From Tasks to Communities: Growing a Research Community through NTCIR

陳重吉 (Chung-Chi CHEN)

Human-Agent Ally Lab (HAA Lab)

National Institute of Informatics, Japan

HAA LAB

haalab.github.io

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

NTCIR

Slides

Outline

- **Stories – From Tasks to Communities**
  - Growing a Research Community through NTCIR
- **Research – History Repeats Itself**
  - **Benchmark**: What Agents Can Replace?
  - From IR to NLP: The Return of Subjectivity in **Evaluation**
  - From Static Metrics to **Verifiable** Outcomes

KEYNOTE 1

Nancy F. Chen

The Long Arc of Language Resources: From Annotation to Alignment to Grounding

Language resources are the backbone of AI: they train models, structure linguistic analysis, and benchmark technological progress. Their evolution mirrors—and actively shapes—the trajectory of computational linguistics, speech technology, natural language processing, and artificial intelligence.

This talk traces the long arc of language resources across successive eras—from curated linguistic annotations to large-scale datasets enabling statistical learning, to representation learning and multimodal pretraining, and to alignment, where data shapes not only what models learn but also how they behave in society. **Building on this arc, we argue that the next phase is grounding: anchoring language technologies to perception, interaction, cultural and social contexts, and domain knowledge.**

KEYNOTE 2

Dan Jurafsky

The Social Failures of Language Models as Conversational Partners

Language models are increasingly used in conversation for information, advice, and emotional support. In this talk I'll summarize studies in our lab showing that models fail in systematic ways as social interlocutors. We find that language models are socially sycophantic, linguistically overconfident, overly anthropomorphic, and epistemically self-centered. We then show that these flaws have real consequences for users: people interacting with models suffer consequences including overreliance, distorted judgment, and reduced personal responsibility. I'll discuss datasets and metrics, explore mitigations, and call for design, evaluation, and accountability mechanisms to protect user well-being.

Human-Agent Ally Lab

HAA LAB

Agent Design / Agentic Framework Design

Information Retrieval for Agents

Investigating how IR techniques can enhance agent capabilities, focusing on retrieval-augmented generation and knowledge grounding.

LLM-based Agent Architectures

Designing and optimizing agentic frameworks powered by large language models for improved reasoning and task execution.

Performance & Efficiency

Benchmarking and enhancing the performance of IR-LLM integrated systems in real-world agent applications.

Human-Agent Teaming

Higher-Order Thinking Augmentation

Enhancing human cognitive capabilities through intelligent agent collaboration, focusing on complex problem-solving and creative thinking.

High-Stakes Decision-Making

Developing frameworks for human-agent collaboration in critical scenarios where decisions have significant consequences.

High-Fidelity Interaction

Creating seamless, intuitive interfaces that enable natural and effective communication between humans and agents.

Human-Agent Society

Societal Transformation

Investigating how the integration of agents into human society reshapes social structures, norms, and relationships.

Governance & Policy Design

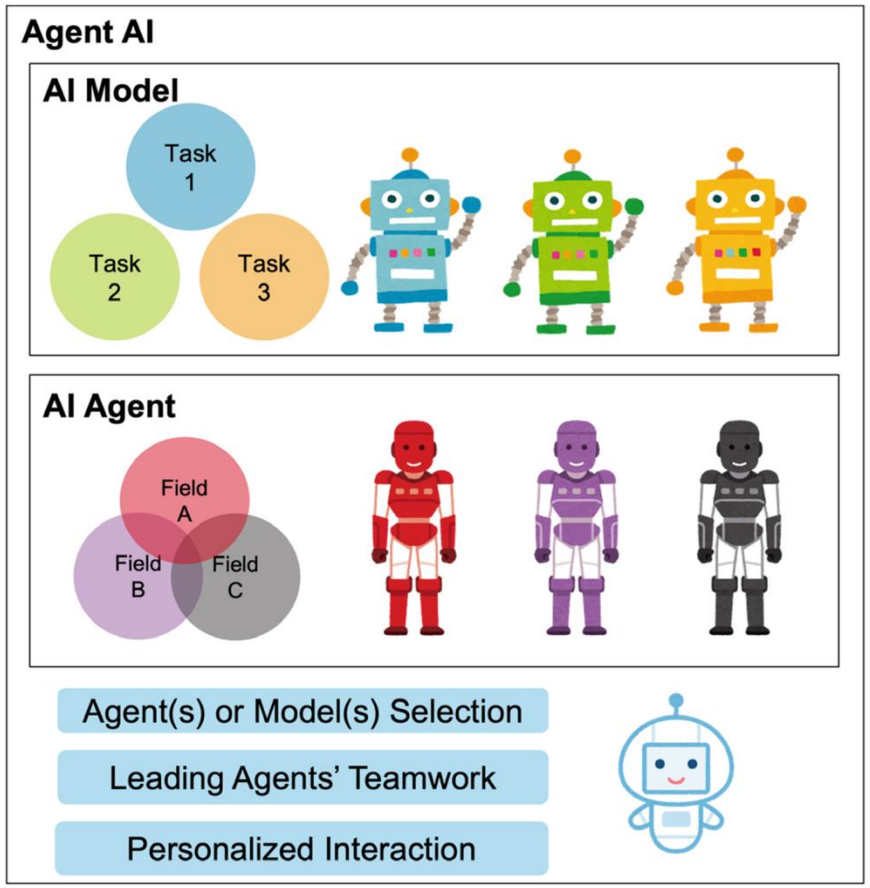
Developing governance, policy, and regulatory approaches for responsible agent deployment at societal scale.

Impact Analysis

Analyzing the economic, cultural, and ethical implications of widespread agent adoption in everyday life.

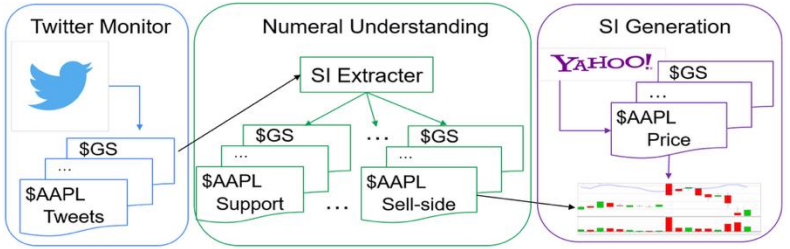
Model as Tool (Before) vs. Agent as Partner (Now & Future)

HAA LAB

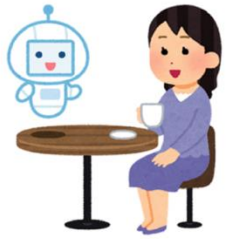


From Data to Signals
(Information Extraction)

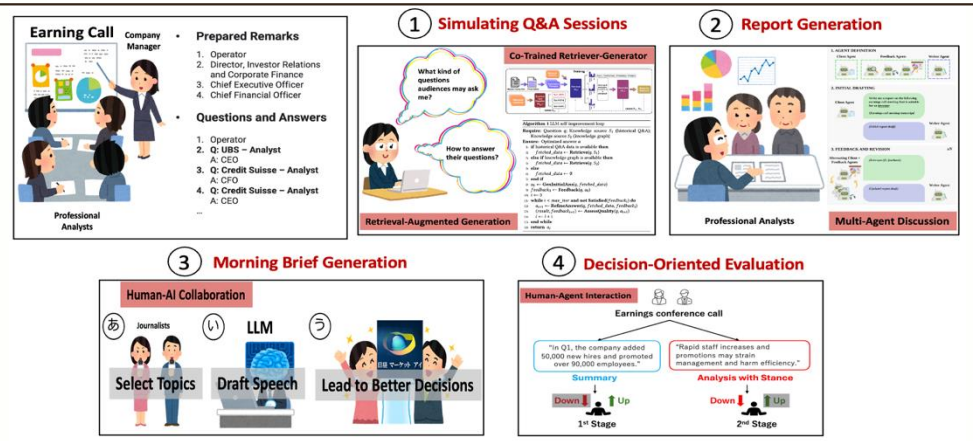
AI as Tool



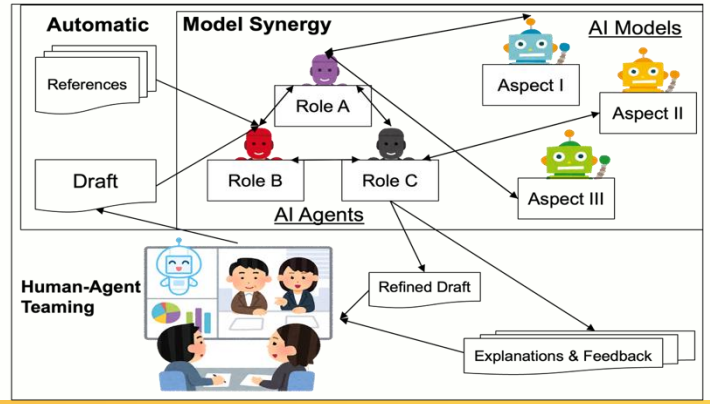
From Signals to Insights
(Human-AI Interaction)



AI as Partner



From Insights to Partnership
(Human-Agent Teaming)



Before We Talk About Agents, Let's Talk About Humans

HAA LAB

2018-2019
FinNum-1

2019-2020
FinNum-2

2021-2022
FinNum-3

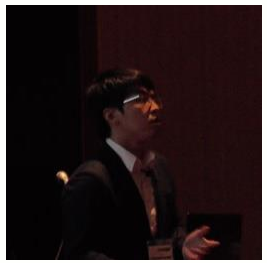
2022-2023
FinArg-1

2024
Program Co-Chair

2024-2025
FinArg-2

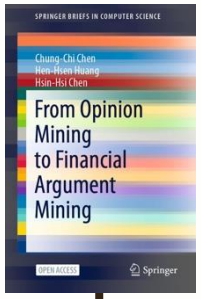
2025-2026
FinArg-3

2025-2026
RegCom



2019
FinNLP
Organizer

2020
AACL
Tutorial



2021
EMNLP
Tutorial

2024
ECAI
Tutorial

2025
SIGIR
Tutorial

2025
AACL
Tutorial

2025
ACL SIG-FinTech
Founder

2021
From Opinion Mining to Financial Argument Mining

2025
Agent AI for Finance: From Financial Argument Mining to Agent-Based Modeling

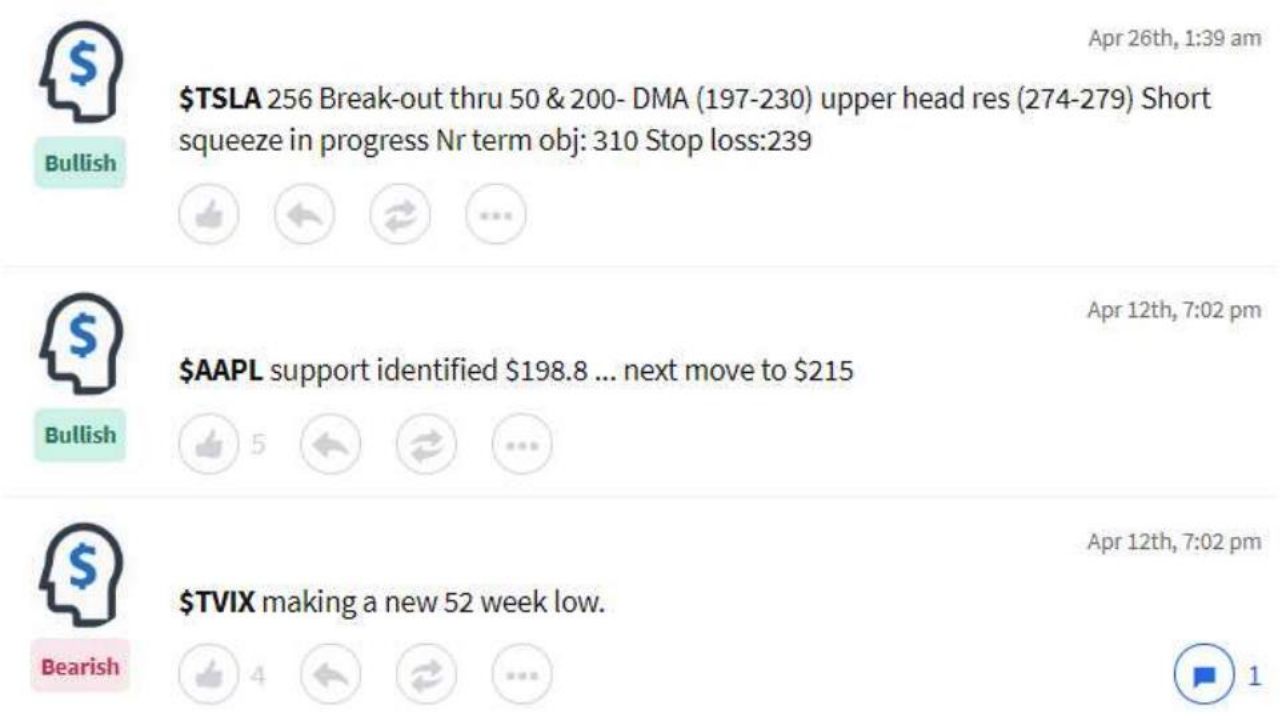
It Starts with a Simple Realization – The Impact of NTCIR on One's Research Journey

FinNum-1 & FinNLP Workshop

2018-2019

FinNum-1: Fine-Grained Numeral Understanding in Financial Tweets

HAA LAB



$TSLA 256 Break-out thru 50 & 200- DMA (197-230) upper head res (274-279) Short squeeze in progress Nr term obj: 310 Stop loss:239. *25 tokens 9 numbers 6 meanings*

We

- propose **fine-grained numeral taxonomy** for financial social media data
- attempt to **leverage the numeral opinions made by the crowd** to mine **additional information** for trading

Oh, You Are also Interesting in This Problem

HAA LAB

Let's Solve This Together and Move Forward Quickly

HAA LAB

- **Part-of-speech (POS) Tags**: Ait Azzi and Bouamor [1] and Liang and Su [7] extracted POS features with CMU ARK Twitter POS Tagger [10] and CoreNLP [8], respectively.
- **Keywords**: Ait Azzi and Bouamor [1] adopted the keywords from Chen et al. [4]. Liang and Su [7] proposed patterns for some (sub)categories.
- **Topic**: Spark [13] used Latent Dirichlet Allocation (LDA) [2] to extract the features for tweets' topics.
- **Position**: The position of the target numeral in the tweet is considered [13].
- **Named Entity**: Named entity extracted by CoreNLP [8] is used [7].
- **Term Frequency**: Term Frequency-Inverse Document Frequency (TF-IDF) is adopted for the context information [15].
- **Format**: Integer (float) format information is also adopted as a feature [13, 16]. Several co-occurrence format information is extracted by the given patterns [16].
- **Numeral Information**: Spark [13] not only used the raw value of the numeral, but also the log of raw value and the normalized raw value.
- **Bag-of-Characters**: The near n characters of the target numeral are considered [13].
- **Prefixes/Suffixes**: Prefixes and Suffixes are used in Wu et al. [16]
- **Brown Cluster**: The j-character prefix of the Brown cluster [3] is considered as features [16].
- **Recognizers-Text Type**: The text type extracted by Microsoft Recognizers is apoted [16].

- Ait Azzi and Bouamor [1] proposed a CNN-based model with enriched word representation, called E-CNN. They used a fusion model to integrate the fine-tuned model for subtask 1 into E-CNN for subtask 2.
- Spark [13] used two-layer rectied linear units (ReLU) as a classifier with both tweet features and number features.
- Liang and Su [7] developed a recurrent neural networks (RNN) model with CNN filter, and made comparison with both CNN and RNN models.
- Wang et al. [15] used SVM model in the formal run, and adopted the BERT model after the evaluation results released.
- Tian and Peng [14] constructed an attention-based LSTM model for the shared task.
- Wu et al. [16] used multi-layer perceptron (MLP) for target numeral and used LSTM for the preceding context and the posterior context of the target numeral.

Subtask 1			Subtask 2		
Submission ID	Micro F1 (%)	Macro F1 (%)	Submission ID	Micro F1 (%)	Macro F1 (%)
Fortia1 - 1	93.94	90.05	Fortia1 - 2	87.17	82.40
Fortia1 - 2	93.70	88.98	Fortia1 - 1	86.53	80.49
DeepMRT - 1	91.87	87.94	DeepMRT - 1	83.03	77.90
DeepMRT - 2	91.16	84.72	DeepMRT - 2	81.27	75.59
ASNLU - 2	89.72	80.93	aiai - 1	80.24	74.11
ASNLU - 1	89.40	79.96	aiai - 2	80.64	73.43
ZHAW - 2	86.45	79.27	ASNLU - 1	79.12	72.51
Fortia2 - 1	89.88	79.26	ASNLU - 2	77.37	70.09
Fortia2 - 2	87.73	78.59	Fortia2 - 2	77.05	68.86
aiai - 1	86.45	78.09	Fortia2 - 1	79.28	68.33
aiai - 2	87.41	78.04	ZHAW - 2	75.54	66.44
ZHAW - 1	84.78	75.40	ZHAW - 1	72.67	64.84
WUST	74.02	63.71	Stark - 1	69.08	56.83
BRNIR - 1	74.27	63.53	WUST	60.88	52.93
Stark - 1	78.01	61.75	BRNIR - 1	63.67	51.90
BRNIR - 2	72.91	58.54	BRNIR - 2	61.99	47.14

The First Workshop on Financial Technology and Natural Language Processing (FinNLP)

HAA LAB

Participants

Microsoft
Fortia Financial Solutions
中央研究院 ACADEMIA SINICA
ORACLE
opt
University of Zurich[UZH]
Lenovo
AUTO1
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
國立臺北大學 National Taipei University
重慶郵電大學
武漢大學


Generated By ChatGPT

A Lunch during NTCIR-2019 Sparked a Collaboration that has Lasted for Eight Years (and Counting)

HAA LAB

Fortia Financial Solutions

- **FinNLP**
  - Sentence Boundary Detection in PDF Noisy Text in the Financial Domain (FinSBD 1-3)
  - Learning Semantic Representations for the Financial Domain (FinSim 1-4)
  - Multi-Lingual ESG (ML-ESG 1-3)

OUTSCALE

- **NTCIR**
  - RegCom: Multinational, Multilingual, Multi-Industry Regulatory Compliance Checking (2026)

The company name has changed, and the team members have changed, but the partnership remains the same

This Was Not Just One Story – A Pattern That Repeats

HAA LAB

Prof. Chenghua Lin
University of Manchester, UK

Generated By ChatGPT

From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls

Tomas Goldsack,[1] Yang Wang,[1] Chenghua Lin,[1,2] Chung-Chi Chen[3]
[1]Department of Computer Science, University of Sheffield, UK
[2]Department of Computer Science, University of Manchester, UK
[3]Artificial Intelligence Research Center, AIST, Japan
{tgoldsack1, Y.Wang4}@sheffield.ac.uk chenghua.lin@manchester.ac.uk
c.c.chen@acm.org

Observing Micromotives and Macrobehavior of Large Language Models

Yuyang Cheng[1,2], Xingwei Qu[1], Tomas Goldsack[4], Chenghua Lin[1], Chung-Chi Chen[3]
[1]University of Manchester, [2]University of Virginia, [3]AIST, Japan, [4]Cohere
jrm9ga@virginia.edu, xingwei.qu@postgrad.manchester.ac.uk, tomas.goldsack@cohere.com
chenghua.lin@manchester.ac.uk, c.c.chen@acm.org

Prof. Iryna Gurevych
Technical University Darmstadt, Germany

SIGIR 2025
Padova
ITALY

Iryna Gurevych

Technical University of Darmstadt, Germany

Commitment Checklist: Auditing Author Commitments in Peer Review

Chung-Chi Chen[1], Iryna Gurevych[2]

[1]AIST, Japan
[2]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science,
TU Darmstadt and National Research Center for Applied Cybersecurity ATHENE, Germany

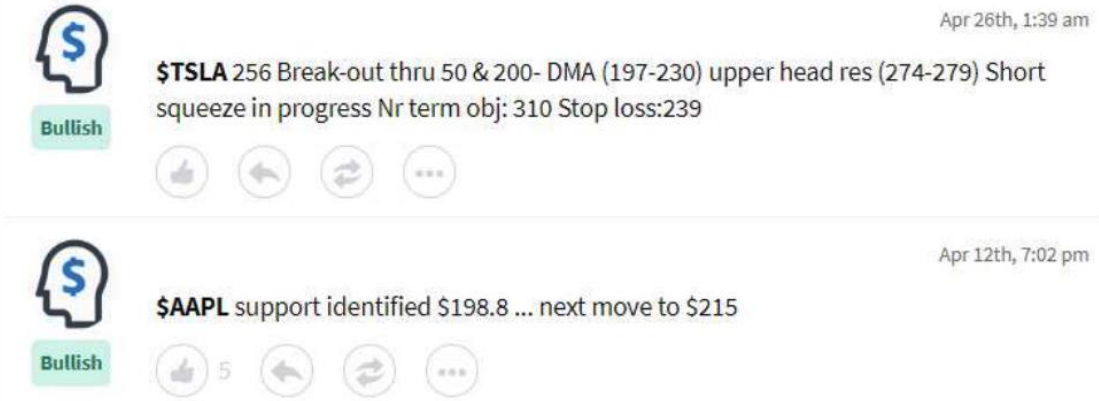
c.c.chen@acm.org, iryna.gurevych@tu-darmstadt.de

