LLM reads the full text and evaluates meaning in context, in principle approximating the judgment of a skilled human reader [9,10]. We use the DeepSeek LLM to score the tone of roughly 20,000 MD&A sections of Chinese A-share listed firms over 2021-2024 and ask three questions. First, does MD&A tone predict future modified audit opinions? Second, what mechanism underlies any predictive power? Third, does the LLM measure add value relative to the conventional dictionary approach?

We find that MD&A tone negatively and robustly predicts next-year modified audit opinions. A one-unit increase in tone (about three standard deviations) is associated with a 1.2-percentage-point decline in the probability of a modified opinion, against an unconditional rate of 3.3%. The effect is present, and indeed stronger, among firms whose current-year opinion is clean, so it is not driven by the persistence of opinions among already-troubled firms. Consistent with a forward-looking-information mechanism, tone predicts next-year return on assets and the incidence of losses after controlling for current profitability. In a horse race, the dictionary-based tone measure has no predictive power for future opinions, whereas the LLM measure remains significant with both measures included; the audit-relevant component of tone is precisely the part that word counting cannot extract. An anonymization experiment that rescores 2,000 MD&As after removing all firm identifiers yields tone scores highly correlated with the originals that retain predictive power, addressing the look-ahead-bias concern specific to LLM-based research. Finally, tone explains neither audit fees nor audit report lag in our sample, a contrast suggesting that soft information enters auditors' risk judgments but not audit pricing in the Chinese market, where fees are sticky and dominated by client size.

This paper contributes to three strands of literature. First, we contribute to the emerging literature applying LLMs to accounting and finance texts [9,10,20] by proposing and implementing a two-part validation framework (a horse race against an established dictionary and an anonymization test for look-ahead bias) that we believe should become standard practice for LLM-based measures constructed over samples that may overlap with model training data. Second, we extend the literature on the economic consequences of disclosure tone [2,4,6] to auditor decisions in an emerging market, documenting that tone predicts audit opinions but not audit fees, a dissociation between risk judgment and pricing that prior single-outcome studies could not reveal. Third, our results are relevant for regulators and practitioners: narrative tone is a machine-readable leading indicator of audit risk that can be incorporated into supervisory screening at negligible cost.

The remainder of the paper proceeds as follows. Section 2 reviews the literature and develops hypotheses. Section 3 describes the sample, tone measurement, and research design. Section 4 presents the main results, mechanism tests, the horse race, and validity checks. Section 5 reports robustness tests. Section 6 concludes.

2. Literature Review and Hypothesis Development

2.1 The Information Content of Disclosure Tone

Loughran and McDonald [2] develop finance-specific English word lists and show that 10-K tone relates to returns and fundamentals, initiating a large literature on disclosure sentiment (see [8] for a survey). Li [1] documents that annual report readability relates to earnings persistence, Feldman et al. [11] find that changes in MD&A tone convey information incremental to accruals and earnings surprises, Davis et al. [3] show that optimistic earnings-press-release language predicts future performance, and Li [5] shows that forward-looking statements in filings carry information about future firm outcomes. For Chinese text, Yao et al. [4] construct a finance-domain sentiment lexicon via dictionary reorganization and deep learning, and show that annual report tone predicts stock returns and crash risk. Xie and Lin [12] find that managerial tone in earnings communication conferences predicts Chinese firms' next-year performance, and Zeng et al. [13] show that annual report tone predicts subsequent insider trading, evidence that tone reflects managers' private information, though it can also be strategically managed for impression purposes [6].

2.2 Tone and Auditing

Auditing standards require the auditor to assess the risk of material misstatement and to evaluate going-concern uncertainty; auditor effort, pricing, and reporting respond to client risk [7,14]. Narrative disclosure is a key input into the auditor's understanding of the entity. U.S.-based studies find that negative or uncertain MD&A language is associated with higher audit fees and longer audit delay [15], consistent with tone proxying for inherent risk that triggers greater audit effort and risk premia. In China's institutional setting, however, audit pricing is heavily influenced by fee stickiness and market competition [16], so fees may respond weakly to soft information. The audit opinion, by contrast, directly reflects the auditor's professional risk judgment: modified opinions in China are informative to the market and responsive to client-level risk factors such as earnings management incentives and regulatory scrutiny [17-19], making the opinion a sharper outcome for detecting whether auditors impound narrative information.

