Reliable Decision Support with LLMs

A Framework for Evaluating Consistency in Binary Text Classification Applications

Fadel M. Megahed, Ying-Ju Chen, L. Allison Jones-Farmer,
Gabe Lee, Jiawei Brooke Wang, Inez M. Zwetsloot

Miami University | University of Dayton | University of Amsterdam
arXiv: 2505.14918 | December 2025

Presented by Ying-Ju (Tessa) Chen

Reliable Decision Support with LLMs: A Framework for Evaluating Consistency in Binary Text Classification Applications

Fadel M. Megahed[1], Ying-Ju Chen[2], L. Allison Jones-Farmer[1], Gabe Lee[1,*], Jiawei Brooke Wang[3], and Inez M. Zwetsloot[3,*]

[1]Farmer School of Business, Miami University, Oxford, OH 45056, USA
[2]College of Arts and Sciences, University of Dayton, Dayton, OH 45469, USA
[3]Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands
*Corresponding authors. Emails: gabelee@miamioh.edu, i.m.zwetsloot@uva.nl

December 23, 2025

Abstract

This study introduces a framework for evaluating consistency in large language model (LLM) binary text classification, addressing the lack of established reliability assessment methods. Adapting psychometric principles, we determine sample size requirements, develop metrics for invalid responses, and evaluate intra- and inter-rater reliability. Our case study examines financial news sentiment classification across 14 LLMs (including `claude-3-7-sonnet`, `gpt-4o`, `deepseek-r1`, `gemma3`, `llama3.2`, `phi4`, and `command-r-plus`), with five replicates per model on 1,350 articles. Models demonstrated high intra-rater consistency, achieving perfect agreement on 90–98% of examples, with minimal differences between expensive and economical models from the same families. When validated against *StockNewsAPI* labels, models achieved strong performance (accuracy 0.76–0.88), with smaller models like `gemma3:1B`, `llama3.2:3B`, and `claude-3-5-haiku` outperforming larger counterparts. All models performed at chance when predicting actual market movements, indicating task constraints rather than model limitations. Our framework provides systematic guidance for LLM selection, sample size planning, and reliability assessment, enabling organizations to optimize resources for classification tasks.

Keywords: Financial text analysis; inter-rater reliability; intra-rater reliability; sentiment analysis; text annotation; text classification; transformer model

1

Scan for paper

Presentation Outline

1. **Motivation & Contributions** — ~6 min
2. **Background** — ~5 min
3. **The Proposed Framework** — ~8 min
4. **Case Study: Stock News** — ~8 min
5. **Results & Key Findings** — ~8 min
6. **Discussion & Implications** — ~5 min

Time Allocation

+ ~5 min Q&A

Motivation

Why do we need a framework for LLM-based classification?

The Problem: Ad-Hoc LLM Classification

LLMs are increasingly used as text classifiers in business and social science.

BUT: Most studies lack systematic evaluation of classification quality.

FOMO-driven model selection: researchers default to the most expensive models without evidence.

?	!	≈
No Standard Protocols	Inconsistent Reporting	Unknown Reliability

Paper Contributions

1

Psychometric-Grounded Framework

Four-phase methodology adapting classical reliability and validity theory to LLM classifiers.

2

Comprehensive Empirical Validation

14 LLMs, 7 providers, 1,350 articles, 5 replicates → 94,500 total classifications.

3

Challenging "Bigger is Better"

Smaller, cheaper models match or outperform expensive ones on binary classification.

4

Dual Validity Assessment

High benchmark accuracy (88%) does NOT guarantee real-world predictive utility (~50%).

Background

LLMs as classifiers and measurement theory

LLMs as Text Classifiers

Traditional classification relies on trained human coders or supervised ML — expensive, slow, task-specific.

LLMs offer a zero-shot alternative: classify text without task-specific training.

The explosion of available models creates a practical challenge: how do you choose?

Proprietary Models

Higher API cost, closed-source
GPT-4o, Claude, Command-R+
Often assumed "better"

Open/Local Models

Low or no cost, run locally
Gemma, Llama, DeepSeek, Phi
Often overlooked

Classical Reliability Concepts

Intra-Rater Reliability

Same rater, same data, different occasions

For LLMs: Does the model give the same label across repeated runs?

Inter-Rater Reliability

Different raters, same data

For LLMs: Do different models agree on the same texts?

Metrics: Gwet's AC1, Brennan-Prediger, Conger's κ, Fleiss' κ, Krippendorff's α

LLMs at temperature 0 are quasi-deterministic — variation arises from floating-point non-determinism.

Validity Concepts

Reliability alone is insufficient — annotations must also be valid.