Oh!! This Actually Works

FinNum Task Series

2018-2022

FinNum-2 – Numeral Attachment



$TSLA 256 Break-out thru 50 & 200- DMA (197-230) upper head res (274-279) Short squeeze in progress Nr term obj: 310 Stop loss:239. *25 tokens 9 numbers 6 meanings*

We

- propose **fine-grained numeral taxonomy** for financial social media data
- attempt to **leverage the numeral opinions made by the crowd** to mine **additional information** for trading

Attached

$NE, last time oil was over $65 you were close to $8

Not Attached

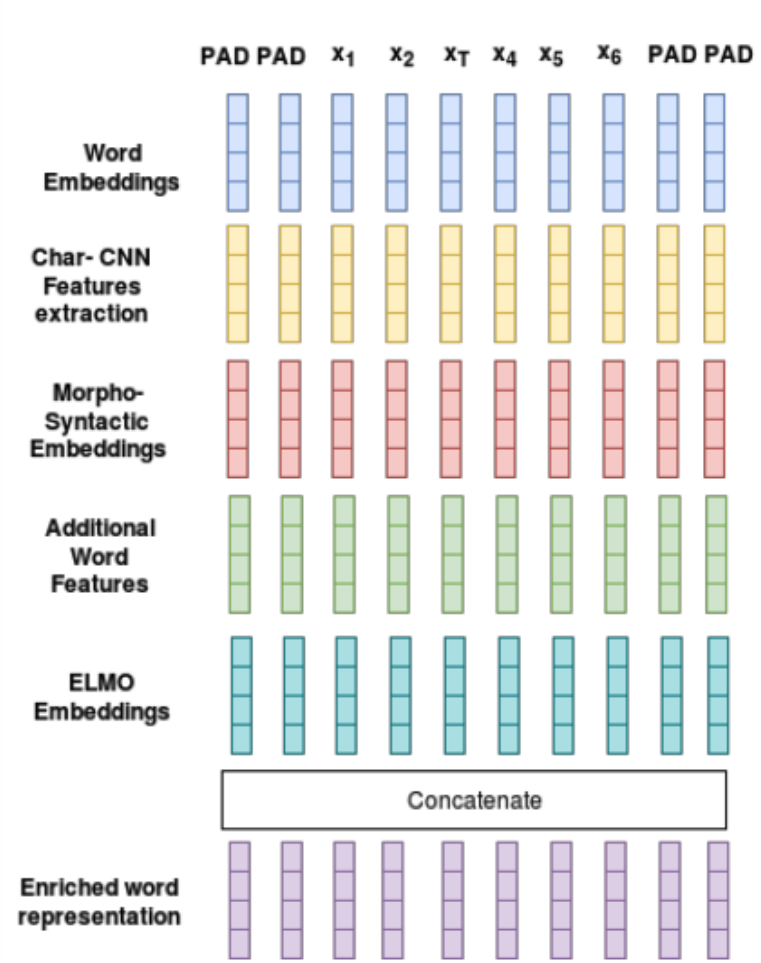
Attached

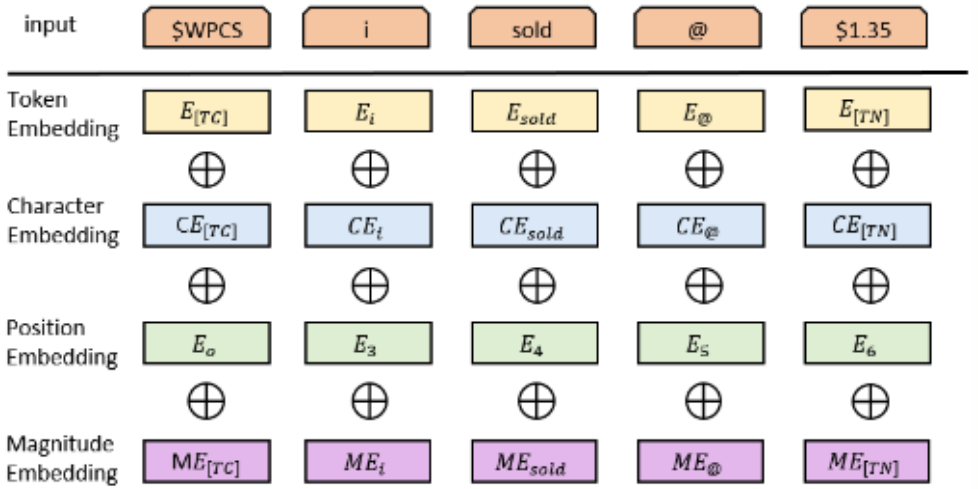
Attached

Guess who sold off about $800 million in $MDLZ after losing about $1 billion on $VRX???

Not Attached

Learn from Participants – FinNum-2 Baseline





Model	Caps-w	Caps-wc	Caps-wcp	Caps-all
Token	v	v	v	v
Character		v	v	v
Position			v	v
Magnitude				v
Macro-F1	60.08%	69.59%	69.73%	73.46%

- Ait Azzi, Abderrahim, and Houda Bouamor. "Fortia1@ the NTCIR-14 FinNum task: enriched sequence labeling for numeral classification." *Proceedings of the 14th NTCIR conference on evaluation of information access technologies*. 2019.
- Chen, Chung-Chi, Hen-Hsen Huang, and Hsin-Hsi Chen. "Numeral attachment with auxiliary tasks." *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019.

A Pattern That Repeats – FinNum-3 to Numeracy

HAA LAB

Format	Example "Fiscal Year 2018 Fourth Quarter"
Mask	Fiscal Year [MASK] Fourth Quarter
Marker	Fiscal Year [NUM] 2018 [NUM] Fourth Quarter
Digit	Fiscal Year [NUM] 2 0 1 8 [NUM] Fourth Quarter
Scientific (sig1)	Fiscal Year [NUM] 2 [EXP] 3 [NUM] Fourth Quarter
Scientific (sig4)	Fiscal Year [NUM] 2 . 0 1 8 [EXP] 3 [NUM] Fourth Quarter

Model	Notation	Tokenized Example
BERT	*Org.*	"147", "##70", "##2"
	Digit	"1", "4", "7", "7", "0", "2"
	SN	"1", ".", "47", "##70", "##200", "##00", "##0", "##e", "+", "05"
RoBERTa	*Org.*	"147", "702"
	Digit	"1", "4", "7", "7", "0", "2"
	SN	"1", ".", "47", "70", "200000", "E", "+", "05"

Model	Notation	QP		QNLI					QQA	Score
		Comment	Headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
BERT	*Original*	70.44%	57.46%	64.40%	59.20%	72.29%	60.42%	99.91%	53.20%	67.17
	Digit-based	65.38%	54.74%	57.86%	56.46%	71.36%	60.11%	99.11%	**53.75%**	64.85
	Scientific Notation	65.31%	55.99%	**64.42%**	**60.73%**	72.23%	59.66%	99.56%	**53.24%**	66.39
CN-BERT	*Digit-based*	69.93%	54.84%	61.07%	**60.27%**	**75.54%**	**65.39%**	99.42%	52.53%	**67.37**
	Scientific Notation	64.87%	56.40%	**66.39%**	54.70%	**75.41%**	**63.94%**	99.42%	51.90%	66.63
LinkBERT	*Original*	68.81%	55.70%	59.94%	56.85%	73.43%	59.01%	99.91%	54.14%	65.97
	Digit-based	63.76%	55.41%	59.54%	**57.42%**	**73.63%**	**60.17%**	99.73%	53.44%	65.39
	Scientific Notation	65.81%	**56.05%**	57.00%	56.78%	**75.51%**	58.51%	99.82%	**54.33%**	65.48
CN-LinkBERT	*Digit-based*	68.61%	54.44%	**63.59%**	55.08%	71.21%	58.99%	**100.00%**	50.44%	65.30
	Scientific Notation	63.48%	53.15%	**62.02%**	**59.39%**	**75.70%**	**62.61%**	99.73%	52.11%	**66.02**

- Onuma, Shunsuke, and Kazuma Kadowaki. "JRIRD at the NTCIR-16 FinNum-3 Task: Investigating the Effect of Numerical Representations in Manager's Claim Detection." Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies. 2022.
- Chen, Chung-Chi, et al. "Improving numeracy by input reframing and quantitative pre-finetuning task." Findings of the Association for Computational Linguistics: EACL 2023. 2023.

More Than Tasks: A Pattern That Shapes a Research Journey

It Grows into a Team –
NTCIR as a Foundation for Team Building

FinNum & FinArg @ NTCIR & FinNLP @ ACL

2019-2024

NTCIR as a Foundation for Team Building – Cross Domain

HAA LAB

FinNum-3

Distilling Numeral Information for Volatility Forecasting

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@iis.sinica.edu.tw

Yu-Lieh Huang
Department of Quantitative Finance, National Tsing Hua University, Taiwan
Center for Research in Econometric Theory and Applications, National Taiwan University, Taiwan
ylihuang@mx.nthu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

Automation of Text-Based Economic Indicator Construction: A Pilot Exploration on Economic Policy Uncertainty Index

Hsiu-Hsuan Yeh
Department of Economics, National Taiwan University
Taipei, Taiwan
r12323011@ntu.edu.tw

Yu-Lieh Huang
Department of Quantitative Finance, National Tsing Hua University
Hsinchu, Taiwan
Center for Research in Econometric Theory and Applications, National Taiwan University
Taipei, Taiwan
ylihuang@mx.nthu.edu.tw

Ziho Park
Department of Economics, National Taiwan University
Taipei, Taiwan
zihopark@ntu.edu.tw

Chung-Chi Chen
National Institute of Advanced Industrial Science and Technology
Tokyo, Japan
c.c.chen@acm.org

Constructing Noise Free Economic Policy Uncertainty Index

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@iis.sinica.edu.tw

Yu-Lieh Huang
Department of Quantitative Finance, National Tsing Hua University, Taiwan
Center for Research in Econometric Theory and Applications, National Taiwan University, Taiwan
ylihuang@mx.nthu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

ELSEVIER

International Review of Economics & Finance

Volume 89, Part A, January 2024, Pages 1286-1302

Semantics matter: An empirical study on economic policy uncertainty index ☆

Chung-Chi Chen [a], Yu-Lieh Huang [b c], Fang Yang [d]

NTCIR as a Foundation for Team Building – Cross Venue

HAA LAB

FinArg

It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, Lionel Brunie

Abstract

With the goal of reasoning on the financial textual data, we present in this paper, a novel approach for annotating arguments, their components and relations in the transcripts of earnings conference calls (ECCs). The proposed scheme is driven from the argumentation theory at the micro-structure level of discourse. We further conduct a manual annotation study with four annotators on 136 documents. We obtained inter-annotator agreement of $lpha_U$ = 0.70 for argument components and lpha = 0.81 for argument relations. The final created corpus, with the size of 804 documents, as well as the annotation guidelines are publicly available for researchers in the domains of computational argumentation, finance and FinNLP.

PDF | Cite | Search | Fix data

Anthology ID: 2022.finnlp-1.22
Volume: Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)
Month: December
Year: 2022
Address: Abu Dhabi, United Arab Emirates (Hybrid)
Editors: Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen
Venue: FinNLP

NTCIR as a Foundation for Team Building – From Participants

FinArg-3

CYUT at the NTCIR-15 FinNum-2 Task: Tokenization and Fine-tuning Techniques for Numeral Attachment in Financial Tweets

Mike Tian-Jian Jiang
Zeals Co, Ltd.
Tokyo, Japan
tmjiang@gmail.com

Yi-Kun Chen
Department of CSIE
Chaoyang University of Technology
Taichung, Taiwan
kun26712930@gmail.com

Shih-Hung Wu†
Department of CSIE
Chaoyang University of Technology
Taichung, Taiwan
shwu@cyut.edu.tw

CYUT at the NTCIR-16 FinNum-3 Task: Data Resampling and Data Augmentation by Generation

Xie-Sheng Hong
s11027602@gm.cyut.edu.tw
Department of CSIE
Chaoyang University of Technology
Taichung, Taiwan

Jia-Jun Lee
s11027603@gm.cyut.edu.tw
Department of CSIE
Chaoyang University of Technology
Taichung, Taiwan

Shih-Hung Wu*
shwu@cyut.edu.tw
Department of CSIE
Chaoyang University of Technology
Taichung, Taiwan

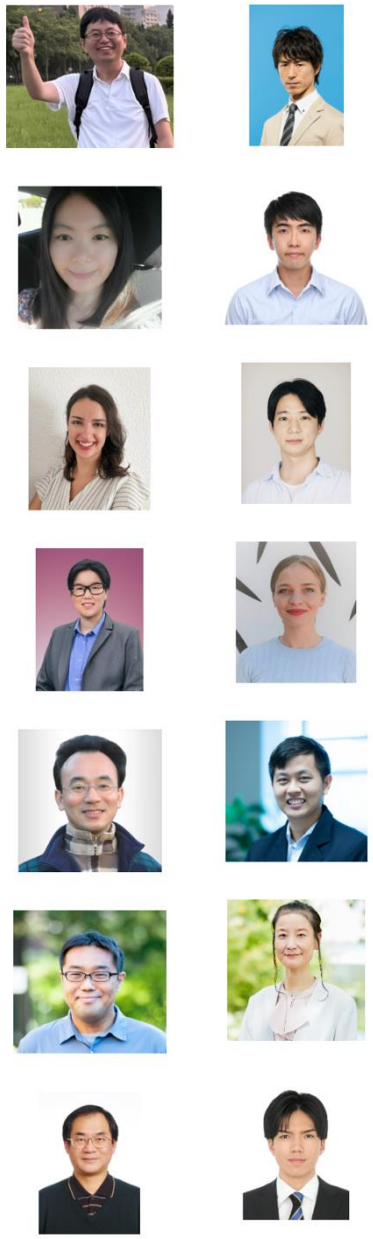
Mike Tian-Jian Jiang
tmjiang@gmail.com
Zeals Co, Ltd
Tokyo, Japan

CYUT at the NTCIR-17 FinArg-1 Task2: A Quantitative Prompt Engineering Approach for Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads

Shih-Hung Wu
Chaoyang University of Technology
Taiwan (R.O.C)
shwu@cyut.edu.tw

TSAI Tsung Hsun
Chaoyang University of Technology
Taiwan (R.O.C)
s11227607@gm.cyut.edu.tw

A Pattern That Repeats – Multi-Lingual ESG Task Series



Multi-Lingual ESG Issue Identification

Chung-Chi Chen,[1] Yu-Min Tseng,[2] Juyeon Kang,[3] Anaïs Lhuissier,[3]
Min-Yuh Day,[4] Teng-Tsai Tu,[5] Hsin-Hsi Chen[2]
[1]AIST, Japan
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[3]3DS Outscale (ex Fortia), France
[4]Graduate Institute of Information Management, National Taipei University, Taiwan
[5]Graduate Institute of International Business, National Taipei University, Taiwan

Multi-Lingual ESG Impact Type Identification

Chung-Chi Chen,[1] Yu-Min Tseng,[2] Juyeon Kang,[3] Anaïs Lhuissier,[3] Yohei Seki,[4]
Min-Yuh Day,[5] Teng-Tsai Tu,[6] Hsin-Hsi Chen[7]
[1]AIST, Japan
[2] Data Science Degree Program, National Taiwan University and Academia Sinica, Taiwan
[3]3DS Outscale, France, [4]University of Tsukuba, Japan
[5]Graduate Institute of Information Management, National Taipei University, Taiwan
[6]Graduate Institute of International Business, National Taipei University, Taiwan
[7]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

Multi-Lingual ESG Impact Duration Inference

Chung-Chi Chen,[1] Yu-Min Tseng,[2] Juyeon Kang,[3] Anaïs Lhuissier,[3]
Yohei Seki,[4] Hanwool Lee,[5] Min-Yuh Day,[6] Teng-Tsai Tu,[7] Hsin-Hsi Chen[8]
[1]AIST, Japan
[2] Data Science Degree Program, National Taiwan University and Academia Sinica, Taiwan
[3]3DS Outscale, France, [4]University of Tsukuba, Japan, [5]NCSOFT, South Korea
[6]Graduate Institute of Information Management, National Taipei University, Taiwan
[7]Graduate Institute of International Business, National Taipei University, Taiwan
[8]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

SemEval-2025 Task 6: Multinational, Multilingual, Multi-Industry Promise Verification

Chung-Chi Chen, Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Juyeon Kang, Hanwool Lee, Min-Yuh Day, Hiroya Takamura

Abstract

While extensive research exists on misinformation and disinformation, there is limited focus on future-oriented commitments, such as corporate ESG promises, which are often difficult to verify yet significantly impact public trust and market stability. To address this gap, we introduce the task of promise verification, leveraging natural language processing (NLP) techniques to automatically detect ESG commitments, identify supporting evidence, and evaluate the consistency between promises and evidence, while also inferring potential verification time points. This paper presents the dataset used in SemEval-2025 PromiseEval, outlines participant solutions, and discusses key findings. The goal is to enhance transparency in corporate discourse, strengthen investor trust, and support regulators in monitoring the fulfillment of corporate commitments.

PDF | Cite | Search | Fix data

ML-Promise: A Multilingual Dataset for Corporate Promise Verification

Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Hanwool Lee, Juyeon Kang, Min-Yuh Day, Chung-Chi Chen

Abstract

Promises made by politicians, corporate leaders, and public figures have a significant impact on public perception, trust, and institutional reputation. However, the complexity and volume of such commitments, coupled with difficulties in verifying their fulfillment, necessitate innovative methods for assessing their credibility. This paper introduces the concept of Promise Verification, a systematic approach involving steps such as promise identification, evidence assessment, and the evaluation of timing for verification. We propose the first multilingual dataset, ML-Promise, which includes English, French, Chinese, Japanese, and Korean, aimed at facilitating in-depth verification of promises, particularly in the context of Environmental, Social, and Governance (ESG) reports. Given the growing emphasis on corporate environmental contributions, this dataset addresses the challenge of evaluating corporate promises, especially in light of practices like greenwashing. Our findings also explore textual and image-based baselines, with promising results from retrieval-augmented generation (RAG) approaches. This work aims to foster further discourse on the accountability of public commitments across multiple languages and domains.

PDF | Cite | Search | Checklist | Fix data

Anthology ID: 2025.emnlp-main.1028
Volume: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing
Month: November
Year: 2025

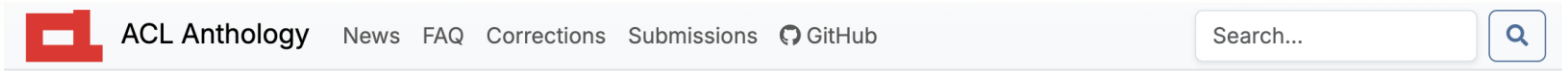
It Evolves into a Community

ACL SIG-FinTech

2025

https://sigfintech.github.io/

Create, Connect, Share



Financial Technology and Natural Language Processing (FinNLP)

Acronym: FinNLP
Venue ID: finnlp

2025
- Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) 49 papers
- Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing 27 papers

2024
- Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing 35 papers
- Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning 20 papers

2023
- Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing 14 papers
- Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting 18 papers

2022
- Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP) 35 papers

2021
- Proceedings of the Third Workshop on Financial Technology and Natural Language Processing 11 papers

2020
- Proceedings of the Second Workshop on Financial Technology and Natural Language Processing 18 papers

2019
- Proceedings of the First Workshop on Financial Technology and Natural Language Processing 23 papers

HAA LAB

- Xiaomo Liu
  - JP Morgan AI Research, US
- Udo Hahn
  - TexKnowlogy, Germany
- Armineh Nourbakhsh
  - JP Morgan AI Research, US
- Zhiqiang Ma
  - JP Morgan AI Research, US
- Charese Smiley
  - JP Morgan AI Research, US
- Véronique Hoste
  - Ghent University, Belgium
- Sanjiv Ranjan Das
  - Santa Clara University, USA
- Manling Li
  - University of Illinois Urbana-Champaign, US
- Mohammad Ghassemi
  - Michigan State University, US
- Antonio Moreno-Sandoval
  - UAM, Spain
- Qianqian Xie
  - The FinAI, Singapore
- Jimin Huang
  - The FinAI, Singapore
- Sophia Ananiadou
  - University of Manchester, UK
  - Archimedes/Athena RC, Greece

A Pattern That Repeats – Workshop → Tutorial

HAA LAB

SIGIR 2025
Padova
ITALY

Chung-Chi Chen (AIST, Japan)

Yongjae Lee (UNIST, South Korea)

Alejandro Lopez-Lira (University of Florida, United States)

Chanyeol Choi (LinqAlpha, United States)

Richard Mccreadie (University of Glasgow, United Kingdom)

Javier Sanz-Cruzado (University of Glasgow, United Kingdom)

NTCIR as a Platform for Starting Human-Human Teaming

HAA LAB

2018-2019
FinNum-1

2019-2020
FinNum-2

2021-2022
FinNum-3

2022-2023
FinArg-1

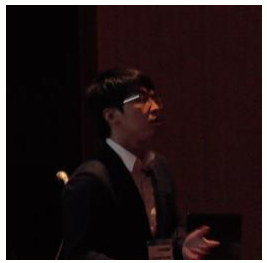
2024
Program Co-Chair

2024-2025
FinArg-2

2025-2026
FinArg-3

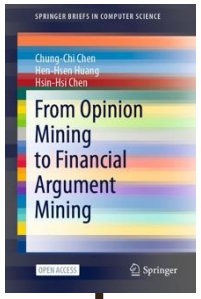
2025-2026
RegCom

LREC 2026 Palma



2019
FinNLP
Organizer

2020
AACL
Tutorial



2021
From Opinion Mining to Financial Argument Mining

2021
EMNLP
Tutorial

2024
ECAI
Tutorial

2025
SIGIR
Tutorial

2025
Agent AI for Finance: From Financial Argument Mining to Agent-Based Modeling

SIG FinTech

2025
ACL SIG-FinTech
Founder

2025
AACL
Tutorial

Not Every Attempt Succeeds

HAA LAB

NTCIR values experimentation.