2.3 LLMs and Textual Measurement

The fundamental limitation of dictionary methods is the bag-of-words assumption: "risks have been effectively controlled" and "facing major risks" receive similar scores [8]. LLMs acquire contextual language understanding through pretraining and can assess the overall tenor of a document. Recent studies report that LLM-based measures of corporate disclosure outperform dictionary and other traditional approaches: sentiment from GPT-3.5-generated summaries of filings explains market reactions better than sentiment from the original text [9], ChatGPT-inferred news sentiment predicts stock returns [10], and GPT-based earnings predictions from anonymized financial statements rival trained human analysts [20]. LLM-based measurement, however, faces a distinctive identification threat, namely look-ahead bias: the model's training corpus may postdate the sample period, so scores for recognizable firms may embed the model's "memory" of subsequent outcomes rather than a genuine reading of the text [21,22]. Any credible design must address this concern directly.

2.4 Hypotheses

The preceding discussion yields three testable hypotheses:

H1: More negative MD&A tone predicts a higher probability of receiving a modified audit opinion in the following year.

H2: The predictive power of tone derives from forward-looking information about fundamentals: firms with more negative tone experience worse future performance.

H3: LLM-measured tone has incremental predictive power relative to dictionary-based tone.

3. Research Design

3.1 Sample and Data

Our initial sample comprises all A-share listed firms over 2021-2024. Annual report full texts (approximately 20,500 documents) are obtained from public disclosure channels (cninfo.com.cn); audit fees, audit opinions, auditor identities, financial data, ultimate-controller type, and Shenwan industry classifications are from the Wind database. We programmatically extract the MD&A section from each report (accommodating historical formats such as "Discussion and Analysis of Operations" and "Report of the Board of Directors"), achieving a 98.9% extraction rate and 20,287 MD&A texts with a mean length of about 59,000 Chinese characters.

Following convention, we exclude financial firms and observations with missing tone or key variables, yielding 18,479 firm-year observations. Continuous variables are winsorized at the 1st and 99th percentiles.

Tests of H1 require matching next-year audit opinions, which yields a prediction sample of 13,503 observations (fiscal years 2021-2023).

3.2 Tone Measurement

LLM tone (Tone_LLM). We use the DeepSeek LLM (deepseek-v4-flash). For each MD&A (first 30,000 characters, which concentrate the qualitative discussion), a uniform prompt instructs the model, acting as a senior financial analyst, to assess the overall tone of the managerial narrative (attending to how performance is characterized, how difficulties and risks are worded, and the confidence expressed about prospects) while explicitly disregarding the magnitude of reported financial figures. The model returns a continuous score from -1 (extremely pessimistic) to 1 (extremely optimistic). Temperature is set to zero for reproducibility.

Dictionary tone (Tone_Dict). We use the formal-text (annual report) sub-lexicon of the Chinese financial sentiment dictionary of Yao et al. [4] (3,592 positive and 1,633 negative words). After segmentation with jieba, we compute Tone_Dict = (Pos - Neg)/(Pos + Neg) for each MD&A.

3.3 Empirical Model

The baseline specification for H1 is:

*ModOpinion(i,t+1) = a + b*Tone(i,t) + c*ModOpinion(i,t) + d'Controls(i,t) + Year & Industry FE + e(i,t)*

ModOpinion equals 1 for any opinion other than a standard unqualified opinion (unqualified with emphasis-of-matter paragraphs, qualified, or disclaimer) and 0 otherwise. Controls include firm size (log of total assets), leverage, ROA, revenue growth, inventory intensity, receivables intensity, operating cash flow scaled by assets, a loss indicator, Big 4 auditor, state ownership (SOE equals one if the ultimate controller is a central or local government agency, a state-owned assets supervision commission, or a state-owned enterprise, and zero otherwise; a small number of collectively-owned and other-type firms are classified as non-SOE), and ST status. Controlling for the current-year opinion isolates tone's incremental predictive content. We estimate logit models in the main tests and linear probability models (LPM) with industry and year fixed effects in robustness checks. Standard errors are clustered at the firm level throughout.

