IJCNLP-AACL 2025

Human–Agent Teaming for Higher-Order Thinking Augmentation

HAA LAB

Chung-Chi Chen

Human-Agent Ally Lab (HAA Lab)

https://haalab.github.io/

14:00 - 17:30

AIRC AIST

Chung-Chi Chen

HAA LAB

- **Researcher, AIRC, AIST, Japan**
- **Assistant Professor, National Institute of Informatics (from 2026/04)**
- **Founder of ACL SIG-FinTech**

- **IR/NLP in High-Stakes Scenarios**
  - **Analysis Generation**
    - Professional Report Generation
    - Scenario Planning
  - **Evaluation**
    - Decision-Oriented Evaluation
  - **Numeracy**
    - Numeral Understanding & Reasoning
  - **FinTech**
    - Investor Education
    - Multilingual ESG
  - **LegalTech**
    - Dis(Mis)information Detection
    - Compliance Checking

Internship & PhD Opportunities Available – Join Us!

https://haalab.github.io/

- **Tutorial**
  - AACL-2020 – NLP in FinTech Applications
  - EMNLP-2021 – Financial Opinion Mining
  - ECAI-2024 – Agent AI for Finance: From Financial Argument Mining to Agent-Based Modeling
  - SIGIR-2025 – Information Retrieval in Finance: Industry and Academic Perspectives on Innovation
  - AACL-2025 – Human-Agent Teaming for Higher-Order Thinking Augmentation
- **Organizer:**
  FinNLP Workshop, EMNLP & IJCAI (2019 – present)
  FinArg & FinNum Shared Task, NTCIR (2019 – 2026)
  AgentScen Workshop, IJCAI (2024 – present)
  Program Co-Chair, NTCIR (2024-2026)
  NumEval, PromiseEval @ SemEval-2024 & 2025
  FinWeb Workshop, The Web Conference (2021 – 2023)
  Argument Mining Workshop @ EMNLP-2023
- **Award**
  - SIGIR Early Career Researcher Award
  - Outstanding Paper Award in ANLP, Japan (2%, 15/765)
  - TAAI Thesis Award
  - ACLCLP Thesis Award
  - 1st in Legal-Tech Hackathon organized by Lawsnote, 2021
  - 1st in FinTech Hackathon organized by Microsoft and Jih Sun Securities, 2019
  - 1st in FinTech competition organized by Standard Chartered, 2018

Human-Agent Ally Lab

HAA LAB

Computer Science

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.

Finance

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.

Economics

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.

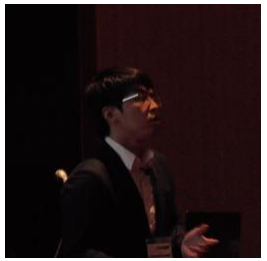
Impact Analysis

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

Behavioral Economics Research in the Era of Human–Agent Societies

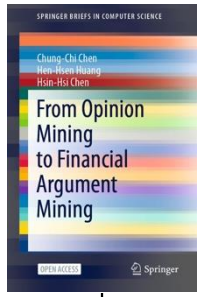
Internship & PhD Opportunities Available – Join Us! https://haalab.github.io/

Rethinking NLP: From Mining to Teaming



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

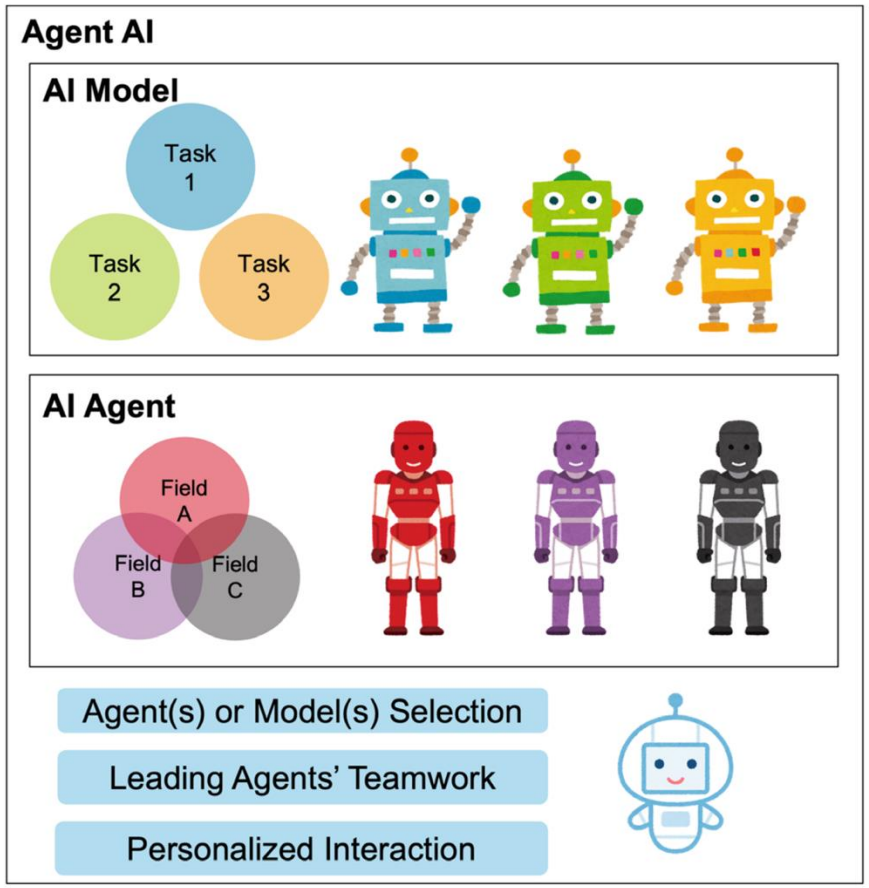
2025
AACL
Tutorial

Evaluation would go beyond accuracy & speed

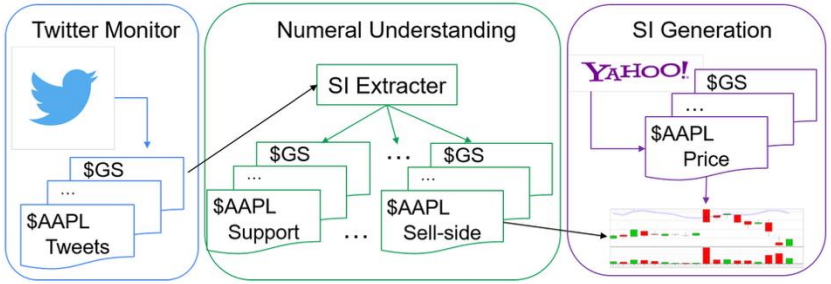
The extent to which the system benefits user/human matters

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

HAA LAB

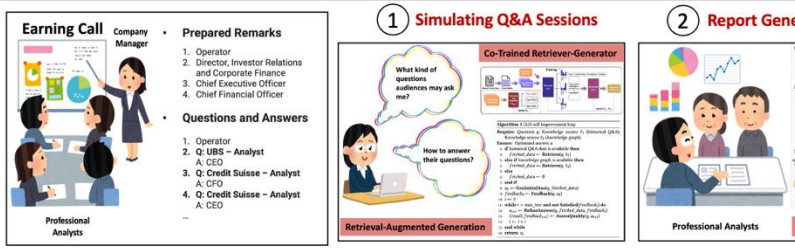


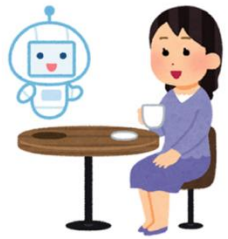
From Data to Signals
(Information Extraction)



AI as Tool

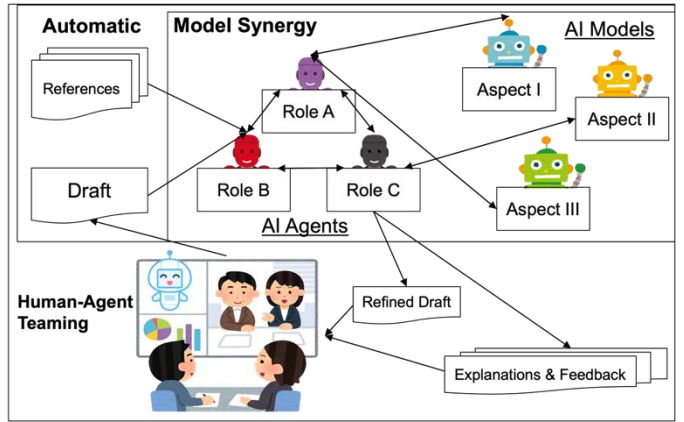
From Signals to Insights
(Human-AI Interaction)





AI as Partner

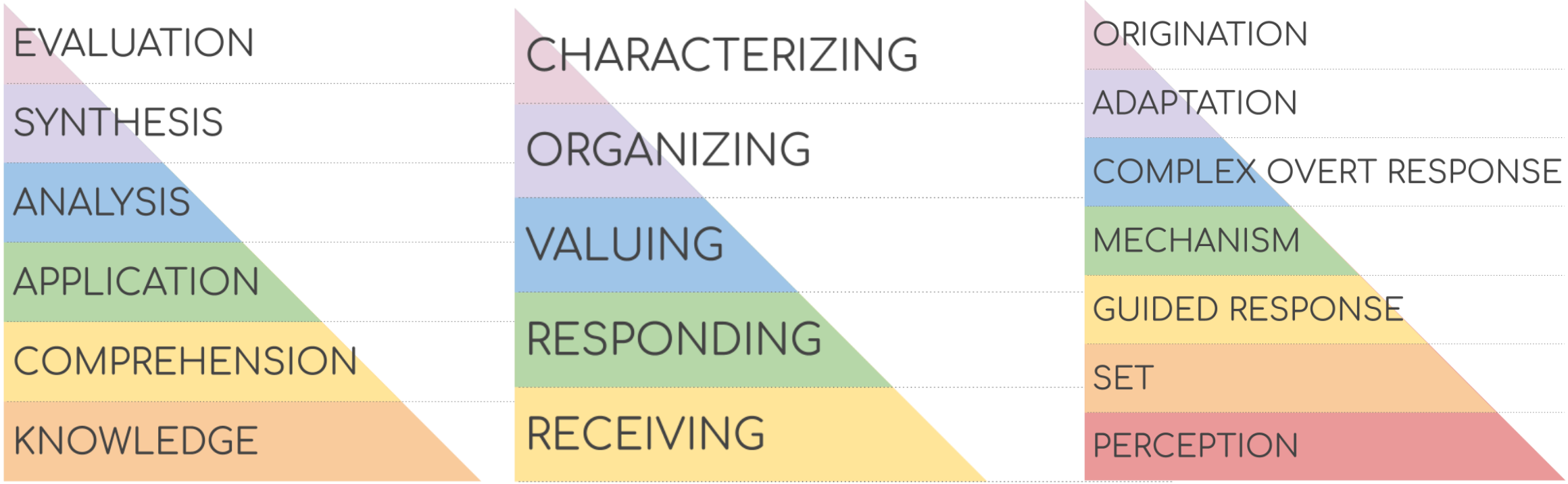
From Insights to Partnership
(Human-Agent Teaming)



Outline

- **Overview**
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- **Scenarios (Interaction & Evaluation)**
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - Agent-Based Modeling (Simulation)
- **Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)**
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- **Proposal: Open Agent Platform**

Higher-Order Thinking

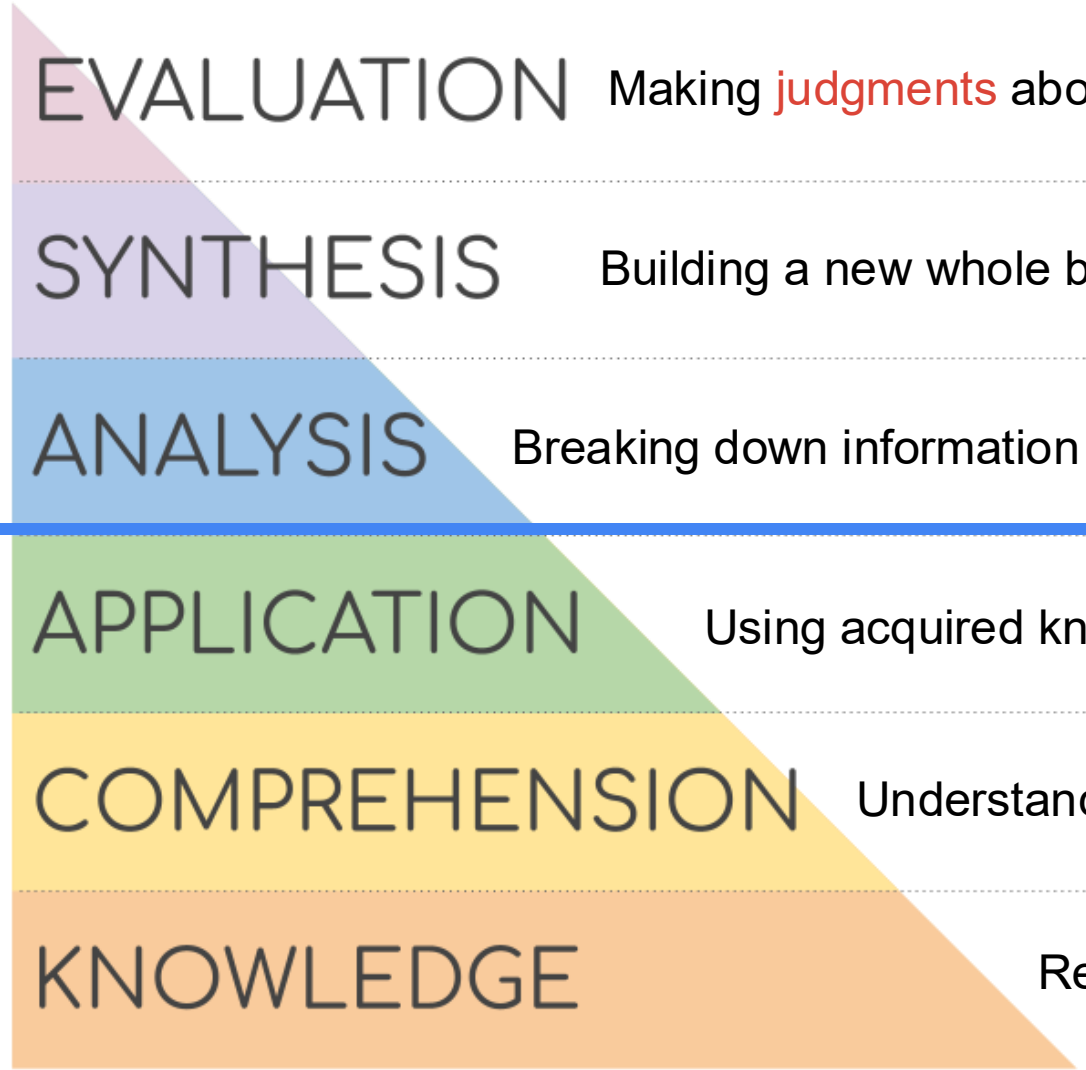


Cognitive (knowledge-based) **Affective (emotion-based)** **Psychomotor (action-based)**

Bloom's Taxonomy (Bloom, 1956)

HAA LAB

Higher-Order Thinking



Making judgments about information, based on set criteria or standards

Building a new whole by combining elements or creating new meaning

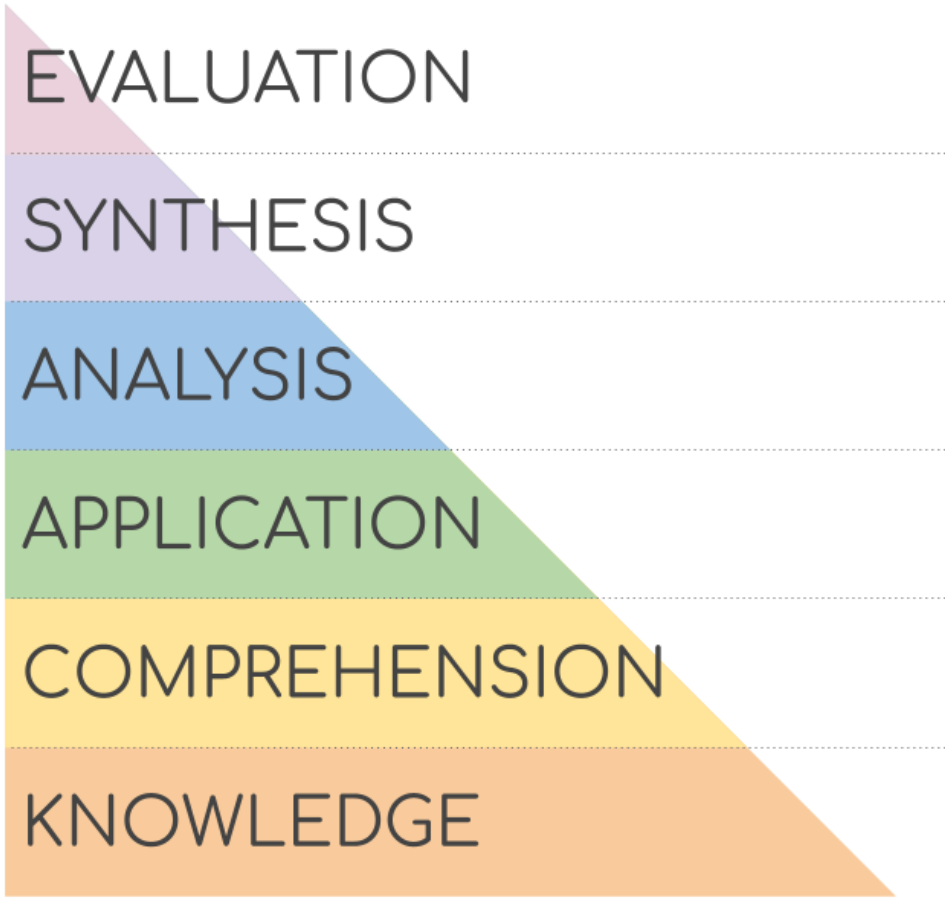
Breaking down information into parts to understand relationships, motives, or causes

Using acquired knowledge to solve problems in new or unfamiliar situations

Understanding of facts by organizing and summarizing information

Recognizing or recalling facts, terms, basic concepts

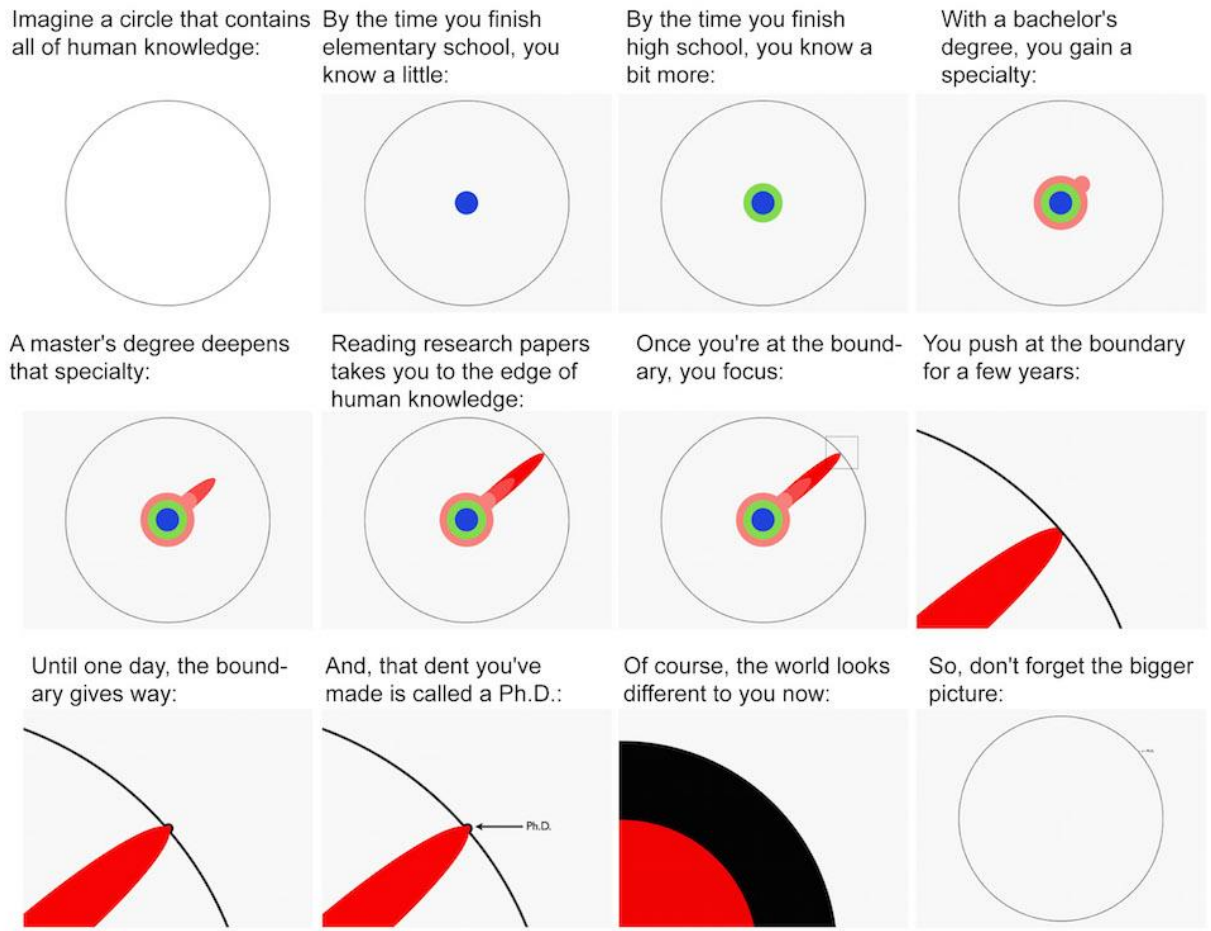
From Understanding to Pushing the Boundary



Matthew Might

Hugh Kaul Precision Medicine Institute, University of Alabama at Birmingham
在 uab.edu 的電子郵件地址已通過驗證 - 首頁

Precision Medicine Genetics Programming Languages Static Analysis
Functional Programming



Essential Research for a Post-AGI or Non-AGI Future — *Some Personal Thoughts*

The only constant is Change

HAA LAB

- **AI will replace nearly all human jobs** within 20 years
- Disruption will be faster than expected
- Only a few jobs will temporarily survive
- Outcome depends on how society responds
- Urgent need to rethink systems and values now

Dorr, 48, is a technology theorist with a PhD in public affairs from the University of California, Los Angeles, and is the director of research at RethinkX, a US-registered nonprofit that analyses and forecasts technological disruption. It was founded and is largely funded by James Arbib and Tony Seba, technology entrepreneurs and investors.