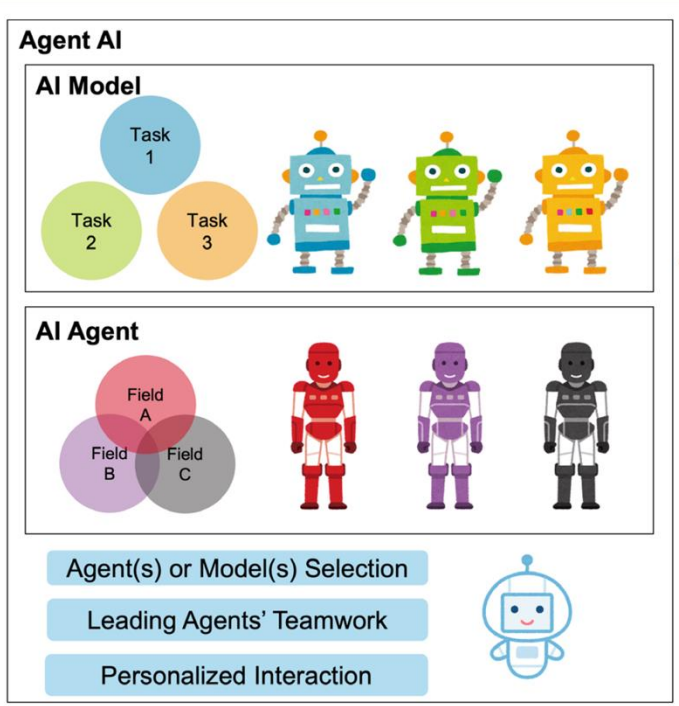
When things don't go as planned, we focus on understanding why and improving for next time.

Outline

HAA LAB

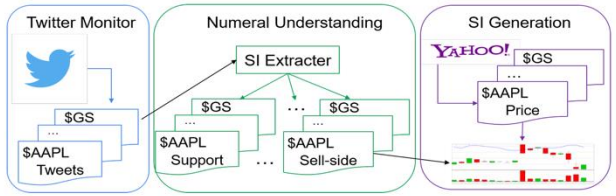
- Stories – From Tasks to Communities
  - Growing a Research Community through NTCIR
- **Research – History Repeats Itself**
  - **Benchmark**: What Agents Can Replace?
  - From IR to NLP: The Return of Subjectivity in **Evaluation**
  - From Static Metrics to **Verifiable** Outcomes

Can Agents Replace This Process?

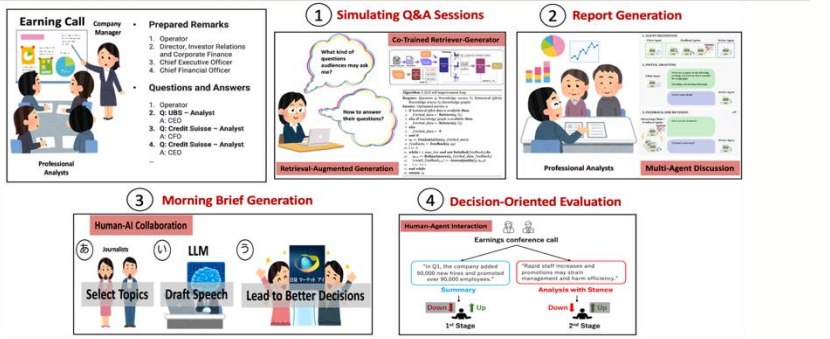


From Data to Signals
(Information Extraction)

AI as Tool

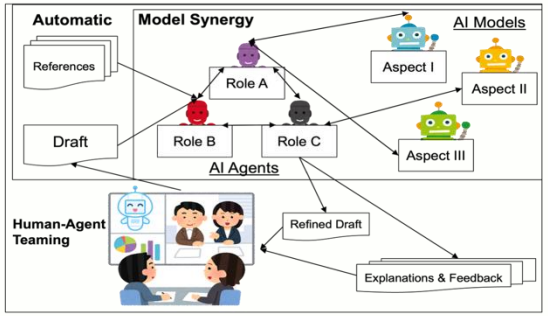


From Signals to Insights
(Human-AI Interaction)



AI as Partner

From Insights to Partnership
(Human-Agent Teaming)



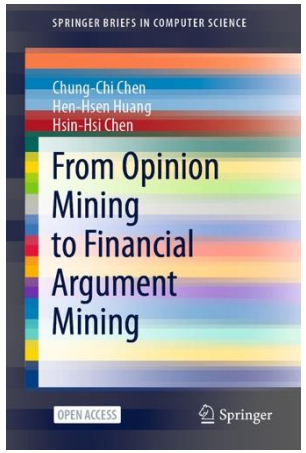
Data to Signal: Dataset Generation Pipeline (FinNum)

HAA LAB

	Earnings calls		Analysis reports [5]		Tweets [8]	
Category	Instances	Ratio	Instances	Ratio	Instances	Ratio
MONETARY: *money*	2,656	19.72%	874	16.99%	736	8.30%
MONETARY: *quote*	–	–	75	1.46%	1,033	11.65%
MONETARY: *change*	753	5.59%	18	0.35%	176	1.98%
MONETARY: *buy price*	–	–	–	–	415	4.68%
MONETARY: *sell price*	–	–	–	–	135	1.52%
MONETARY: *forecast*	–	–	–	–	355	4.00%
MONETARY: *stop loss*	–	–	–	–	35	0.39%
MONETARY: *support or resistance*	–	–	–	–	302	3.41%
PERCENTAGE: *relative*	3,040	22.57%	708	13.76%	767	8.65%
PERCENTAGE: *absolute*	969	7.19%	810	15.75%	346	3.90%
TEMPORAL: *date*	2,647	19.65%	2,134	41.49%	2,653	29.92%
TEMPORAL: *time*	8	0.06%	3	0.06%	365	4.12%
OPTION: *exercise price*	–	–	–	–	132	1.49%
OPTION: *maturity date*	–	–	–	–	70	0.79%
INDICATOR	–	–	–	–	216	2.44%
QUANTITY	2,199	16.33%	278	5.40%	982	11.07%
PRODUCT/VERSION	349	2.59%	136	2.64%	150	1.69%
RANKING	50	0.37%	3	0.06%	–	–
OTHER	798	5.92%	105	2.04%	–	–
	13,469	100.00%	5,144	100%	8,868	100%

Can Agents Replace This Process?

1. **Signal Definition (Taxonomy) – Ideation**
2. **Annotation Guideline Design**
3. **Raw Data Collection**
4. **Data Cleaning & Filtering**
5. **Human Annotation**
6. **Quality Review & Validation**
7. **Dataset Finalization**



Ideation

Research Idea

Business Idea

Can Agents Replace This Process?

1. **Signal Definition (Taxonomy) – Ideation**
2. Annotation Guideline Design
3. Raw Data Collection
4. Data Cleaning & Filtering
5. Human Annotation
6. Quality Review & Validation
7. Dataset Finalization

Man-Computer Symbiosis (Licklider, 1960)

HAA LAB

“The question is not ‘What is the answer?’

The question is ‘What is the question?’”

— J. C. R. Licklider (1960)

Human Role (Goals / Intuition / Judgment)

- **Sets goals**
- **Asks meaningful questions**
- **Provides intuition and creativity**
- **Evaluates results and makes decisions**

Computer Role (Computation / Search / Simulation)

- Performs routinizable work
- Searches and retrieves information
- Transforms and visualizes data
- Tests models and runs simulations

Historical Trajectory

1960: Vision of time-sharing and interactive computing

1960s: Rise of time-sharing systems and interactive computation

1990s: The Internet turns “thinking centers” into reality

Today: LLMs, copilots, and ChatGPT as thinking partners

Core Idea	Man and computer should form a **symbiotic relationship**, working together to solve problems neither could solve alone
Primary Focus	**Human–computer collaboration** and real-time interactive computing
Role of the Computer	A **thinking partner** that complements human cognitive strengths
Role of the Human	Provides goals, intuition, creativity, and judgment
Approach	Conceptual and visionary
Scope	Individual human–computer interaction
Key Contributions	Introduced the concept of interactive computing and cognitive symbiosis
Historical Impact	Influenced AI, HCI, human–AI collaboration

Divide Work based on what Each Human/Agent is Good at

HAA LAB

- **How do humans and LLM-based agents differ in research idea generation?**

- What AI Agents Do Better
  - Higher novelty: AI-generated ideas are rated significantly more novel by expert reviewers
  - Scalability: Can generate and explore a large space of candidate ideas quickly
  - Creative recombination: Effective at combining existing concepts in unexpected ways

- What Humans Do Better
  - Feasibility & grounding: Human ideas tend to be more practical and execution-aware
  - Use of domain intuition: Better alignment with established research practices and constraints
  - Judgment & evaluation: Humans are more reliable at assessing idea quality and feasibility

- Takeaway: Complementary Strengths
  - **AI excels at idea generation and novelty**
  - Humans excel at selection, refinement, and execution
  - Effective research agents should combine AI ideation with human judgment

Product Business Idea Generation from Patents

- **Goal**

  Generate a realistic product business idea from a real-world patent.
- **Input**
  - Full patent document

    (abstract, claims, technical description)
- **Output**

  For each patent, generate:
  - **Product Title**
  - **Product Description**
  - **Implementation**
  - **Differentiation**

AI shows Promise in Moving from Patent to Product

HAA LAB

1. **LLMs Can Generate Plausible Product Ideas**
   - Strong performance in **NLP** and **Computer Science** domains
   - **Human and LLM-based evaluations largely agree**
2. **Domain Expertise Still Matters**
   - In **Material Chemistry**, **human experts often disagreed with LLM judges**
   - Technical depth and feasibility require specialized knowledge
3. **Specificity Is Critical**
   - More concrete ideas consistently score higher
   - Vague ideas fail early in evaluation
4. **Business Reasoning Remains Challenging**
   - **Market size and competitive advantage** are harder than idea generation
   - Creativity alone is not enough ➔ In Business: Ideas are cheap; execution is everything

Agents Are Beyond What 1960 Could Have Imagined

HAA LAB

	Licklider (1960)	Si et al. (2024)	Hirota et al. (2025)
Core Question	Humans ask the questions	LLMs can generate novel research questions	LLMs can generate product business ideas from patents
Human Strength	Goals, intuition, judgment	**Judgment**, feasibility, selection	Market validation, execution, business strategy
Computer / AI Role	Computation and search	Large-scale research idea generation	**Product ideation and concept expansion**
Creativity	Primarily human	AI ideas rated more novel	AI can generate plausible and concrete business ideas
Limitation	Limited by human cognitive capacity	Feasibility and grounding remain challenging	Business reasoning and competitive analysis remain difficult
Division of Labor	Human thinks, computer computes	AI generates; **humans decide** & execute	AI proposes; **humans validate**, refine, and commercialize

The organizer determines the research direction (humans decide), while the proposal is used to persuade the reviewers (humans validate).

HAA LAB

* PROPOSAL TYPES:

We will accept two types of task proposals:

- Proposal of a Core task:
This is for fostering research on a particular information access problem by providing researchers with a common ground for evaluation. New test collections and evaluation methods may be developed through the collaboration between task organizers (proposers and task participants. At NTCIR-18, the core tasks are AEOLLM, FairWeb-2, FinArg-2, Lifelog-6, MedNLP-CHAT, RadNLP, and Transfer-2. Details can be found at http://research.nii.ac.jp/ntcir/NTCIR-18/tasks.html.

- Proposal of a Pilot task:
This is recommended for organizers who propose to focus on a novel information access problem, and there are uncertainties either in task design or organization. It may focus on a sub-problem of an information access problem and attract a smaller group of participating teams than core tasks. However, it may grow into a core challenging task in the next round of NTCIR. At NTCIR-18, the pilot tasks are HIDDEN-RAD, SUSHI, and U4. Details can be found at http://research.nii.ac.jp/ntcir/NTCIR-18/tasks.html.

FinArg-1 Proposal 2023

Short Name	Language	Source	Task
FinArg-1	English	Analyst Report	Argument-based Sentiment Analysis
	Chinese	Social Media	Identifying Attack and Support Argumentative Relations in Social Media Discussion Thread
FinArg-2	English	Analyst Report	Premise's Influence Period Assessment
	Chinese	Social Media	Claim's Validity Period Assessment
FinArg-3	English	Analyst Report	High Forecasting Skill Report Retrieval
	Chinese	Social Media	High Forecasting Skill Opinion Retrieval

Table 1: Overview of FinArg task series.

Task	Subtask
1. Argument-based Sentiment Analysis	1. Argument Classification 2. Premise Sentiment Analysis 3. Claim Sentiment Analysis
2. Identifying Attack and Support Argumentative Relations in Social Media Discussion Thread	-

Table 2: Overview of FinArg-1.

Guideline Design

Social Science

Can Agents Replace This Process?

1. Signal Definition (Taxonomy) – Ideation
2. **Annotation Guideline Design**
3. Raw Data Collection
4. Data Cleaning & Filtering
5. Human Annotation
6. Quality Review & Validation
7. Dataset Finalization

Economic Policy Uncertainty Index

HAA LAB

Economic Policy Uncertainty

Home Methodology Research & Applications About Us

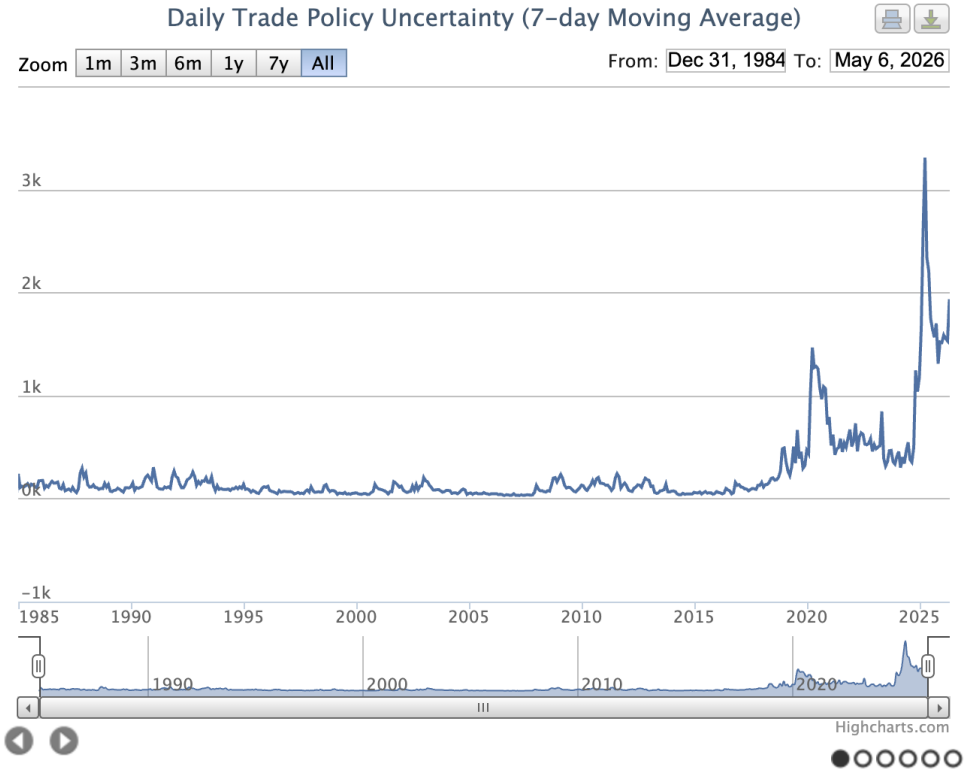
EPU Indices

All Country-Level Data

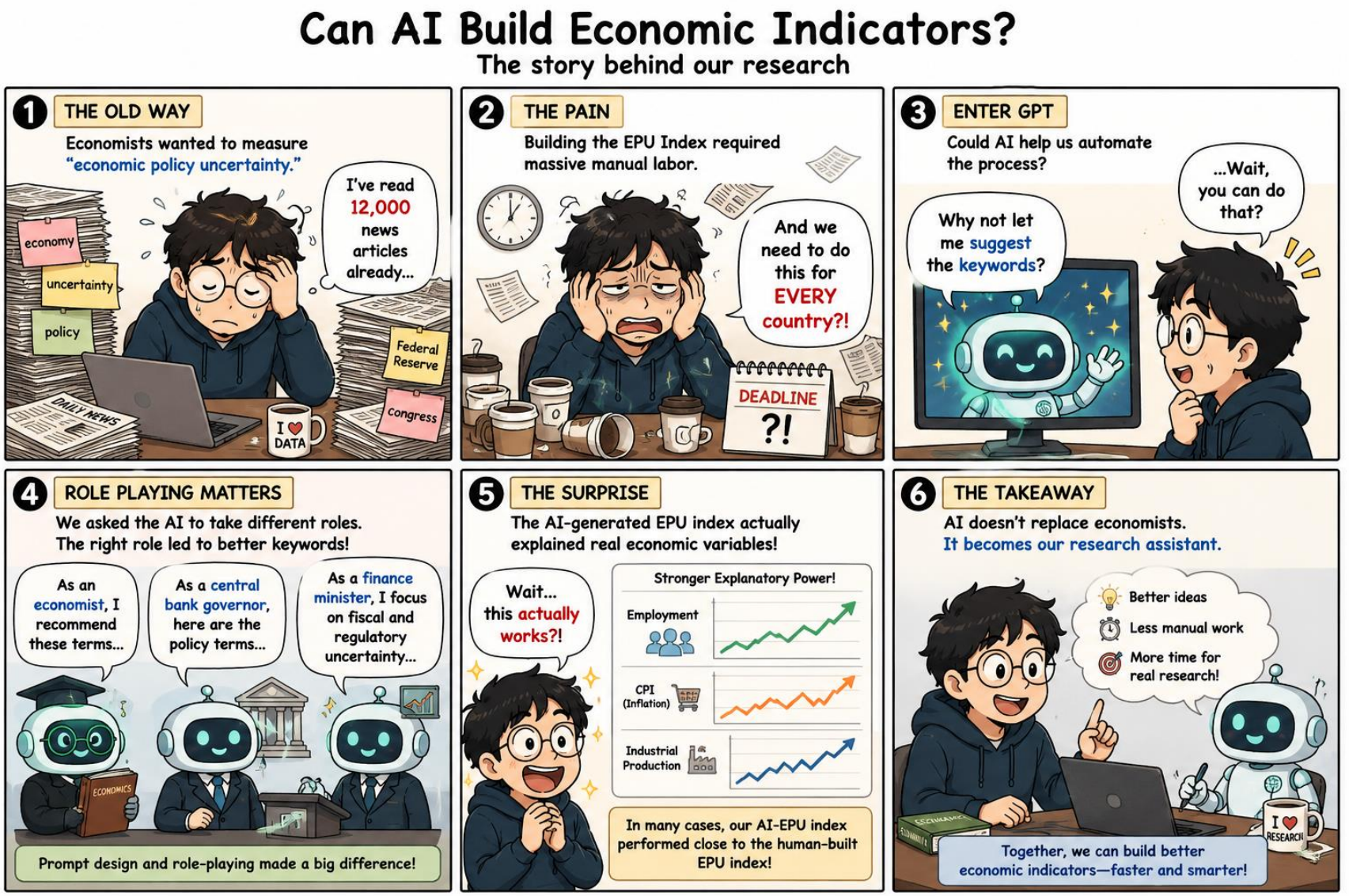
Global	USA
Argentina	Australia
Belgium	Bolivia
Brazil	Canada
Chile	China
Colombia	Costa Rica
Croatia	Denmark
Dominican Republic	Ecuador
El Salvador	France
Germany	Greece
Guatemala	Honduras
Hong Kong	India
Ireland	Italy
Japan	Mexico
Morocco	Netherlands
New Zealand	Nicaragua
Nigeria	Pakistan

Economic Policy Uncertainty Index

We develop indices of economic policy uncertainty for countries around the world.



Automation ➔ Yes



Human/Agent Annotation

Can Agents Replace This Process?