4. Empirical Results

4.1 Descriptive Statistics

Table 1 presents descriptive statistics. Tone_LLM has a mean of 0.449 and a median of 0.600, indicating that annual report narratives are systematically optimistic, consistent with prior evidence; its standard deviation of 0.333 reflects ample cross-sectional variation. Tone_Dict has a mean of 0.794 and a much tighter distribution (standard deviation 0.075), a first indication of the dictionary measure's limited discriminatory power. Modified opinions occur in 3.3% of firm-years; 20.9% of observations report losses; 28.0% are state-owned; 6.3% are audited by Big 4 firms. The correlation between the two tone measures is 0.396, suggesting they capture related but distinct textual attributes.

Table 1: Descriptive statistics

	N	Mean	SD	Min	P25	Median	P75	Max
Tone_LLM	18479	0.449	0.333	-0.95	0.2	0.6	0.7	1
Tone_Dict	18479	0.794	0.075	0.56	0.751	0.804	0.849	0.93
ModOpinion	18479	0.033	0.178	0	0	0	0	1
Ln(AuditFee)	18479	13.955	0.618	11.775	13.528	13.864	14.286	18.11
AuditLag (days)	18479	106.938	13.785	18	99	113	117	148
Size	18479	22.316	1.3	17.879	21.397	22.07	23.013	28.791
Leverage	18479	0.41	0.208	0.053	0.242	0.398	0.557	0.933
ROA	18479	0.03	0.071	-0.228	0.006	0.033	0.067	0.22
Growth	18479	0.098	0.31	-0.565	-0.071	0.063	0.214	1.527
Inventory	18479	0.122	0.102	0	0.052	0.101	0.161	0.556
Receivables	18479	0.129	0.101	0.001	0.05	0.108	0.186	0.454
OCF	18479	0.045	0.067	-0.157	0.008	0.045	0.084	0.24
Loss	18479	0.209	0.406	0	0	0	0	1

	N	Mean	SD	Min	P25	Median	P75	Max
Big4	18479	0.063	0.242	0	0	0	0	1
SOE	18479	0.28	0.449	0	0	0	1	1
ST	18479	0.036	0.185	0	0	0	0	1

Notes: The sample comprises 18,479 firm-year observations of Chinese A-share listed firms over 2021-2024, excluding financial firms. Continuous variables are winsorized at the 1st and 99th percentiles. Variable definitions are provided in Section 3.

4.2 Tone Predicts Modified Audit Opinions (H1)

Table 2 reports the baseline results. In column (1), the logit coefficient on Tone_LLM is -0.618 (z = -3.15), significant at the 1% level: firms with more negative MD&A tone are more likely to receive a modified opinion the following year. Column (2) adds year and industry fixed effects, and column (3) switches to the LPM; the conclusion is unchanged (coefficients -0.705 and -0.019, both significant at 1%). Column (4) restricts the sample to firms with a clean current-year opinion, a pure "clean-to-modified" prediction test, and the coefficient strengthens to -0.820 (z = -3.47), indicating the result is not driven by opinion persistence among already-troubled firms.

In economic terms, the average marginal effect implied by column (1) is approximately -0.012: a one-unit increase in tone (about three standard deviations) lowers the probability of a next-year modified opinion by 1.2 percentage points, sizable against the 3.3% unconditional rate. Control variables behave as expected: leverage and ST status load positively; ROA, inventory intensity, and operating cash flow load negatively. H1 is supported.