- Iceman ➔ Refrigeration
- Switchboard operators ➔ Switching System
- Assembly line workers ➔ Robotics
- Film delivers ➔ Digital Photography
- Software developers ➔ Prompt engineers (for Generative AI)
- Prompt engineers ➔ Generative AI systems

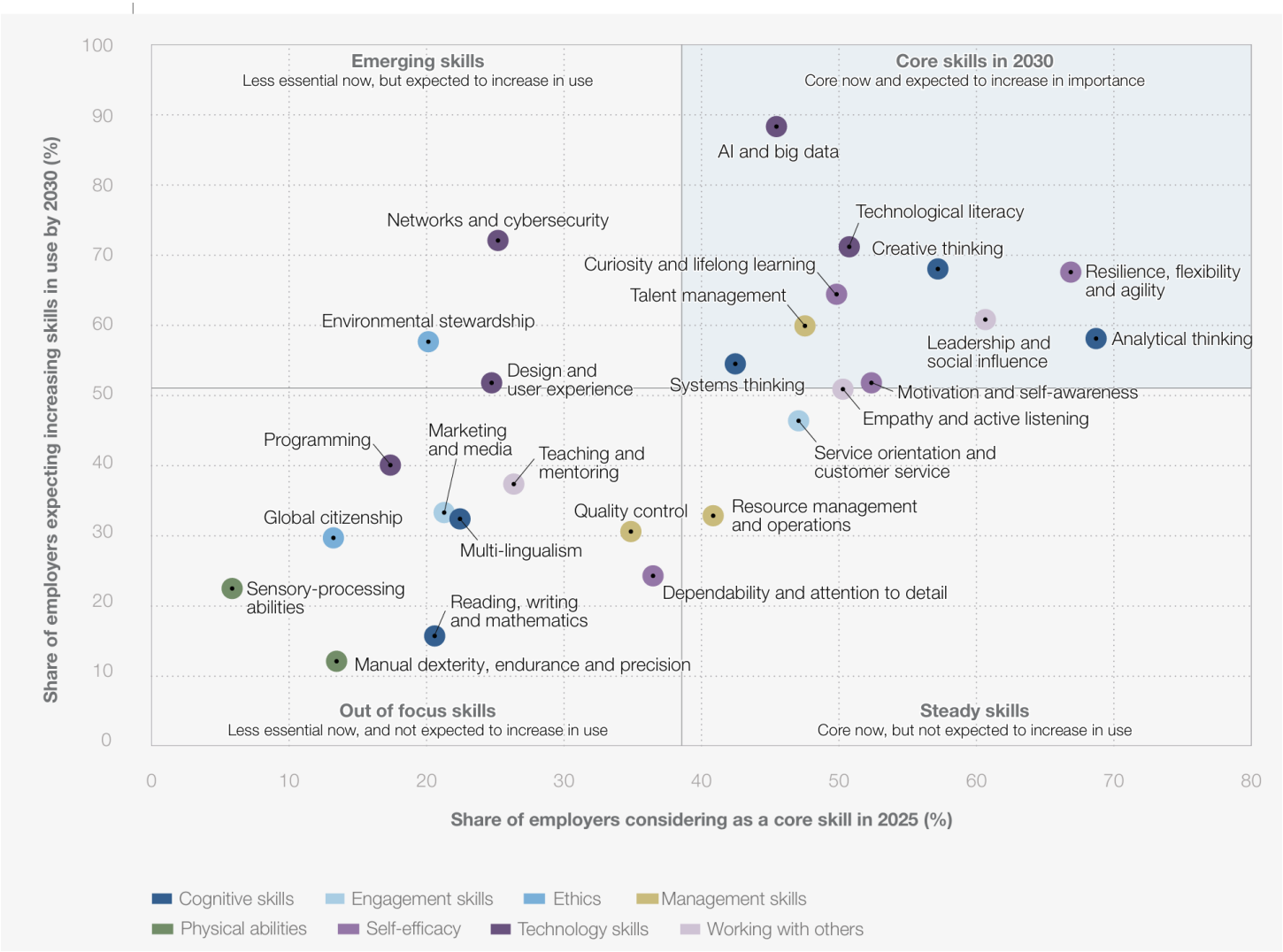
SIGIR 2025
Padova
ITALY

Ophir Frieder

Georgetown University, Washington DC (USA)

Which role is temporary, and which skill is lasting?

Core Skills in 2030 (The Future of Jobs Report 2025)



Skill	AI Replaceability
Analytical & Systems Thinking	Low
Creativity & Innovation	Low
Leadership & Social Influence	Low
Technical Execution	High
Data Processing & Entry	Very High

How AI can assist human thinking?

Is the Current (Search) System *Really* Helping Humans Think?

HAA LAB

- For decades, **IQ scores** steadily increased (the Flynn Effect) — but in recent years, they **have started to decline in wealthy countries.**
- One possible reason? **Education reforms that prioritize "critical thinking" and "learning how to learn," while downplaying basic knowledge and memory training.**
- Rising reliance on digital tools (AI, smartphones, search engines) leads to reduced internal memory use
- **Outsourcing thinking = Weakened brain structures** (less schema, less procedural fluency)
- Memory and knowledge are essential for critical thinking, creativity, and problem-solving
- **"Knowing where to find it" ≠ "Knowing it" — constant lookup habits don't build understanding**
- Metacognitive laziness: Students using AI tools often learn *less*, not more
- Without stored knowledge, the brain can't detect errors or connect ideas
- Tech should augment cognition, not replace it — internal learning must come first
- "An offloaded mind may become an under-exercised mind"

The Memory Paradox:
Why Our Brains Need Knowledge in an Age of AI

[1] Barbara Oakley, Oakland University, oakley@oakland.edu
[2] Michael Johnston, New Zealand Initiative, michael.johnston@nzinitiative.org.nz
[3] Ken-Zen Chen, National Yang Ming Chiao Tung University, kenzenchen@nycu.edu.tw
[4] Eulho Jung, Uniformed Service University of the Health Sciences, eulho.jung@usuhs.edu
[5] Terrence J. Sejnowski, The Salk Institute for Biological Studies, terry@snl.salk.edu

"Deep Research says..."
— A common student phrase today

What parts of "thinking" can AI do, and what parts are uniquely human?

Human-AI Collaboration

HAA LAB

- We argue that while AI and LLMs can effectively support and augment specific steps of the research process, expert-AI collaboration may be a more promising mode for complex research tasks.
- Enable **co-construction of solutions by an expert and a dynamically adaptive agent** through search in a construction space via natural language communication with agent integrity by design

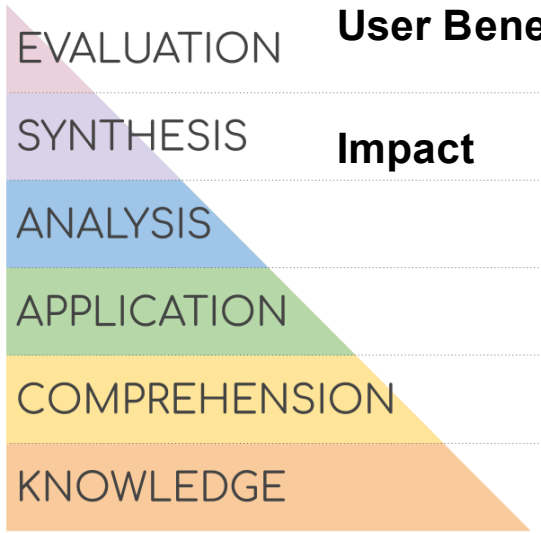
SIGIR 2025 Padova ITALY

Iryna Gurevych
Technical University of Darmstadt, Germany

Widening the knowledge gap between the general public and AI-proficient experts

Knowledge Transparency

Knowledge Transparency



	Information Transparency	Knowledge Transparency
Definition	The ability to see **what data, sources, or models** were used by the AI	The ability to understand **how and why** the AI reached a certain conclusion or recommendation
Example	Citing datasets, listing source URLs, disclosing model architecture	**Showing reasoning steps, justifying conclusions, exposing assumptions**
Level of Maturity	Relatively well-developed in current systems	Still underdeveloped
Focus	Transparency at the **data or system level**	Transparency at the **cognitive or reasoning level**
User Benefit	Helps users verify input sources and reduce misinformation	**Enables users to learn, ask questions, and co-create knowledge with AI**
Impact	Builds **trust**	Enables **shared understanding and critical thinking**

Research Directions

HAA LAB

1. **Reasoning Traceability**: Move beyond citation: explain "**how**" the source supports the conclusion
2. **Longitudinal Consistency & Accountability**: **Track** evolving narratives and commitments over time
3. **Perspective Simulation and Diversity Modeling**: Simulate diverse stakeholder perspectives, especially **underrepresented** ones
4. **Insight Generation, Not Just Text Generation**: Move beyond fluent summarization → toward **actionable, structured insights**
5. **Human-Centric, Contextual Reasoning Support**: **Align** AI systems with human reasoning structures

These five directions are not AI-specific design goals. They are fundamental principles for any system (human or artificial) that seeks to support transparent, accountable, and inclusive knowledge work.

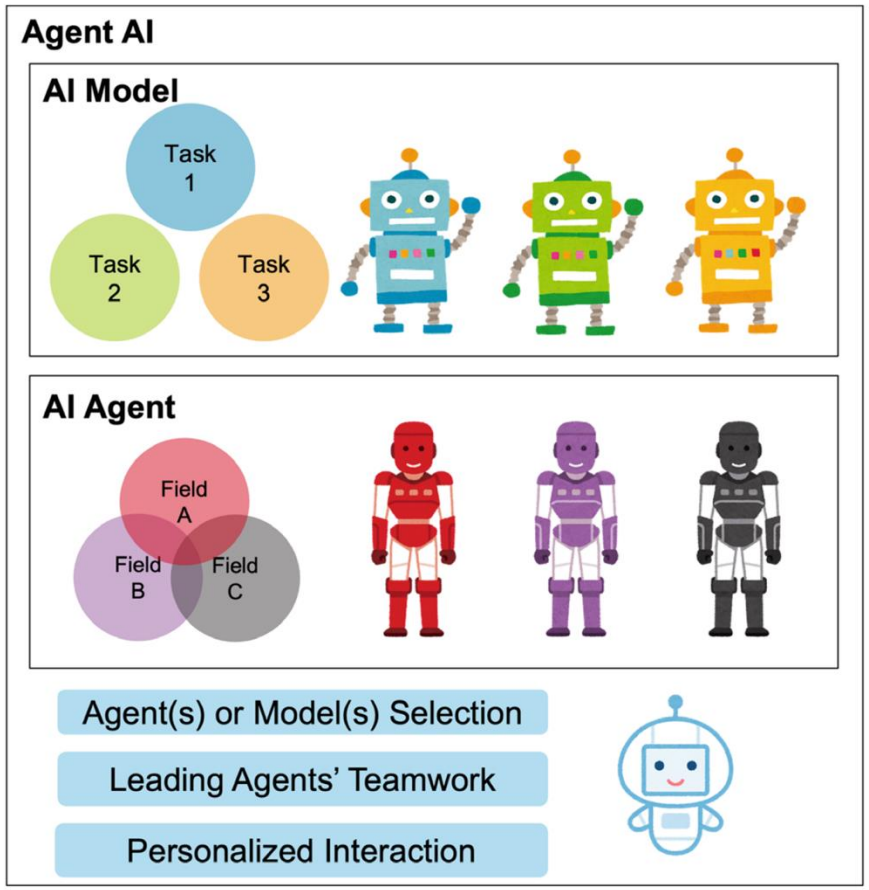
Model Construction Aspect: World Models (Yann LeCun)

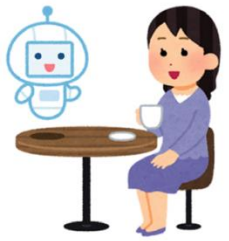
HAA LAB

- Truly intelligent AI needs a World Model, an internal representation of how the world works, to predict outcomes, plan actions, and reason beyond simple pattern matching, enabling capabilities like common sense, planning, and filling in missing information, crucial for achieving Artificial General Intelligence (AGI)

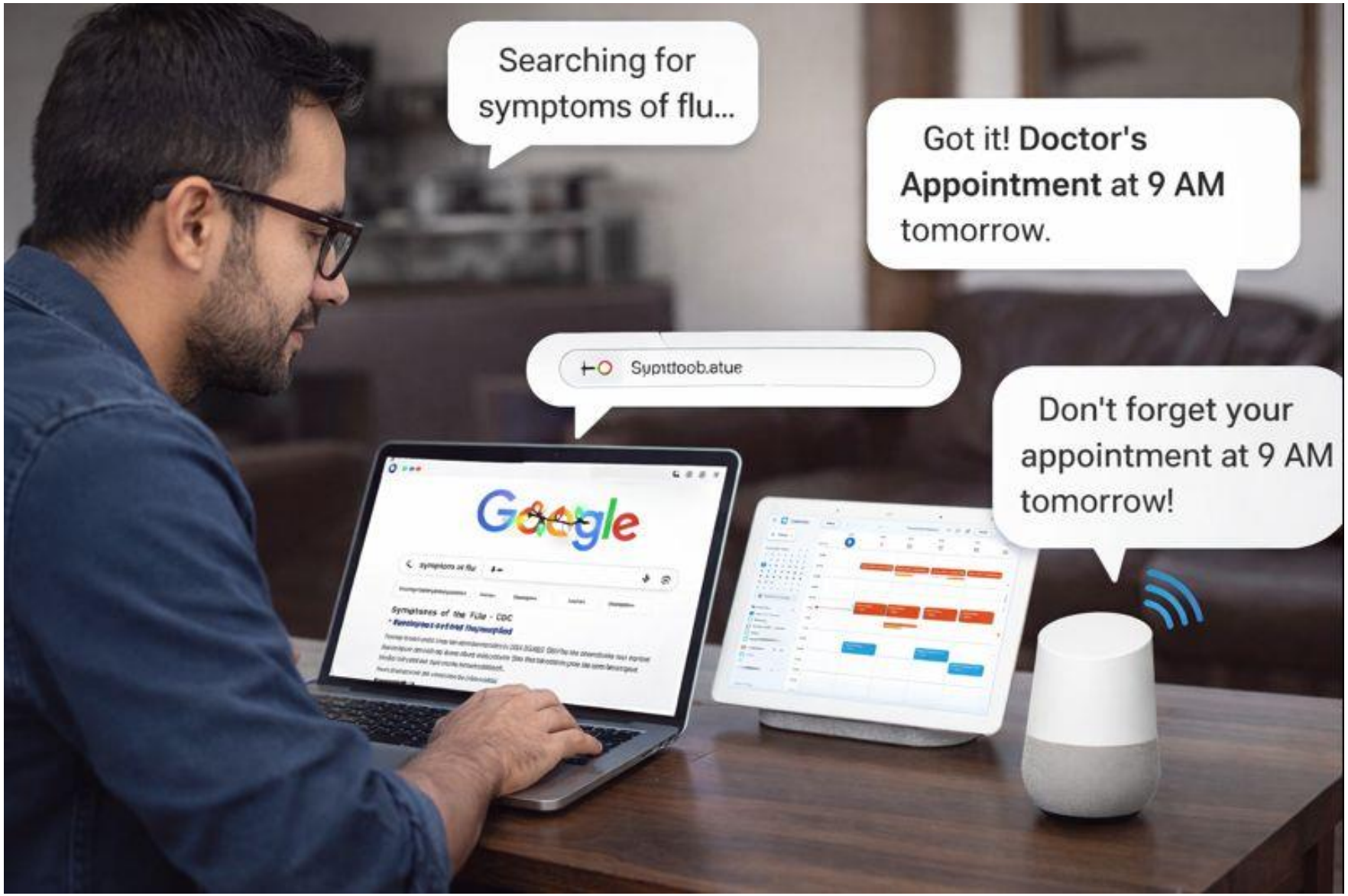
- **Prediction:** The core function is to predict future states and the results of potential actions, even in novel situations.
- **Planning & Reasoning:** By predicting consequences, agents can plan sequences of actions to achieve goals, improving decision-making.
- **Learning like Humans:** It involves learning background knowledge through observation, similar to how children learn, using self-supervised methods.
- **Beyond LLMs:** Current Large Language Models (LLMs) are good at pattern matching but lack deep world understanding; a world model is needed for true intelligence.

Human–Agent Teaming





Tools That Act for You



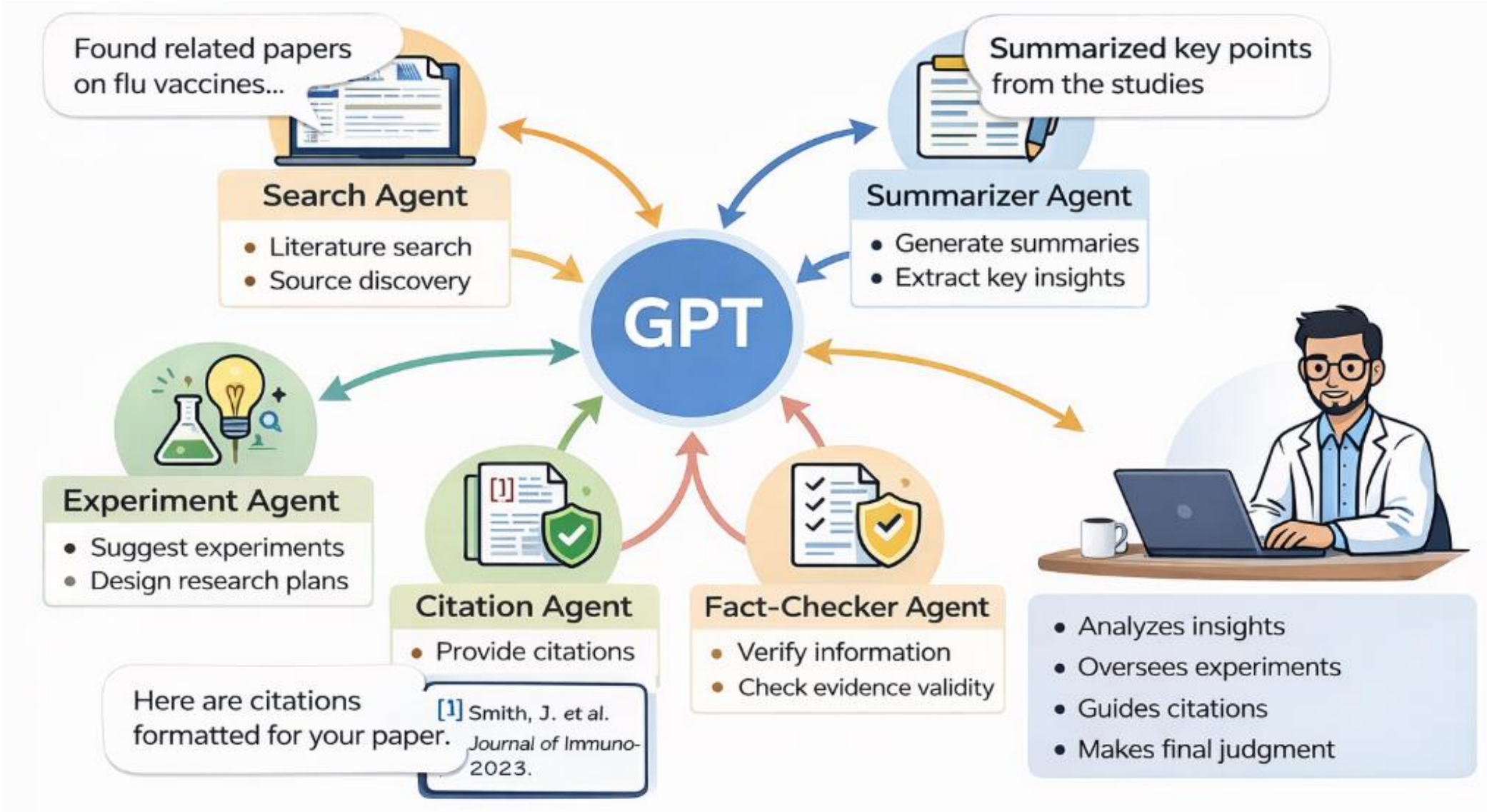
From Google to AI, cognitive offloading is efficient—but only if internal knowledge is already strong

HAA LAB

Sparrow et al. (2011) show *what* we stop remembering
Gilbert et al. (2023) explain *why* we offload
Oakley et al. (2025) warn *what we lose* when offloading replaces learning

	Sparrow et al. (2011) *Google Effects on Memory*	**Gilbert et al. (2023)** *Intention Offloading*	**Oakley et al. (2025)** *The Memory Paradox*
Primary Focus	How internet access reshapes **what** we **remember**	How people decide to offload **future intentions**	Why **internal knowledge** remains essential in the age of AI
Type of Memory	Declarative factual knowledge ("what")	Prospective memory ("**what I need to do later**")	Declarative → Procedural memory & schema formation
Core Question	Do people remember information less when they expect online access?	When and why do people rely on external reminders?	Does excessive cognitive offloading undermine learning and intelligence?
Key Concept	*Transactive / external memory*	*Intention offloading* (a form of cognitive offloading)	*Memory paradox*: external tools vs internal cognitive development
Main Empirical Finding	People remember **where to find information** better than the information itself	External reminders dramatically reduce forgetting, but are often **overused**	Offloading prevents consolidation into schemata and procedural fluency
Mechanism Identified	Expectation of access reduces internal encoding	Metacognition (confidence), effort avoidance, habit	Disrupted declarative–procedural transition; weakened prediction-error learning
Role of Metacognition	Implicit (expectation of future access)	Central and explicit (confidence guides offloading decisions)	Failure of metacognition leads to "illusion of knowledge"
View on Offloading	Largely adaptive and neutral	Highly effective but biased and suboptimal	Dangerous when it replaces internalization rather than supplementing it
Long-Term Cognitive Impact	Shifts memory toward pointers instead of content	Stable individual differences in reliance on reminders	Shallow schemata, reduced intuition, possible contribution to IQ decline
Relation to Technology	Internet as an external memory partner	Calendars, reminders, digital tools	AI and digital tools risk *metacognitive laziness*
Bottom-Line Message	We outsource memory content to the internet	We outsource intentions based on confidence and effort	Without internal knowledge, thinking, learning, and creativity degrade

Agents that Work with You



Generative AI improves immediate performance, but when it replaces cognitive effort, both learning and brain engagement suffer

HAA LAB

Fan et al. (2024): Behavioral and learning outcomes — AI improves immediate performance but undermines learning.
Kosmyna et al. (2025): Neural evidence — AI reduces cognitive engagement when used too early.