1. Signal Definition (Taxonomy) – Ideation
2. Annotation Guideline Design
3. Raw Data Collection
4. Data Cleaning & Filtering
5. **Human Annotation**
6. Quality Review & Validation
7. Dataset Finalization

Argument Mining for Forward-Looking Statements (FinArg)

HAA LAB

J.P.Morgan

Michaels

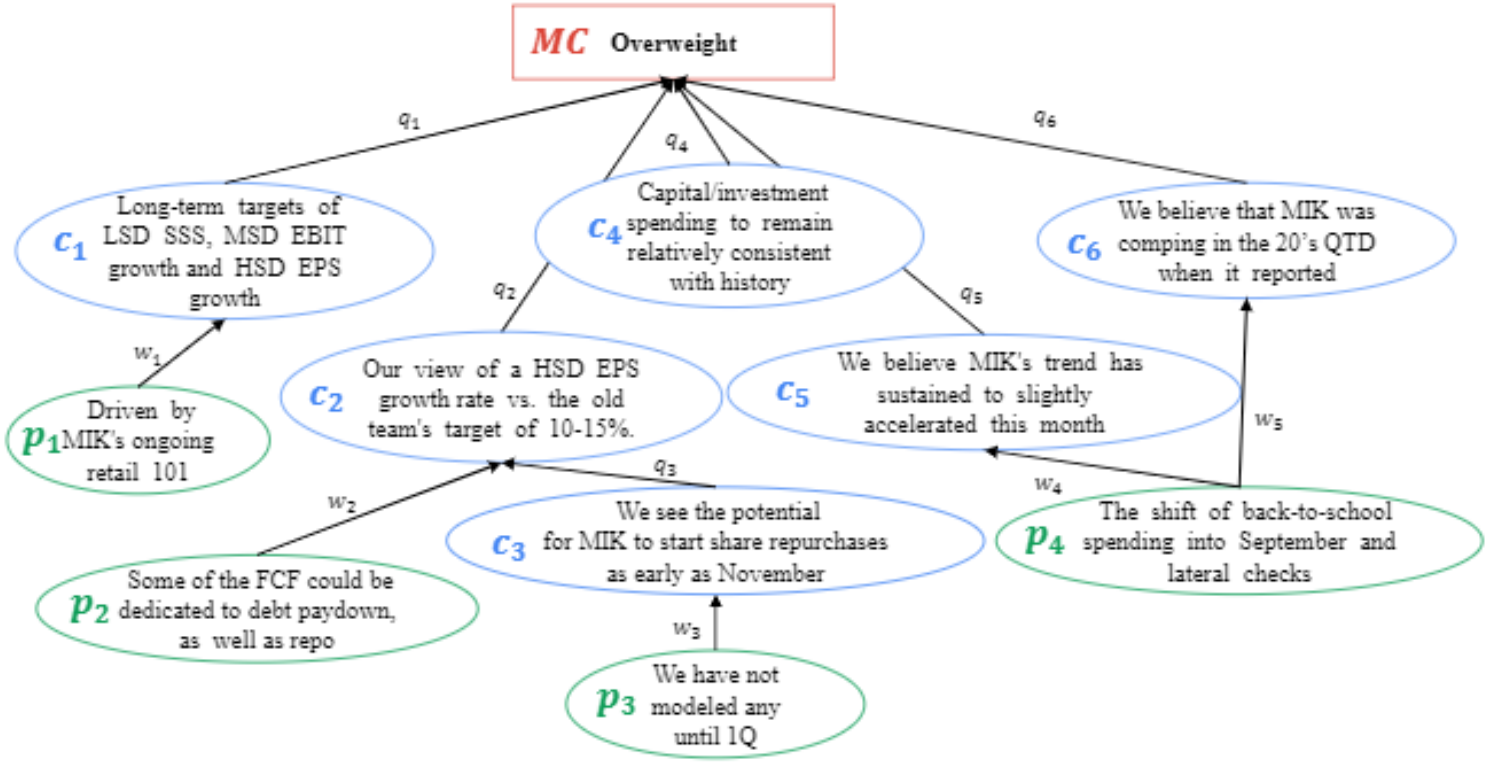
Overweight

MIK, MIK US

Price (16 Sep 20): $10.18

Price Target (Dec-20): $16.00

We expect the following: (1) Long-term targets of LSD SSS, MSD EBIT growth and HSD EPS growth driven by MIK's ongoing retail 101, omni-channel, and makers/Pro initiatives to drive topline/share (see bullet below) with the opportunity to improve margins through labor efficiency, merchandising rigor, inventory flow disciplines, cost leverage, and sourcing/private label expansion. **(2)** Capital/investment spending to remain relatively consistent with history given modest new store growth and a highly manageable omni-channel investment cycle (i.e., no need for a big supply chain or tech stack buildout); MIK targeted 2.5-3.0% of sales for capex on its last analyst day. **(3)** In terms of capital allocation, at the last analyst day MIK also targeted excess free cash flow solely to share repurchases (and we highlight its current FCFE yield of 23% and FCFF yield of 13%). However, new management has rightfully acknowledged that the company's financial leverage (5.5x gross debt to EBITDAR on our '21 estimates) is holding back its valuation given algorithmic trading and some value investors' aversion to leverage. This suggests some of the FCF could be dedicated to debt paydown, as well as repo, and hence our view of a HSD EPS growth rate vs. the old team's target of 10-15%. Notably, with MIK currently refinancing its term loan and the peak holiday inventory build happening now, we see the potential for MIK to start share repurchases as early as November, although we have not modeled any until 1Q (with 2021 embedding a total repo of 16MM shares for ~$250MM). **(4)** Recall, we believe that MIK was comping in the 20's QTD when it reported on September 3rd. Given the shift of back-to-school spending into September and lateral checks, we believe MIK's trend has sustained to slightly accelerated this month, although it remains unclear if management will speak to QTD.



Notation	Denotation
MC	Main Claim
C	Investor's claims
P	Premises
w	Weighting of the premise to the supported claim
q	Claim's quality

LLMs Perform Well on Semantic Understanding Tasks

HAA LAB

	Sentiment Label	Training	Development	Test
Claim	Bullish	3,831	426	439
	Bearish	2,397	267	320
	Neutral	1,348	150	170
Premise	Positive	5,058	562	1,965
	Negative	4,120	458	1,387
	Neutral	1,456	162	149
Scenario	Continued Growth	2,431	270	629
	Steady State	504	56	110
	Collapse	1,927	214	417
	Transformation	453	50	52

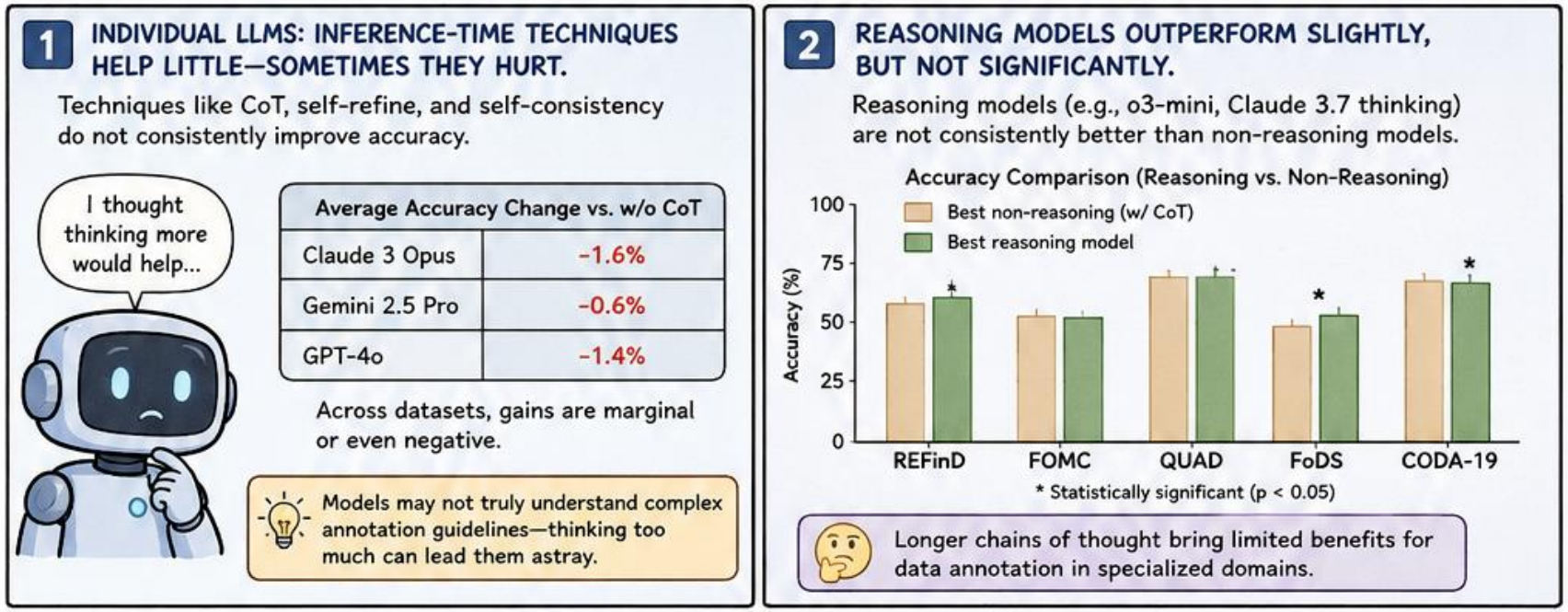
	Argument Unit Identification			
	Accuracy	Precision	Recall	F1 Score
ChatGPT(Zero-Shot)	0.723	0.739	0.723	0.725
ChatGPT(Few-Shot)	0.754	0.830	0.766	0.779
GPT-4(Few-Shot)	0.774	0.824	0.819	0.812
BERT	0.902	0.901	0.903	0.902
FinBERT	0.899	0.901	0.898	0.901
RoBERTa	**0.905**	**0.907**	**0.905**	**0.906**

	Claim				**Premise**				**Scenario**			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
ChatGPT(Zero-shot)	0.819	0.862	0.825	0.833	0.875	**0.944**	0.882	0.901	0.774	0.891	0.773	0.817
ChatGPT(Few-Shot)	0.883	0.886	0.883	0.884	0.883	0.941	0.883	0.906	0.840	**0.915**	0.840	0.870
GPT-4(Few-Shot)	0.919	0.923	0.919	0.920	0.923	0.918	0.923	0.912	0.755	0.882	0.755	0.809
BERT	0.927	0.932	0.929	0.930	0.912	0.901	0.918	0.911	0.866	0.883	0.870	0.871
FinBERT	**0.929**	0.930	0.930	0.930	0.904	0.906	0.901	0.903	0.872	0.861	0.875	0.862
RoBERTa	0.922	**0.933**	**0.932**	**0.931**	**0.925**	0.919	**0.925**	**0.920**	**0.884**	0.904	**0.885**	**0.893**

Agent Annotators – Expert-Annotated Datasets

HAA LAB

- High benchmark scores do NOT mean reliable expert annotation ability
- Multi-Agent Discussion Works Better
- Reasoning Models Can Be Too Stubborn
- Current LLMs are NOT reliable replacements for human expert annotators



NTCIR encourages the development of domain-specific datasets and expert-annotated dataset for new tasks

HAA LAB

- NTCIR-19
  - **Finance**
    - FinArg-3 (financial argument assessment)
  - **Biomedical**
    - MedNLP-CALL (emergency medical NLP)
    - HIDDEN-RAD2 (radiology reasoning)
  - **Legal**
    - RegCom (regulatory compliance)
  - **Science**
    - SciClaimEval (scientific claim verification)

Dataset Quality

Can Agents Replace This Process?

1. Signal Definition (Taxonomy) – Ideation
2. Annotation Guideline Design
3. Raw Data Collection
4. Data Cleaning & Filtering
5. Human (Expert) Annotation
6. Quality Review & Validation
7. Dataset Finalization

Argument Mining for Forward-Looking Statements (FinArg)

HAA LAB

J.P.Morgan

Michaels

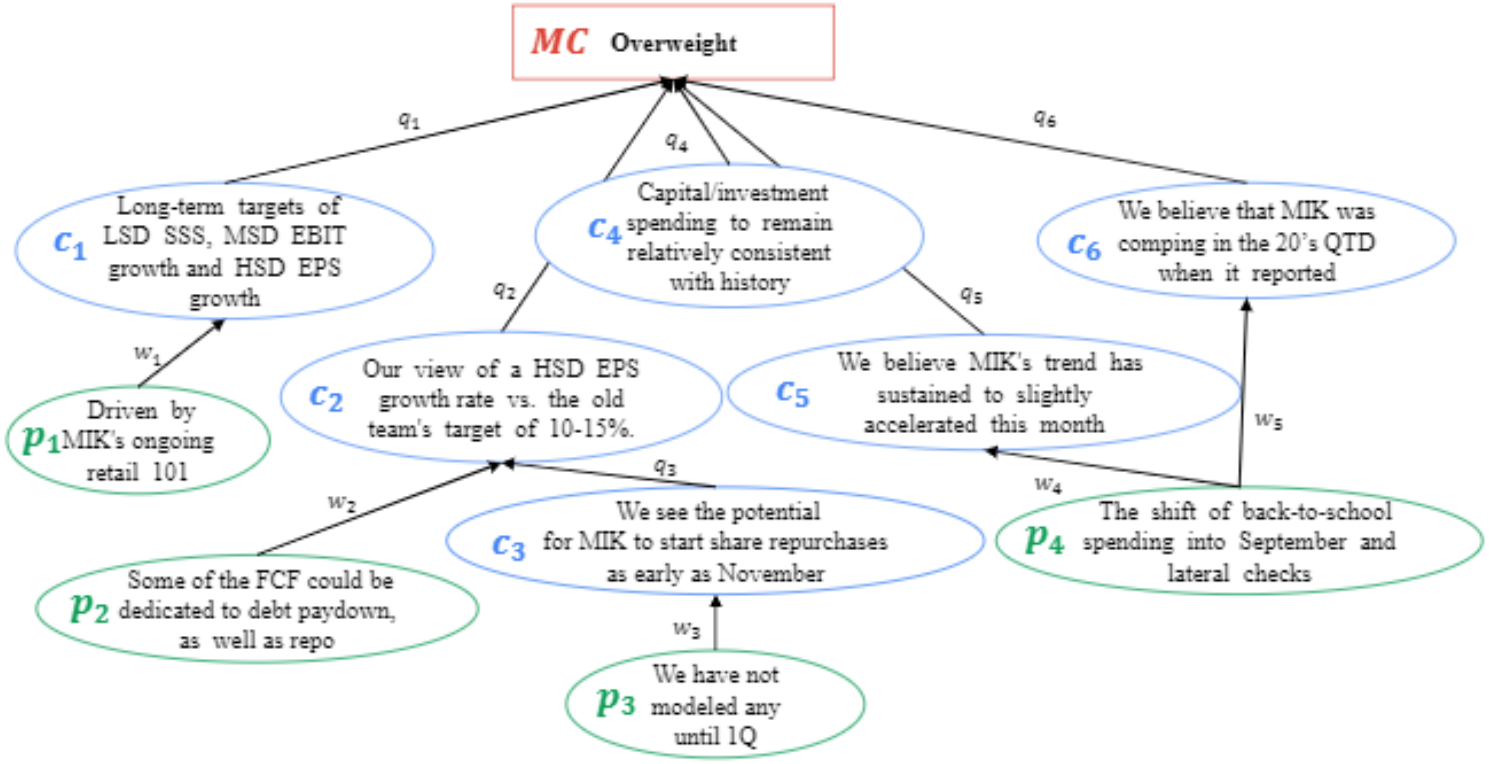
Overweight

MIK, MIK US

Price (16 Sep 20): $10.18

Price Target (Dec-20): $16.00

We expect the following: (1) Long-term targets of LSD SSS, MSD EBIT growth and HSD EPS growth driven by MIK's ongoing retail 101, omni-channel, and makers/Pro initiatives to drive topline/share (see bullet below) with the opportunity to improve margins through labor efficiency, merchandising rigor, inventory flow disciplines, cost leverage, and sourcing/private label expansion. **(2)** Capital/investment spending to remain relatively consistent with history given modest new store growth and a highly manageable omni-channel investment cycle (i.e., no need for a big supply chain or tech stack buildout); MIK targeted 2.5-3.0% of sales for capex on its last analyst day. **(3)** In terms of capital allocation, at the last analyst day MIK also targeted excess free cash flow solely to share repurchases (and we highlight its current FCFE yield of 23% and FCFF yield of 13%). However, new management has rightfully acknowledged that the company's financial leverage (5.5x gross debt to EBITDAR on our '21 estimates) is holding back its valuation given algorithmic trading and some value investors' aversion to leverage. This suggests some of the FCF could be dedicated to debt paydown, as well as repo, and hence our view of a HSD EPS growth rate vs. the old team's target of 10-15%. Notably, with MIK currently refinancing its term loan and the peak holiday inventory build happening now, we see the potential for MIK to start share repurchases as early as November, although we have not modeled any until 1Q (with 2021 embedding a total repo of 16MM shares for ~$250MM). **(4)** Recall, we believe that MIK was comping in the 20's QTD when it reported on September 3rd. Given the shift of back-to-school spending into September and lateral checks, we believe MIK's trend has sustained to slightly accelerated this month, although it remains unclear if management will speak to QTD.



Notation	Denotation
MC	Main Claim
C	Investor's claims
P	Premises
w	Weighting of the premise to the supported claim
q	Claim's quality

Can LLM-generated labels replace human annotations?

HAA LAB

	Claim				Premise				Scenario			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
RoBERTa (Human)	**0.922**	**0.933**	**0.932**	**0.931**	**0.925**	0.919	**0.925**	**0.920**	**0.884**	**0.904**	**0.885**	**0.893**
BERT (ChatGPT)	0.769	0.860	0.770	0.790	0.897	0.920	0.900	0.910	0.781	0.830	0.780	0.790
FinBERT (ChatGPT)	0.777	0.855	0.777	0.793	0.888	0.920	0.889	0.901	0.803	0.852	0.804	0.821
RoBERTa (ChatGPT)	0.801	0.856	0.802	0.811	0.909	**0.931**	0.910	0.919	0.821	0.847	0.822	0.828

	Training Set	Argument Unit Identification				Impact Duration Inference			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
RoBERTa	Human	**0.905**	**0.907**	**0.905**	**0.906**	**0.780**	**0.778**	**0.780**	**0.777**
BERT	ChatGPT	0.651	0.729	0.652	0.655	0.451	0.463	0.462	0.420
FinBERT		0.657	0.735	0.658	0.661	0.452	0.458	0.459	0.433
RoBERTa		0.686	0.745	0.686	0.689	0.466	0.466	0.471	0.441

What Can (and Cannot Yet) Be Replaced by Agents in the Dataset Preparation Pipeline

HAA LAB

Pipeline Stage	Can Agents Replace It?	Key Observation
Signal Definition / Ideation	**Partially**	Agents are strong at generating novel ideas and exploring large hypothesis spaces, but humans still define meaningful research goals, feasibility, and real-world relevance.
Annotation Guideline Design	**Partially**	Agents can assist drafting and refining guidelines, but expert judgment is still required to resolve ambiguity, domain nuance, and social-science considerations.
Raw Data Collection	**Mostly Yes**	Data retrieval, crawling, organization, and preprocessing are increasingly automatable.
Data Cleaning & Filtering	**Mostly Yes**	Agents perform well on scalable filtering, normalization, and semantic preprocessing tasks.
Human / Expert Annotation	**Not Yet (Expert)**	LLMs show strong semantic understanding, but current agents are still unreliable as expert annotators, especially in specialized domains.
Quality Review & Validation	**Not Yet (New Tasks)**	Benchmark performance does not guarantee annotation reliability. Human experts remain essential for consistency checking and final validation.
Dataset Finalization	**Partially**	Agents can support formatting and documentation, but humans still determine dataset standards, scope, and release quality.

Outline

- Stories – From Tasks to Communities
  - Growing a Research Community through NTCIR
- **Research – History Repeats Itself**
  - **Benchmark**: What Agents Can Replace?
  - From IR to NLP: The Return of Subjectivity in **Evaluation**

KEYNOTE 1

Nancy F. Chen

The Long Arc of Language Resources: From Annotation to Alignment to Grounding

Language resources are the backbone of AI: they train models, structure linguistic analysis, and benchmark technological progress. Their evolution mirrors—and actively shapes—the trajectory of computational linguistics, speech technology, natural language processing, and artificial intelligence.

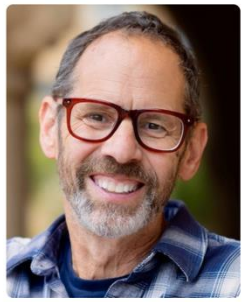
This talk traces the long arc of language resources across successive eras—from curated linguistic annotations to large-scale datasets enabling statistical learning, to representation learning and multimodal pretraining, and to alignment, where data shapes not only what models learn but also how they behave in society. **Building on this arc, we argue that the next phase is grounding: anchoring language technologies to perception, interaction, cultural and social contexts, and domain knowledge.**

KEYNOTE 2

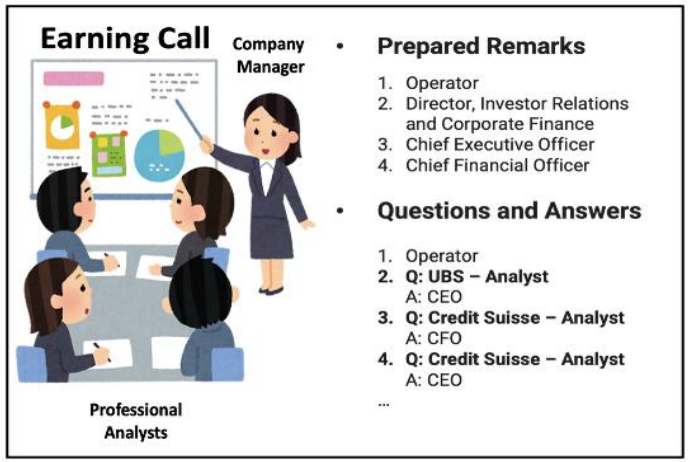
Dan Jurafsky

The Social Failures of Language Models as Conversational Partners

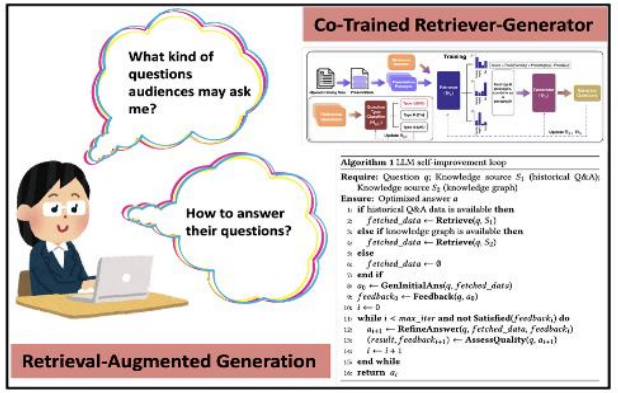
Language models are increasingly used in conversation for information, advice, and emotional support. In this talk I'll summarize studies in our lab showing that models fail in systematic ways as social interlocutors. We find that language models are socially sycophantic, linguistically overconfident, overly anthropomorphic, and epistemically self-centered. We then show that these flaws have real consequences for users: people interacting with models suffer consequences including overreliance, distorted judgment, and reduced personal responsibility. I'll discuss datasets and metrics, explore mitigations, and call for design, evaluation, and accountability mechanisms to protect user well-being.