Table 2: MD&A tone and next-year modified audit opinions

	(1) Logit	(2) Logit + FE	(3) LPM + FE	(4) Clean current opinion
Tone_LLM	-0.6184***	-0.7051***	-0.0187***	-0.8200***
Tone_LLM (t/z)	(-3.15)	(-3.54)	(-3.09)	(-3.47)
ModOpinion_t	3.2663***	3.2170***	0.4367***	
ModOpinion_t (t/z)	(17.69)	(17.24)	(18.09)	
Size	0.0383	0.0550	0.0004	0.0601
Size (t/z)	(0.62)	(0.90)	(0.31)	(0.84)
Leverage	0.9816***	1.1825***	0.0346***	2.0105***
Leverage (t/z)	(2.83)	(3.22)	(3.17)	(4.57)
ROA	-2.6971**	-2.3023*	-0.0869**	-4.2988***
ROA (t/z)	(-2.14)	(-1.79)	(-2.06)	(-2.65)
Growth	0.4365**	0.4363**	0.0139**	0.2559
Growth (t/z)	(2.23)	(2.12)	(2.29)	(1.09)
Inventory	-1.6466***	-2.2038***	-0.0621***	-2.1879***
Inventory (t/z)	(-2.69)	(-3.38)	(-4.34)	(-2.88)
Receivables	0.4328	0.5997	0.0123	0.6447
Receivables (t/z)	(0.74)	(0.92)	(0.77)	(0.86)
OCF	-2.3291**	-2.5254**	-0.0298	-3.1152**
OCF (t/z)	(-2.38)	(-2.47)	(-1.16)	(-2.49)
Loss	0.3277	0.3060	0.0063	0.1356
Loss (t/z)	(1.52)	(1.41)	(1.14)	(0.52)
Big4	-0.4325	-0.4957	-0.0055	-0.2858
Big4 (t/z)	(-1.23)	(-1.35)	(-1.33)	(-0.77)
SOE	-0.3472**	-0.3409**	-0.0089***	-0.6477***
SOE (t/z)	(-2.12)	(-2.00)	(-2.78)	(-3.29)
ST	2.3522***	2.3416***	0.2243***	2.6280***
ST (t/z)	(14.87)	(14.74)	(11.07)	(13.84)
Year & Industry FE	No	Yes	Yes	No
N	13503	13503	13503	13066
Adj. R2 / Pseudo R2	0.417	0.426	0.342	0.214

*Notes: The dependent variable is ModOpinion(t+1). Columns (1), (2), and (4) report logit estimates; column (3) reports the linear probability model. Column (4) restricts the sample to firms with a standard unqualified opinion in year t. t/z statistics based on firm-clustered standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.*

4.3 Mechanism: Forward-Looking Information in Tone (H2)

Why does tone predict opinions? Our explanation is that tone reflects managers' private information about the firm's true condition; that information subsequently materializes in fundamentals, to which the auditor's risk assessment responds. Table 3 provides direct evidence. Controlling for current ROA, Tone_LLM predicts next-year ROA with a coefficient of 0.013 (t = 7.37) and next-year loss incidence with a logit coefficient of -0.668 (z = -8.02), both significant at the 1% level. Optimistic-tone firms genuinely perform better in the following year; negative-tone firms are more likely to slide into losses. H2 is supported.

Table 3: Forward-looking information content of tone

	(1) ROA_t+1 (OLS)	(2) Loss_t+1 (Logit)
Tone_LLM	0.0134***	-0.6682***
Tone_LLM (t/z)	(7.37)	(-8.02)
ROA	0.5645***	-12.1438***
ROA (t/z)	(33.91)	(-12.68)
Loss	-0.0031	0.9970***
Loss (t/z)	(-1.57)	(10.19)
Year & Industry FE	Yes	No
N	13503	13503
Adj. R2 / Pseudo R2	0.463	0.252

*Notes: Column (1) reports OLS estimates with ROA(t+1) as the dependent variable; column (2) reports logit estimates with Loss(t+1) as the dependent variable. t/z statistics based on firm-clustered standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.*

4.4 Horse Race: LLM Tone Versus Dictionary Tone (H3)

Table 4 pits the two measures against each other. Used alone, Tone_Dict does not predict next-year modified opinions (coefficient -0.159, z = -0.19); used alone, Tone_LLM does. With both included, Tone_LLM remains significant at the 1% level (-0.707, z = -3.23) while Tone_Dict remains insignificant. The audit-relevant component of tone is therefore precisely the contextual, semantic component that word frequencies cannot capture; it is recoverable only by a reader, human or machine, that understands language in context. H3 is supported, and the finding shows that our results could not have been obtained with the conventional method.