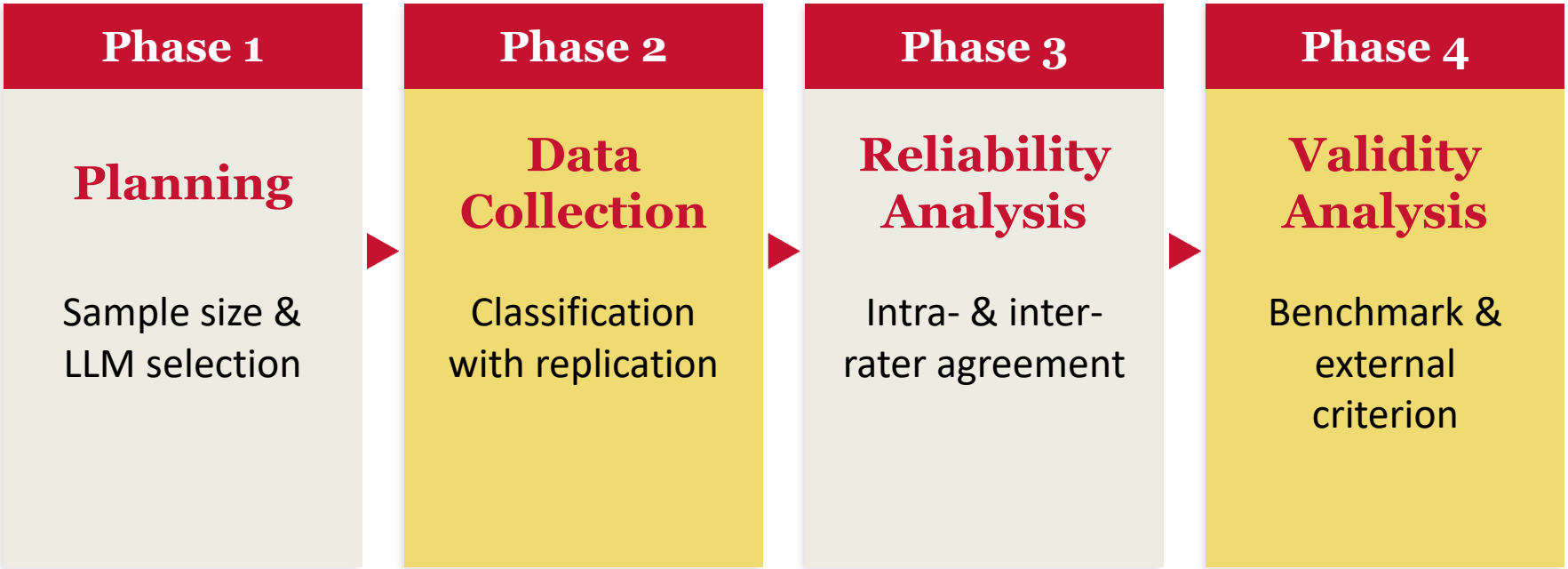
Benchmark Validity	Compare LLM labels against a known ground truth. Metrics: accuracy, sensitivity, specificity, F1.
External Criterion Validity	Do the LLM labels predict real-world outcomes? E.g., does sentiment predict actual stock movements?

Key insight: High benchmark accuracy does NOT guarantee external validity

The Proposed Framework

A four-phase methodology for reliable LLM classification

Four-Phase Framework Overview



Each phase builds on the previous — together they provide a comprehensive quality assurance pipeline

Phase 1: Planning — Sample Size

How many texts must an LLM classify before reliability estimates stabilize?

Gwet's psychometric approach (adapted for sequential sampling):

$n_o = z^2\alpha \cdot C / E^2_o$ where C = 1/A (from agreement coefficient tables)

Target precision: $E_o = 0.10$, confidence = 90%

Šidák correction applied for multiple comparisons

Conservative: use largest n across all metrics → n = 1,350

Sample Sizes by Metric

Gwet's AC1: n = 216

Percent Agreement: n = 847

Brennan-Prediger: n = 1,317

→ Final: n = 1,350 (conservative)

Phase 1: Planning — LLM Selection

Pilot multiple LLMs across cost tiers to avoid provider-specific biases.

Anthropic	OpenAI	Microsoft	DeepSeek
claude-3-7-sonnet claude-3-5-haiku	gpt-4o gpt-4o-mini	phi4:latest phi4-mini	deepseek-r1:7B deepseek-r1:1.5B

Google	Meta	Cohere
gemma3:27B gemma3:1B	llama3.2:3B llama3.2:1B	command-r-plus command-r7b

14 LLMs from 7 providers — both proprietary APIs and local open-source models via Ollama

Phase 2: Data Collection Design

K = 5 Replicates	Each text classified 5 times per model for intra-rater analysis
Temperature = 0	Minimizes randomness; residual variation from floating-point arithmetic
Manual CoT Prompt	Same prompt with 2 examples across all 14 models (2-shot CoT)
Invalid Tracking	Track when LLMs produce labels outside defined categories