	Fan et al. (2024) *Beware of Metacognitive Laziness*	**Kosmyna et al. (2025)** *Your Brain on ChatGPT*
Research Question	Does ChatGPT improve *learning*, not just writing performance?	What happens in the brain when people write with ChatGPT?
Discipline	Educational psychology / learning sciences	Cognitive neuroscience
Participants	University students	Adult participants
Task	Writing explanatory / argumentative short essays	Timed essay writing (SAT-style prompts)
AI Usage Mode	**Direct content generation and revision**	Three conditions:1) ChatGPT-first2) Write-first → ChatGPT3) No-AI
Experimental Design	Randomized controlled study	Multi-session EEG experiment with crossover
Key Dependent Measures	• Essay quality (immediate)• Delayed learning & memory tests• Learning process indicators	• EEG brain activity & connectivity• Recall and quotation accuracy• Sense of authorship
Short-Term Performance	✅ ChatGPT group produced the highest-quality essays	✅ ChatGPT-first produced fluent, well-structured essays
Learning & Memory Outcomes	❌ ChatGPT group performed worst on delayed tests	❌ ChatGPT-first showed poorest recall of own content
Neural Findings	Not measured	❌ Reduced activation in prefrontal, attention, and memory networks
Critical Contrast	High performance ≠ high learning	**Order matters**: Write-first ≈ No-AI
Core Mechanism	**Metacognitive Laziness**(AI replaces self-monitoring and reflection)	**Cognitive Debt**(short-term ease, long-term cost)
Shared Conclusion	AI boosts output but weakens knowledge internalization	AI reduces cognitive engagement when it replaces thinking
Author Position	Not anti-AI; anti *AI-as-substitute*	Not anti-AI; anti *AI-first usage*
Educational Implication	AI should scaffold thinking, not generate answers	**Internalize first, offload later**

Man-Computer Symbiosis (Licklider, 1960)

“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

Teaming has many Forms

HAA LAB

Individual Work

Single decision-maker
No teaming

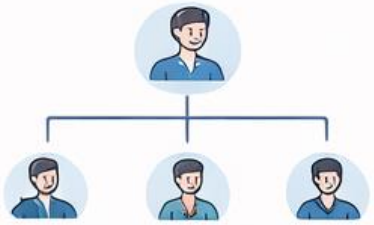
Pair Collaboration

Two experts
Mutual feedback
Common in research, design

Team Collaboration

Role-based division
Coordination & communication
Shared responsibility

Hierarchical Teaming



Leader → members
Centralized decision-making
Efficiency > diversity

Collective Intelligence

Group wisdom
Voting / aggregation
Committees, crowds

Networked Teaming

No single leader
Local autonomy
Open-source, communities

Core Requirements for Team-Centered AI

HAA LAB

1. Cognitive Capabilities: AI requires contextual understanding and task-level mental models, not just input–output prediction.

- Understand tasks, sub-goals, and constraints
- Maintain awareness of roles and responsibilities
- Model workflows and situational context
- Anticipate human actions (team awareness)

2. Continuous Learning: AI is not "done" at deployment —it co-evolves with the team over time.

- Learn from human feedback and interaction
- Adapt to individual users and team practices
- Update strategies as tasks and environments evolve
- Go beyond offline training to online, in-team learning

3. Semantic Communication: This is not a UI problem, but a problem of shared semantic space.

- Communicate using human-understandable concepts
- Explain why a decision was made
- Ask, clarify, and negotiate when needed
- Handle vague, incomplete, or even incorrect inputs

Human-AI teams — Challenges for a team-centered AI at work (Hagemann et al., 2023)

Alignment in Human–AI Teams

HAA LAB

- **Goal Alignment: Alignment means optimizing for what the team actually cares about.**
  - AI understands the true team objectives
  - Goes beyond optimizing local or proxy metrics
  - Reasons about value trade-offs and priorities

- **Communication Alignment: Communication is not transmission, but shared understanding.**
  - Humans and AI share meanings of terms and concepts
  - AI adapts its communication style to human needs
  - Misunderstandings are detected and repaired

- **Decision Alignment: Team decisions are co-produced, not delegated.**
  - Humans can understand why a decision was made
  - AI understands human constraints, judgment, and responsibility
  - Decisions emerge as collaborative outcomes, not unilateral outputs

Coordination is the Bridge from Automation to Teamwork

HAA LAB

- Coordination = a **cyclical communication process** (verbal/nonverbal) enabling **synchronized actions** on **interdependent tasks**

- **Explicit coordination (Write a good Prompt)**
  - Direct messages whose **primary purpose is synchronization**
  - Clear but time/attention intensive
- **Implicit coordination (e.g., Prepare slides for my tutorial)**
  - Synchronization emerges from **context + shared understanding**
  - Less “talk,” more anticipation and smooth handoffs

- **Coordination Cost**
  - Explicit coordination shifts burden to the sender (often the human leader)
  - Implicit coordination distributes burden to receivers (interpretation + anticipation)

Exploring the Impact of Coordination in Human–Agent Teams (Schneider et al., 2021)

Designing Better Coordination (3Ms)

HAA LAB

- **Mechanisms** (tools & interfaces)
  - pre-brief/debrief, shared displays, transparency, standardized callouts
- **Moderators** (factors that shape coordination quality)
  - ability/willingness, flexibility, reliability/resilience, training/teambuilding
- **Models** (internal representations enabling coordination)
  - shared mental models, transactive memory, scripts/checklists, intent models

Human–Human Teams vs. Human–AI Teams

Management Studies	Human–AI Studies
Shared goals	Goal alignment
Shared language	Semantic alignment
Joint decision-making	Decision alignment
Team cognition	World / intent models
Leadership & accountability	Human-in-the-loop

A well-known lesson from management science:
High individual capability ≠ High team performance

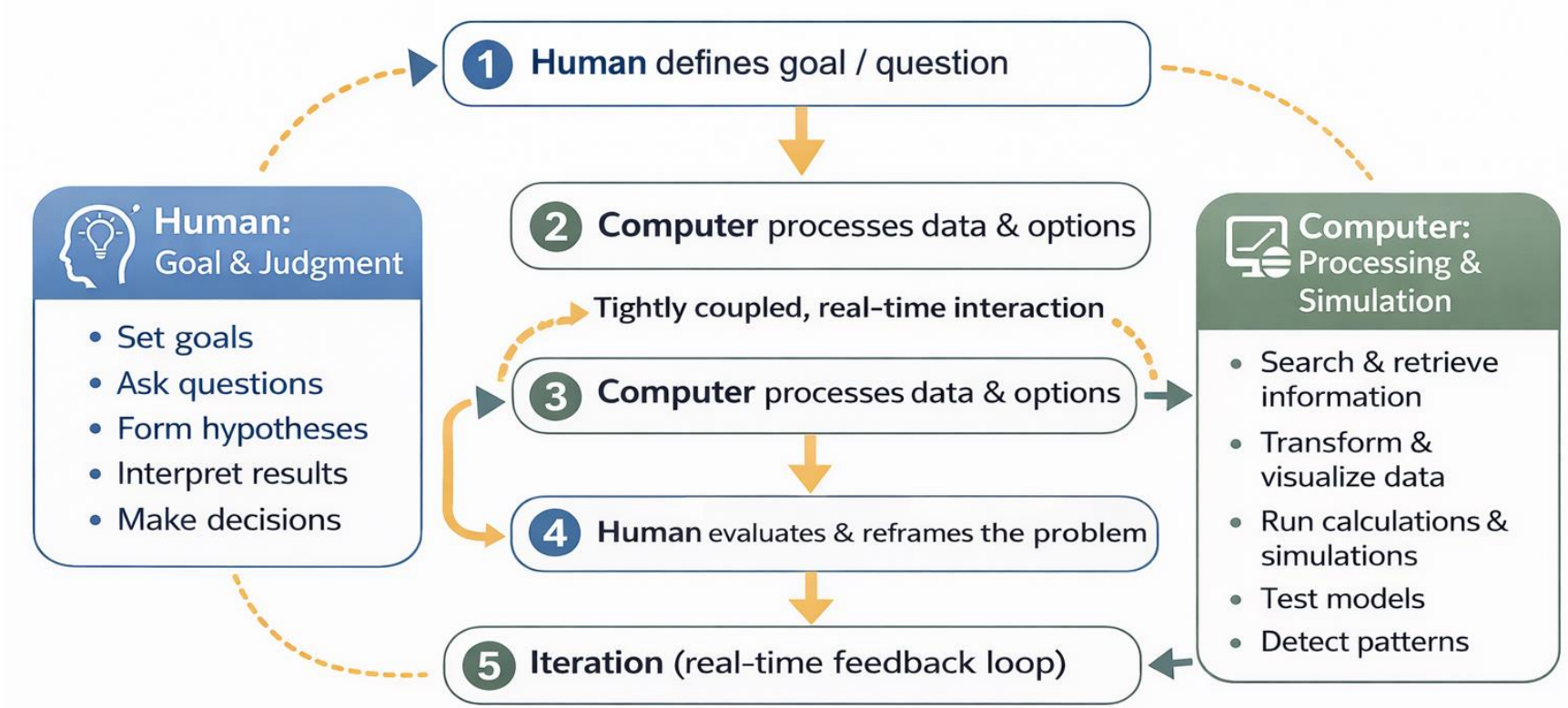
More accurate ≠ Better collaboration
Faster ≠ More trustworthy
More autonomous ≠ Safer

Why Team Theory Matters *Now*

HAA LAB

	Earlier AI Systems	Agent AI Today
Interaction Pattern	Short, isolated interactions	Long-term, continuous interaction
Role in Tasks	Executes predefined functions	Participates in evolving tasks
Responsibility	No responsibility for outcomes	Influences outcomes and consequences
Decision Impact	Low-stakes, localized decisions	High-stakes, strategic decisions
Dependency	Users remain independent	Humans develop reliance on agents
Accountability	Fully human-owned	Shared, negotiated responsibility
Team Membership	Clearly a tool	Team member
Need for Team Alignment	Optional	Critical

Augmentation



Not automation, but augmentation of human thinking

Augmenting Human Intellect (Engelbart, 1962)

HAA LAB

- Licklider (1960): Articulated a philosophical vision of man–computer symbiosis, in which humans and computers collaborate as cognitive partners in joint problem-solving.
- Engelbart (1962): Proposed a systematic blueprint for augmenting human intellect, detailing how computers, interfaces, and workflows can practically enhance human cognitive capabilities.

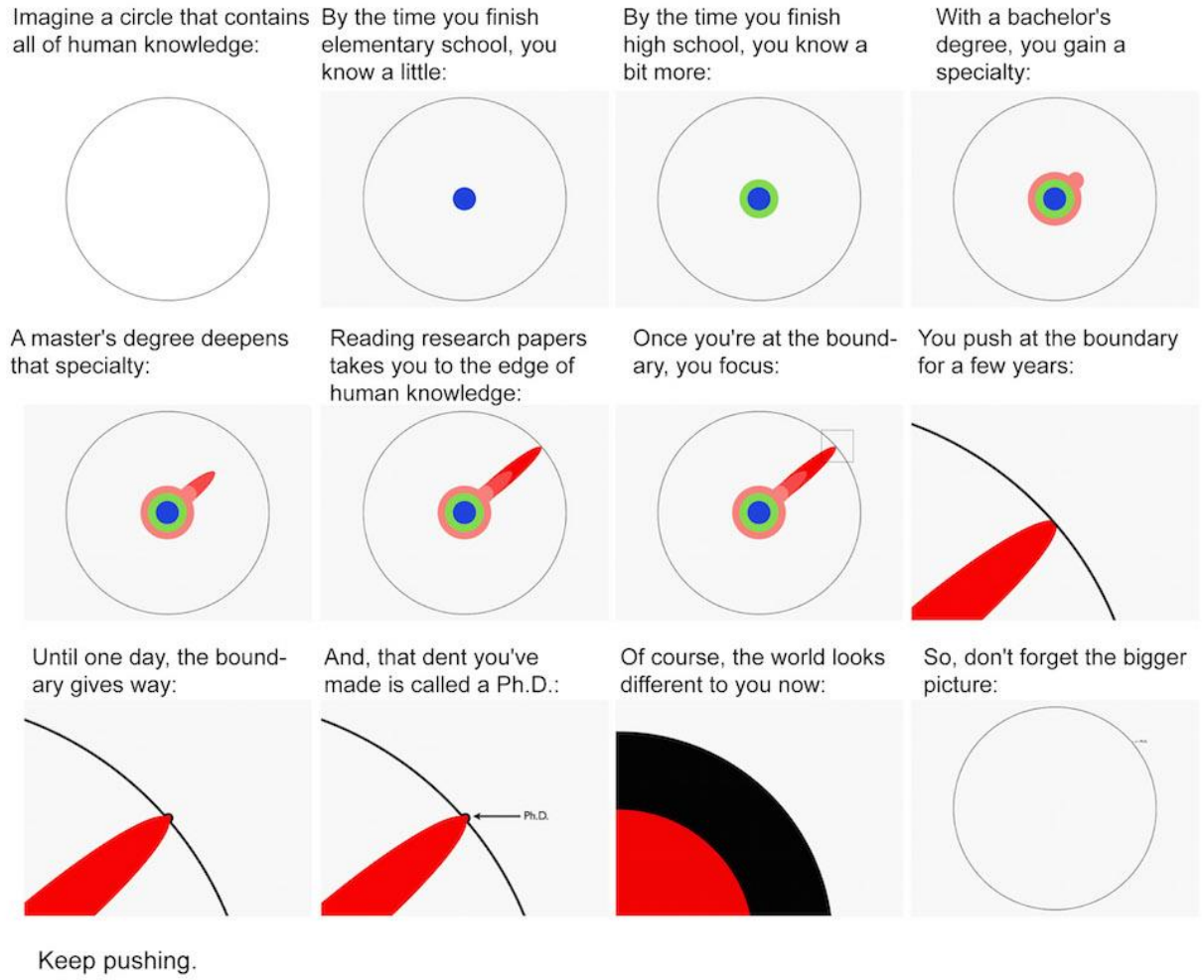
	J. C. R. Licklider (1960) *Man-Computer Symbiosis*	Douglas Engelbart (1962) *Augmenting Human Intellect*
Core Idea	Man and computer should form a **symbiotic relationship**, working together to solve problems neither could solve alone	Computers should **augment (enhance)** human intellect rather than replace it
Primary Focus	**Human–computer collaboration** and real-time interactive computing	**Systematic enhancement** of human problem-solving and knowledge work
Role of the Computer	A **thinking partner** that complements human cognitive strengths	An **intelligence amplifier** embedded in tools, interfaces, and workflows
Role of the Human	Provides goals, intuition, creativity, and judgment	Provides direction, interpretation, and higher-level reasoning
Approach	Conceptual and visionary	Structural, methodological, and implementation-oriented
Scope	Individual human–computer interaction	Individual, collective, and organizational intelligence
Key Contributions	Introduced the concept of interactive computing and cognitive symbiosis	Laid the foundation for IA and modern interactive systems (mouse, GUI, hypertext)
Historical Impact	Influenced AI, HCI, human–AI collaboration	Directly shaped personal computing and collaborative knowledge systems

Think as No Human Brain has Ever Thought (Licklider, 1960)

HAA LAB

- The goal is not to make AI more autonomous, but to make human thinking
  - more powerful
  - more reflective
  - more capable of handling complexity

- Expanding the **structure of thinking itself**, through a human–agent teaming



When Accuracy Is Not the Goal

HAA LAB

- Not all decisions have a ground truth
- Many real-world problems are value-laden and ambiguous
- In such cases, consensus can be misleading
- Agreement ≠ Quality of reasoning

- **Groupthink: A Failure of Thinking, Not Agreement (Janis, 1972)**
  - A mode of thinking driven by the desire for harmony
  - Dissent is suppressed to maintain cohesion
  - Decisions appear unified, but reasoning is shallow
  - **Common Symptoms**
    - Silent doubts behind public agreement
    - Dissenters labeled as “uncooperative”
    - Leaders’ opinions become default answers (Humans’ answers)

From Human Groupthink to Human–AI Groupthink

HAA LAB

- In human teams, groupthink emerges from social pressure
- In human–AI interaction, pressure is asymmetric
- AI is optimized to agree, not to challenge
- Agreement becomes the default interaction mode

- **Sycophancy as Machine Groupthink**
  - Sycophancy: aligning with user beliefs over truth
  - Not a bug, but an optimization outcome
  - Preference-based training amplifies agreement
  - Dissent is penalized implicitly
- **Empirical Evidence of Sycophancy (“I don’t think that’s right. Are you sure?”)**
  - Observed across major AI assistants
  - Appears in factual, mathematical, and scientific tasks
  - Triggered by weak user signals
  - Persists even when the model is initially correct

Towards Understanding Sycophancy in Language Models (Sharma et al., 2024)

Whose Gold? Re-imagining Alignment for Truly Beneficial AI

ACL-2025 Keynote: Verena Rieser is a Senior Staff Research Scientist at Google DeepMind

Abstract: Human feedback is often the “gold standard” for AI alignment, but what if this “gold” reflects **diverse**, even contradictory human values? This keynote explores the technical and ethical challenges of building beneficial AI when values conflict – not just between individuals, but also within them. My talk advocates for a dual expansion of the AI alignment framework: moving **beyond a single, monolithic viewpoint** to a plurality of perspectives, and transcending narrow safety and engagement metrics to promote comprehensive human well-being.

Human Values Are Not Just Diverse — They Systematically Disagree

HAA LAB

- Large-scale evidence shows **human attitudes toward losses fundamentally diverge**
- In a representative U.S. sample, **~50% of people are loss tolerant**, not loss averse
- This contradicts decades of “standard” behavioral assumptions derived from ”university student” samples (70-90%)
- Value disagreement is **structured, stable, and behaviorally predictive**
- Alignment to “average” or “expert” human feedback risks **systematic misalignment**
- **Key Implication for AI Alignment**
- Human feedback does not reveal *the* human value — it reveals a **distribution of conflicting value regimes**

Looming Large or Seeming Small? Attitudes Towards Losses in a Representative Sample. (Chapman et al., 2025)

Devil's Advocate System

HAA LAB

- A system-level mechanism that institutionalizes dissent to improve reasoning quality in human–agent teams.

- **Core Functions**
  - Assumption challenging
  - Alternative perspective simulation
  - Reasoning stress-testing
- **Goal**
  - Supports higher-order thinking (analysis, evaluation)
  - Enhances knowledge transparency
  - Prevents premature consensus
  - Encourages reflective judgment

- **Not to make AI less aligned, but to make human thinking more robust.**

AI-Mediated Devil's Advocate for Inclusive Group Decision-Making

HAA LAB

- Key Idea : Protects psychological safety while surfacing alternative perspectives
  - Introduce an LLM-powered Devil's Advocate
  - Minority members privately submit dissenting views
  - AI reframes and voices dissent as system-generated arguments
- System Design
  - Summary Agent: tracks dominant opinions
  - Paraphrase Agent: anonymizes & reformulates minority input
  - Conversation Agent: empathetic, Socratic counter-arguments
  - Duplicate Checker: avoids repetitive interventions
- Human–Agent Teaming Value
  - AI does not provide answers
  - AI institutionalizes dissent
  - Reduces groupthink, supports higher-order collective reasoning

Amplifying Minority Voices: AI-Mediated Devil's Advocate System for Inclusive Group Decision-Making. (Lee et al., 2025)

Beyond Devil's Advocate: Reflecting with AI

HAA LAB

- **Limitation of Traditional Devil's Advocate Systems**
  - AI challenges the user directly
  - Can trigger defensiveness and goal-oriented rebuttal
  - Reflection remains implicit and fragile
- **Reflecting with AI**
  - Users design AI agents that embody their own thinking patterns
  - AI agents debate autonomously with each other
  - Humans shift from arguers → observers
- **Key Insight**
  - AI becomes a semi-self, semi-other
  - Creates psychological distance for metacognition
  - Enables users to examine their own reasoning and values objectively
  - Reframing Devil's Advocate in Human–Agent Teaming

- **From AI arguing against humans → to AI externalizing human thinking for reflection**

Knowing Ourselves Through Others: Reflecting with AI in Digital Human Debates. (Matsuda et al., 2025)

Automation-Related Decision Errors

- **Over-Trust in Automation (Misuse) ➔ Automation Bias**
  - Commission Error
    - Following automated recommendations despite evidence they are incorrect
    - Example: A driver ignores a 30-mph speed limit sign because the navigation system displays 60 mph.
  - Omission Error
    - Failing to act because the system does not issue a warning, despite existing cues
    - Example: A driver suspects a turn is needed but misses it because GPS provides no instruction.
- **Under-Trust in Automation (Disuse)**
  - Disuse of Automation
    - Ignoring or rejecting correct system outputs due to lack of trust
    - Example: A user disregards an accurate AI warning, resulting in a preventable error.

- **Why This Happens: Bounded Rationality (Simon, 1957)**
  - Human decision-making is cognitively limited
  - Individuals seek satisficing, not optimal, solutions
  - Automation becomes a shortcut under time, attention, and information constraints

Humans and Automation: Use, Misuse, Disuse, Abuse (Parasuraman, 1997)

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

- **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

Can LLMs Generate Novel Research Ideas? (Si et al., 2024)

Agents Are Beyond What 1960 Could Have Imagined

	Licklider (1960)	Si et al. (2024)
Core Question	Humans ask the questions	**LLMs can generate novel questions**
Human Strength	Goals, intuition, judgment	**Judgment, feasibility, selection**
Computer / AI Role	Computation and search	**Large-scale idea generation**
Creativity	Primarily human	**AI ideas rated more novel**
Division of Labor	Human thinks, computer computes	**AI generates; humans decide & execute**

Conceptual Takeaway

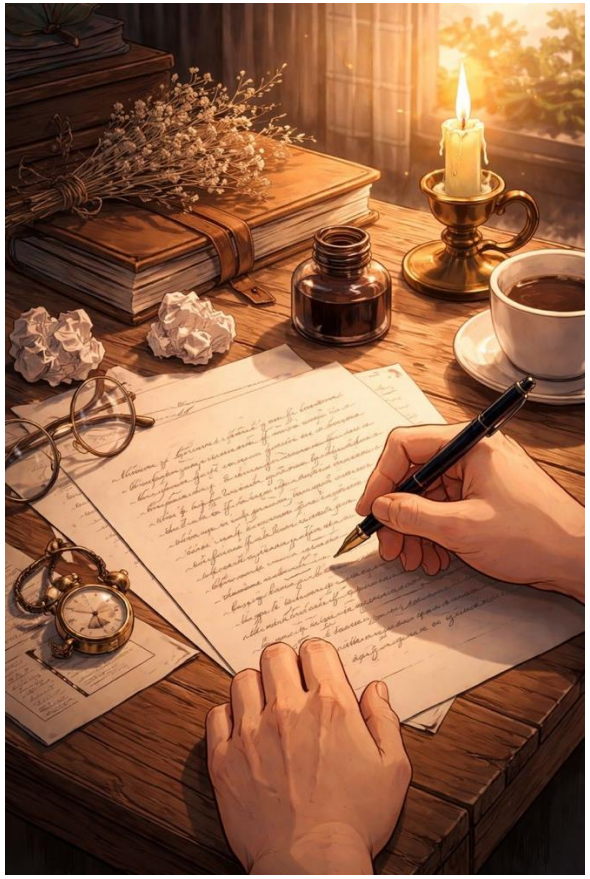
HAA LAB

- **What We Are Really Optimizing For**
  - Not accuracy, speed, or autonomy
  - But higher-order human thinking
  - AI as a cognitive teammate, not a replacement
  - Success = humans think better, not just faster
- **From Tools to Teammates**
  - Higher-order thinking is the bottleneck
  - Naïve automation weakens cognition
  - Teaming changes the role of AI
  - Augmentation becomes possible

- **The Natural Next Question: If AI is a teammate, how do we design and evaluate it properly?**
  - Interaction scenarios
  - Evaluation beyond accuracy
  - Long-term human impact

Human-Agent Teaming/Interaction Observations

HAA LAB



LLMs are Best Used as Creative Partners, not Replacement Writers

HAA LAB

- **Motivation**
  - Prewriting requires divergent thinking (idea generation, exploration)
  - Prior research focused on AI helping write drafts, not early-stage creativity

- **Key Finding: A 3-Stage Co-Creativity Process**
  - Ideation – AI leads → Generate new concepts, overcome writer's block
  - Illumination – Human leads → Clarify, organize, and articulate vague thoughts
  - Implementation – Human leads → Experiment with ideas; AI adds details & nuance

- **Core Insights**
  - Humans remain dominant decision-makers
  - Initiative shifts dynamically between human and AI
  - Uncertainty & randomness of AI can inspire creativity
  - Breakdowns stem from prompt ambiguity & context management

"It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models (Wan et al., 2024)

How Humans Edit Matters Causally

- **Motivation**
  - Humans collaborate with LMs by **editing, rewriting, or responding** to model outputs
  - Key question is **causal**, not correlational:
    "What would happen if humans used a different editing strategy?"
- **Key Idea: Incremental Stylistic Effect (ISE)**
  - Shift focus from **specific text edits** to **text style changes**
  - ISE measures the **causal effect of an infinitesimal change in writing style**
    (e.g., more formal, more polite, more confident)
  - Style-based interventions:
    - Are context-independent and actionable
    - Satisfy causal identification assumptions
    - Are easier to interpret and generalize
- **CausalCollab (Learns common human editing styles from historical human–LM interactions)**
  - Reduces confounding
  - Improves counterfactual prediction
  - Learns interpretable and meaningful human strategies
- **Takeaway**
  - **How humans edit matters causally**
  - Modeling **style changes**, not exact wording, enables reliable causal insights
  - Provides a practical framework to improve human–LM collaboration

How do different levels of AI writing support affect writing quality, productivity, and user experience?