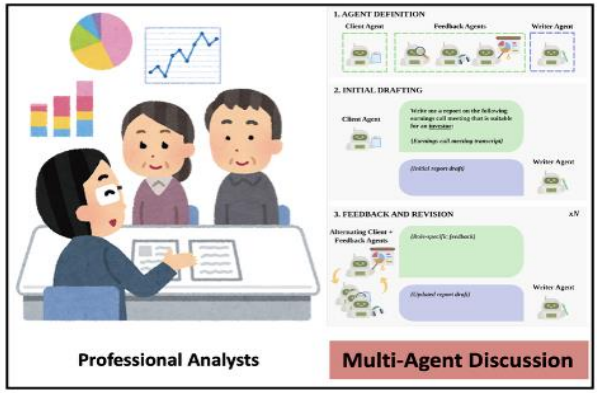
From Signals to Insights – Evaluation



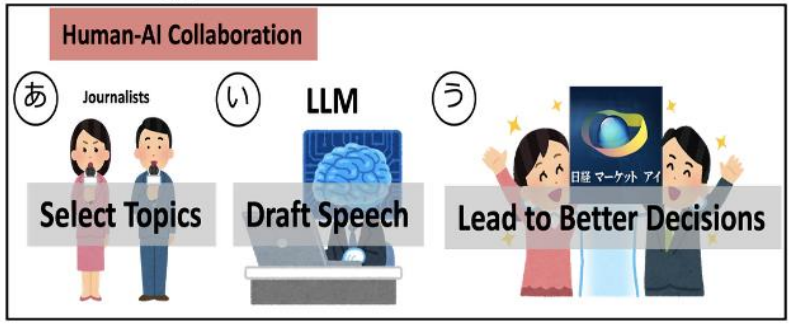
1 Simulating Q&A Sessions



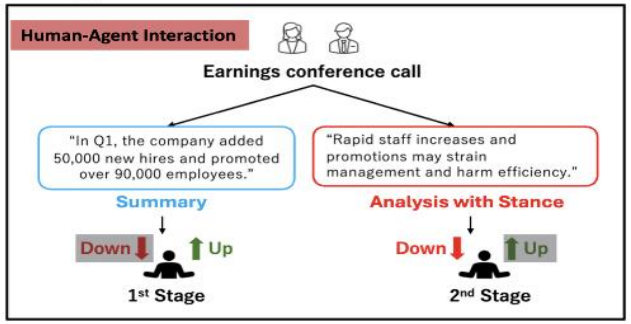
2 Report Generation



3 Morning Brief Generation

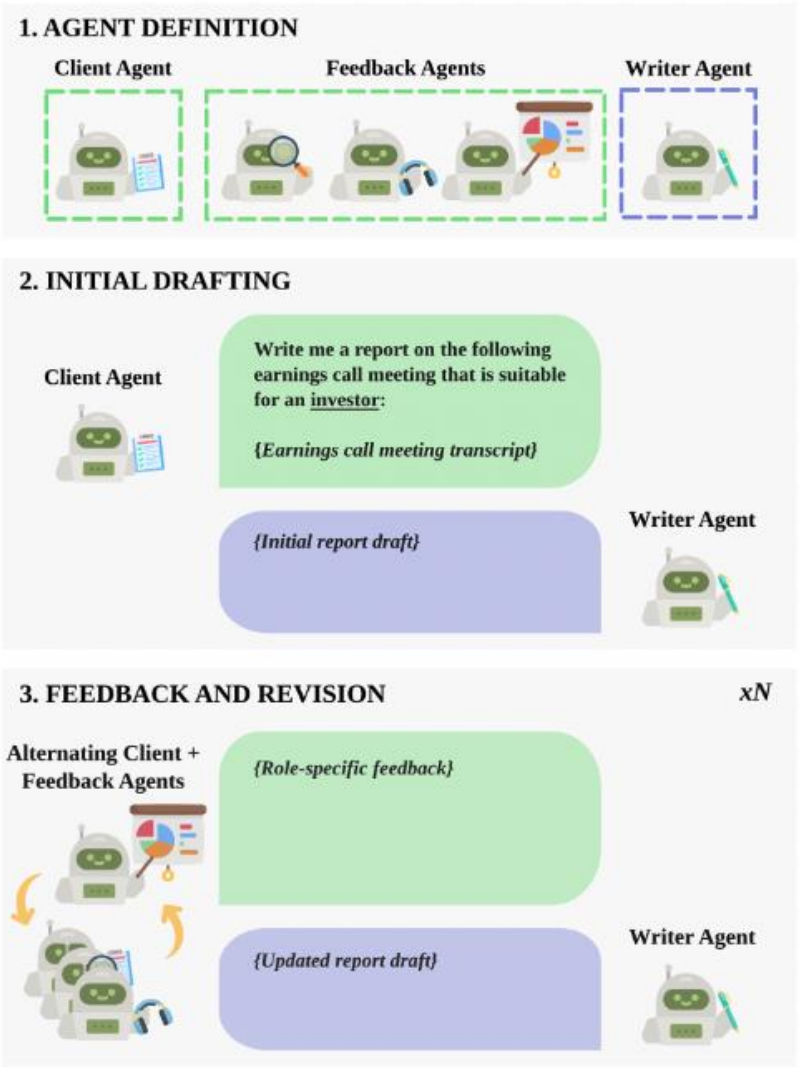


4 Decision-Oriented Evaluation



Report Generation (Audience Feedback/Reaction)

Different Aspects & Different Roles



Individual Thoughts or Collaboration

Agent	Initialisation Prompt
Writer ✍️	You are a Writer who is responsible for drafting the requested output text and making adjustments based on other agents' suggestions. Note that, unless otherwise specified, you should avoid completely rewriting the report and focus on making smaller targeted changes or additions based on other agent's feedback. You should only respond with updated versions of the report.
Client (Investor) 📋	You are an Investor who requires accurate investment and market analysis data to build investment strategies. You are responsible for ensuring the report contains the information that is relevant to you by providing feedback to the Writer. If you are happy with the report, respond with "TERMINATE".
Analyst 📈	You are an Analyst, a financial expert who is responsible for determining what past financial data might be relevant to the report and explaining this data to the Writer.
Psychologist 🎧	You are a Psychologist who is responsible for using data derived from the audio recording to identify notable features (e.g., that may express confidence, doubt, or other emotional giveaways) in audio-derived statistics of management's answers in the Q&A session that might be relevant to the report and explaining these features to the Writer.
Editor 🔍	You are an Editor who is responsible for ensuring that the output text is suitable for the intended audience (in terms of content, style, and structure) and that important information from previous revisions of the report is not lost by providing feedback to the Writer.

Expert vs. LLMs

HAA LAB

Readability

Agents	# Sents	FKGL	CLI	ARI	Abst
✍️	24.35	12.88	16.42	16.87	41.74
✍️🔍	22.90	13.67	17.55	17.83	48.03
✍️📈	21.43	13.44	17.32	17.24	49.46
✍️🎧	20.03	15.71	19.03	20.26	57.95
✍️📈🔍	19.65	14.76	18.33	19.10	53.40
✍️🎧🔍	19.68	15.69	19.18	20.11	56.87
✍️🎧📈🔍	18.58	15.11	18.98	19.46	56.72
JPMorgan (Expert)	19.25	7.26	8.54	8.85	47.14

More Agents, Greater Complexity

Preference

Report	An. 1	An. 2	An. 3	Avg.
✍️📈🔍	0.0	8.33	41.67	16.67
Expert	100.0	91.67	58.33	83.33

Report	GPT-4		Gemini-pro		Mistral	
	#1	#2	#1	#2	#1	#2
✍️📈🔍	100.0	70.83	87.5	100.0	91.67	16.67
Expert	0	29.17	12.5	0.0	8.33	83.33

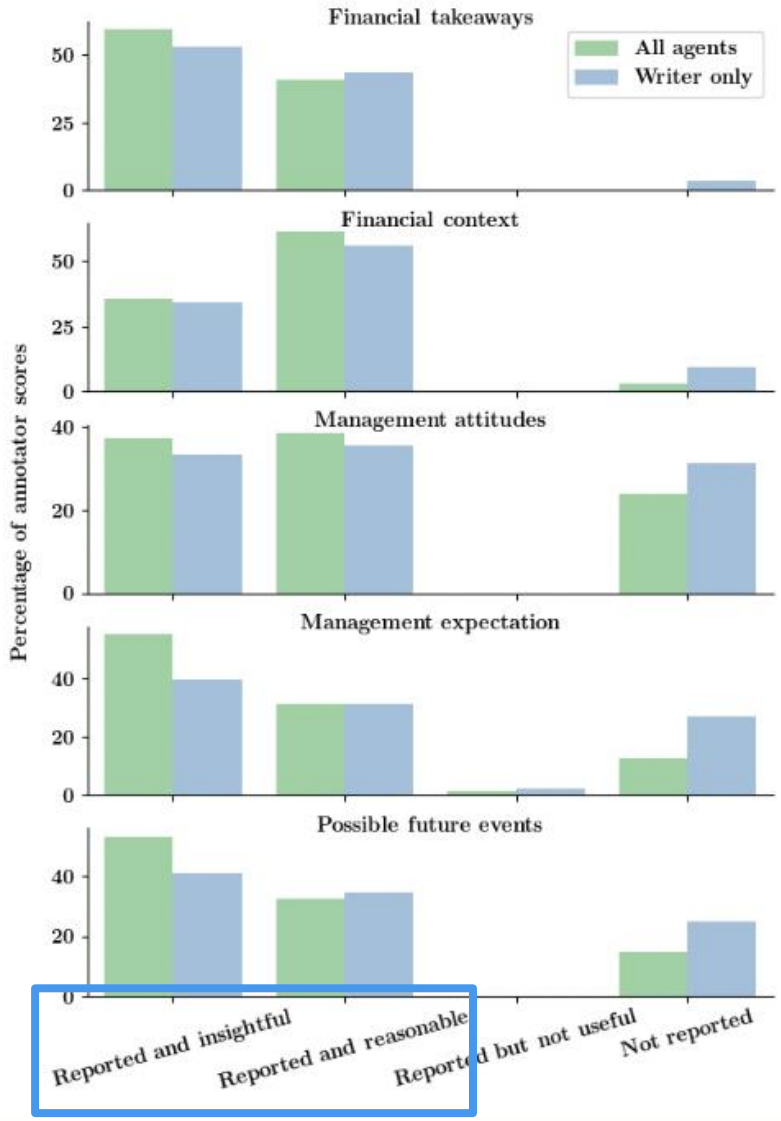
- **Expert-written reports better than agent-written**
- **LLMs have preference to agent-written reports**
- **Mistral is influenced by the order**

To P.16

Evaluation (Human vs. LLMs)

Report characteristic	Description
Financial takeaways	The key financial details from the meeting (i.e., numerical statistics relating to company performance for the quarter).
Financial context	Any additional information (e.g., financial details from previous quarters) that helps to contextualize the current financial performance.
Management attitudes	Information on how management (e.g., CEO, CFO, etc..) feels about the companys financial performance.
Management expectation	Details about how the company is expected to perform in the future/next quarter.
Possible future events	Details surrounding any noteworthy events/scenarios that are likely to occur in the future.

Characteristic	GPT-4			Gemini-pro			Mistral-medium		
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ
Financial Takeaways	0.375	0.160	0.412	0.156	0.018	0.014	0.139	0.205	0.192
Financial Context	0.597	0.455	0.397	0.341	0.330	0.292	0.758	0.437	0.397
Management Attitudes	0.570	0.524	0.463	0.248	0.301	0.266	0.463	0.558	0.492
Management Expectation	0.529	0.511	0.441	0.643	0.598	0.521	0.670	0.661	0.581
Future Events	0.472	0.379	0.327	0.179	0.194	0.167	0.422	0.382	0.330
Average	0.509	0.405	0.408	0.313	0.288	0.252	0.490	0.449	0.398



Evaluate Based on Human Decision Accuracy

HAA LAB

- **Two subsets**, total of **64 earnings call transcripts**:
  - **ECTSum Subset** (40 transcripts): Includes optional reference summaries (“ref”)
  - **Professional Subset** (24 transcripts): Only transcripts provided; analyst comparisons done later by organizers
  - **Submission Requirement**: Must generate reports for **all 64 transcripts**
- **Evaluation Criteria**
  - Participants may use LLM-based or custom evaluation methods
  - **Official ranking is based on human evaluation**:
    - Judges make investment decisions (Long/Short) based on the report
    - Timeframes: **Next day**, **Next week**, **Next month**
    - **Final score**: Average decision accuracy across the 3 timeframes

High Likert Scores do not imply High Decision Accuracy

HAA LAB

Team	Average	Clarity	Logic	Persuasiveness	Readability	Usefulness
LangKG	**5.96**	**6.02**	**5.92**	5.90	**5.81**	**6.13**
Jetsons	5.90	6.00	5.89	5.81	**5.81**	6.01
DKE	5.74	5.71	5.89	**5.95**	5.17	5.98
SigJBS	5.67	5.76	5.68	5.59	5.61	5.72
SI4Fin	5.56	5.52	5.84	5.60	5.06	5.80
DataLovers	5.50	5.56	5.45	5.32	5.73	5.47
Bgreens	5.49	5.51	5.61	5.51	5.09	5.74
KrazyNLP	5.29	5.15	5.49	5.21	5.01	5.59
iiserb	5.19	5.01	5.51	5.14	4.72	5.57
Finturbo	5.11	5.02	5.39	4.90	4.86	5.40
bds-LAB	4.99	4.91	5.21	5.03	4.55	5.27
PassionAI	4.70	4.64	4.74	4.39	4.88	4.86

Table 2: Average Likert scores across five qualitative dimensions.

Team	Average	Day	Week	Month
DKE	**0.581**	0.596	0.577	**0.570**
DataLovers	0.579	0.597	**0.611**	0.529
Jetsons	0.571	0.607	0.555	0.552
SigJBS	0.545	**0.609**	0.513	0.512
iiserb	0.537	0.576	0.558	0.477
PassionAI	0.537	0.588	0.557	0.466
Finturbo	0.524	0.504	0.568	0.500
Bgreens	0.522	0.469	0.581	0.516
LangKG	0.518	0.589	0.542	0.424
SI4Fin	0.515	0.525	0.524	0.497
KrazyNLP	0.471	0.514	0.525	0.375
bds-LAB	0.462	0.478	0.434	0.474

Table 1: Average accuracy of financial decisions across time horizons.

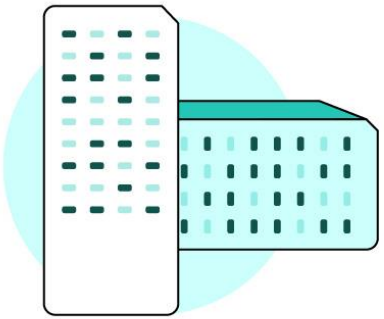
Human-Agent Teaming Era

HAA LAB

Evaluation would go beyond accuracy & speed

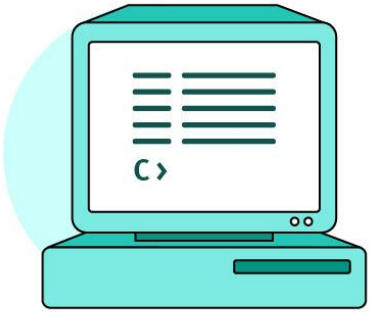
The extent to which the system benefits user/human matters

User-Interface Paradigms of Computing



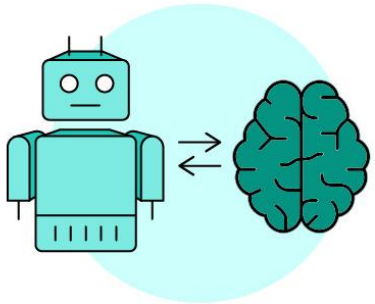
Paradigm 1

Batch Processing



Paradigm 2

Command-Based Interaction



Paradigm 3

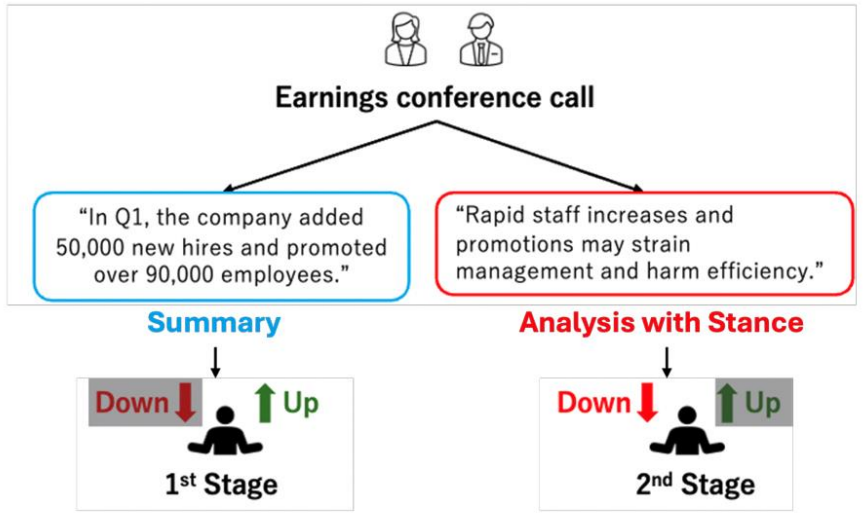
Intent-Based Outcome Specification

NNGROUP.COM NN/g

https://www.nngroup.com/articles/ai-paradigm/

LLM Opinions Sway Human Decisions

- We use GPT-4 to generate (1) a summary, (2) an analysis (given stance), and (3) a promotional analysis (given stance) based on the transcript of an earnings call.
- We invite participants from three categories: amateurs, experts (working in the financial industry), and veterans (with over 10 years of experience in the financial industry).
- The decision-making process consists of two rounds. In the first round, participants make a three-day trading decision based on the provided summary. In the second round, they receive a (promotional) analysis with stance and decide whether to modify their initial decision.
- Participants receive an hourly salary that is 1.5 times their original rate if they make correct decisions for over 50% of instances.



GPT-4 can Influence Expert Decisions, but in a Wrong Direction

HAA LAB

- GPT-4's analysis has only a **small impact on human decisions**, with the smallest influence on veterans.
- Decision changes among amateurs are double that of veterans.
- Promotional analysis is seen as more convincing, logical, and useful by all participants.
- In the financial market, promotion of investment products requires caution due to strict regulatory requirements across different regions.
- GPT-4-generated analysis **negatively impacts the accuracy of decisions** made by both amateurs and experts.
- GPT-4 produces persuasive analysis, but it may not necessarily help humans in making better decisions.
- **This raises a research issue about evaluating the effectiveness of generated analysis in improving decision-making. (Challenge)**

	Amateur	Expert	Veteran
Frequency	31.30%	24.70%	15.60%
Decrease of Accuracy	15.40%	16.60%	11.10%

Outline

- Stories – From Tasks to Communities
  - Growing a Research Community through NTCIR
- **Research – History Repeats Itself**
  - **Benchmark**: What Agents Can Replace?
  - From IR to NLP: The Return of Subjectivity in **Evaluation**
    - Subjectivity
    - Reproducible Evaluation (One of NTCIR's Goal)

KEYNOTE 1

Nancy F. Chen

The Long Arc of Language Resources: From Annotation to Alignment to Grounding

Language resources are the backbone of AI: they train models, structure linguistic analysis, and benchmark technological progress. Their evolution mirrors—and actively shapes—the trajectory of computational linguistics, speech technology, natural language processing, and artificial intelligence.