Total: 1,350 articles × 14 LLMs × 5 replicates = 94,500 classifications

Phase 3: Reliability Analysis

Intra-Rater

Pairwise comparison of K=5 replicates per model

Five coefficients computed: Gwet's AC1, Brennan-Prediger, Conger's κ, Fleiss' κ, Krippendorff's α

High self-agreement = consistent decision boundary

Inter-Rater

Agreement across different LLMs

Same five metrics, computed across models

Analyses: all 14, top-N subsets, cost-tier groupings

Both forms needed — a model can be self-consistent but systematically wrong

Case Study Results

Financial news sentiment classification

Case Study: Financial News Classification

Dataset

StockNewsAPI: 90 tickers

~10,000 articles collected

Downsampled to 1,350 (675 pos, 675 neg)

Binary: Positive vs. Negative

Linked to actual stock price movements

Design

14 LLMs from 7 providers

5 replicates per article per model

Total: 94,500 classifications

Temperature = 0, manual CoT (2-shot)

Local via Ollama; proprietary via APIs

Majority Vote: Final label = mode across K=5 runs. Invalid labels in minority replicates are naturally outvoted.

What Are We Evaluating?

Three Evaluation Questions

Consistency	Does the same *LLM* produce the same label across repeated runs?
Agreement	Do different *LLMs* agree when classifying the same text?
Validity	Do the classifications align with a bench sentiment label Or real-world outcomes (stock returns)?

A reliable model should be **consistent, agree with others, and produce meaningful predictions.**

Results: Intra-Rater Reliability (Top 7)

Finding: 12 of 14 models achieve >88% perfect agreement (NA-penalized) across 5 replicates.

Model	% Perfect Agree (NA-penalized)	BP Coeff.
gemma3:27B	98.4%	0.99
phi4:latest	97.2%	0.98
claude-3-5-haiku	97.2%	0.98
claude-3-7-sonnet	96.1%	0.96
llama3.2:1B	95.9%	0.97
command-r7b	95.4%	0.98
gpt-4o-mini	93.9%	0.95

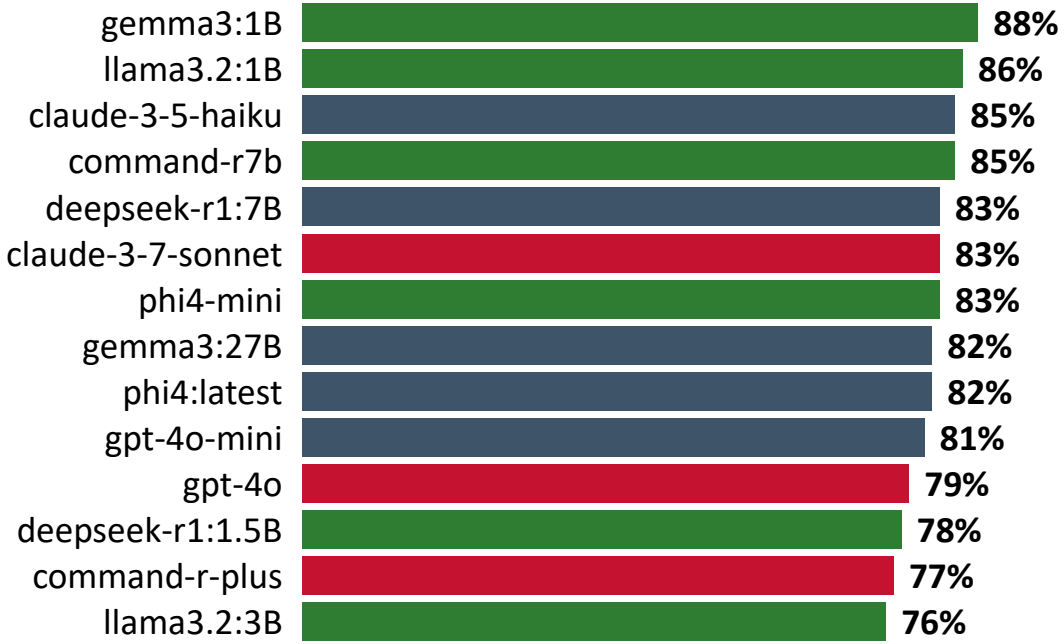
Results: Intra-Rater Reliability (Bottom 7)

Model	% Perfect Agree (NA-penalized)	BP Coeff.
gemma3:1B	93.6%	0.98
gpt-4o	92.3%	0.93
command-r-plus	89.9%	0.92
llama3.2:3B	89.4%	0.99
phi4-mini	88.1%	0.97
deepseek-r1:7B	86.0%	0.93
deepseek-r1:1.5B	**76.2%**	**0.85**

Key: NA-penalized agreement counts articles with invalid responses as non-agreeing. BP coefficients (computed from valid pairwise comparisons) remain high (≥0.92) for all except deepseek-r1:1.5B.