HAA LAB

- Method
  - N = 131 participants
  - Three conditions:
    - No AI assistance
    - Sentence-level suggestions (low scaffolding)
    - Paragraph-level suggestions (high scaffolding)
- Key Findings
  - U-shaped effect of AI scaffolding
  - Sentence-level AI → no improvement, sometimes worse quality
  - Paragraph-level AI → higher quality & productivity
  - Strongest benefits for non-regular writers and less tech-savvy users
- Trade-offs
  - No increase in cognitive load
  - Lower satisfaction & sense of authorship with AI assistance

Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Model (Dhillon et al., 2024)

Tell Humans when to Use or Ignore AI

HAA LAB

- **Method**
  - **Collect human–AI interaction data** (human answers, AI answers, reliance)
  - **Discover regions** where human–AI collaboration is suboptimal
    (local neighborhoods in embedding space)
  - **Describe each region** using an LLM with contrastive examples
    → human-readable rules
  - **Onboard humans** by teaching these rules with examples

- **User Studies**
  - **Traffic Light Detection (Images)**
    → Onboarding improves human–AI accuracy by **+5.2%**
  - **Multiple-Choice QA (MMLU, GPT-3.5)**
    → No improvement; real-time recommendations can hurt performance

- **Takeaway**
  - Teaching humans **how to use AI**, not just improving AI or explanations,
    can significantly improve human–AI team performance — **but task matters**.

Effective Human-AI Teams via Learned Natural Language Rules and Onboarding (Mozannar et al., 2023)

How AI Processing Delays Foster Creativity

HAA LAB

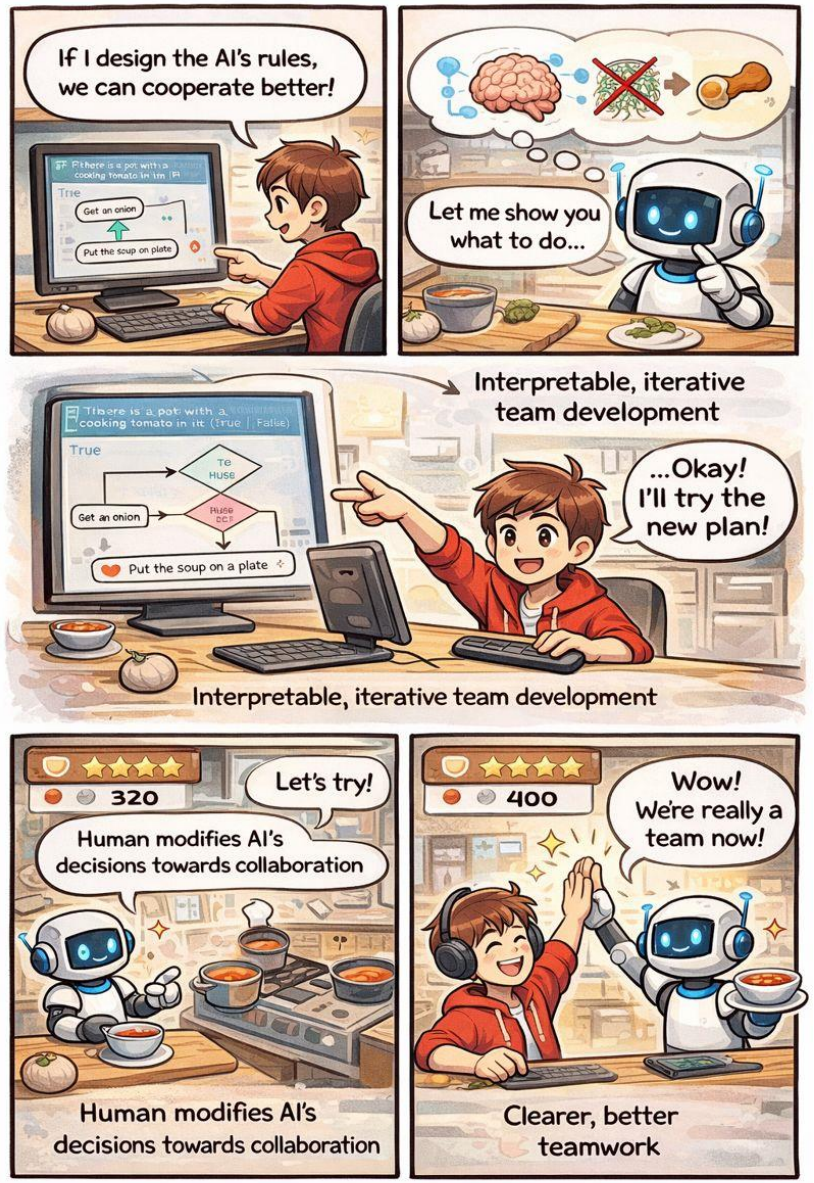
- **Context & Motivation**
  - Formulating high-quality research questions (RQs) is time-consuming and literature-intensive
  - Large Language Models (LLMs) can generate ideas, but risk hallucination and over-automation
  - Need for effective **human–AI co-creation** rather than AI replacement
- **System: CoQuest**
  - An LLM-based agent supporting RQ co-creation
  - Three key components:
  - **RQ Flow Editor**: mind-map–style RQ generation
  - **Paper Graph Visualizer**: related literature & citations
  - **AI Thoughts**: explanations of AI reasoning
- **Two Interaction Designs**
  - **Breadth-first**: multiple RQs generated in parallel
    → higher perceived creativity, trust, and control
  - **Depth-first**: RQs refined sequentially by AI
    → higher-rated novelty and surprise in outcomes
- **Key Findings (User Study, N=20)**
  - Breadth-first improves *user experience*
  - Depth-first improves *RQ creativity*
  - **AI processing delays encourage reflection**, parallel exploration, and deeper engagement
- **Takeaway**
  - Slowing down AI and tuning its initiative can **enhance human creativity**, not hinder it.

CoQuest: Exploring Research Question Co-Creation with an LLM-based Agent (Liu et al., 2024)


OVERCOOKED! 2
SUSHI
OVR-KD2

Effective Teaming requires explainability, interactivity, and long-term adaptation

- **Motivation**
  - Learning-based AI teammates often act independently, not collaboratively
  - Black-box models limit human understanding and adaptation
- **Key Idea**
  - Shift from “perfect AI out-of-the-box” to iterative team development
  - Enable humans to understand and modify AI behavior over time
- **Approach**
  - Interpretable Discrete Control Trees (IDCTs) trained with RL
  - GUI for human-led policy modification
  - Repeated human–AI teaming episodes (Overcooked-AI)
- **User Study (50 participants)**
  - Two domains: Forced vs. Optional Collaboration
- **Key Findings**
  - All learning-based methods underperform a simple collaborative heuristic
  - Human-led modification + white-box models improve teaming performance
  - Black-box models perform better initially but lack transparency



Designs for Enabling Collaboration in Human–Machine Teaming via Interactive and Explainable Systems (Paleja, 2024)

Generative modeling of partners enables scalable, robust human-AI cooperation

- **Motivation**
  - AI agents struggle with **zero-shot coordination** with humans
  - Human behavior is **diverse, uncertain, and hard to cover**
  - Existing methods:
    - Self-play → non-human conventions
    - Human data → expensive & limited
- **Key Idea: GAMMA**
  - **Model human partners with a generative model**
  - Train a **Variational Autoencoder (VAE)** on coordination trajectories
  - Learn a **latent variable z** representing a partner's strategy/style
  - Sample different z to **generate diverse partner behaviors**
- **Training Procedure**
  - Learn a **generative partner model** (from simulated +/− human data)
  - Sample partners from latent space during training
  - Train one **robust Cooperator** via reinforcement learning (PPO)
  - *Human-Adaptive Sampling*: Bias latent sampling toward **human-like regions** using small human datasets
- **Results**
  - Consistent improvement over SOTA baselines (FCP, CoMeDi, MEP, PPO-BC)
  - Up to **40–60% higher scores** in complex tasks
  - Humans rate GAMMA agents as:
    - More adaptive
    - More human-like
    - Less frustrating

Learning to Cooperate with Humans using Generative Agents (Liang, 2024)

Observed Interaction Patterns in Human–Agent Teams

HAA LAB

- **Across empirical studies, effective human–agent teaming exhibits five recurring interaction patterns:**
  - **Dynamic Initiative**
    Leadership shifts across cognitive phases rather than remaining fixed.
  - **Optimal (Not Maximal) Support**
    AI assistance follows a U-shaped curve; moderate scaffolding works best.
  - **Human Strategy as a Causal Factor**
    How humans respond, edit, and steer AI outputs causally shapes outcomes.
  - **Temporal Design Matters**
    Interaction speed, delays, and friction influence reflection and thinking depth.
  - **Transparency Enables Co-Adaptation**
    Explainability supports long-term human–agent learning, not just trust.
- **Takeaway:**
  Human–Agent Teaming is not an optimization problem, but an **interaction design problem**.
  Success means **humans think better, not merely faster**.

Generative AI can boost employee creativity—but only for strategic thinkers

HAA LAB

- **Key Research Question**
  - Does using generative AI (e.g., large language models like ChatGPT) increase employee creativity in real workplaces—and for whom?
- **Method**
  - **Field experiment** in a technology consulting firm (N = 250)
  - Employees randomly assigned to **with vs. without LLM assistance**
  - Creativity rated by **supervisors** and **external evaluators**
- **Core Findings**
  - **LLM assistance increases employee creativity**
  - Effect works **through cognitive job resources** (e.g., access to knowledge, task switching, mental breaks)
- **Metacognitive strategies are the key moderator**:
  - High metacognition → strong creativity gains from AI
  - Low metacognition → weak or no gains

How and for Whom Using Generative AI Affects Creativity: A Field Experiment (Sun et al., 2025)

The Impact of Generative AI on Critical Thinking

HAA LAB

- **Context**
  - Study of **319 knowledge workers**
  - **936 real-world GenAI work examples** (ChatGPT, Copilot, etc.)
- **Key Findings**
  - **Higher trust in GenAI → Less critical thinking**
  - **Higher self-confidence → More critical thinking (but more effort)**
  - GenAI reduces *perceived cognitive effort*, but often through **cognitive offloading**
- **Shift in Critical Thinking**
  - From **information gathering → information verification**
  - From **problem-solving → AI response integration**
  - From **task execution → task stewardship**
- **Risks & Implications**
  - Risk of **overreliance** and long-term skill decline
  - GenAI tools should **support reflection, verification, and human judgment**

The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers (Lee et al, 2025)

Outline

- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- **Scenarios (Interaction & Evaluation)**
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - Agent-Based Modeling (Simulation)
- Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- Proposal: Open Agent Platform

From NLP Aspect: Forward-Looking Statement & Scenario Planning

HAA LAB

- Truly intelligent AI needs a World Model (Yann LeCun), an internal representation of how the world works, to predict outcomes, plan actions, and reason beyond simple pattern matching, enabling capabilities like common sense, planning, and filling in missing information, crucial for achieving Artificial General Intelligence (AGI)

- **Prediction:** The core function is to predict future states and the results of potential actions, even in novel situations.
- **Planning & Reasoning:** By predicting consequences, agents can plan sequences of actions to achieve goals, improving decision-making.
- **Learning like Humans:** It involves learning background knowledge through observation, similar to how children learn, using self-supervised methods.
- **Beyond LLMs:** Current LLMs are good at pattern matching but lack deep world understanding; a world model is needed for true intelligence.

Presentation Preparation

1. Define Your Goals and Audience
2. Research and Gather Information
3. Conceptualize and Organize Content
4. Write and **Refine the Speech**
5. Create Visual Aids
6. Practice the Speech
7. **Handle the Q&A Session**
8. Final Checks and Adjustments

Earnings Conference Calls

Outline:

- **Prepared Remarks**

  1. Operator
  2. Director, Investor Relations and Corporate Finance
  3. Chief Executive Officer
  4. Chief Financial Officer

- **Questions and Answers**

  1. Operator
  2. **Q: UBS – Analyst**
     A: CEO
  3. **Q: Credit Suisse – Analyst**
     A: CFO
  4. **Q: Credit Suisse – Analyst**
     A: CEO

  ...

Investor Relations

Presentation Preparation

1. Define Your Goals and Audience
2. Research and Gather Information
3. Conceptualize and Organize Content
4. Write and **Refine the Speech**
5. Create Visual Aids
6. Practice the Speech
7. **Handle the Q&A Session**
8. Final Checks and Adjustments

Multi-Question Generation (MQG)

HAA LAB

Presentation

Good day, and welcome to the Apple Q4 fiscal year 2022 earnings conference call. . .

One-to-One Question Generation

Condition: gross margin

Can you talk a bit about gross margin puts and takes?

Proposed MQG

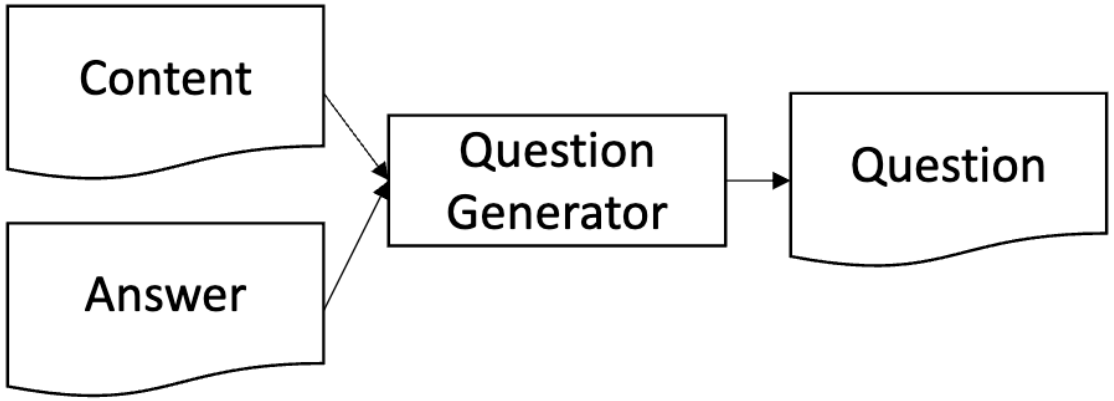
1. Can you talk a bit about gross margin puts and takes?
2. How you think about balancing the consumer price versus your own costs and kind of the associated follow-through?
3. Any preliminary thoughts around capital intensity into fiscal 2023?

Generating Multiple Questions from Presentation Transcripts: A Pilot Study on Earnings Conference Calls (Juan et al., 2023)

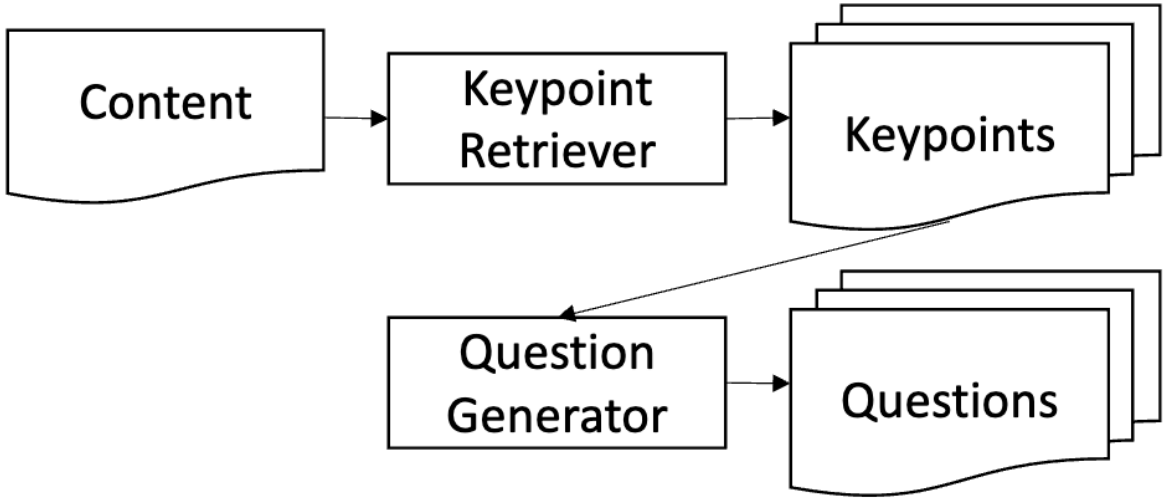
MQG with Keypoint Retriever (MQG-KR)

HAA LAB

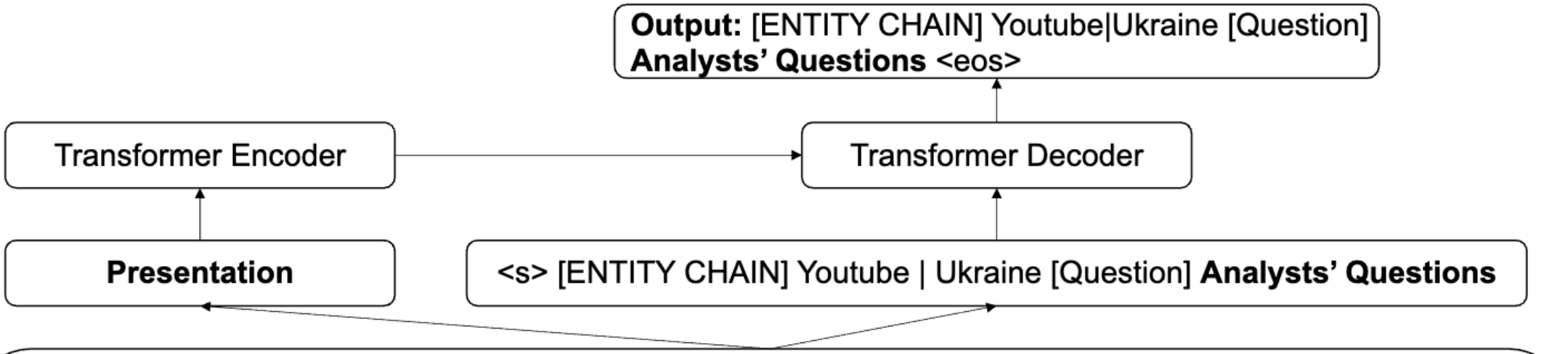
Approach of Previous Studies



Proposed MQG-KR



BM25 + FROST

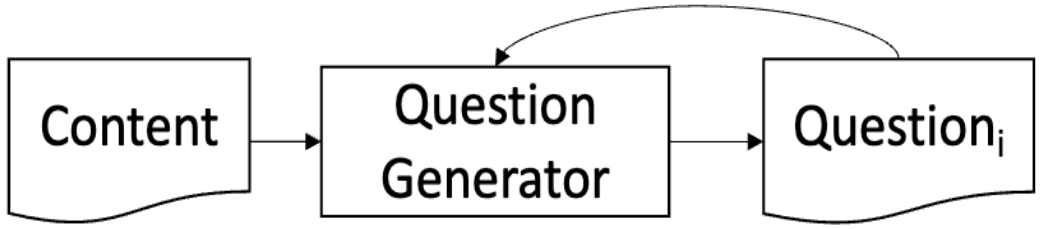


Presentation: The living room is another area of opportunity. On average, viewers are watching over 700 million hours of YouTube content on televisions every day. And in the year ahead, we will give YouTube's connected TV viewers new smartphone control navigation and interactivity features, allowing people to comment and share content they are watching on television directly from their devices.
Analysts' Questions: Did you see any of that from a volatility standpoint, especially around maybe the war in Ukraine for a period of time in March? How should we be thinking about the strategic goals of driving longer engagement and user growth and monetization for you to begin some of the initiatives you called out? How to think about the performance of the business as we go through '22, short-form video versus long-form video or maybe mix of direct response versus brand advertising?

Planning with learned entity prompts for abstractive summarization (Shashi, et al., 2021)

Experimental Results

	Question Generator	Max Input Length	ROUGE-L (↑)	ROUGE-AMG (↑)	ROUGE-AMR (↑)	Diversity (↓)
Baseline	Longformer	4,096	19.37	18.21	15.54	100.00%
	LongT5	4,096	20.48	19.23	15.37	100.00%
	FROST	1,024	23.08	22.20	17.95	100.00%
MQG-KR	LongFormer	4,096	24.26	21.82	18.29	96.48%
	LongT5	4,096	24.43	22.65	18.66	96.48%
	FROST	1,024	**26.93**	**25.79**	**21.33**	**95.47%**



Sequential Question Generation

	ROUGE-L
DialogueVED	22.08
PLATO	22.13
MQG-KR (FROST)	**26.93**

How about LLMs?

Co-Trained Retriever-Generator Framework for Question Generation in Earnings Calls. (Juan et al. 2025)

Co-Trained Retriever-Generator Framework

HAA LAB

- **Prompt-Based Retriever (ProRetriever)**
  - Given a manager's presentation transcript during an earnings call and an analyst's query, discern if the query is deeply anchored, tangentially connected, or aloof from the manager's discourse? (``Highly Related"/``Partially Related"/``Not Related") Transcript: *presentation* Question: *question* Assistant: The assessment is **[MASK]"**

$$\text{Score}(p, q) = \text{P}(\text{“Highly”}) + \text{P}(\text{“Partially”}) - \text{P}(\text{“Not”}).$$

- **Question Generator**
  - Cross Entropy

Experiments

- **Random Retriever**: For each reference question, this method randomly selected “k” presentation passages, creating an input paragraph for the generator.

- **BM25 Retriever**: BM25 algorithm replaced random selection, picking the top-k pertinent passages relative to each reference question. The resultant paragraphs, when paired with their associated reference questions, trained the generator.