This talk traces the long arc of language resources across successive eras—from curated linguistic annotations to large-scale datasets enabling statistical learning, to representation learning and multimodal pretraining, and to alignment, where data shapes not only what models learn but also how they behave in society. **Building on this arc, we argue that the next phase is grounding: anchoring language technologies to perception, interaction, cultural and social contexts, and domain knowledge.**

KEYNOTE 2

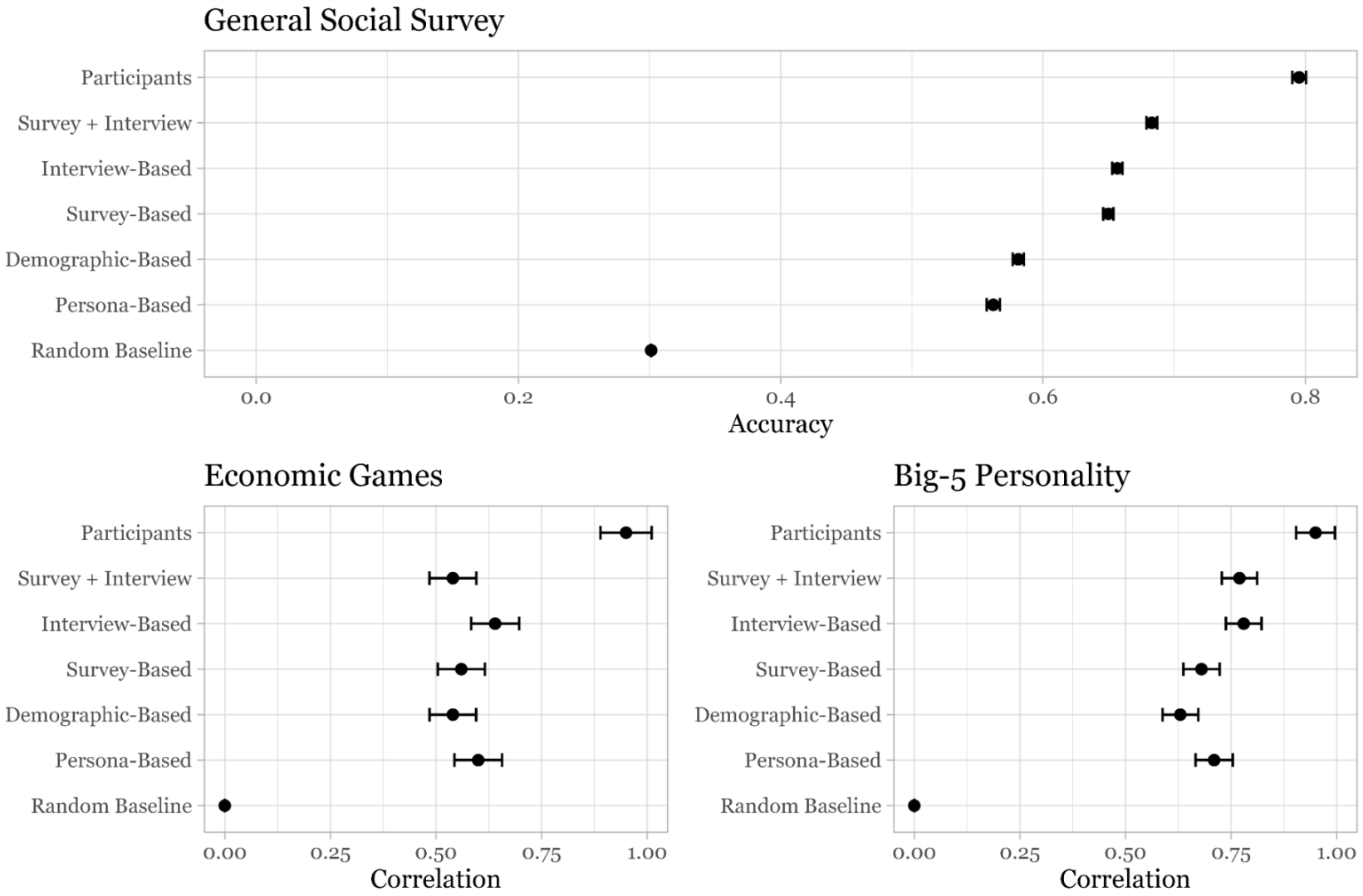
Dan Jurafsky

The Social Failures of Language Models as Conversational Partners

Language models are increasingly used in conversation for information, advice, and emotional support. In this talk I'll summarize studies in our lab showing that models fail in systematic ways as social interlocutors. We find that language models are socially sycophantic, linguistically overconfident, overly anthropomorphic, and epistemically self-centered. We then show that these flaws have real consequences for users: people interacting with models suffer consequences including overreliance, distorted judgment, and reduced personal responsibility. I'll discuss datasets and metrics, explore mitigations, and call for design, evaluation, and accountability mechanisms to protect user well-being.

Make the Agent Replicate and Reproduce Human Decisions

HAA LAB



Make the Agent Replicate and Reproduce Expert Decisions

HAA LAB

Expert	Prompting		Fine-tune	Model Editing
	Zero-shot	Few-shot	LoRA	AlphaEdit (Background)
A	33.78	**41.81**	34.15	33.78
B	23.19	**43.54**	**43.54**	31.88
C	2.33	10.91	**29.09**	23.26
D	9.38	30.41	**45.92**	20.31
E	17.54	22.78	21.51	**22.81**
F	15.32	22.51	27.75	**28.23**
G	15.00	19.05	30.95	**31.67**
H	14.44	30.25	32.77	**34.44**
I	14.29	22.45	33.67	**38.57**

Table 2: Overall performance (%) of different alignment strategies across nine experts.

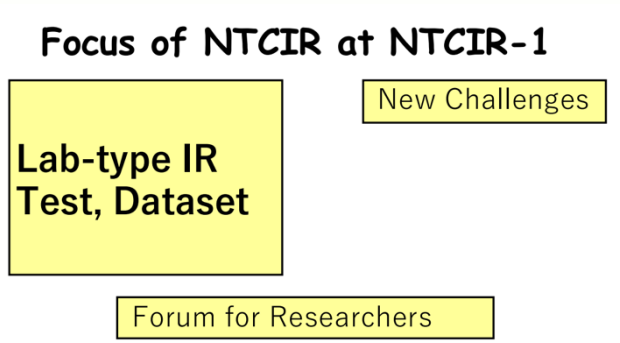
Expert	Background	+Criteria (Δ)	+Reasoning (Δ)
A	33.78	+2.03	+1.63
B	31.88	-2.89	-5.79
C	23.26	-2.33	0.00
D	20.31	0.00	-4.69
E	22.81	-3.51	-3.51
F	28.23	-0.81	-3.23
G	31.67	-13.34	-10.00
H	34.44	-3.33	-7.77
I	38.57	-4.28	-12.86

Table 3: Effect of additional expert information in model editing (%).

From IR to NLP: The Return of Subjectivity in Evaluation

HAA LAB

- **Early NLP**
  - Benchmarks made evaluation appear objective
  - One input → one correct answer
  - Accuracy became the dominant paradigm
- **Generative AI / Agents**
  - Multiple plausible outputs
  - Human preference matters
  - Persuasive ≠ Helpful
  - Good generation ≠ Good decision support
- **History Repeats Itself**
  - **IR has faced this problem for decades**
  - Relevance is inherently subjective
  - The challenge is NOT removing subjectivity
  - The challenge is reproducible evaluation under subjectivity



- We have enjoyed such challenged and overcome with the power of Community

Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park[1]*, Carolyn Q. Zou[1,2], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Robb Willer[6], Percy Liang[1], Michael S. Bernstein[1]

Affiliations:

[1]Computer Science Department, Stanford University; Stanford, CA, 94305, USA.

KEYNOTE 2

Dan Jurafsky

The Social Failures of Language Models as Conversational Partners

Language models are increasingly used in conversation for information, advice, and emotional support. In this talk I'll summarize studies in our lab showing that models fail in systematic ways as social interlocutors. We find that language models are socially sycophantic, linguistically overconfident, overly anthropomorphic, and epistemically self-centered. We then show that these flaws have real consequences for users: people interacting with models suffer consequences including overreliance, distorted judgment, and reduced personal responsibility. I'll discuss datasets and metrics, explore mitigations, and call for design, evaluation, and accountability mechanisms to protect user well-being.

A Possible Direction: Verification-First AI

From Static Metrics to Verifiable Outcomes

Traditional Evaluation	→	Verification-Oriented Evaluation
Accuracy / F1	→	Decision Quality
Preference Scores	→	Human Outcomes
Single-turn Outputs	→	**Longitudinal Tracking**
Static Benchmarks	→	**Real-world Verification**
Persuasiveness	→	**Accountability**
One Correct Answer	→	**Defensible / Auditable Claims**

Human-Agent Ally Era
Verification Will Become the Primary Bottleneck

HAA LAB

Preventing the Collapse of Peer Review Requires Verification-First AI

Lei You[1,2], **Lele Cao**[2], **Iryna Gurevych**[3,4]

[1]Technical University of Denmark, [2]CSPaper @ Scholar7, [3]Technical University Darmstadt, [4]MBZUAI

When claims outpace verification capacity, systems inevitably drift from truth to proxies.

when current decisions still appear reliable. These results motivate actions for tool builders and program chairs: deploy AI as an adversarial auditor that generates auditable verification artifacts and expands effective verification bandwidth, rather than as a score predictor that amplifies claim inflation.

Some Simple Economics of AGI*

Christian Catalini (MIT)
Xiang Hui (WashU)
Jane Wu (UCLA)

February 26, 2026

Extended Abstract

For three hundred thousand years, human cognition was the primary engine of

Execution scales exponentially, but verification remains biologically bottlenecked.

an economy where autonomous agents (L_a) act with broad agency rather than narrow instructions, the binding constraint on growth is no longer intelligence. It is *human verification bandwidth*: the scarce capacity to validate outcomes, audit behavior, and underwrite meaning and responsibility when execution is abundant.

We model the transition toward AGI as the collision of two racing cost curves: an exponentially decaying Cost to Automate (c_A), driven by compute and accumulated knowledge, and a biologically bottlenecked Cost to Verify (c_H), bounded by human time and embodied experience. This structural asymmetry widens a Measurability Gap (Δm) between what agents can execute and what humans can afford to verify, and determines the verifiable share of deployment (s_v) that separates truly productive agentic output from merely *plausible* output. It also drives a shift from skill-biased to *measurability-biased* technical change and a radical bifurcation of economic value: as measurable execution commoditizes toward the marginal cost of compute, rents migrate to what remains scarce—verification-grade ground truth, cryptographic provenance, and liability underwriting (the ability to insure outcomes rather than merely generate them).

We Could Verify the Present, But Not the Future

HAA LAB

- Current verification research focuses on static claims
  - Verifying consistency within a document
  - Verifying consistency across documents
  - Verifying alignment between representations (e.g., paper ↔ code)

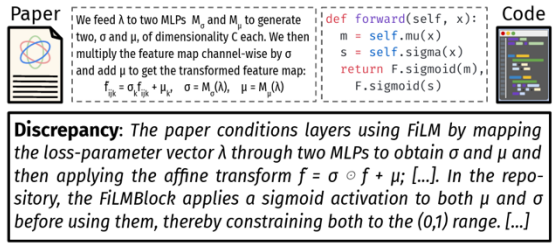
SciCoQA: Quality Assurance for Scientific Paper–Code Alignment

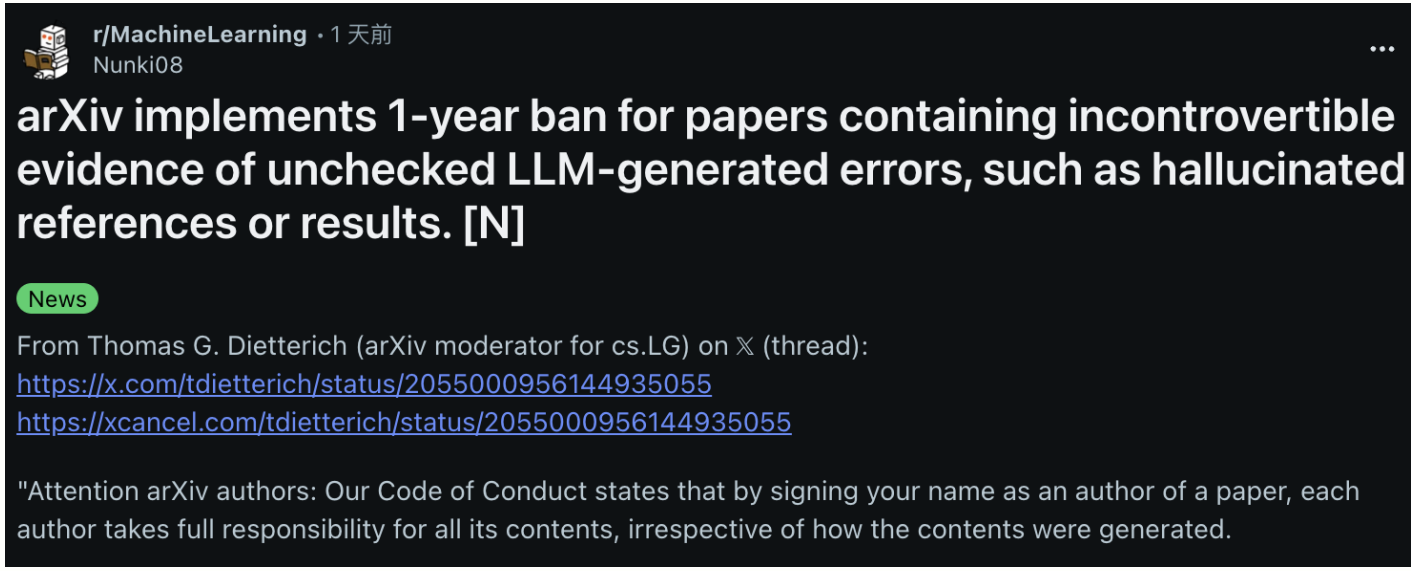
Tim Baumgärtner and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science,
TU Darmstadt and National Research Center for Applied Cybersecurity ATHENE

Code: https://github.com/ukplab/scicoqa
Data: https://hf.co/datasets/ukplab/scicoqa
Blog: https://ukplab.github.io/scicoqa

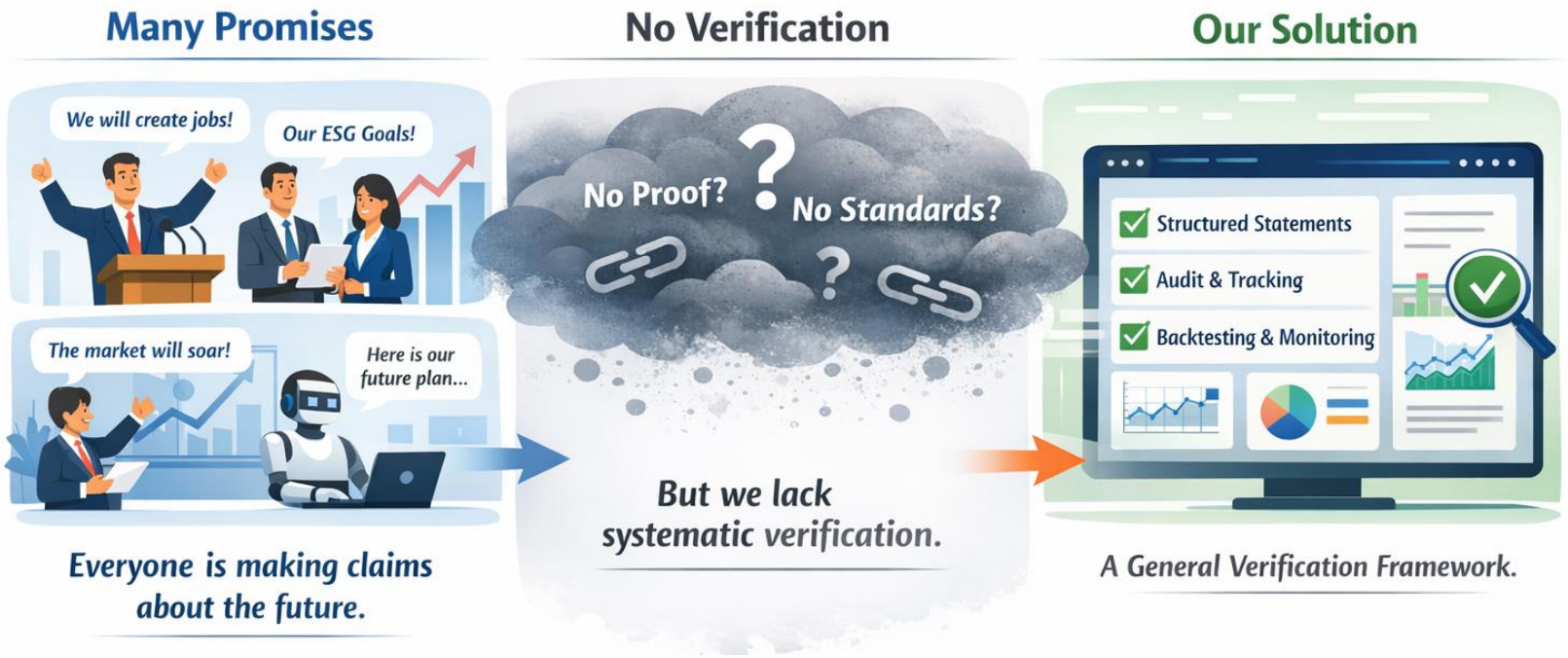
Abstract

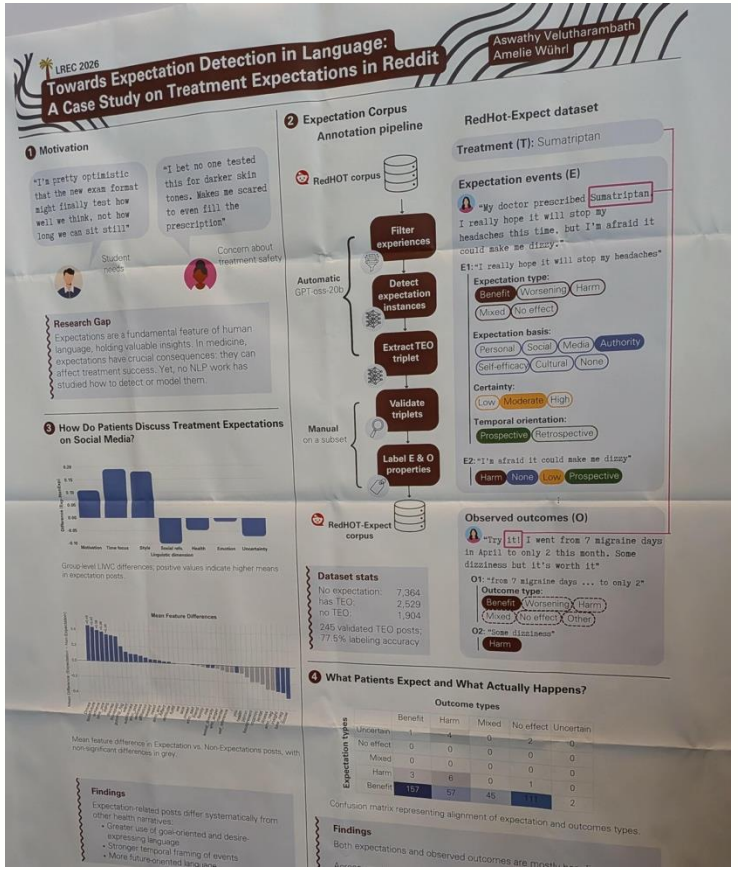
We present SciCoQA, a dataset for detecting *discrepancies* between scientific publications and their codebases to ensure faithful implementations. We construct SciCoQA from GitHub issues and reproducibility papers, and to scale our dataset, we propose a synthetic data





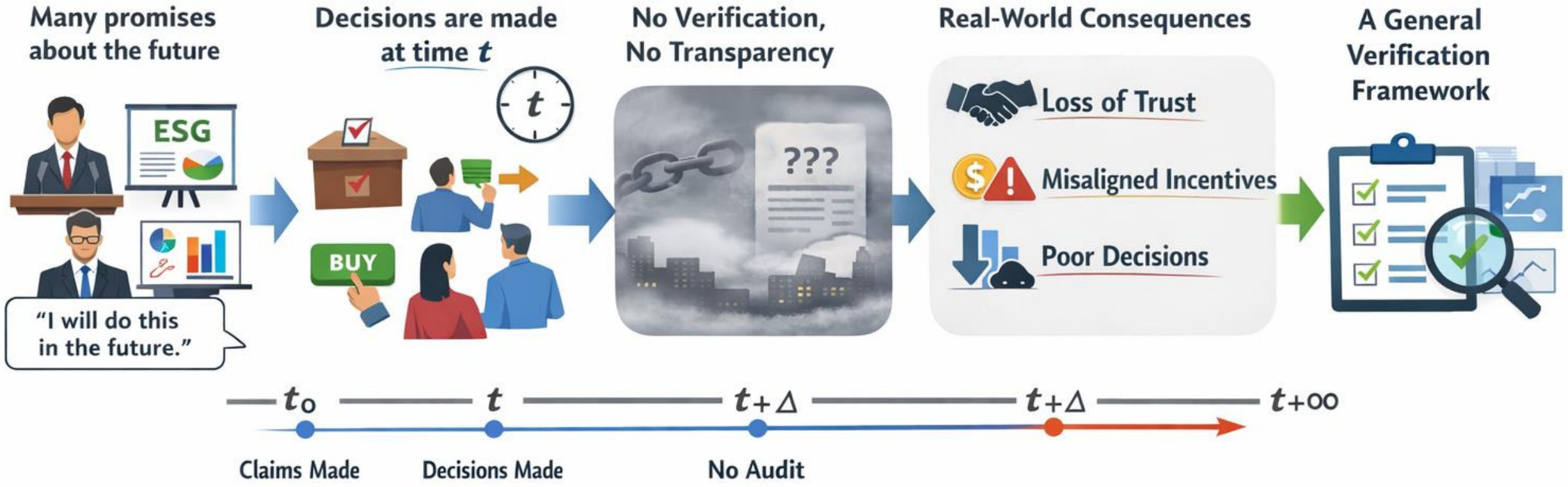
Verifying and Auditing Forward-Looking Statements





The Impact of Forward-Looking Statements

HAA LAB



False Information vs. Failed Commitments

HAA LAB

- **Past/Current**

Type	Is the information accurate?	Does the sender **know it's false**?	Is there malicious **intent**?
Misinformation	❌ No	❌ No	❌ No
Disinformation	❌ No	✅ Yes	✅ Yes

- **Forward-Looking Arguments & Commitments**

Situation	Type	Explanation
Genuinely **tried but failed**	❌ Not mis/disinformation	It's a failure, not deception
Overestimated or **exaggerated**	⚠️ Possibly misinformation	**Misleading**, but not necessarily with bad intent
Knew it was **impossible** from the start	✅ Disinformation	A deceptive **promise** made to gain trust or favor

Our goal is to

- **(Now) Assist writers to avoid potential overstatements as they write**
- **(Now) Help readers assess the strength of reasoning**
- **(Later) Allow for future review and accountability**

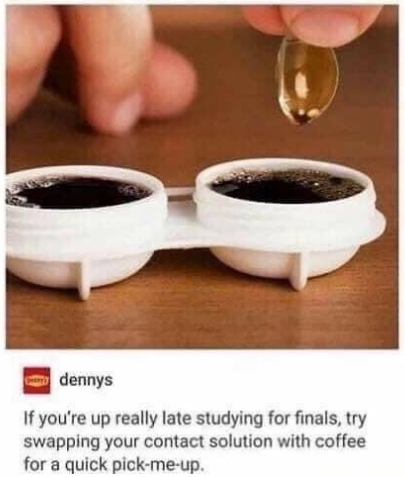
From Society-Undermining Disinformation to Promises

HAA LAB

Type	Example
Society-Undermining Disinformation (Punishable)	Sharing a video of a bank robbery from another country and claiming, "This happened in Taipei."
Disinformation	A company knowingly publishing fake success metrics to attract investors
Misinformation	A relative sharing a false health tip on social media believing it's true
Forward-Looking Scenario (Prediction)	An analyst projecting "20% revenue growth next year" based on **weak evidence**
Corporate Promise	A company pledging carbon neutrality by 2030 with **no actual** implementation

Society-Undermining Disinformation or Misinformation?