Table 4: Horse race between LLM tone and dictionary tone

	(1) LLM only	(2) Dictionary only	(3) Both
Tone_LLM	-0.6184***		-0.7073***
Tone_LLM (t/z)	(-3.15)		(-3.23)
Tone_Dict		-0.1590	1.0119
Tone_Dict (t/z)		(-0.19)	(1.04)
Year & Industry FE	No	No	No
N	13503	13503	13503
Adj. R2 / Pseudo R2	0.417	0.414	0.417

*Notes: The dependent variable is ModOpinion(t+1); all columns report logit estimates with the full control set of Table 2. t/z statistics based on firm-clustered standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.*

4.5 Validity of the LLM Measure

4.5.1 Look-Ahead Bias: An Anonymization Test

The central identification concern for LLM-based measures is look-ahead bias: the model may have "seen" these firms' subsequent outcomes in its training corpus, so its scores could encode memory rather than reading [21,22]. To address this, we draw 2,000 firm-year observations (all 446 next-year-modified cases plus a random control group), strip each MD&A of all identifying information (company names, name variants, and security codes are replaced with generic placeholders), and have the model rescore the anonymized texts under the identical prompt. Table 5 shows that anonymized tone correlates 0.788 (Pearson; 0.787 Spearman) with the original scores, with nearly identical means and standard deviations. More importantly, within this subsample the anonymized tone retains significant predictive power for next-year modified opinions (coefficient -0.668, z = -2.44), close in direction and magnitude to the original measure (-0.855, z = -3.32).

The model identifies audit risk from the text alone, without knowing which firm it is reading. Look-ahead bias cannot explain our findings. The slightly larger coefficient on the original score admits the possibility that firm identity contributes a modest increment of information, but the predictive power of tone derives predominantly from the text itself.

Table 5: Anonymization test for look-ahead bias

Item	Value	
Matched observations	1996	
Pearson corr. (original vs. anonymized)	0.788	
Spearman corr.	0.787	
Original tone: mean / SD	0.398 / 0.370	
Anonymized tone: mean / SD	0.426 / 0.368	
	(1) Original tone	(2) Anonymized tone
Tone_LLM	-0.8548***	
Tone_LLM (t/z)	(-3.32)	
Tone_Anon		-0.6676**
Tone_Anon (t/z)		(-2.44)
Year & Industry FE	No	No
N	1996	1996
Adj. R2 / Pseudo R2	0.476	0.474

*Notes: The anonymization subsample comprises 1,996 firm-year observations (all next-year-modified cases plus a random control group). The lower panel reports logit estimates with ModOpinion(t+1) as the dependent variable and the full control set of Table 2. t/z statistics based on firm-clustered standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.*

4.5.2 Human Validation

We draw a stratified random sample of 100 MD&A texts (20 from each quintile of the LLM tone distribution, order randomized, all firm identifiers withheld) and have two trained coders independently and blindly classify each text as optimistic, neutral, or pessimistic under a common coding manual. We follow a two-stage protocol. In the first stage, independent coding yields 51% raw agreement (Cohen's kappa = 0.02), with disagreements concentrated on the neutral-optimistic boundary, an indication of how difficult tone judgment is for non-expert readers, and itself a motivation for machine-based measurement. In the second stage, the coding manual is refined to resolve the sources of disagreement (separating firm-authored narrative from quoted industry statistics; classifying "difficulty-response-achievement" framing), and the two coders jointly adjudicate each disputed case to a consensus label, with a handful of deadlocked cases resolved by a third adjudicator applying the manual. The final consensus labels comprise 57 optimistic and 43 neutral texts (no pessimistic labels, consistent with the well-documented positive skew of annual report narratives). Validation results are strong: mean LLM tone is 0.594 for texts labeled optimistic versus 0.355 for neutral ($t = 4.37$, $p < 0.001$), and the Spearman rank correlation between LLM scores and consensus labels is 0.442 ($p < 0.001$). The LLM's scores align significantly with vetted human judgment, providing direct evidence of measurement validity.

4.6 A Contrasting Result: Tone, Audit Fees, and Audit Report Lag

Prior U.S. studies typically use audit fees as the outcome [15]. Table 6 shows that in our sample the univariate relation between tone and log audit fees is significantly positive, but the coefficient collapses to zero once firm size and other controls are added (-0.003, $t = -0.19$). Tone is likewise unrelated to audit report lag ($t = -0.75$). Alternative specifications, including abnormal tone (the residual from regressing tone on fundamentals), first differences, firm fixed effects, and lagged tone, yield the same null (available on request). This null result is informative when juxtaposed with the opinion results: audit fees in China are sticky and dominated by hard client characteristics such as size, leaving little room for soft information in pricing; yet auditors' risk judgments, expressed in opinions, do impound tone. Soft information shapes what auditors conclude, not what they charge.