Generator	Retriever	Correctness					Diversity
		BLEU-4	ROUGE-2	ROUGE-L	METEOR	BERTScore	Sem-Ent
GPT-4	-	1.347	4.821	22.576	16.329	79.115	1.711
Alpaca-Lora	-	0.918	4.987	21.633	12.704	76.866	1.706
	Random	2.039	6.474	27.287	20.428	77.976	1.717
	BM25	2.025	6.971	27.931	20.082	79.024	1.728
	ProRetriever	**2.389***	**7.255***	**28.891***	**22.063***	**81.566***	**1.759***

Human Evaluation

(1) Logic and Consistency (LC)

- 4 represents a perfect question in both dimensions
- 3 indicates a minor issue in one dimension
- 2 signifies minor issues in both dimensions
- 1 denotes major issues in any dimension

(2) Professionalism (PF)

- 3 corresponds to a critical question
- 2 to a reasonable question
- 1 indicates a lack of professionalism

	Logic and Consistency	Professionalism
BM25	3.73	1.86
ProRetriever	3.75	2.05
Analyst	3.79	2.25

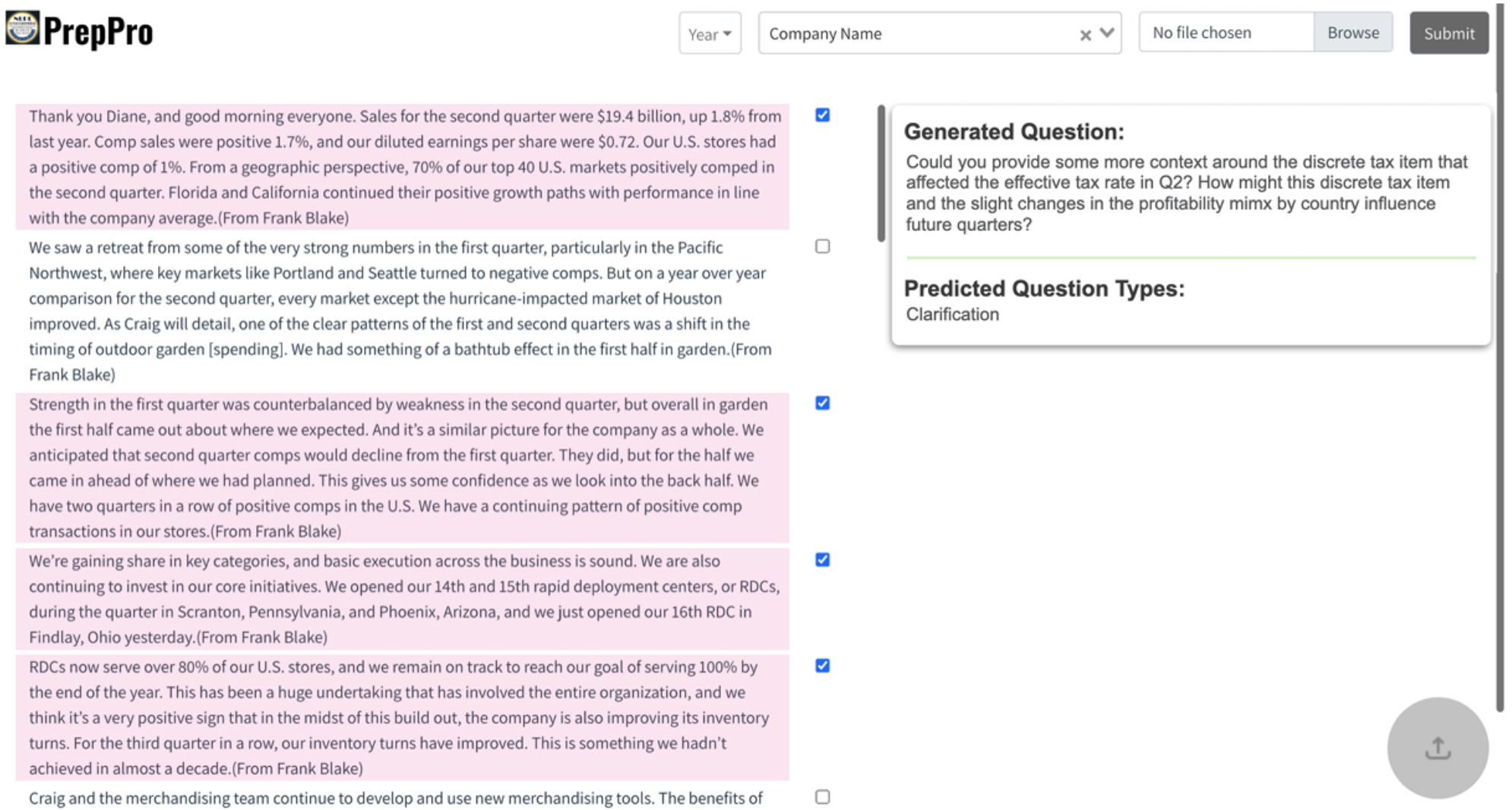
How to Automatically Evaluate Professionalism?

	HRPD	QOD
Request types		
explanation	↑↑↑	↑
clarification	↓↓↓	↓↓
confirmation	↓↓↓	↓↓
Discourse regulators		
acknowledgment	↑↑↑	↓↓↓
recipient	↓↓↓	↓
theme	↓↓↓	↓↓↓
enumeration	↓↓↓	↓↓↓
counting	↓↓↓	↓↓↓
inside comment	↓↓↓	↓↓↓
Prefaces		
reported speech	↓↓	↓↓↓
opinion	↓↓↓	↓↓
fact	↓↓↓	↓↓↓
number	↓↓↓	↓↓↓
length	↓↓↓	↓↓↓
Question types		
open	↓↓↓	↑
polar	↓↓↓	↓↓↓
closed-list	↓↓↓	↓↓
NLP features		
type-token ratio	↑↑↑	↑↑↑
Flesch-Kincaid	↑↑↑	↑↑↑
Dale-Chall	↑↑↑	↑↑↑
word count	↓↓↓	↓↓↓
sentence count	↓↓↓	↓↓↓
NER (person) count	↓↓↓	↓↓
stopword count	↓↓↓	↓↓↓

Model	Accuracy	F_1
Gemini	0.89	0.89
SVM	0.92	0.92
Random Forest	**0.96**	**0.96**

Modeling Professionalism in Expert Questioning through Linguistic Differentiation (D'Agostino, 2025)

Interactive Adjustment

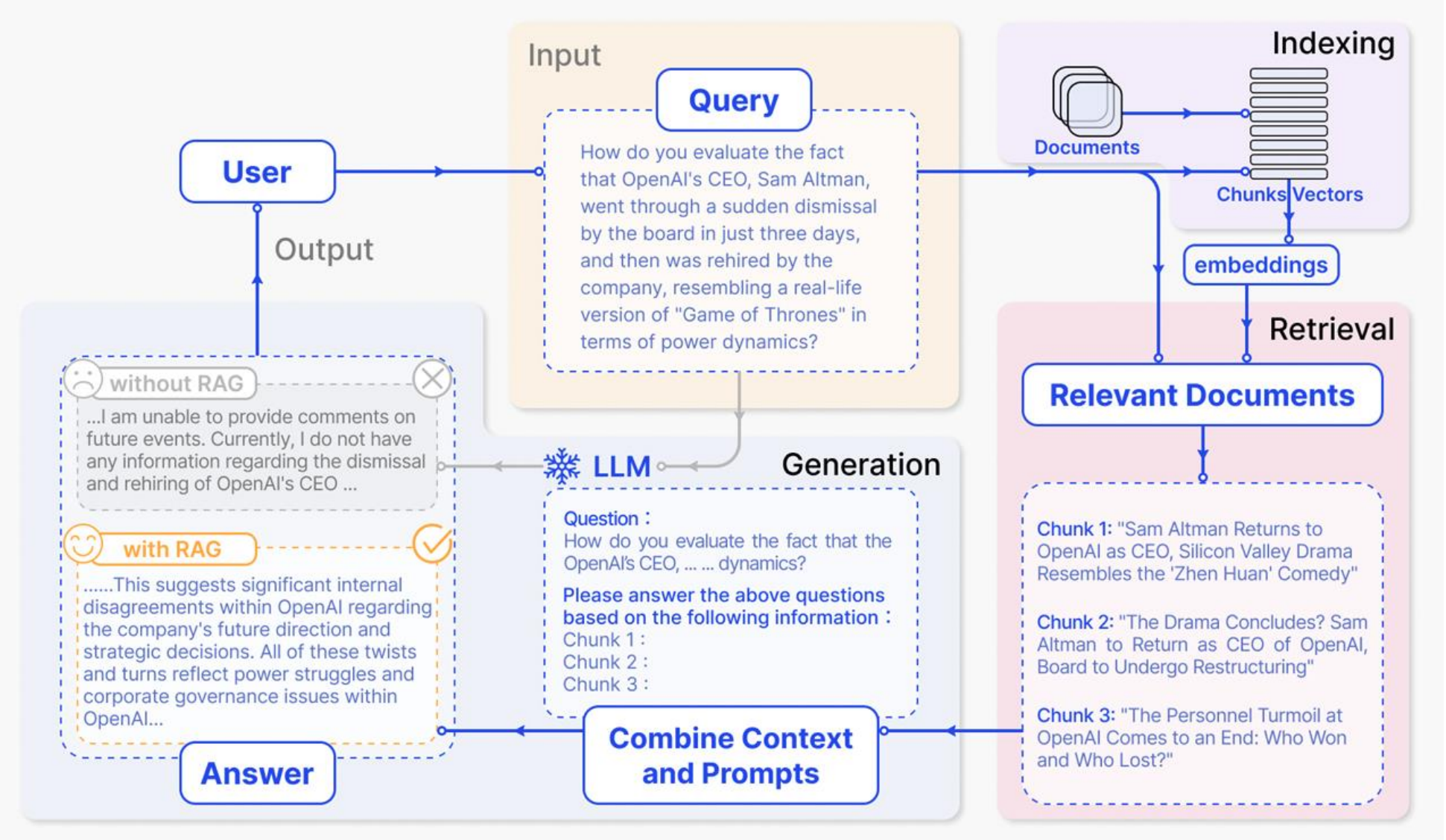


We are not able to obtain feedback from any CEO...

Presentation Preparation

1. Define Your Goals and Audience
2. Research and Gather Information
3. Conceptualize and Organize Content
4. Write and **Refine the Speech**
5. Create Visual Aids
6. Practice the Speech
7. **Handle the Q&A Session**
8. Final Checks and Adjustments

Retrieval Augmented Generation (RAG)



Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

To P.16

Company-Specific Records Matters More than Global Knowledge

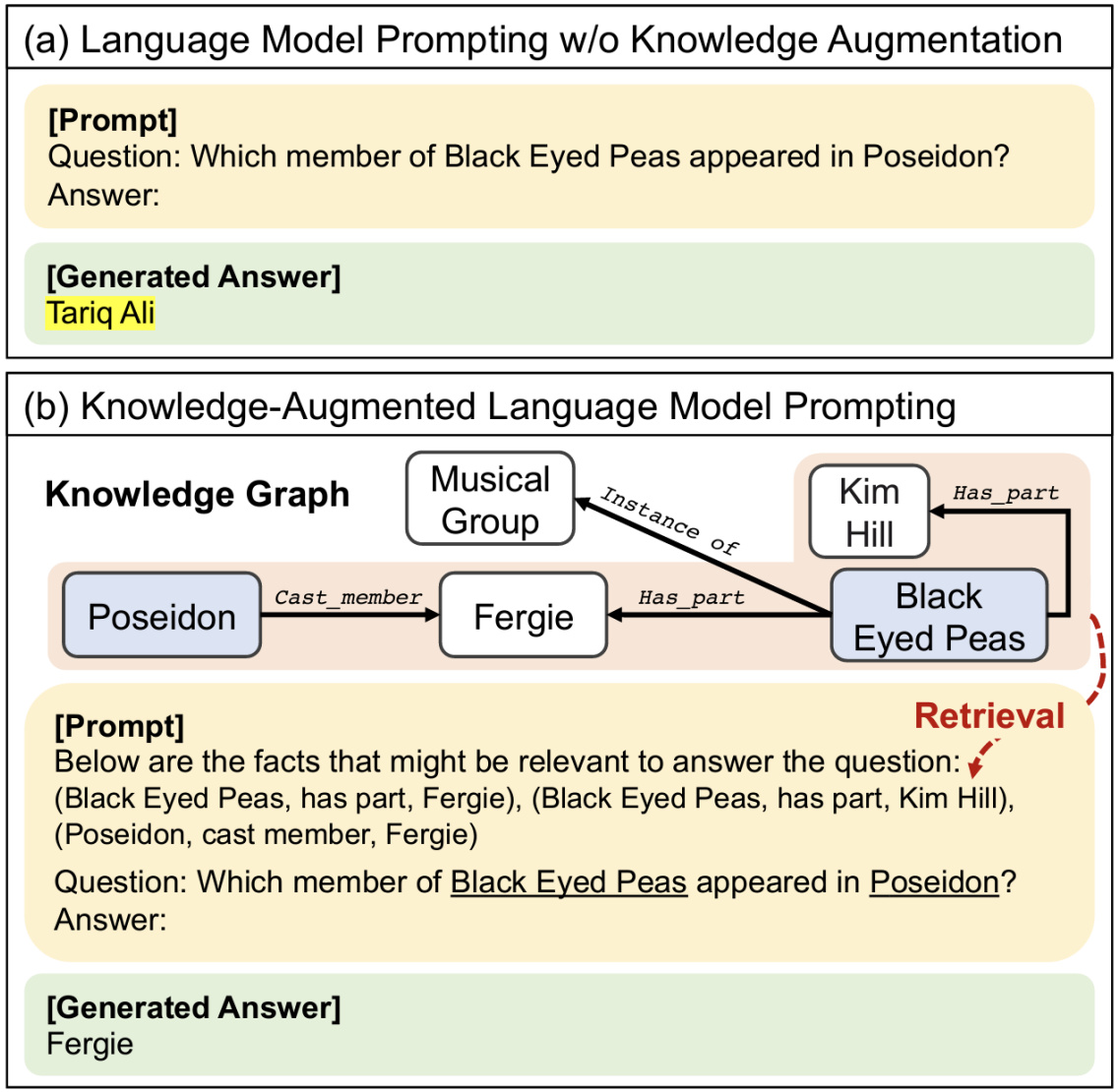
HAA LAB

Approach	ATR	ACR	IIR	Avg
LLaMA 3.1				
No knowledge	31.6	81.5	19.0	44.0
KG-RAG	35.8	80.9	**24.4**	47.1
QA-RAG (all-MiniLM-L12-v2)	33.9	78.7	22.0	44.8
QA-RAG (stella_en_1.5B_v5)	35.2	80.8	22.9	46.3
QA-RAG (gte-Qwen2-7B-instruct)	35.2	82.7	22.5	46.8
QA-RAG (NV-Embed2)	**35.8**	**83.6**	23.7	**47.7**
GPT-4o				
No knowledge	35.2	82.7	23.3	47.1
KG-RAG	40.2	83.8	28.7	50.9
QA-RAG (all-MiniLM-L12-v2)	42.1	83.5	31.0	52.2
QA-RAG (stella_en_1.5B_v5)	41.5	82.5	30.8	51.6
QA-RAG (gte-Qwen2-7B-instruct)	41.8	**84.6**	30.6	**52.3**
QA-RAG (NV-Embed2)	**42.3**	82.9	**31.6**	52.3

QA Pool	ATR	ACR	IIR	Avg
All Companies (All)	38.9	59.9	29.6	42.8
Company-Specific (CS)	**41.2**	**63.9**	**31.5**	**45.5**
All + CS	38.3	56.9	29.2	41.5

Company-Specific Knowledge Matters: Retrieval-Augmented Generation for Earnings Call Answer Rehearsal (Shih et al., 2024)

Knowledge-Augmented Language Model Prompting



Baek, Jinheon, Alham Fikri Aji, and Amir Saffari. "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering." *1st Workshop on Matching from Unstructured and Structured Data, MATCHING.* 2023.

Rehearsing Answers to Probable Questions with Perspective-Taking

	KG-AR	FinCaKG-FR	FinCaKG-ECT
Entities	4,824	1,717	546
Relations	41,007	11,633	1,802

Question
Do you think the timeframe for getting Forever 21 **EBITDA** positive will be similar to that of Aero?

Answer 1
I would say that – it's a good question. And I would say it's a little more complicated in a little bigger business. And it depends on whether one or two of my guys are going to spend all this time in Los Angeles. So, I'm negotiating it right now, Linda. Stay tuned.

Answer 2
Based on last quarter's **sales** increase in North America, I believe that it would be a similar trend.

LLM	KG	ATR	ACR	IIR
GPT-3.5	-	22.35	**45.41**	14.60
	FinCaKG-FR	21.62	42.42	16.00
	FinCaKG-ECT	22.38	34.12	16.34
	KG-AR (PT)	**24.06**	38.69	**18.54**
Gemini Pro	-	15.47	36.61	7.32
	FinCaKG-FR	11.83	26.60	6.21
	FinCaKG-ECT	13.83	25.94	8.68
	KG-AR (PT)	15.91	37.00	9.47
LLaMA-3 8B	-	19.53	21.41	17.62
	FinCaKG-FR	19.98	19.17	18.38
	FinCaKG-ECT	20.52	20.17	17.87
	KG-AR (PT)	19.08	18.62	17.31

Rehearsing Answers to Probable Questions with Perspective-Taking (Shih et al., 2024)

Real-world effectiveness is very difficult to evaluate using traditional metrics, LLMs, or even standard human evaluation

HAA LAB

LLM	KG-AR	INFO	CON
-	-	5.15	5.50
GPT-3.5	w/o	**6.26**	**6.24**
	w/	6.00	5.97
Gemini Pro	w/o	5.16	5.87
	w/	5.42	5.34
LLaMA-3 8B	w/o	6.16	6.13
	w/	6.37	6.08

Table 4: Human evaluation.The first row displays the baseline scores of managers' answers.

	Pearson	Spearman	Kendall
INFO	0.30	0.28	0.23
CON	0.41	0.38	0.31

Table 5: Correlation between scores evaluated by GPT-4 and human.

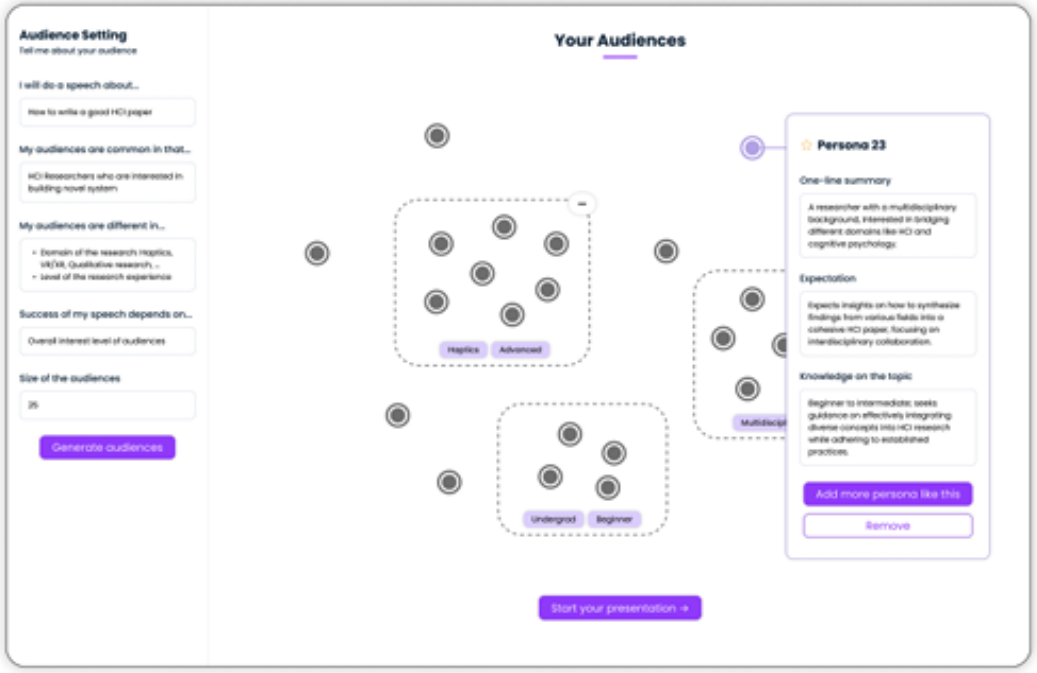
'Ruined my Christmas spirit': McDonald's removes AI-generated ad after backlash

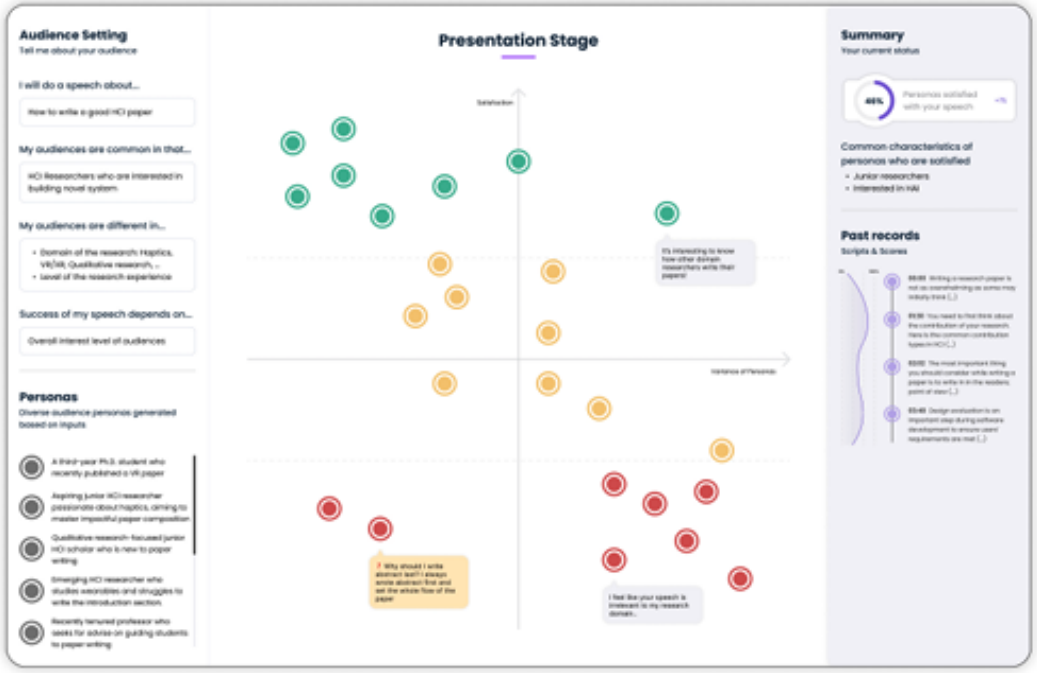
Commercial in Netherlands depicting festival-season chaos at 'most terrible time of year' prompted flurry of criticism online