Humor or Misinformation?



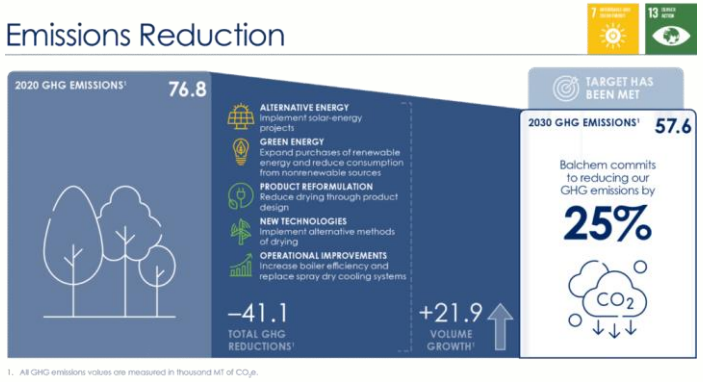
Forward-Looking Scenario

Vornado Realty Trust *(Underweight; Price Target: $40.00)*

Investment Thesis

We maintain our Underweight rating on VNO's shares. Our concerns over the NYC office and street retail markets existed prior to COVID and are now only heightened. We think there is risk of multi-year headwinds to lease economics that will land VNO's growth below that of other REITs. We also believe the company remains more complex than other REITs and carries above-average leverage. Longer-term development and re-development efforts should improve cash flows, though we may be a couple years away from having visibility on the full impact of projects like the PENN district.

Corporate ESG Promise



- These are examples, but it does **not** imply that these are (dis)misinformation.
  - 20220523_JP-Morgan_-Delayed--Vornado-Realty-Trust--Updated-_1.pdf
  - https://balchem.com/responsibility/sustainability/2030-esg-goals/

Intent, Accuracy, and Impact – Challenges

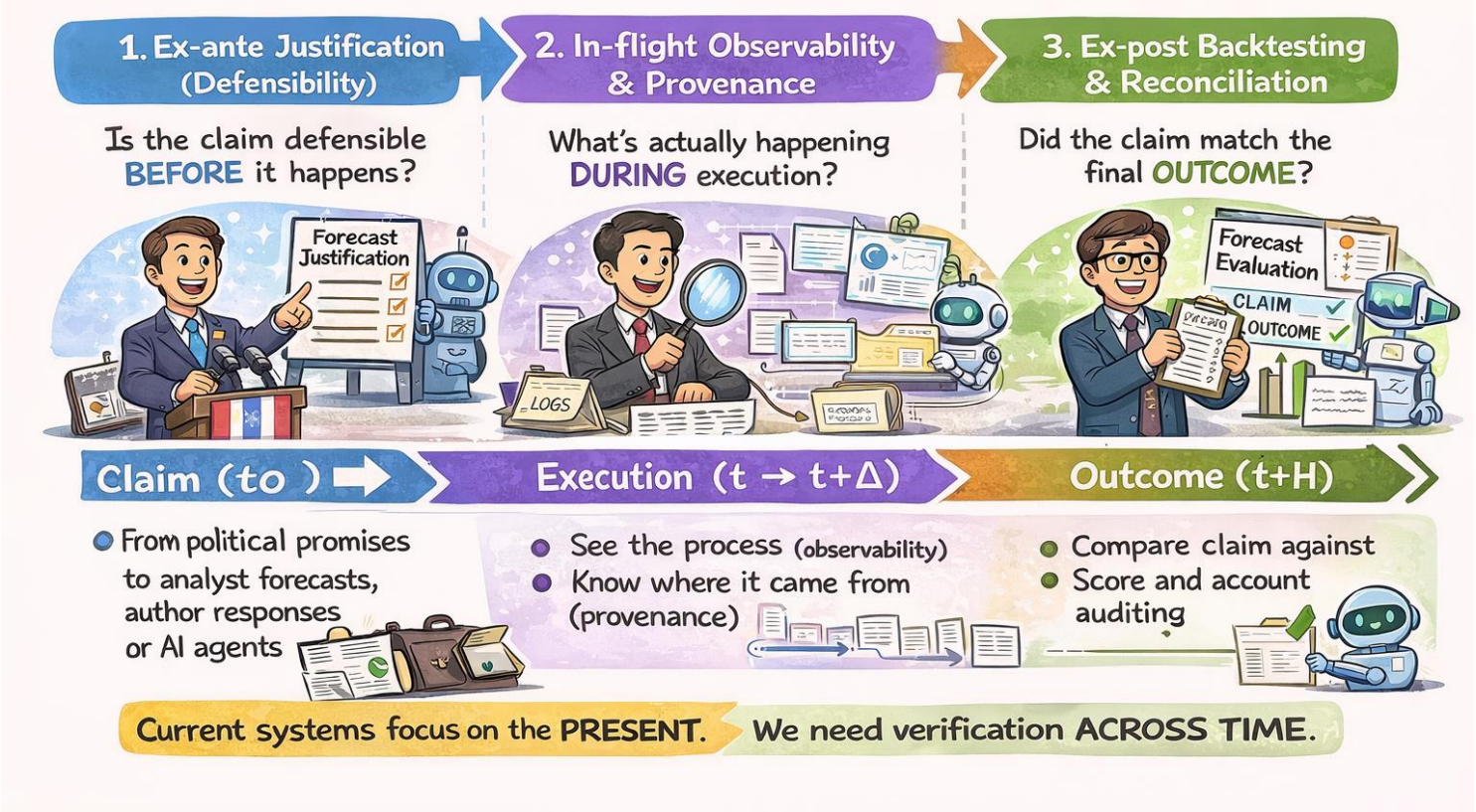
HAA LAB

Type	Intentional?	Factual Accuracy	Legality	Potential Impact
Society-Undermining Disinformation	✅ Yes (Malicious intent) **(Verifiable?)**	❌ False	❌ Illegal (court bans)	🛑 **Undermines Society**
Disinformation	✅ Yes	❌ False	⚠️ Sometimes regulated	📉 Misleads public, distorts perception or policy
Misinformation	❌ No	❌ False	✅ Legal	🤦 Confuses people, spreads unintentionally
Forward-Looking Scenario (Prediction)	❓ Uncertain	🔄 Hypothetical **(Verifiable?)**	✅ Legal	📊 Can **mislead** expectations or markets
Corporate ESG Promise	✅ Often Yes	🔮 Future-Oriented **(Verifiable?)**	⚠️ May fall under greenwashing laws	🌱 May **mislead** stakeholders, reputation & legal risks

Forward-Looking Verification: A Three-Stage Framework

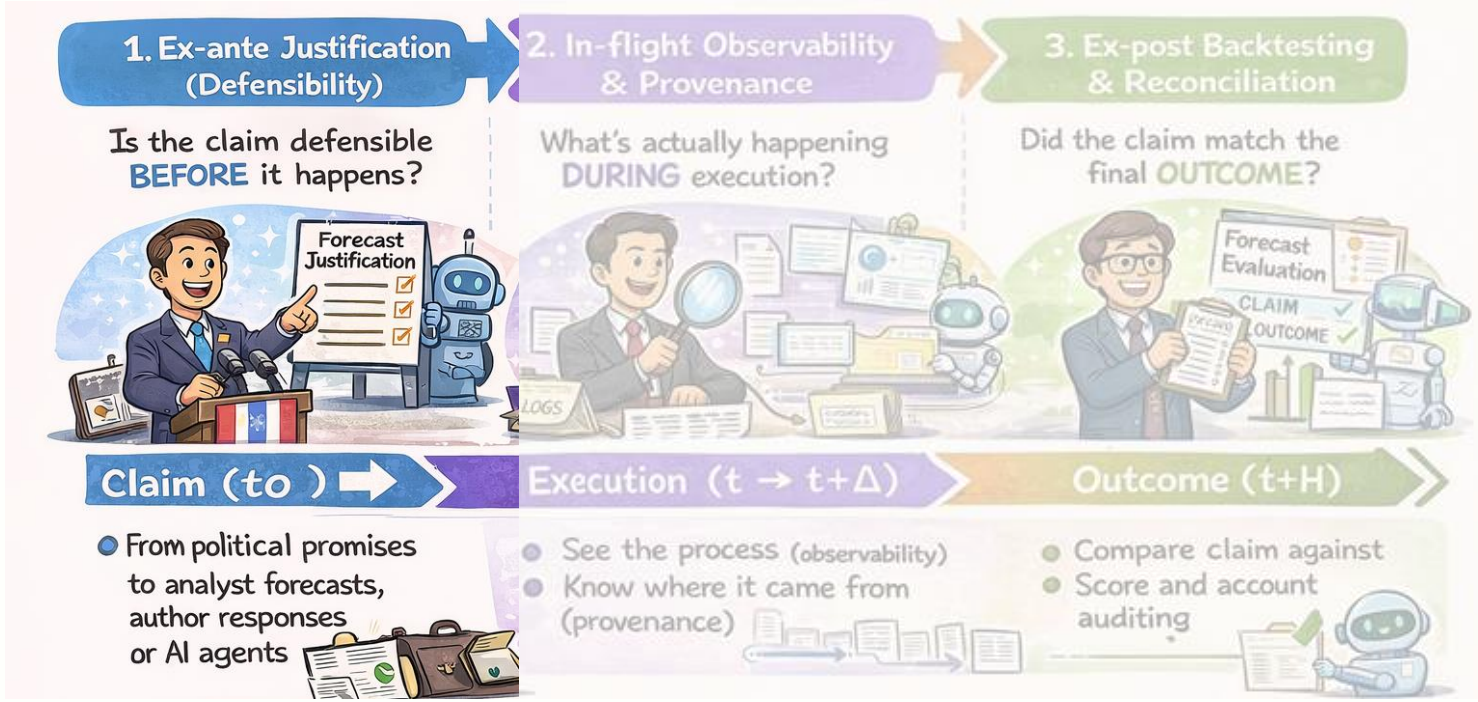
HAA LAB

- **Ex-ante Justification (Defensibility)**
  - Can we justify the claim *before* it happens?
  - Forecast justification / credibility
  - Assumptions, reasoning, plausibility
  - Causal arguments and supporting evidence
- **In-flight Observability & Provenance**
  - Can we track what happens *during execution*?
  - Observability → can we see the process?
  - Provenance → do we know where it comes from?
  - Logs, traces, intermediate decisions
- **Ex-post Backtesting & Reconciliation**
  - Did the claim match what actually happened?
  - claim → wait → outcome → compare
  - Forecast evaluation / scoring rules
  - Accountability and auditing



Ex-ante Justification (Defensibility)

- **Ex-ante Justification (Defensibility)**
  - Can we justify the claim *before* it happens?
  - Forecast justification / credibility
  - Assumptions, reasoning, plausibility
  - Causal arguments and supporting evidence
- **In-flight Observability & Provenance**
  - Can we track what happens *during execution*?
  - Observability → can we see the process?
  - Provenance → do we know where it comes from?
  - Logs, traces, intermediate decisions
- **Ex-post Backtesting & Reconciliation**
  - Did the claim match what actually happened?
  - claim → wait → outcome → compare
  - Forecast evaluation / scoring rules
  - Accountability and auditing



Quality Assessment

HAA LAB

1. **Investment Opinion**
   - **Professionalism: Are the claims reasonable and domain-grounded?**
     - Analysts are expected to base their forward-looking projections on sound reasoning, domain knowledge, and data integrity.
     - ➢ Professionalism-Aware Pre-Finetuning for Profitability Ranking. CIKM-2024
     - ➢ Evaluating the Rationales of Amateur Investors. WWW-2021
   - **Argument Mining: Are the premises valid, and reasoning complete?**
     - Forward-looking statements are often composed of conclusions and their supporting premises — and these can be mined, evaluated, and scored.
     - ➢ Enhancing Investment Opinion Ranking through Argument-Based Sentiment Analysis. AACL-2025
     - ➢ Argument-Based Sentiment Analysis on Forward-Looking Statements. ACL-2024
   - **Exaggerated Information: Is the projection reasonable?**
     - A 10% difference can often lead to significantly different decisions in many situations.
     - ➢ Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments. ACL-2019

2. **ESG Report**
   - ➢ ML-Promise: A Multilingual Dataset for Corporate Promise Verification. EMNLP-2025

The Global Sustainability Commitment Crisis: Systemic Failures from Nations to Corporations

HAA LAB

Germany Likely to Miss 2030 Climate Goal, Council of Experts on Climate Change Says

by ESG News • June 5, 2024 Share:

Features

Fashion brands lagging on living wages will miss 2030 SDG goals

The deadline for global Sustainable Development Goals (SDGs) is 2030, but a new report suggests fashion brands are still behind on the broad implementation of living wages and incomes across their supply chains.

Laura Husband | January 22, 2025

Walmart to Miss 2025, 2030 Climate Targets

Mark Segal | December 20, 2024

Walmart does not expect to hit its interim climate goals, including its targets to reduce operational greenhouse gas (GHG) emissions by 35% by 2025 and by 65% by 2030, according to a new post on the company's website, citing "factors beyond our control," including a lack of low carbon refrigeration and mobility technologies, and clean energy policy and infrastructure.

The retail giant said that it will continue to work towards its "aspirational goal of zero emissions by 2040," but warned that "progress will not be linear."

Walmart set its 2040 net zero goal in 2020, outlining at the time a series of initiatives it would pursue to achieve its goal, including sourcing 100% renewable energy to power its facilities with by 2035, electrifying and zeroing out emissions from all of its vehicles – including long-haul trucks – by 2040, and transitioning to low-impact refrigerants for cooling and electrified equipment for heating in its stores, clubs, and data and distribution centers by 2040.

- https://esgnews.com/germany-likely-to-miss-2030-climate-goal-council-of-experts-on-climate-change-says/
- https://www.just-style.com/features/fashion-brands-lagging-on-living-wages-will-miss-2030-sdg-goals/
- https://www.esgtoday.com/walmart-to-miss-2025-2030-climate-targets/

PromiseEval: Multinational, Multilingual, Multi-Industry Promise Verification (SemEval-2025 Task 6 @ ACL-2025)

HAA LAB

- Broken promises may not be lies — but they can still mislead investors, regulators, and the public

- **Promises are forward-oriented** and often vague.

- We ask:
  - Is this a **promise**?
  - Is there **evidence**?
  - Is the link **clear** or misleading?
  - **When** should this be verified?

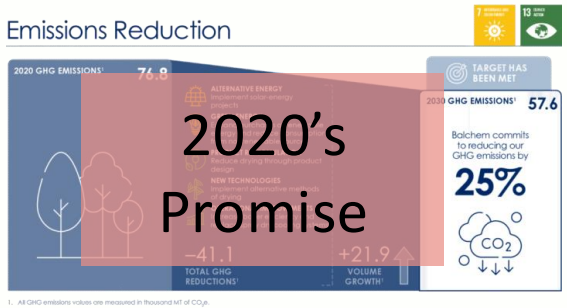
Task	Label	English	French	Chinese	Japanese	Korean
Promise Identification	Yes	755	764	464	898	155
	No	245	236	635	102	45
Actionable Evidence	Yes	549	646	267	621	146
	No	451	354	832	277	47
Clarity of Promise-Evidence Pair	Clear	327	440	147	365	128
	Not Clear	212	197	75	233	7
	Misleading	10	9	1	23	0
	Other	451	354	876	-	-
Timing for Verification	Within 2 years	76	64	187	48	65
	2-5 years	150	166	26	55	12
	Longer than 5 years	105	95	81	104	25
	Other	245	236	805	0	41
	Already	424	439	-	691	-

- Dataset
  - 5 Languages: English, French, Chinese, Japanese, Korean
  - 8+ Industries: Energy, Finance, Technology, Luxury, Biomedical...
  - 12+ Countries: UK, US, France, Canada, Taiwan, Japan, Korea...

Modeling Results & Next Steps

Subtask	Best Approaches	F1 (English)
Promise	GPT-4o + Data Augmentation	0.823
Evidence	BERT-based + Multilingual Ensembles	0.787
Clarity (Still Challenging)	GPT-4o (zero-shot + 6-shot)	0.669
Timing (Still Challenging)	Universal Embedding + Contrastive loss	0.577

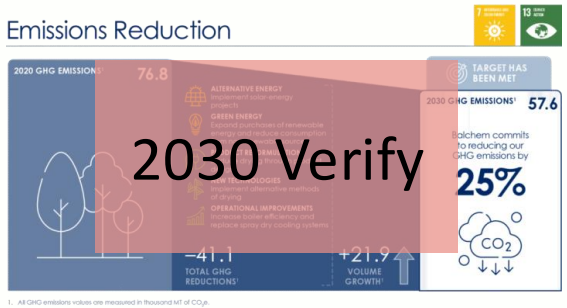
- **Greenwashing Risk**: Detect vague, feel-good claims that lack concrete support (argument mining)
- **Stakeholder Impact**: Assess who actually benefits from the promise (and how) (Intent)
- Scenario Verification, **Now with Promise**
  - Can we retrieve updated reports and check whether there's any trace of follow-up action?



2024 Identify Actions

balchem

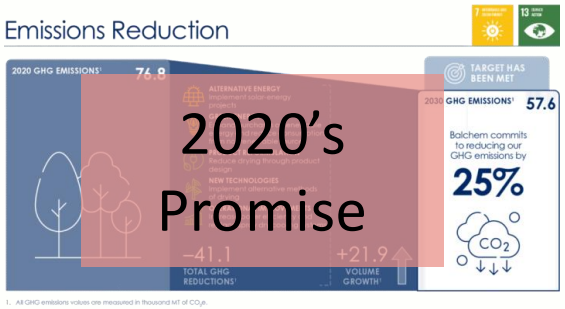
Our 2024 Sustainability Report



In-flight Observability & Provenance

- **Ex-ante Justification (Defensibility)**
  - Can we justify the claim *before* it happens?
  - Forecast justification / credibility
  - Assumptions, reasoning, plausibility
  - Causal arguments and supporting evidence
- **In-flight Observability & Provenance**
  - Can we track what happens *during execution*?
  - Observability → can we see the process?
  - Provenance → do we know where it comes from?
  - Logs, traces, intermediate decisions
- **Ex-post Backtesting & Reconciliation**
  - Did the claim match what actually happened?
  - claim → wait → outcome → compare
  - Forecast evaluation / scoring rules
  - Accountability and auditing

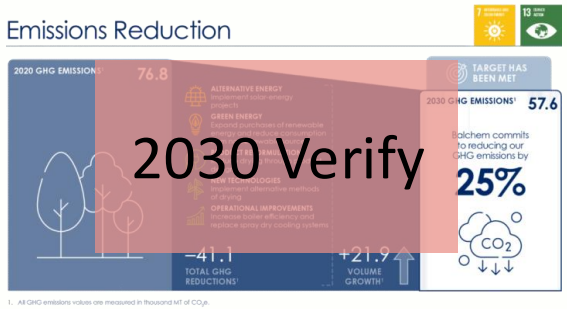
Make sure we can Track and Verify them



2024 Identify Actions

balchem
Solve Today. Shape Tomorrow.

Our 2024 **Sustainability Report**



- **Broken promises may not be lies — but they can still mislead investors, regulators, and the public**
- **Promises are forward-oriented** and often vague.

- **We ask in PromiseEval-2025:**
  - Is this a **promise**?
  - Is there **evidence**?
  - Is the link **clear** or misleading?
  - **When** should this be verified?