Results: Benchmark Validity

Finding: Accuracy ranges 76–88% vs. StockNewsAPI labels. Smaller models often outperform.



Key Finding

- Small/local
- Mid-tier
- Large/costly

gemma3:1B (smallest!) achieves highest accuracy at 88%

Smaller models outperform in 6 of 7 provider families

Challenges the FOMO assumption

Challenging "Bigger is Better"

Model size and cost do NOT predict classification quality.

Google	**Smaller: gemma3:1B (88%)**	Larger: gemma3:27B (82%)	←
Meta	**Smaller: llama3.2:1B (86%)**	Larger: llama3.2:3B (76%)	←
Cohere	**Smaller: command-r7b (85%)**	Larger: command-r-plus (77%)	←
Anthropic	**Smaller: claude-3-5-haiku (85%)**	Larger: claude-3-7-sonnet (83%)	←
Microsoft	**Smaller: phi4-mini (83%)**	Larger: phi4:latest (82%)	←
OpenAI	**Smaller: gpt-4o-mini (81%)**	Larger: gpt-4o (79%)	←
DeepSeek	Smaller: deepseek-r1:1.5B (78%)	Larger: deepseek-r1:7B (83%)	→

Smaller model wins in 6 of 7 provider families. DeepSeek is the sole exception.

Results: External Criterion Validity

Finding: ALL models perform near chance (~50%) at predicting actual stock price movements.

Market Prediction

Accuracy vs. stock movements: ~49–52%

F1 scores: ~0.49–0.54

No model outperforms a coin flip

Interpretation

Not a model failure. Reflects:

EMH: sentiment already priced in

Stock prediction ≠ sentiment classification

Validates external criterion testing

88% benchmark accuracy ≠ 88% real-world predictive power. Always test external validity.

Discussion

Implications for research and practice

Theoretical Contributions

1. **Bridging Psychometrics and LLMs**
   First framework adapting classical reliability metrics to LLM classification.

2. **Reconceptualizing LLM Reliability**
   Variation at temp=0 from floating-point arithmetic; thresholds need reinterpretation.

3. **Dual Validity Framework**
   Benchmark accuracy and external validity can diverge dramatically (88% vs. ~50%).

4. **Evidence Against "Bigger is Better"**
   Model size/cost does not predict binary classification performance.

Practical Recommendations

1. **Pilot Multiple Models**
   Test 2–3 models from different providers and cost tiers.
2. **Use Multiple Replicates**
   Classify each text K=3–5 times; aggregate via majority vote.
3. **Report Reliability Metrics**
   Include agreement coefficients, not just accuracy.
4. **Test External Validity**
   Beyond benchmarks, test if labels predict real-world outcomes.
5. **Track Invalid Labels**
   Monitor out-of-category responses as a data quality indicator.

Limitations & Future Directions

Limitations

- Binary classification only
- Single domain (financial news)
- Single prompt design tested
- Temperature fixed at 0
- Models tested at a snapshot in time

Future Directions

- Multi-class and multi-domain
- Alternative prompting strategies
- Multilingual evaluation
- Longitudinal tracking
- Open-source R/Python toolkit

Conclusion

- LLM-based text classification needs rigorous quality assurance — just like human annotation.
- The four-phase framework provides systematic guidance for reliable LLM classification.
- Most LLMs achieve high intra-rater reliability (88–98% agreement) at temperature 0.
- Smaller, cheaper models matched or outperformed larger alternatives in 6 of 7 families.
- High benchmark accuracy (88%) does NOT translate to real-world predictive utility (~50%).
- Always test external validity and report reliability metrics — accuracy alone is insufficient.

Key Takeways

- Reliability can be systematically evaluated
  Our framework adapts **psychometric reliability theory** to LLM classification.

- **Bigger models are not necessarily better**
  Smaller, cheaper models often match or outperform expensive flagship models.

- **Reliability ≠ Real-world usefulness**
  High agreement with benchmark labels does **not guarantee predictive validity**.

The main contribution is not choosing the best LLM —
but providing a rigorous framework for evaluating LLM classifiers.

Thank You!

Questions & Discussion

Paper: arXiv 2505.14918 (Dec 2025)

Megahed, Chen, Jones-Farmer, Lee, Wang, Zwetsloot

Miami University | U. of Dayton | U. of Amsterdam

Presented by: Ying-Ju (Tessa) Chen

Scan to read the full paper

arxiv.org/pdf/2505.14918

AI Acknowledgment:
An initial draft of these presentation slides was generated using Claude.ai. The authors subsequently reviewed, revised, and verified all content.