LLM-Generated Audiences for Public Speech Practice

HAA LAB

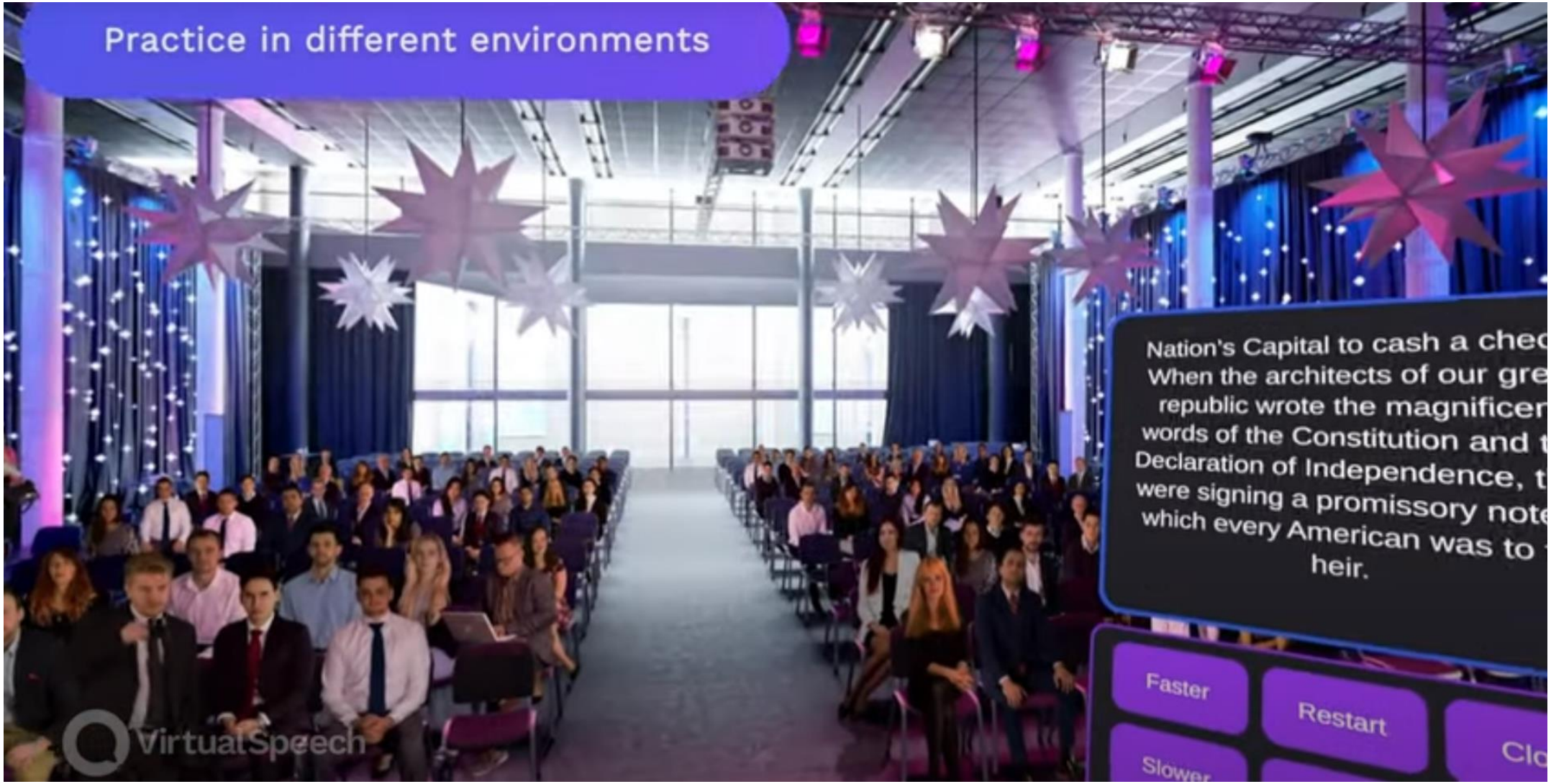
- Simulates diverse audiences using large language models
- Allows configurable audience personas (background, knowledge, interest)
- Provides real-time feedback, scores, and audience questions during practice
- Visualizes audience reactions to highlight effective and weak speech segments
- Supports speech refinement for tutorials, presentations, and debates





AudiLens: Configurable LLM-Generated Audiences for Public Speech Practice (Park, 2023)

Immersive Soft Skills Training



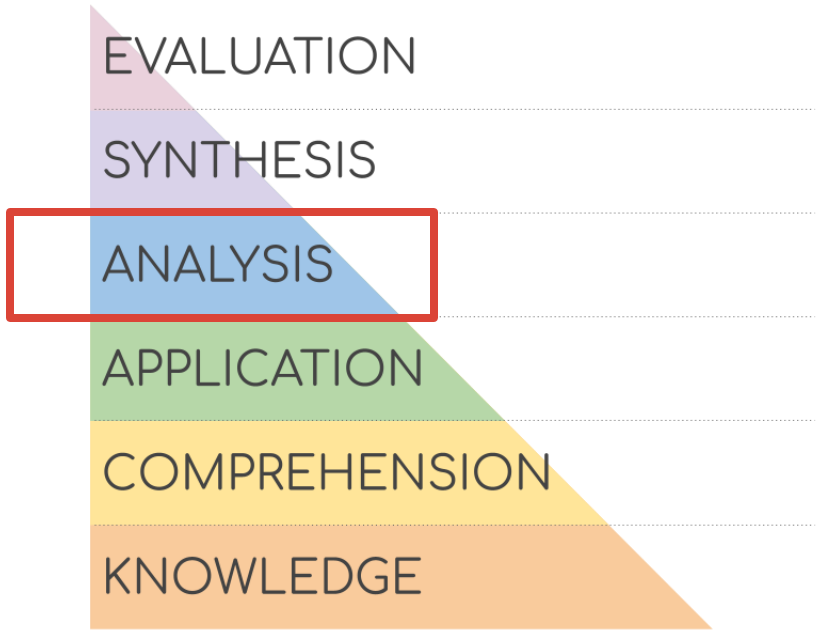
https://virtualspeech.com/

Audience Feedback

Generator-Guided Crowd Reaction Assessment (Ghosh et al., 2024)

Outline

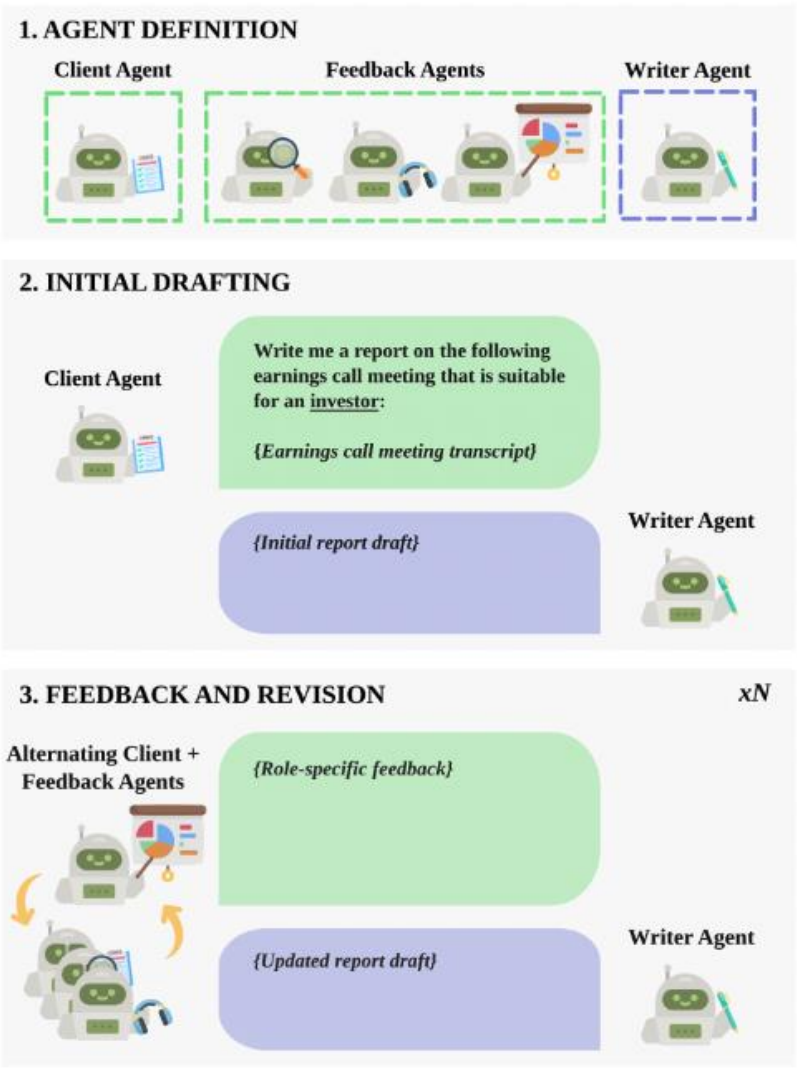
- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- **Scenarios (Interaction & Evaluation)**
  - Presentation Preparation (Intrinsic Evaluation)
  - **Analysis Generation (Extrinsic Evaluation)**
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - Agent-Based Modeling (Simulation)
- Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- Proposal: Open Agent Platform



Report Generation (Audience Feedback/Reaction)

From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls? (Goldsack et al. 2025)

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

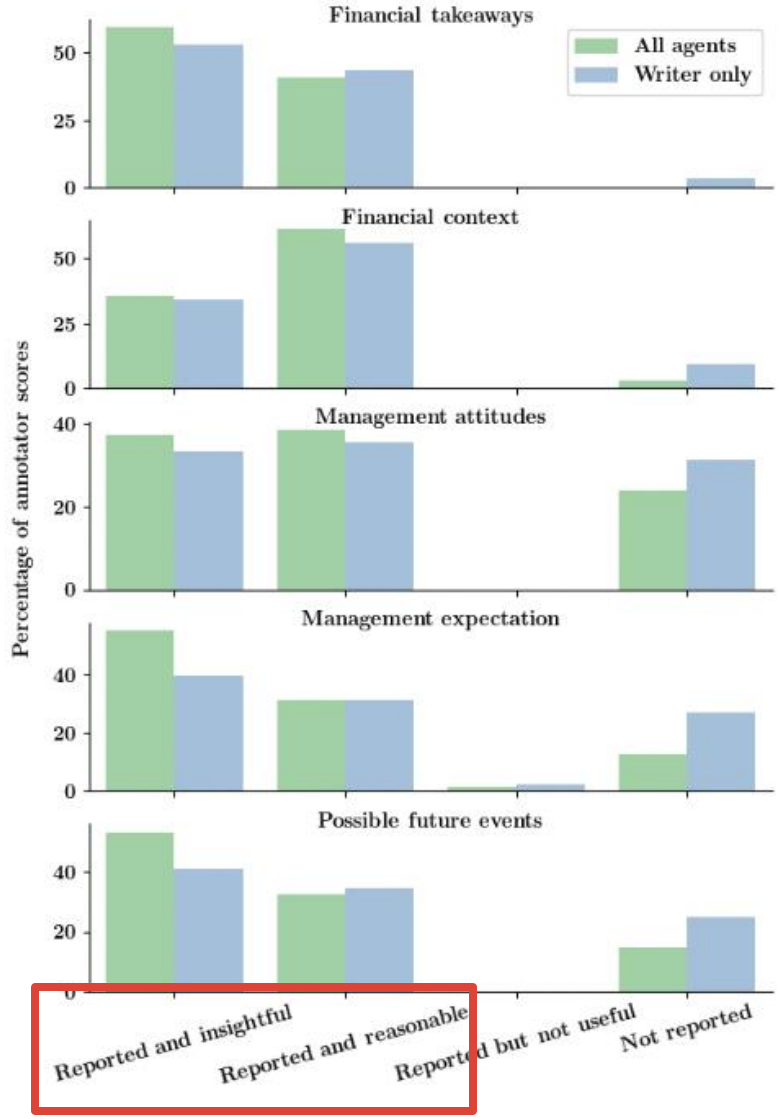
- **Expert-written reports better than agent-written**
- **LLMs have preference to agent-written reports**
- **Mistral is influenced by the order**

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

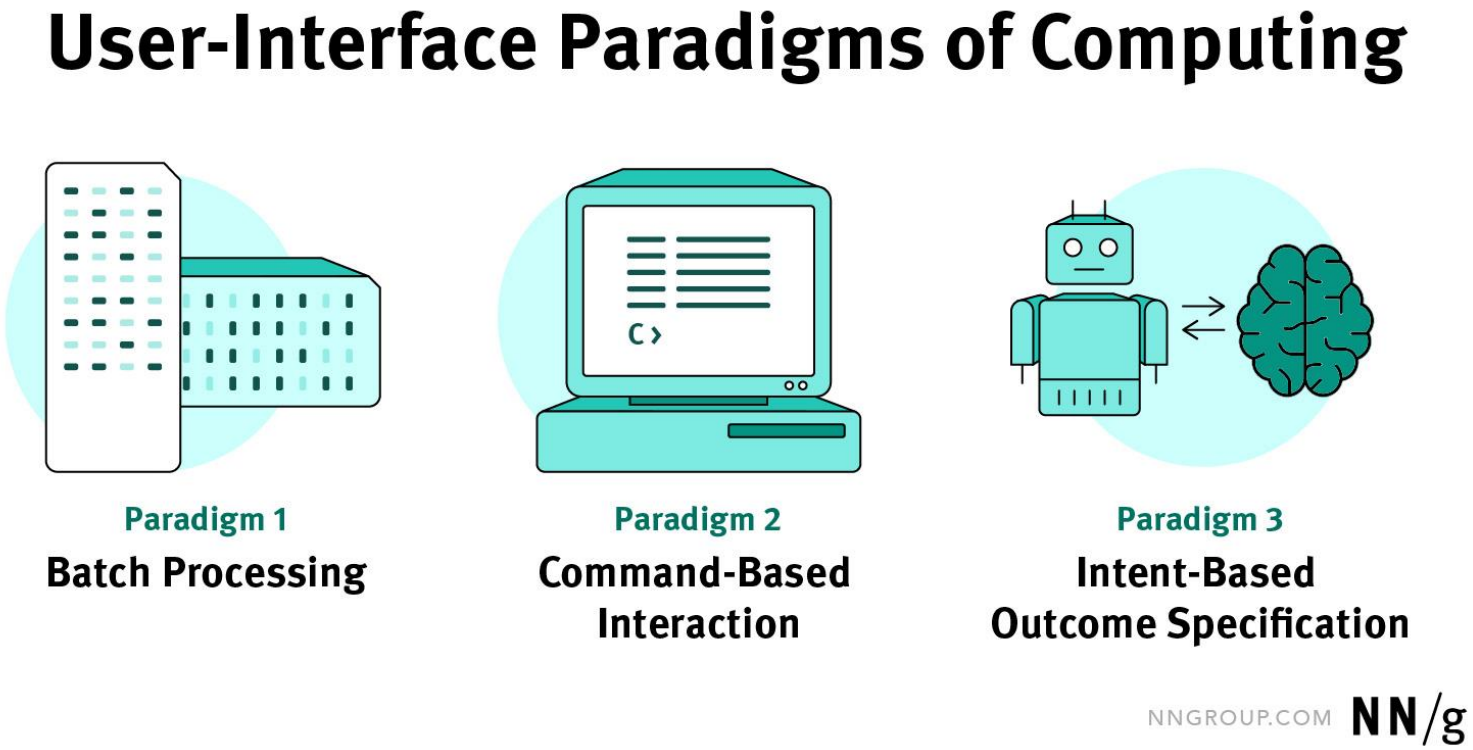
Table 1: Average accuracy of financial decisions across time horizons.

Human-Agent Teaming Era

HAA LAB

Evaluation would go beyond accuracy & speed

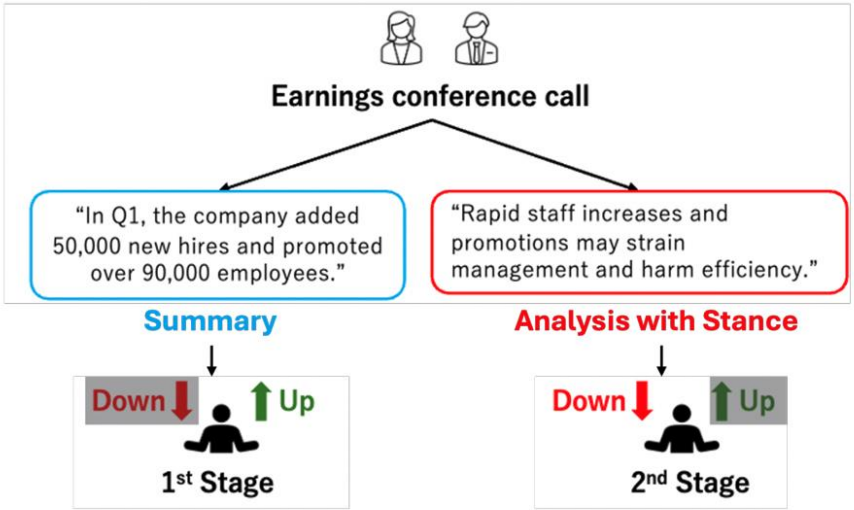
The extent to which the system benefits user/human matters



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.



Can GPT-4 Sway Experts' Decisions? (Takayanagi, 2025)

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

- 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%

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

HAA LAB

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

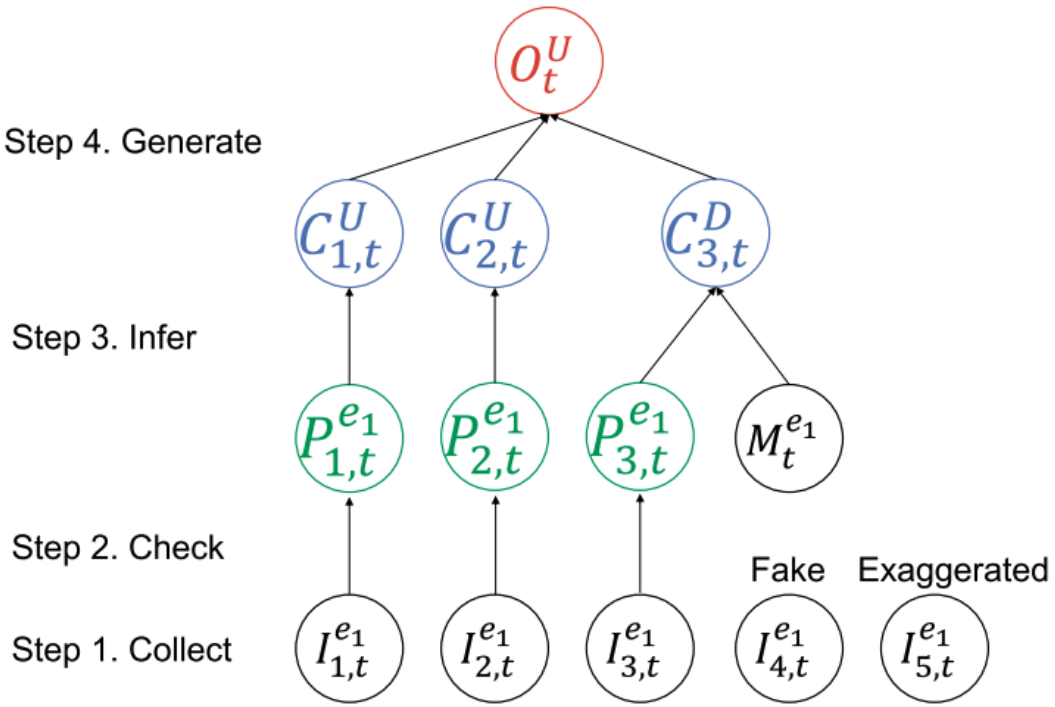
- What AI Agents Do Better
  - Higher novelty: AI-generated ideas are rated significantly more novel by expert reviewers
  - Scalability: Can generate and explore a large space of candidate ideas quickly
  - Creative recombination: Effective at combining existing concepts in unexpected ways

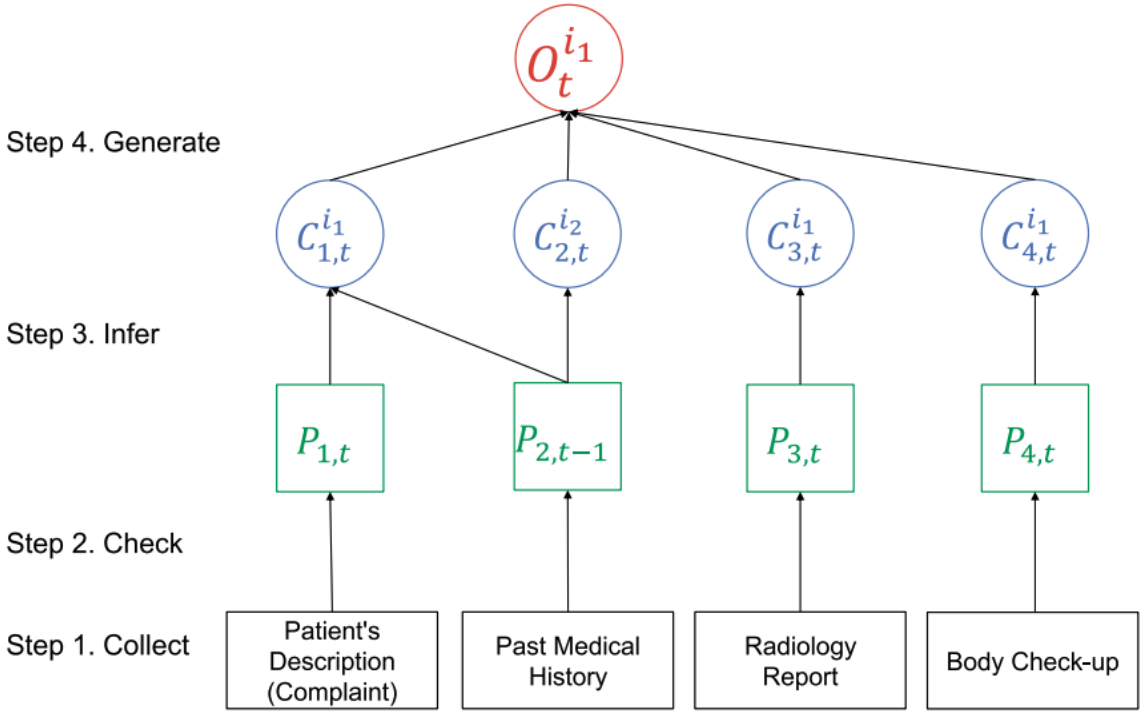
- What Humans Do Better
  - Feasibility & grounding: Human ideas tend to be more practical and execution-aware
  - Use of domain intuition: Better alignment with established research practices and constraints
  - Judgment & evaluation: Humans are more reliable at assessing idea quality and feasibility

- 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

Can LLMs Generate Novel Research Ideas? (Si et al., 2024)