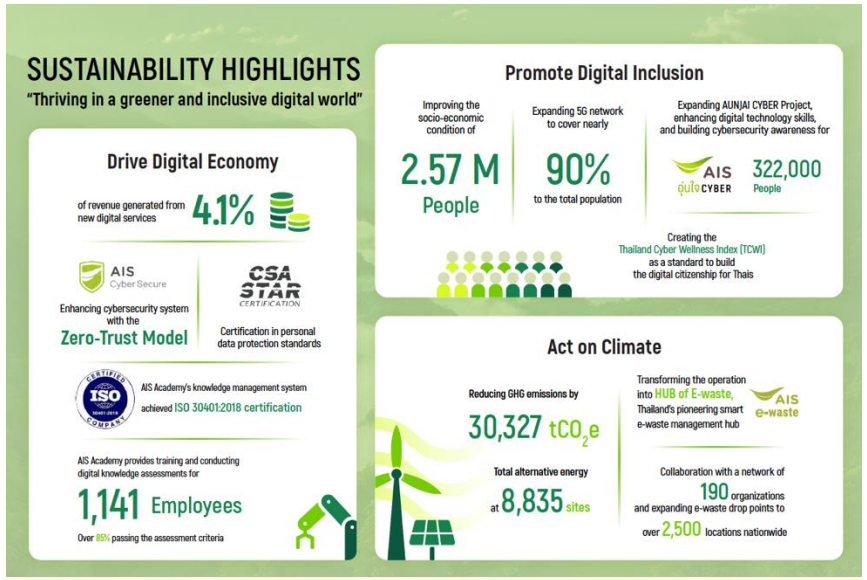
We ask in RegCom-2026:

- **Whether companies write down actions or achievements for tracing?**
- RegCom: Multinational, Multilingual, Multi-Industry Compliance Checking (NTCIR-2026)
  - http://regcom.nlpfin.com/

RegCom: Multinational, Multilingual, Multi-Industry Compliance Checking (NTCIR-2026)

HAA LAB

- **Whether companies write down actions or achievements for tracing?**
- **Cross-Domain Task ➔** Finance & Legal (Regulations)
- **Multilingual ➔** English, French, Korean, Chinese, Japanese, Thai
- **Multinational ➔** UK, USA, Jordan, Sonth Africa, Switzerland, Canada, France, Luxembourg, Taiwan, Japan, South Korea, Thailand, Australia
- **Multi-Industry ➔** Energy, Finance, Luxury, Semiconductor, Technology, Biomedical, Automotive, Trading



SUSTAINABILITY DISCLOSURE TOPICS & METRICS

Table 1. Sustainability Disclosure Topics & Metrics

TOPIC	METRIC	CATEGORY	UNIT OF MEASURE
Transparent Information & Fair Advice for Customers	(1) Number and (2) percentage of licensed employees and identified decision-makers with a record of investment-related investigations, consumer-initiated complaints, private civil litigations, or other regulatory proceedings [1]	Quantitative	Number, Percentage (%)
	Total amount of monetary losses as a result of legal proceedings associated with marketing and communication of financial product-related information to new and returning customers [2]	Quantitative	Presentation currency
	Description of approach to informing customers about products and services	Discussion and Analysis	n/a

Ex-post Backtesting & Reconciliation

- **Ex-ante Justification (Defensibility)**
  - Can we justify the claim *before* it happens?
  - Forecast justification / credibility
  - Assumptions, reasoning, plausibility
  - Causal arguments and supporting evidence
- **In-flight Observability & Provenance**
  - Can we track what happens *during execution*?
  - Observability → can we see the process?
  - Provenance → do we know where it comes from?
  - Logs, traces, intermediate decisions
- **Ex-post Backtesting & Reconciliation**
  - Did the claim match what actually happened?
  - claim → wait → outcome → compare
  - Forecast evaluation / scoring rules
  - Accountability and auditing

2027 → Cross-Temporal Cross-Report Multinational, Multilingual, Multi-Industry Promise Verification

HAA LAB

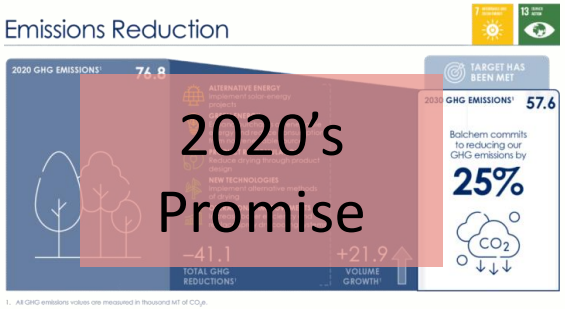
- **Alignment Between Actions and Promises**
  - Which **Promise** from the previous reports is supported by the **Actions (RegCom-2026)** mentioned in the new report?
  - To what extent has this Promise been **fulfilled**?
  - What is the **level of achievement**?
  - What **follow-up actions** are planned?

- **Review of Past Promises**
  - For each **past Promise**, has it reached the stage where it can be **verified**?
  - Has the Promise been **achieved**? If so, to what degree?
  - Is the fulfillment of the Promise **explicitly mentioned** in the new report?

Sustainability Reports

- 2024 Sustainability Report
- 2023 Sustainability Report
- 2022 Sustainability Report
- 2021 Sustainability Report
- 2013 Corporate Social Responsibility Report
- 2020 Corporate Social Responsibility Report
- 2019 Corporate Social Responsibility Report
- 2018 Corporate Social Responsibility Report
- 2017 Corporate Social Responsibility Report

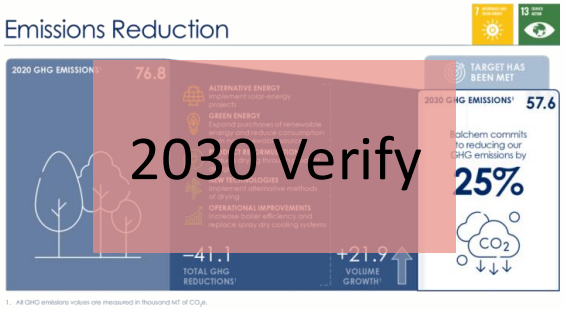
Single Document



2024 Identify Actions

balchem

Our 2024 Sustainability Report



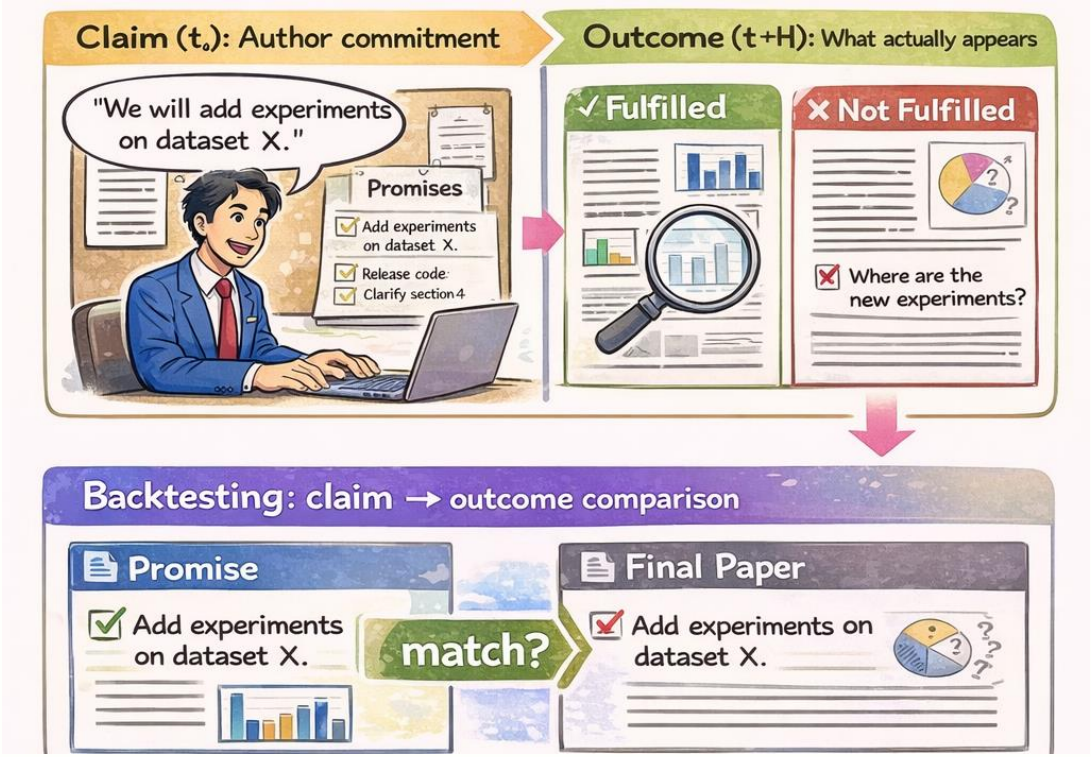
Cross-Document – Auditing Author Commitments in Peer Review

HAA LAB

Commitment Checklist: Auditing Author Commitments in Peer Review

Chung-Chi Chen[1], Iryna Gurevych[2]

[1]AIST, Japan
[2]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science,
TU Darmstadt and National Research Center for Applied Cybersecurity ATHENE, Germany



r/MachineLearning · 1 天前
Nunki08

arXiv implements 1-year ban for papers containing incontrovertible evidence of unchecked LLM-generated errors, such as hallucinated references or results. [N]

News

From Thomas G. Dietterich (arXiv moderator for cs.LG) on 𝕏 (thread):
https://x.com/tdietterich/status/2055000956144935055
https://xcancel.com/tdietterich/status/2055000956144935055

"Attention arXiv authors: Our Code of Conduct states that by signing your name as an author of a paper, each author takes full responsibility for all its contents, irrespective of how the contents were generated.

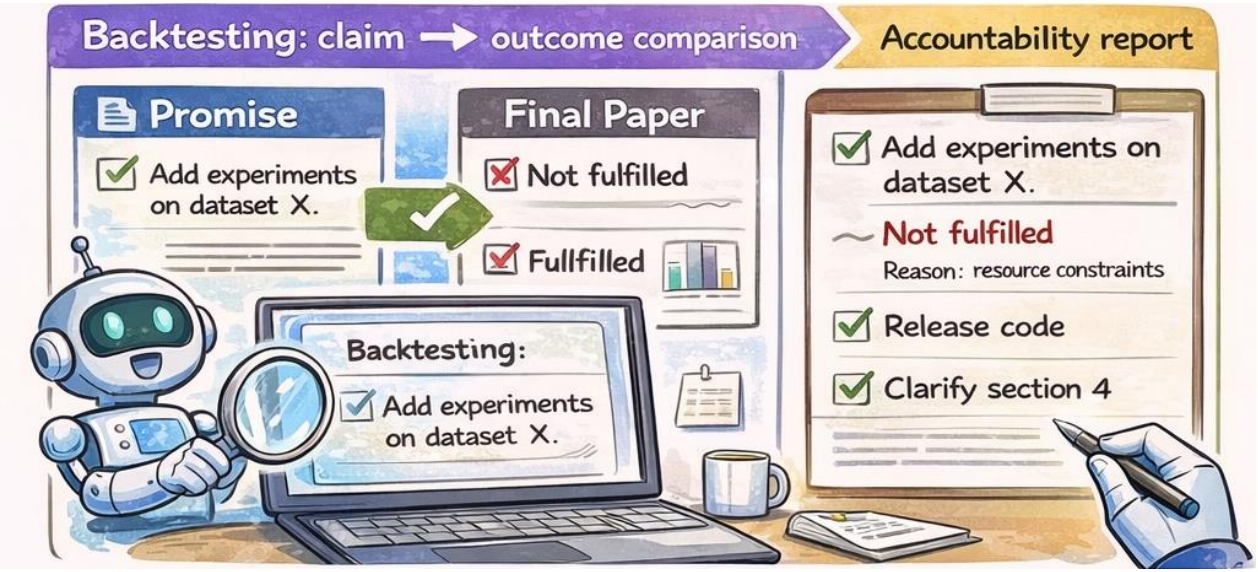
Toward Natural-Language-Based Verification

HAA LAB

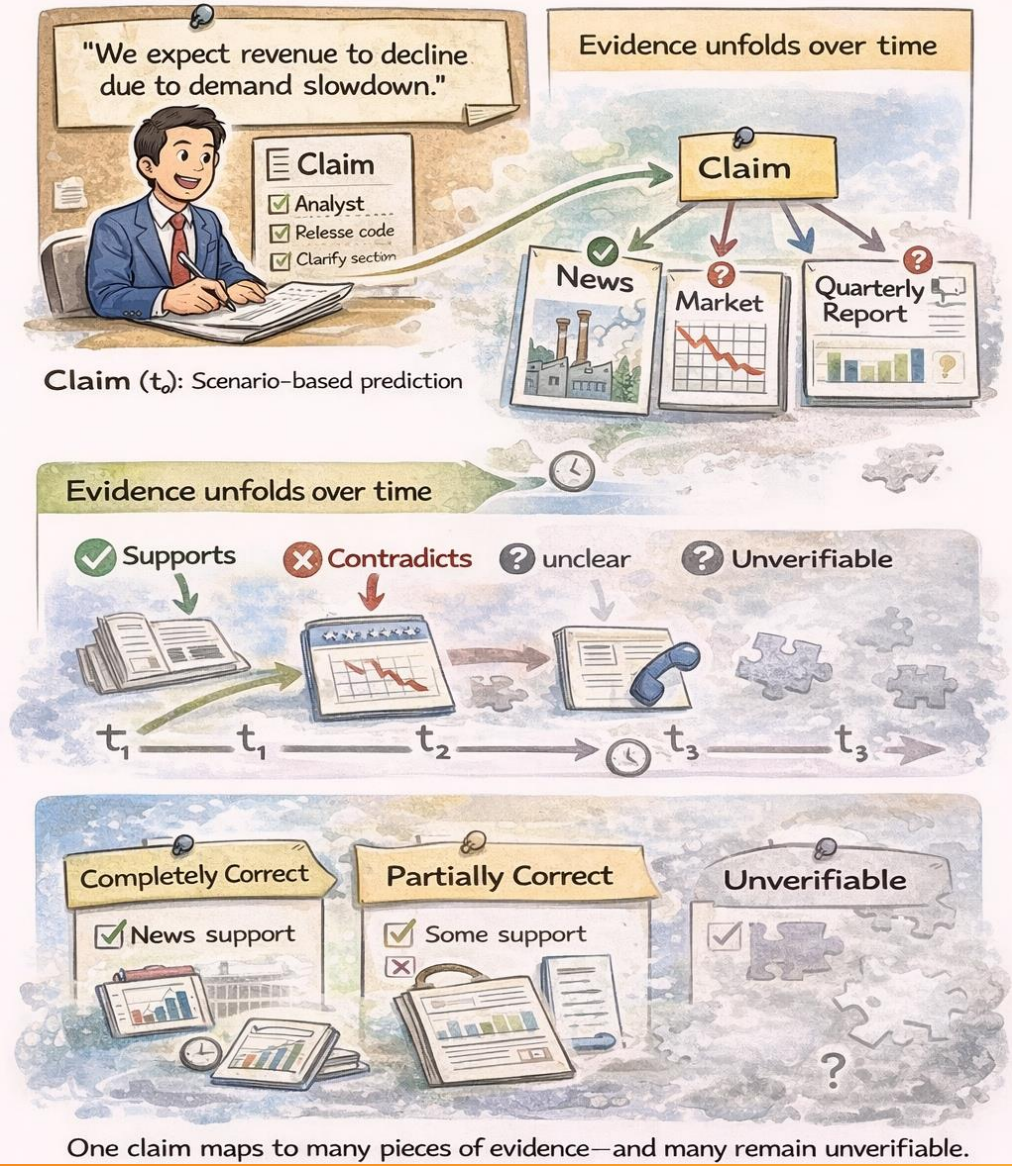
- **We Can Now Verify Natural-Language Commitments**
- Authors make many commitments (up to ~10+ per paper)
- ~25% of commitments are not fulfilled
- Missing experiments are among the most common failures
- LLM-based verification achieves ~75% agreement with humans

	ICLR 2025	EMNLP 2024
Total papers collected	7,192	1,050
Total commitments extracted	84,739	4,320
Avg. commitments per paper	11.78	4.11
Median commitments	11	3
Max commitments (single paper)	59	21

	Promise Identification	Commitment Verification
Avg. input tokens	2,875.82	5,384.24
Avg. output tokens	713.56	525.27
Input cost (USD)	0.0036	0.0108
Output cost (USD)	0.0071	0.0063
Total cost per paper: $0.0278 USD		



Another Perspective: Verifying Analyst Forward-Looking Statements



- **Human Annotation – Given Scenario (English), Find Evidence on the Web**
  - Can Verify
    - 51.45% Correct, 10.14% Incorrect
  - Cannot Verify: **38.41%**

- **Automatic Approach**
  - It's easier to find **supporting news** than disconfirming evidence

- **Open-World Retrieval** (Grounding Agents)
  - GPT-4o Grounding Agent
  - Only **around 22%**

We Can Verify — But Not Everything Can Be Verified

HAA LAB

- **Verification Is Becoming Feasible**
  - LLM + retrieval enables **end-to-end verification workflows**
- Human–AI collaboration:
  - improves evidence coverage
  - significantly reduces verification cost
- Makes **large-scale verification practical**

- **The Real Bottleneck Is NOT Models**
  - The main limitation is **missing evidence**
    - Evidence is:
      - delayed
      - incomplete
      - sometimes never disclosed

- **Implication:** We Need Self-Tracking & Disclosure
  - Verification cannot rely only on external observers
  - Analysts / organizations should:
    - track their own claims
    - disclose outcomes over time
    - report unmet assumptions

Verification Workflow	Avg. Time (min/scenario)	Speed-up vs. Human
Human-involved Workflows		
Human	12.89	–
w/ AI Browser	6.50	~50% faster
w/ LLM Web UI	1.11	~91% faster
Fully Automated Pipelines		
LLM (Closed-Book)	0.57	~95% faster
Grounded LLM (Retrieval-Augmented)	3.34	~74% faster

Harness alone is not enough.
The bottleneck is no longer intelligence — it is **institutional design**.

From Promises to Accountability
Toward Verifiable Forward-Looking Systems

HAA LAB

1. Forward-looking statements are everywhere

- Political promises
- Corporate ESG commitments
- Analyst scenarios
- These shape real decisions — today

2. But verification is fundamentally missing

- Existing systems focus on **static, present claims**
- Forward-looking claims:
- unfold over time
- lack immediate ground truth
- are often partially or never verifiable

3. We are entering a new era

- LLM + retrieval enable:
- claim extraction
- cross-document tracking
- ex-post verification workflows
- Natural-language verification is now feasible

4. But the real limitation is structural

- Many claims remain **unverifiable**
- Not because models fail
- But because **evidence is missing or never disclosed**

5. The key shift

Verification is not just a modeling problem
— it is a system design problem

6. Call to action

- Move from:
  - one-time statements

    → to

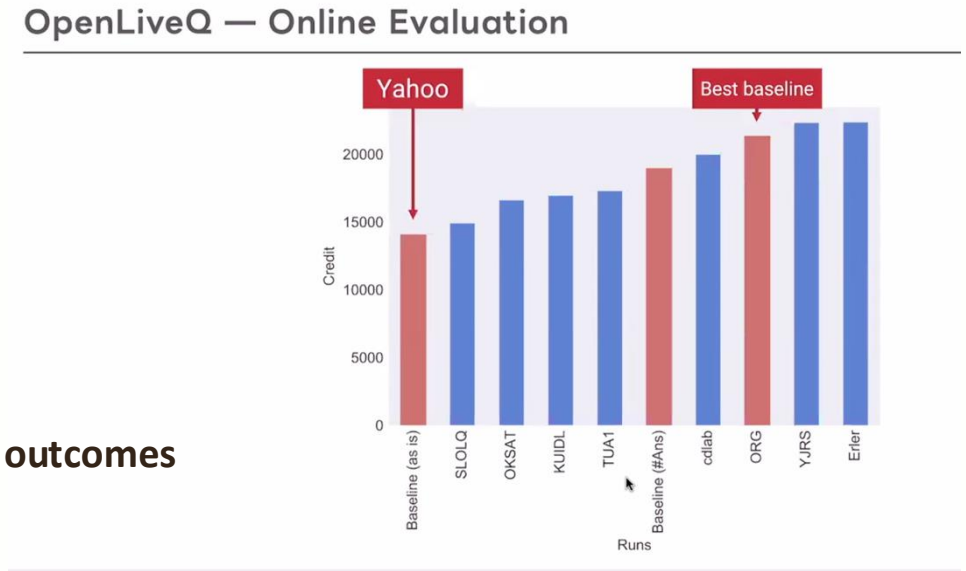
  - **trackable, auditable commitments**
- Require:
  - self-tracking
  - transparency
  - follow-up disclosure

**If the future cannot be immediately verified,
we must design systems that make it accountable.**

Conclusion – From Benchmarks to Verifiable Human-Agent Systems

HAA LAB

- **NTCIR Is More Than Benchmark Collection**
  - Tasks create long-term research directions
  - Communities grow through repeated collaboration
  - **Human-Human Teaming remains foundational in the Agent Era**
- **Evaluation Is Returning to Subjectivity**
  - High benchmark scores do NOT guarantee human benefit
  - Persuasive ≠ Helpful
  - Static metrics are insufficient for Human-Agent systems
  - **Evaluation must consider decision quality, interaction, and long-term outcomes**
- **Verification Will Become the Next Bottleneck**
  - Execution scales faster than verification
  - Future-oriented claims require:
    - justification
    - observability
    - backtesting
  - **The challenge is no longer only modeling but designing verifiable systems**



The challenge is no longer whether AI can do it — but whether our institutions are ready to support it.

Thank You!

HAA LAB

Let's Build the Next Generation of Benchmarks Together

We are building systems for verifying forward-looking statements.

PhD & Interns Welcome

Upcoming Events

- **Workshop** FinNLP @ EMNLP-2026 finnlp.nlpfin.com
- **Shared Task** NTCIR-2026 FinArg-3 finarg.nlpfin.com | RegCom regcom.nlpfin.com

Join us @haalab.github.io