Table 6: Tone, audit fees, and audit report lag

	(1) Ln(AuditFee), FE only	(2) Ln(AuditFee), full controls	(3) AuditLag, full controls
Tone_LLM	0.1021***	-0.0025	-0.2959

Tone_LLM (t/z)	(5.55)	(-0.19)	(-0.75)
Year & Industry FE	Yes	Yes	Yes
N	18479	18479	18479
Adj. R2 / Pseudo R2	0.099	0.605	0.105

*Notes: The dependent variable is the natural logarithm of audit fees in columns (1)-(2) and audit report lag (days from fiscal year-end to the audit report date) in column (3). All columns include year and industry fixed effects. t statistics based on firm-clustered standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.*

4.7 Which Type of Modified Opinion Does Tone Predict?

A modified opinion in China spans two economically distinct categories: an unqualified opinion with an emphasis-of-matter paragraph, and a severe modification (qualified opinion, disclaimer, or adverse opinion). The two arise for different reasons. An emphasis-of-matter paragraph is the auditor's device for flagging going-concern doubt or significant uncertainty without qualifying the opinion, precisely the kind of soft, forward-looking risk that managerial tone would reveal. Severe modifications, by contrast, stem from concrete GAAP departures or scope limitations, that is, hard reporting failures with no intrinsic connection to how management narrates the year. We therefore expect tone to predict the former but not the latter.

Table 7 confirms this. Using next-year emphasis-of-matter (1.9% of firm-years) as the outcome, Tone_LLM is strongly predictive (coefficient -0.789, $z = -3.45$, $p < 0.001$); using next-year severe modifications (1.4% of firm-years), tone has no predictive power (coefficient -0.142, $z = -0.49$). The baseline result is thus driven entirely by emphasis-of-matter paragraphs. This decomposition sharpens the interpretation: tone predicts auditors' concern about uncertainty, not their detection of hard reporting failures, consistent with our thesis that tone conveys soft, forward-looking information about the firm's condition.

Table 7: Which type of modified opinion does tone predict?

	(1) Emphasis-of-matter	(2) Severe modification
Tone_LLM	-0.7890***	-0.1419
Tone_LLM (t/z)	(-3.45)	(-0.49)
ModOpinion_t	2.7008***	2.4406***
ModOpinion_t (t/z)	(10.31)	(9.86)
Year & Industry FE	No	No
N	13503	13503
Adj. R2 / Pseudo R2	0.329	0.326

*Notes: The dependent variable is an indicator for a next-year unqualified opinion with an emphasis-of-matter paragraph in column (1), and an indicator for a next-year severe modification (qualified, disclaimer, or adverse) in column (2). Both columns report logit estimates with the full control set of Table 2. t/z statistics based on firm-clustered standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.*

5. Robustness Tests

In addition to the checks reported above, our results are robust to : (i) excluding 2021 or any single year from the sample; (ii) replacing the continuous tone measure with an indicator for tone below the industry-year median; and (iii) augmenting the mechanism regressions with additional controls for size and leverage.

6. Conclusion

Using an LLM to measure the tone of roughly 20,000 MD&A sections of Chinese listed firms over 2021-2024, we show that narrative tone negatively predicts next-year modified audit opinions; that this effect is driven by emphasis-of-matter paragraphs, the auditor's signal of going-concern doubt and uncertainty, rather than by severe modifications rooted in hard reporting failures; that it operates through forward-looking information about fundamentals; that the predictive component of tone is invisible to a state-of-the-art sentiment dictionary; and that an anonymization experiment rules out look-ahead bias. Tone does not, in contrast, explain audit fees or report lag.

The implications are threefold. For researchers, LLMs provide a measurement technology for (Chinese) financial text that materially outperforms dictionaries, but LLM-based measures should be accompanied by validity checks, and we offer a replicable two-part framework (dictionary horse race plus anonymization test). For auditors and regulators, narrative tone is an inexpensive, machine-readable leading indicator of audit risk; firms whose tone is abnormally negative, or inconsistent with reported fundamentals, warrant incremental scrutiny, and exchanges could incorporate tone screening into inquiry-letter targeting. For investors, how management speaks contains information beyond what it reports, and is worth integrating into fundamental analysis.