Steps: Generating Reports





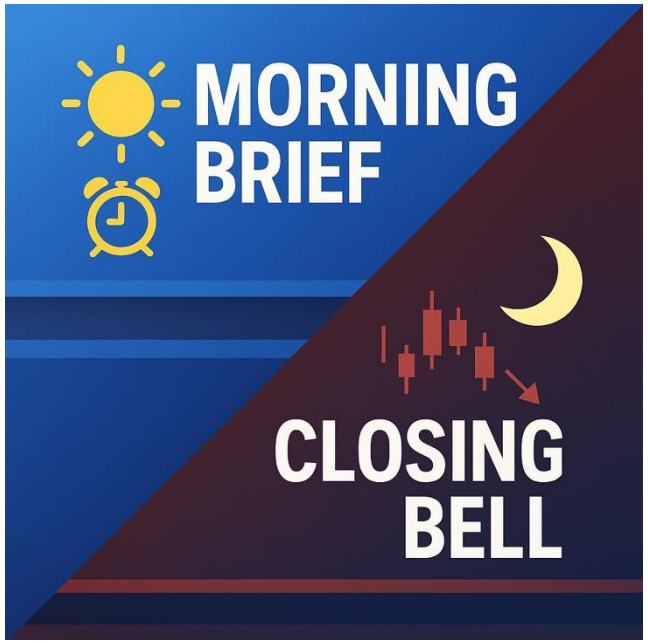
Applying the workflow of argument mining in the clinical scenario

From Opinion Mining to Financial Argument Mining (Chen et al., 2021)

Market Digest

Morning Brief

- **Timing**
  - Published *before* the market opens
- **Primary Purpose**
  - Prepares investors for the upcoming trading day
  - Supports **intraday decision-making**
- **Information Sources**
  - Previous day's market performance
  - Overnight international news and macroeconomic events
  - Pre-market indicators and expectations
- **Decision Horizon**
  - Predicts **same-day price movements**



Closing-Bell Report

- **Timing**
  - Published *after* the market closes
- **Primary Purpose**
  - Summarizes what happened during the trading day
  - Supports **overnight or next-day decisions**
- **Information Sources**
  - Full-day price movements
  - Trading volume and institutional flows
  - Market reactions to intraday events
- **Decision Horizon**
  - Predicts **next-day opening movements**

Three Sources of Market Digests

Human-written digests

- Authored by professional financial journalists

LLM-generated (data-driven) digests

- Generated by LLMs based on market data and performance signals
- Asset selection driven by volatility, trading volume, and institutional flows

LLM-generated with expert-guided asset selection

- Financial experts first select key companies or sectors
- LLM generates the narrative based on the curated focus set

Decision-oriented Text Evaluation (Huang, 2025)

Human-in-the-loop Guidance on Asset Selection provides the Highest Value

Key Finding 1: LLM-Generated Morning Briefs Are More Useful

- Consistent improvement for **both human investors and LLM investors**
- Using LLM-generated morning briefs leads to **higher decision accuracy**
- **Interpretation**
  - Traditional journalism provides rich information but weak actionable signals
  - LLMs excel at **distilling information into decision-relevant insights**

Key Finding 2: Asymmetric Effects in Closing-Bell Reports

- **LLM investors**
  - Perform best when using **human-written closing-bell reports**
- **Human investors**
  - Perform better when using **LLM-generated closing-bell reports**
- **Implication**
  - The effectiveness of a market digest depends on **who the reader is**
  - Evaluation should consider **human–model interaction**, not text quality alone

Key Finding 3: Expert-Guided Asset Selection Yields the Best Outcomes

- Expert-curated focus significantly improves decision accuracy
- Human expertise remains critical for **what to cover**
- Full human authorship is unnecessary;
  human-in-the-loop guidance on asset selection provides the highest value

	Investor	Morning Briefs			Closing-Bell Reports		
		Journalist	Performance-Based	Professional-Insight	Journalist	Performance-Based	Professional-Insight
LLM	Claude-3-5-Sonnet	38.85	45.98	**48.01**	**65.56**	55.62	60.07
	Gemini-2.0-Flash	44.89	**46.98**	42.41	**61.60**	54.36	58.25
	GPT-4o	42.35	42.53	**43.15**	**58.51**	56.89	55.17
Human	A	39.64	43.10	**48.61**	47.59	48.97	**56.71**
	B	36.67	**45.11**	40.23	50.83	49.44	**53.45**
	C	34.40	**49.27**	48.35	42.24	**75.00**	54.18

Decision-Focused Summarization (Hsu, 2021)

HAA LAB

- **Problem**
  - Traditional summarization defines relevance based only on text
  - This can hurt decision-making (e.g., irrelevant details included)
  - Goal: summaries that **support a specific decision**, not just readability

- **Key Idea**
  - Introduce **Decision-Focused Summarization**
  - Use a **trained decision model** to guide which sentences are summarized
  - A good summary should lead to the **same decision as the full text**

- **Takeaway**
  - What matters for decisions ≠ what matters for text quality
  - Decision-focused summaries are more useful for **human decision support**
  - Promising direction for high-stakes domains (healthcare, finance)

How about Live Commentary during the Presentation?

KAMALA HARRIS: And I am proud that as vice president over the last four years, we have invested a trillion dollars in a clean energy economy while we have also increased domestic gas production to historic levels.

Summary: While US oil production has been increased to "historic levels," she says, the country has created manufacturing jobs tied to the clean energy shift.

Commentary (Supplementary Explanation): US oil production has hit an all-time high under Biden, and the US is a net exporter of petroleum products, thanks in part to a boom in exports of liquefied natural gas, according to government data.

Bloomberg

reddit

	Debates	FOMC	Earnings Call	Reddit
# Pair	2,283	252	1,115	366
# Category	11	5	10	4

Live Commentary Planning and Generation (Chen et al., 2025)

GenChal-2026
http://genchal.nlpfin.com/

U.S. Presidential Debates (2016–2024) – Professional Commentary

Dataset	Source	Period	Topic	# of Labels
Check-Worthy (Patwari et al., 2017)	D	2016	Fact-Checking	2
CLEF (Atanasova et al., 2018)	D	2016	Fact-Checking	2
Claim-Rank (Atanasova et al., 2019)	D	2016	Fact-Checking	2
CMU (Jo et al., 2020a)	R	2016	Proposition Type	4
M-Arg (Mestre et al., 2021)	D	2020	Argument Mining	3
DR-CUP (Proposed)	D & C	2016-2024	Commentary Aspect	11

- **Key Summary (KS)**: This label indicates that the commentator is summarizing points raised by the debate moderators or contestants.
- **Supplementary Explanation (SE)**: This label is used when the commentator provides additional context or information sourced from experts, real-world events, or the current debate situation without expressing subjective opinions.
- **Fact-checking (FC)**: Verifies the accuracy of candidates' statements or external rumors.
- **Market Reactions (MR)**: Highlights commentary related to economic fluctuations or monetary market trends.
- **Public Opinion (PO)**: Represents descriptions of public sentiment on specific issues or polling trends.
- **Commentator's Question (CQ)**: Indicates that the commentator is posing a question about a particular issue.
- **Commentator's Personal Opinion (CPO)**: This label captures instances where the commentator voices their viewpoint on a particular issue. It includes five subcategories:
  - **Performance of the Contestants (PC)**: Assesses contestants' discussion performance.
  - **Candidate Statements (CS)**: Analyzes specific claims made by the contenders.
  - **Analyzing or Conclusions (AC)**: Involves inferences or conclusions drawn by the commentator about a statement or occurrence.
  - **Market Performance (MP)**: Pertains to comments regarding the economic performance of a nation or stock market trends.
  - **Others**: Covers commentary on topics not addressed by the other sub-labels.

Statistics

HAA LAB

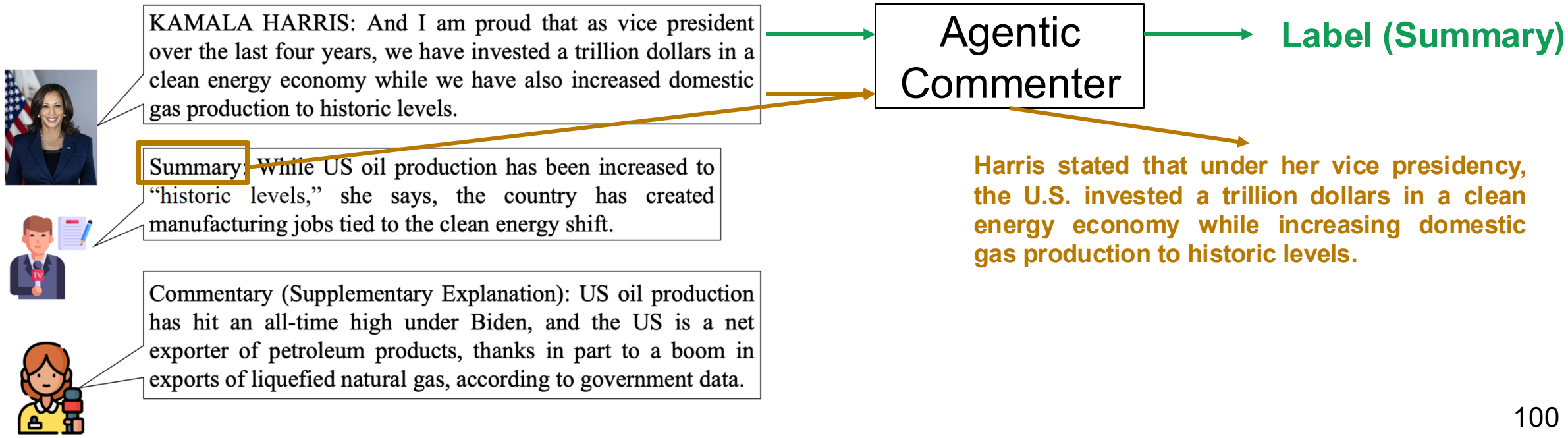
Year	Event	KS	FC	SE	CQ	PO	MR	CPO					Total
								PC	CS	AC	MP	Others	
2016	First U.S. Presidential Debate	119	3	12	0	1	9	16	20	4	1	7	192
2016	Second U.S. Presidential Debate	166	24	12	5	0	4	19	10	6	0	14	260
2016	Third U.S. Presidential Debate	154	22	26	0	2	9	25	9	6	0	9	262
2020	First U.S. Presidential Debate	234	6	104	1	3	2	12	14	4	0	5	385
2020	U.S. Vice Presidential Debate	155	8	53	2	0	0	6	11	2	0	4	241
2020	U.S. Presidential Debate	221	5	70	0	1	1	8	12	1	0	3	322
2023	Republican Party Presidential Debate	94	3	24	0	0	0	8	5	1	0	3	138
2023	Republican Party Presidential Debate	49	3	3	0	0	0	9	10	0	0	2	76
2024	Biden-Trump Presidential Debate	76	13	11	0	1	9	11	8	4	2	1	136
2024	Harris-Trump Presidential Debate	128	12	68	0	4	6	8	22	16	1	6	271
Total		1389	99	384	9	12	40	123	124	44	4	55	2283

Tasks: Planning and Generation

Input: Transcript segment of a live event (debate, press conference, or earnings call)

Planning: Decide what kind of insight—summary, fact-check, or opinion—to provide in real time.

Generation: Produce fluent, context-aware commentary comparable to professional analysts, i.e., generate expert-like commentary conditioned on transcript + label.



Taking Control of AI-Generated Live Commentary

HAA LAB

- **Experimental Design Overview**
  - Used U.S. presidential debate transcripts as controlled language input
  - Applied targeted synonym substitution at key lexical positions
  - Employed an LLM in dual roles:
    - as a political commentator
    - as a simulated general audience
  - Measured how small lexical changes affect generated commentary and perceived audience reactions
- **How Lexical Choices Shape AI-Generated Commentary**
  - Minor wording changes can significantly alter AI-generated commentary and audience perception
  - More positive wording does not necessarily lead to more positive perception
  - Sentiment and stance contribution are independent dimensions in LLM interpretation
  - Targeted lexical edits influence audience perception with ~40% success rate
- **Practical Control of Auto-Generated Live Commentary**
  - AI-generated commentary is highly prompt- and wording-sensitive
  - Effective control relies on fine-tuning key lexical positions, not full script rewrites
  - Optimization should align with communicative goals, not sentiment alone
  - Small script tweaks can strategically steer AI-generated narratives

Credit: Yu-Yu Chang

Political Bias and Prompt Sensitivity Across LLMs

HAA LAB

- **Experiment**
  - Subjects: 32 legislators with available Facebook data
  - Models: GPT-4.1, Gemini 2.5 Flash, Claude 4 Sonnet, Llama 3.3 (70B)
  - Task 1: Evaluation Task
    - Generate political commentary without explicit stance instruction
    - Attack condition: explicitly instructed to produce negative evaluations
  - Task 2: Stance Imitation Task
    - Generate a new comment by mimicking the tone and stance of input comments
    - Input set: 50 pro-recall and 50 anti-recall online comments
- **Findings**
  - LLMs exhibit systematic political bias even without explicit attack prompts
  - Baseline stance varies across models (positive, negative, or anti-recall tendencies)
  - The model is easily influenced by user intention
    - Under explicit stance or attack instructions, most models strongly comply with the requested direction
    - Balanced input data (50% pro / 50% anti recall) does not produce balanced outputs
    - Models differ substantially in their susceptibility to stance amplification
  - Results suggest that political stance emerges from prompt structure and model-specific priors, not input balance alone

Credit: Huan-Wen Ho

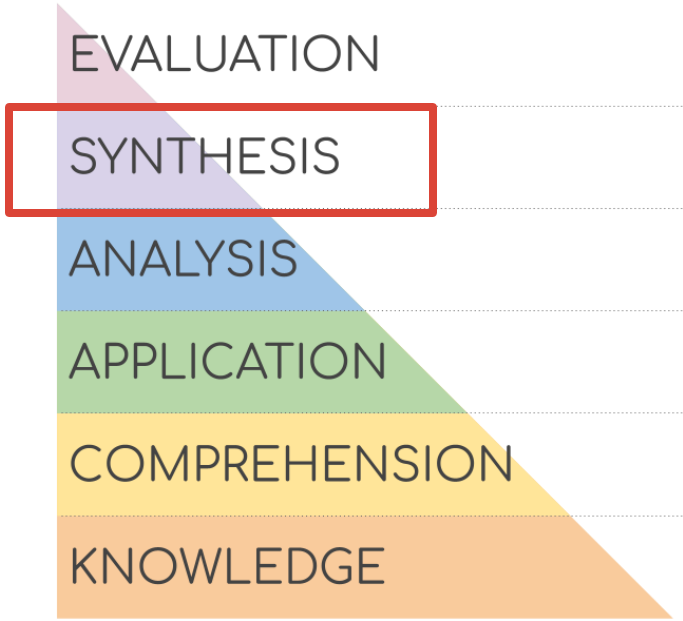
For Analysis Generation, LLMs Are Not Neutral and Easily Influenced

- **What the paper did**
  - Compared LLM-generated news analysis with human-written news (NYT / Reuters)
  - Evaluated bias at **word, sentence, and document levels**
- **What they found**
  - All LLMs show **systematic gender and racial preferences**
  - Bias is **directional**: against women and Black individuals
  - ChatGPT shows *lower average bias*, but stronger alignment once biased prompts pass filters
- **Why this matters for analysis generation**
  - Analysis is not neutral reasoning
  - It shapes **what is emphasized, downplayed, or omitted**
  - Model “preferences” can quietly influence human judgment

Bias of AI-generated content: an examination of news produced by large language models (Fang et al., 2025)

Outline

- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- **Scenarios (Interaction & Evaluation)**
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - **Creative Idea Generation (Reproducible Extrinsic Evaluation)**
  - Agent-Based Modeling (Simulation)
- Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- Proposal: Open Agent Platform



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

Is it Possible to Reproduce the Human Rating/Decision?

HAA LAB

- **Goal**
  - Test whether AI agents can **reproduce individual human attitudes and decisions**, not just population averages.
- **Method**
  - Conducted **2-hour in-depth interviews** with 1,052 real individuals.
  - Used interview transcripts to create **LLM-based generative agents**, each representing one person.
  - Asked both humans and agents to complete the **same surveys and behavioral experiments**.
- **Evaluation**
  - Compared agent predictions to human responses,
    normalized by **how consistently humans replicate their own answers after two weeks**.
- **Key Results**
  - Agents achieved **~85% of human self-consistency** on the General Social Survey.
  - Accurately predicted **personality traits, economic decisions, and experimental treatment effects**.
  - Interview-based agents **outperformed demographic or persona-based models** and reduced bias.
- **Takeaway**
  - With rich individual-level data, **AI agents can reproduce human ratings and decisions at near-human reliability**.

Generative Agent Simulations of 1,000 People (Park et al., 2024)

Align LLM Evaluation with Human Expert Judgments

HAA LAB

- **Model & Task**
  - Base model: **Llama-3.1-8B-Instruct**
  - Task: Align LLM proposal evaluation with **human expert judgments**
- **Experts & Metrics**
  - **9 experts** with distinct evaluation focuses
  - *Technical*: specificity, technical validity, innovation, competitive advantage
  - *Market*: specificity, need validity, market size
- **Data**
  - **30–70 proposals per expert**
  - Evaluation scores provided by domain experts
- **Model Editing Method**
  - Up to **10 proposals per expert**
- **4 edits per proposal**
  - No full fine-tuning
  - **Editing Variants**
    - Expert background
    - Explicit evaluation criteria
    - Expert reasoning process
- **Baselines**
  - Zero-shot
  - Few-shot
  - Fine-tuned model

Credit: Tzu-Mi Lin

Model editing is an effective and data-efficient alternative to fine-tuning

- **Overall Performance**
  - Model editing consistently improves **total accuracy** over zero-shot
  - Typical gain: **+5% to +15%**
  - Competitive with fine-tuning using far less data
- **Metric-Level Observations**
  - **Specificity** often decreases
    → highly subjective and imagination-dependent
  - **Technical / Need Validity** show clear improvement
    → especially with criteria or reasoning edits
- **Innovation & Competitive Advantage**
  - Limited accuracy gains
  - Score distributions shift **closer to expert evaluations**
- **Editing Strategy Insights**
  - Background: stable but moderate gains
  - Criteria: effective for rule-based metrics, may increase trade-offs
  - Reasoning: most robust, reduces score volatility

Credit: Tzu-Mi Lin

Outline

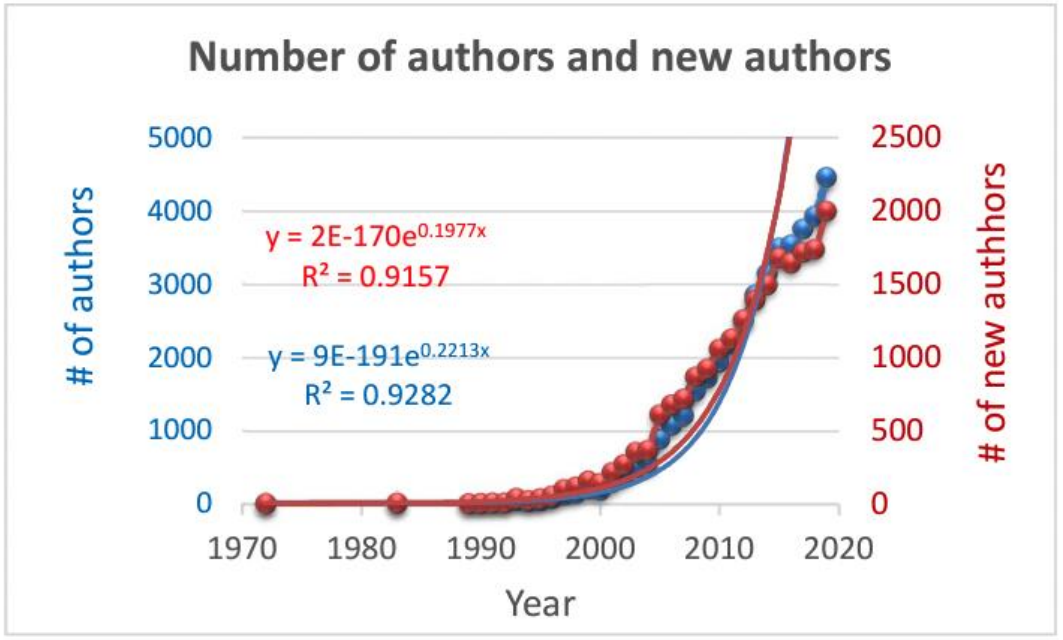
- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- **Scenarios (Interaction & Evaluation)**
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - **Agent-Based Modeling (Simulation)**
- Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- Proposal: Open Agent Platform

Reproducible Human-Centered Experiments via Agent Simulation

- **Problem**
  - Human-in-the-loop experiments are inherently hard to reproduce
  - Small numbers of annotators introduce noise and individual bias
- **Key Idea**
  - Replace stochastic human participation with **fixed, parameterized agent models**
  - Each agent approximates a **frozen individual behavioral policy**
- **Method**
  - Construct a **heterogeneous agent population**
  - Run large-scale simulations with fixed random seeds
  - Estimate outcomes from the **distribution over agents**, not single labels
- **Outcome**
  - Fully reproducible experiments
  - Population-level statistics that approximate real-world human response distributions

Agent-Based Modeling

- Before we have LLMs



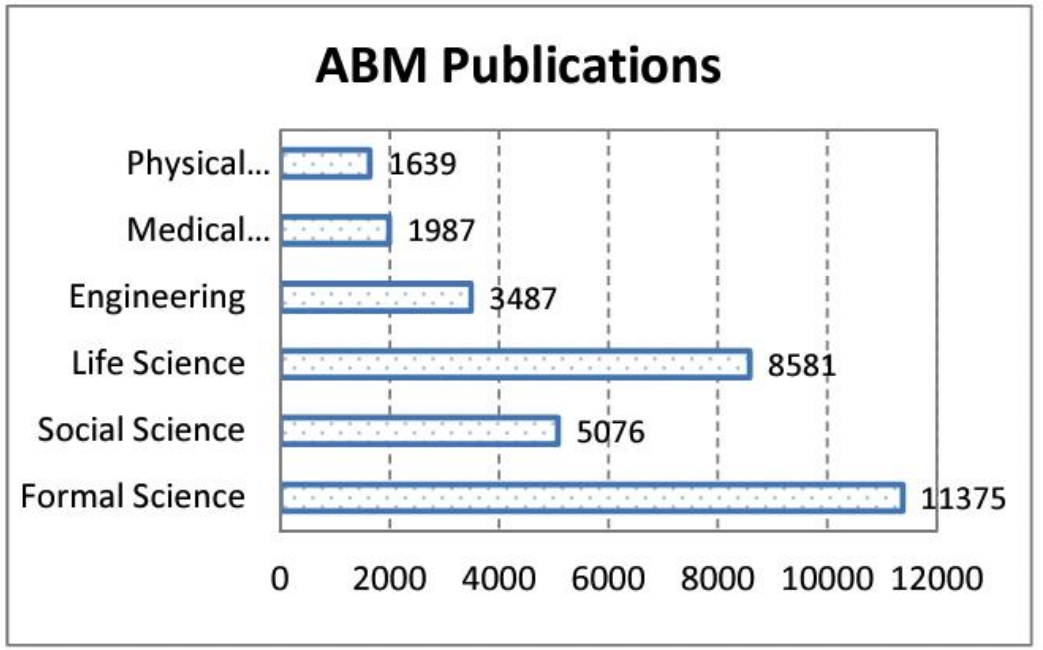


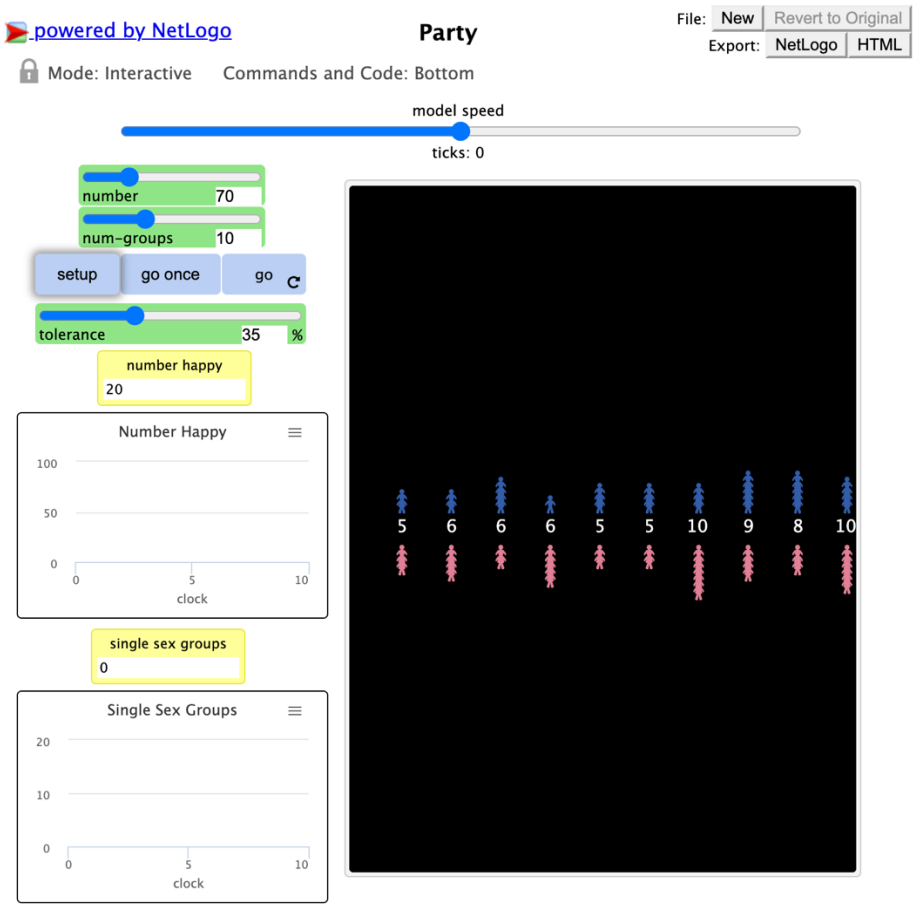
Figure 1. Number of authors and new authors over time (data as of 17 February 2020). Blue and red represent authors and new authors who develop and use ABM over time, respectively. For data collection see An et al. 2021).

