



# From Similarity to Consequences: Decision-Oriented Evaluation of Market Digest Generation

Yu-Shiang Huang<sup>1,2</sup>, Chuan-Ju Wang<sup>1</sup>, Chung-Chi Chen<sup>3</sup>

F09946004@ntu.edu.tw, cjwang@citi.sinica.edu.tw, c.c.chen@acm.org

<sup>1</sup>Academia Sinica, Taiwan

<sup>2</sup>National Taiwan University, Taiwan,

<sup>3</sup>National Institute of Advanced Industrial Science and Technology, Japan

## TL; DR

- We introduce market digest generation as a practical, retail-investor-oriented NLG task, and benchmark performance-conditioned and professional-insight baselines.
- We show that traditional reference-based metrics fail to capture the real utility of financial text.
- To address this gap, we introduce a consequence-driven evaluation that measures how digests actually influence the trading decisions of both humans and LLM agents.
- Our results reveal that LLM-generated briefs can surpass human references in decision accuracy, while expert-curated asset selection further improves outcomes—highlighting the need for decision-oriented evaluation in high-stakes finance.

## Motivation Scenario



## Generation Task & Dataset Construction

We study market digest generation, focusing on two practitioner-relevant formats:

- Morning Briefs ( $M_t$ ) – summarize overnight global events.
- Closing-Bell Reports ( $C_t$ ) – recap the day's movements using intraday statistics.

Each digest is generated per trading day  $t$ , conditioned on structured market data + news.

We collect 30 days of aligned financial media transcripts:

- News articles:  $N_t$  (~2,400/day)
- Market data:  $P_t$  (price time series)
- Listed entities:  $E_t$
- Supplementary stats:  $S_t$  (e.g., institutional flows, volumes)
- Professional announcer transcripts → human references
  - $M_t^{\text{human}}$ : Morning Calls
  - $C_t^{\text{human}}$ : Closing-bell Reports

## Decision-Oriented Evaluation

Table 1: Automatic Evaluation Results

\*SentM. (Sentiment Match) is the proportion of cases where human and generated commentaries receive the same sentiment label (positive, negative, neutral) from an LLM-based classifier.

Scenario	Method	Len.	ROUGE-Lsum	BERTScore-F1	SentM.
Morning	$M_t^{\text{perf}}$	478.74	0.0479	0.6365	0.4831
	$M_t^{\text{pro}}$	12013.16	0.1790	0.6353	0.3820
Closing	$C_t^{\text{perf}}$	527.20	0.0975	0.6432	0.6067
	$C_t^{M_{\text{human}}}$	546.13	0.1086	0.6462	0.6404
	$C_t^{\text{pro}}$	8742.03	0.1396	0.6346	0.5169

Table 2: Decision-oriented Evaluation Results

Scenario	Method	LLM Investor			Human Investor			Average
		Claude	Gemini	GPT-4o	A	B	C	
Morning	$M_t^{\text{human}}$	38.85	44.89	42.35	37.97	36.67	34.40	39.19
	$M_t^{\text{perf}}$	45.98	<b>46.98</b>	42.53	42.92	<b>45.11</b>	<b>49.27</b>	<b>45.47</b>
	$M_t^{\text{pro}}$	<b>48.01</b>	42.41	<b>43.15</b>	<b>46.28</b>	40.23	48.35	44.74
Closing	$C_t^{\text{human}}$	<b>65.56</b>	<b>61.60</b>	<b>58.51</b>	48.13	50.83	42.24	54.48
	$C_t^{\text{perf}}$	49.42	49.57	43.29	51.32	45.11	65.06	50.63
	$C_t^{M_{\text{human}}}$	55.62	54.36	56.89	48.44	49.44	<b>75.00</b>	<b>56.63</b>
	$C_t^{\text{pro}}$	60.07	58.25	55.17	<b>54.05</b>	<b>53.45</b>	54.18	55.86
Investor's Average		51.93	51.15	48.84	47.02	45.84	52.64	

### Key Observations

- Reference-based metrics miss the decision value of digests.
- LLM texts may diverge from journalism yet offer clearer trading signals. → Evaluate digests by decision impact, not textual similarity.
- LLM-generated morning briefs improve decision accuracy for both humans and LLM investors.
- Closing-bell reports show a human-LLM divergence: LLMs perform best with journalist texts, humans with LLM texts.
- Expert-guided asset selection boosts performance, though gains diminish for the most capable investors.
- Human decisions exhibit high variability, limiting reproducibility.
- LLM investors provide stable baselines for decision-oriented evaluation.
- Digest value lies in decision utility, not resemblance to human references.

## Generation and Evaluation Pipeline