Limitations suggest avenues for future work. Our sample period is short, constrained by text availability; scores depend on a specific model and prompt, and stability across models and prompts merits systematic study; and the method extends naturally to earnings-call transcripts, inquiry-letter responses, and other corporate texts.

References

[1] Li F. Annual report readability, current earnings, and earnings persistence[J]. Journal of Accounting and Economics, 2008, 45(2-3):221-247.

[2] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J]. Journal of Finance, 2011, 66(1):35-65.

[3] Davis A K, Piger J M, Sedor L M. Beyond the numbers: Measuring the information content of earnings press release language[J]. Contemporary Accounting Research, 2012, 29(3):845-868.

[4] Yao J, Feng X, Wang Z, et al. Tone, sentiment and market impacts: The construction of a Chinese sentiment dictionary in finance[J]. Journal of Management Sciences in China, 2021, 24(5):26-46. (In Chinese)

[5] Li F. The information content of forward-looking statements in corporate filings—A naive Bayesian machine learning approach[J]. Journal of Accounting Research, 2010, 48(5):1049-1102.

[6] Huang X, Teoh S H, Zhang Y. Tone management[J]. The Accounting Review, 2014, 89(3):1083-1113.

[7] DeFond M L, Zhang J. A review of archival auditing research[J]. Journal of Accounting and Economics, 2014, 58(2-3):275-326.

[8] Loughran T, McDonald B. Textual analysis in accounting and finance: A survey[J]. Journal of Accounting Research, 2016, 54(4):1187-1230.

[9] Kim A, Muhn M, Nikolaev V. Bloated disclosures: Can ChatGPT help investors process information?[R]. Working paper, University of Chicago, 2023.

[10] Lopez-Lira A, Tang Y. Can ChatGPT forecast stock price movements? Return predictability and large language models[R]. Working paper, University of Florida, 2023.

[11] Feldman R, Govindaraj S, Livnat J, et al. Management's tone change, post earnings announcement drift and accruals[J]. Review of Accounting Studies, 2010, 15(4):915-953.

[12] Xie D, Lin L. Can management tone predict firms' future performance? Evidence from the earnings communication conferences of Chinese listed companies[J]. Accounting Research, 2015(2):20-27. (In Chinese)

[13] Zeng Q, Zhou B, Zhang C, et al. Annual report tone and insider trading: Consistent words and deeds or saying one thing and doing another?[J]. Management World, 2018, 34(9):143-160. (In Chinese)

[14] Simunic D A. The pricing of audit services: Theory and evidence[J]. Journal of Accounting Research, 1980, 18(1):161-190.

[15] Abernathy J L, Guo F, Kubick T R, et al. Financial statement footnote readability and corporate audit outcomes[J]. Auditing: A Journal of Practice & Theory, 2019, 38(2):1-26.

[16] DeFond M L, Wong T J, Li S. The impact of improved auditor independence on audit market concentration in China[J]. Journal of Accounting and Economics, 2000, 28(3):269-305.

[17] Chen C J P, Su X, Zhao R. An emerging market's reaction to initial modified audit opinions: Evidence from the Shanghai Stock Exchange[J]. Contemporary Accounting Research, 2000, 17(3):429-455.

[18] Chen C J P, Chen S, Su X. Profitability regulation, earnings management, and modified audit opinions: Evidence from China[J]. Auditing: A Journal of Practice & Theory, 2001, 20(2):9-30.

[19] Hu N, Xu J, Xue S. Regulatory risk and auditors' reporting conservatism: Evidence from Chinese comment letters[J]. Journal of Accounting and Public Policy, 2022, 41:106997.

[20] Kim A, Muhn M, Nikolaev V. Financial statement analysis with large language models[R]. Working paper, University of Chicago, 2024.

[21] Glasserman P, Lin C. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis[R]. Working paper, Columbia University, 2023.

[22] Sarkar S K, Vafa K. Lookahead bias in pretrained language models[R]. Working paper, Harvard University, 2024.

Funding

This research was supported by the College Students' Innovation and Entrepreneurship Training Program of Nanjing Audit University [grant number: X2025112870050].

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.