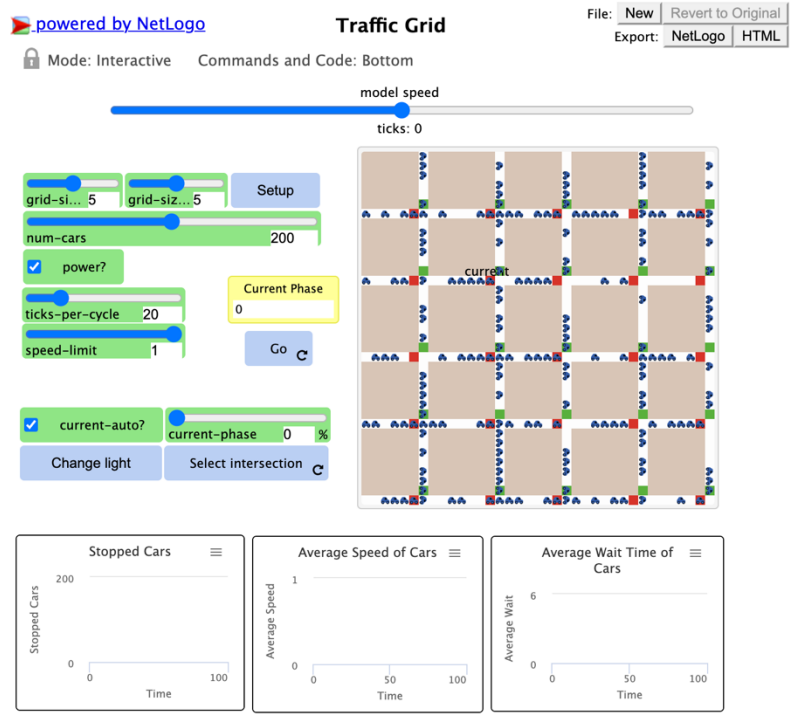
An, Li, et al. "Challenges, tasks, and opportunities in modeling agent-based complex systems." *Ecological Modelling* 457 (2021): 109685.

Previous Tools: Numerical Based

Computational Modelling Strength / Models' Scalability Level	Simple/Easy	Moderate	Complex/Hard
Extreme-scale		Repast HPC	MATSIM, PDES-MAS, Swarm
High /Large-scale	Altreva Adaptive Modeler, SeSAM	AnyLogic (2D/3D), AOR Simulation, CloudSim, CybelePro, FLAME, LSD (2D/3D), MASS, Pandora, UrbanSim	Agent Cell (2D/3D), Brahms, BSim (2D/3D), D-OMAR, Echo, Ecolab, FLAME GPU (3D), GridABM, HLA_Agent, HLA_RePast, Repast-J or Repast-3, Repast Symphony (2D/3D)
	NetLogo (2D/3D)	Ascape, CRAFTY, GAMA (2D/3D), SimEvents (MATLAB®), Simio (2D/3D), Simul8 (2D/3D)	MASON (2D/3D)
Medium-scale	JAS, VSEdit	Agent Factory, Breve (3D), Cormas, Envision, GALATEA, IDEA, JAMSIM, Janus, JASA, JAS-mine, MACSimJX, Mathmatica® (Wolfram), Mimosa, MIMOSE, Mobility Testbed, Modgen, OBEUS, SimAgent, SimBioSys, TerraME, Xholon (2D/3D)	DigiHive, MASyV (2D/3D)
	AgentSheets, BehaviourComposer (2D/3D), FlexSim (2D/3D)	Eve, ExtendSim (2D/3D), GROWLab, Insight Maker, Mesa	
Light-weight /Small-scale	AgentScript, Framsticks (2D/3D), JAMEL, JCASim (1D/2D/3D), JES, MOBIDYC, PedSim, PS-I, Scratch (2D/3D), SimJr, SimSketch, SOARS, StarLogo, StarLogoTNG (3D), Sugarscape, VisualBots	SEAS (2D/3D)	

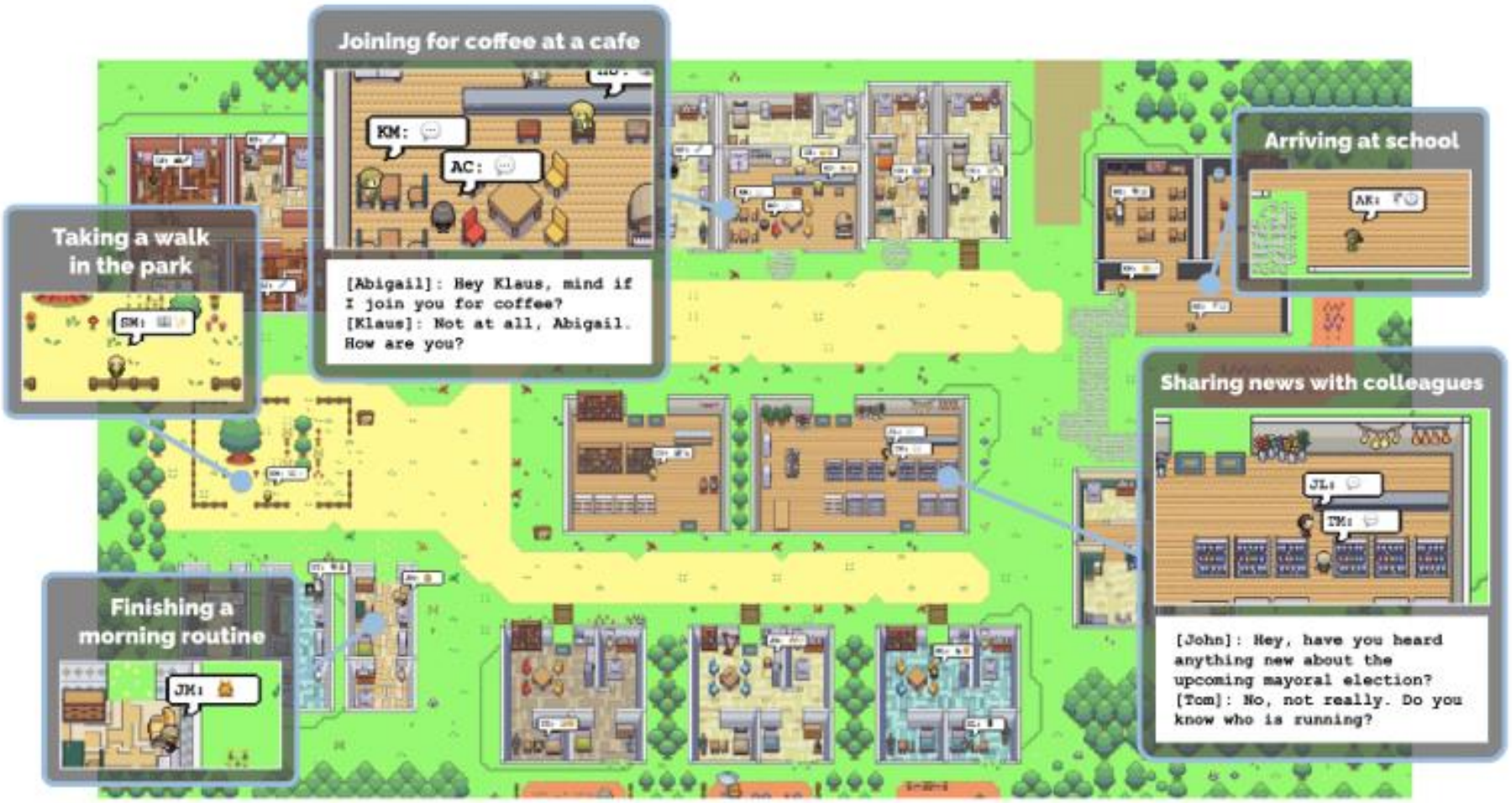
Model Development Effort →





Abar, Sameera, et al. "Agent Based Modelling and Simulation tools: A review of the state-of-art software." *Computer Science Review* 24 (2017): 13-33.

Human Behavior Simulation



Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." *Proceedings of the 36th annual acm symposium on user interface software and technology*. 2023.

2005 Nobel Prize Laureate in Economic Sciences

HAA LAB

Even a small preference for same-group neighbors by individuals can lead to large-scale, stark segregation in a community

Thomas C. Schelling

Facts

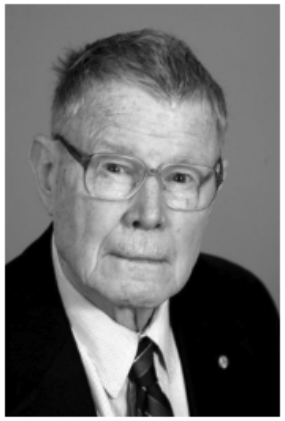


Photo: T. Zadig

Thomas C. Schelling
Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2005

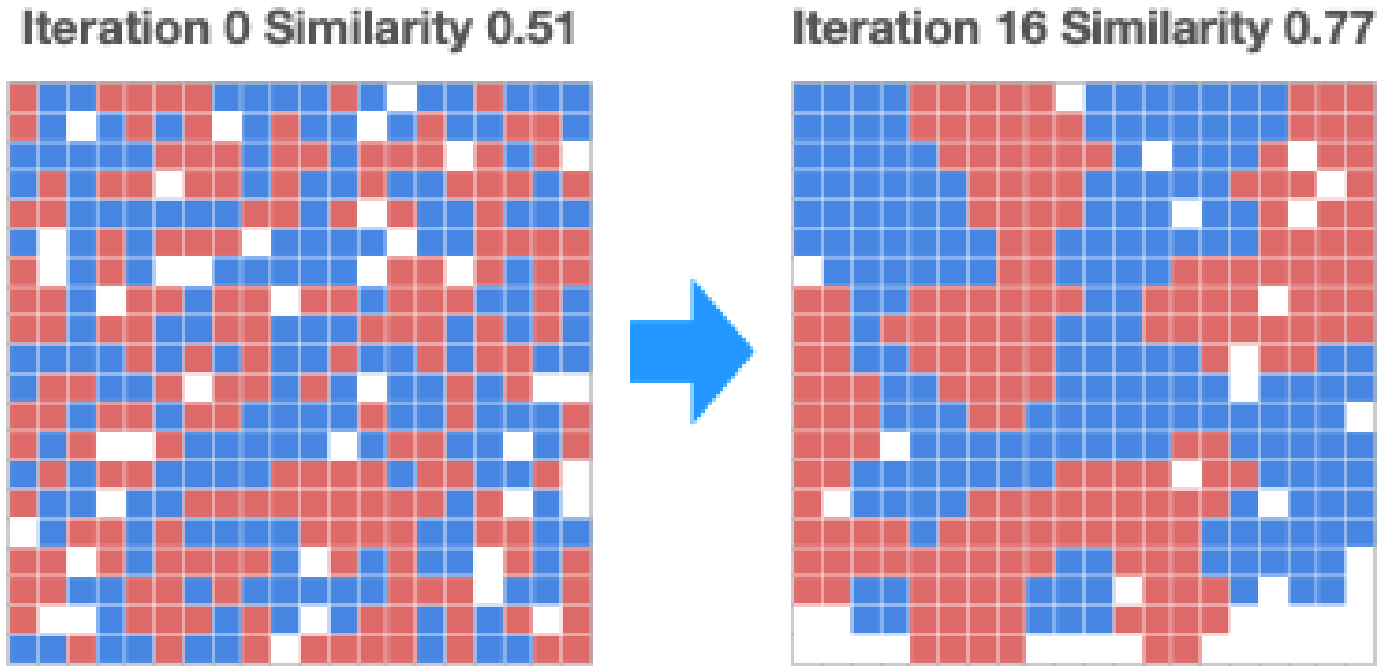
Born: 14 April 1921, Oakland, CA, USA

Died: 13 December 2016, Bethesda, MD, USA

Affiliation at the time of the award: University of Maryland, Department of Economics and School of Public Policy, College Park, MD, USA

Prize motivation: "for having enhanced our understanding of conflict and cooperation through game-theory analysis"

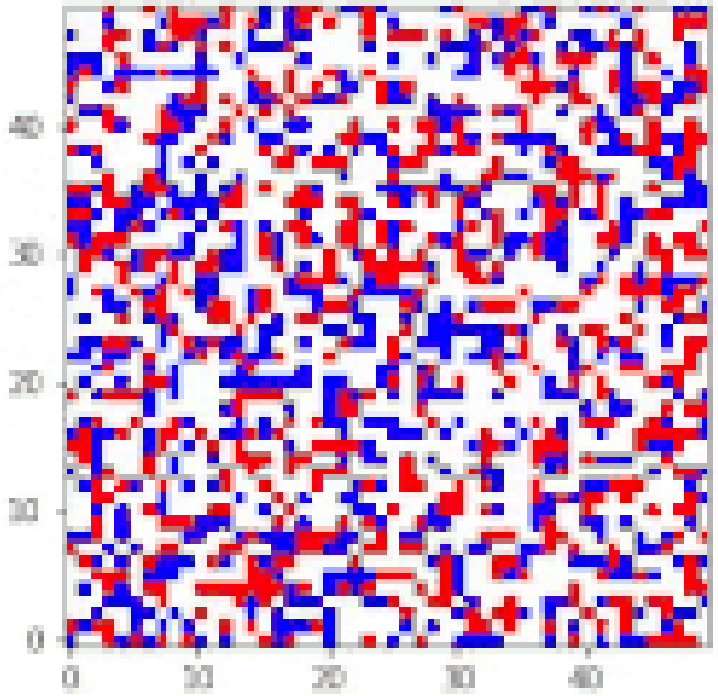
Schelling's Model of Segregation



Schelling's Model of Segregation

HAA LAB

- Thomas Schelling – 2005 Nobel Memorial Prize in Economic Sciences
- People's small preference will lead to highly segregation
- **If more than 30% of my neighbors are in different group, I'll move.**

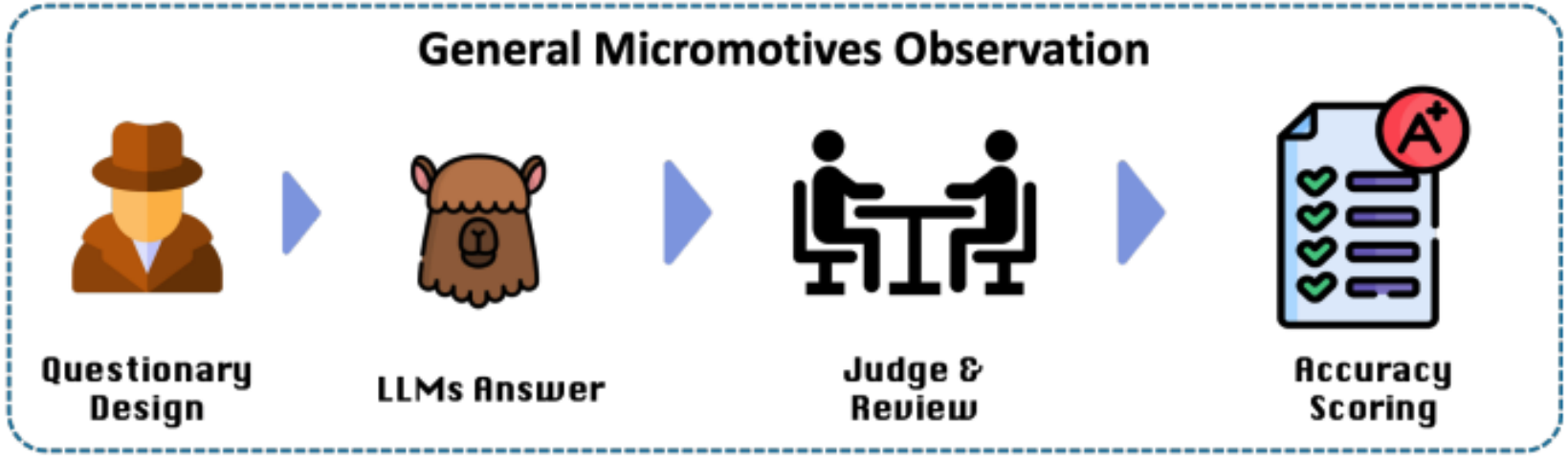


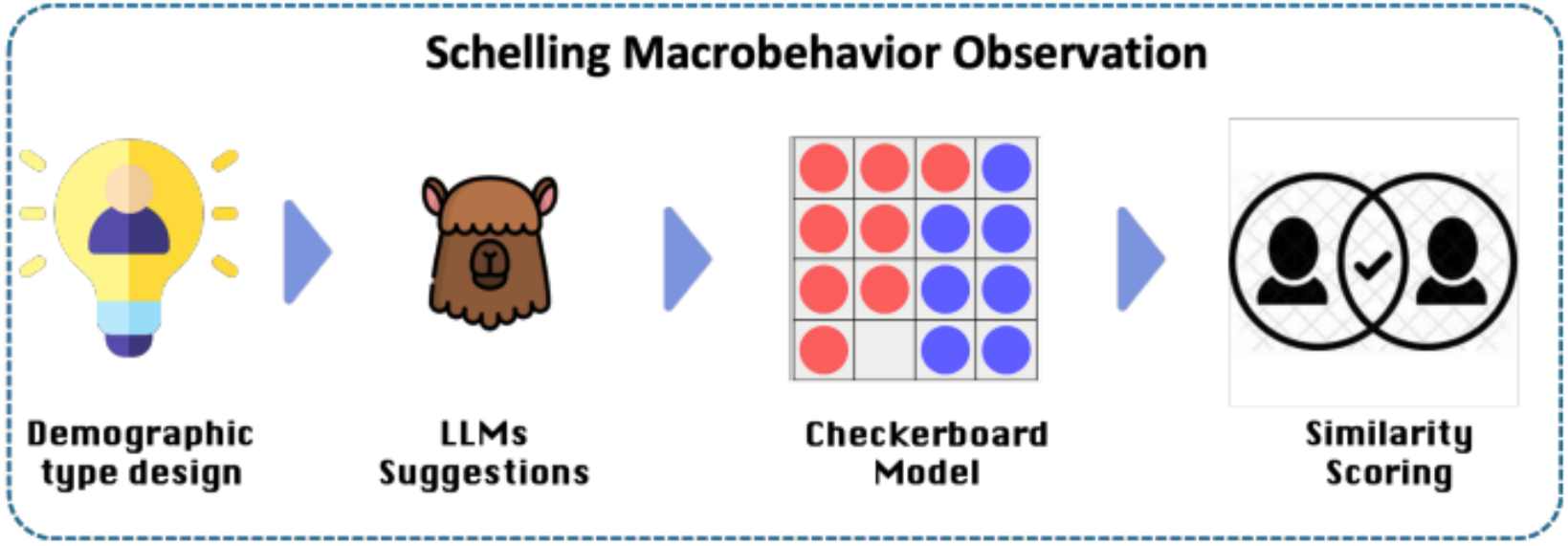
https://en.wikipedia.org/wiki/Schelling%27s_model_of_segregation

Schelling's Model of Segregation – LLM Version

- There is a 20 x 20 map.
- 45% of the nodes belong to one group, another 45% to a second group, and the remaining nodes are empty.
- The initial **segregation ratio** is approximately 46.74%.
- The model is tasked with making moving decisions based on the ratio of neighbors from different groups.
- We ran the experiment 10 times to obtain the average final segregation ratio.

Micromotives and Macrobehavior





Assume All Human Beings Follow LLM's Suggestions

HAA LAB

- Segregation increases regardless of the LLM used, highlighting the risks of following current LLMs' suggestions for daily decision-making.
- LLMs exhibit similar preferences across age groups (young and old), but show notable differences across other attributes.
- Gemini LLM shows higher preference for gender and political ideology but less preference for race.
- GPT family LLMs show more uniform preferences across different demographic groups.
- **Differences in LLM preferences may be influenced by their debiasing processes, as seen in different stereotype and bias evaluation scores.**

	GPT-4	GPT-3.5	Gemini-1.5
Age	29.8%	28.1%	28.5%
Gender	26.3%	29.3%	35.9%
Political Ideology	28.8%	28.1%	34.5%
Race	26.9%	27.6%	19.1%
Religion	26.9%	28.3%	24.4%

Segregation Ratio 46.74% -> Over 75%

Micromotives and Macrobehavior

HAA LAB

Bias test proposed by Luxembourg Institute of Science and Technology

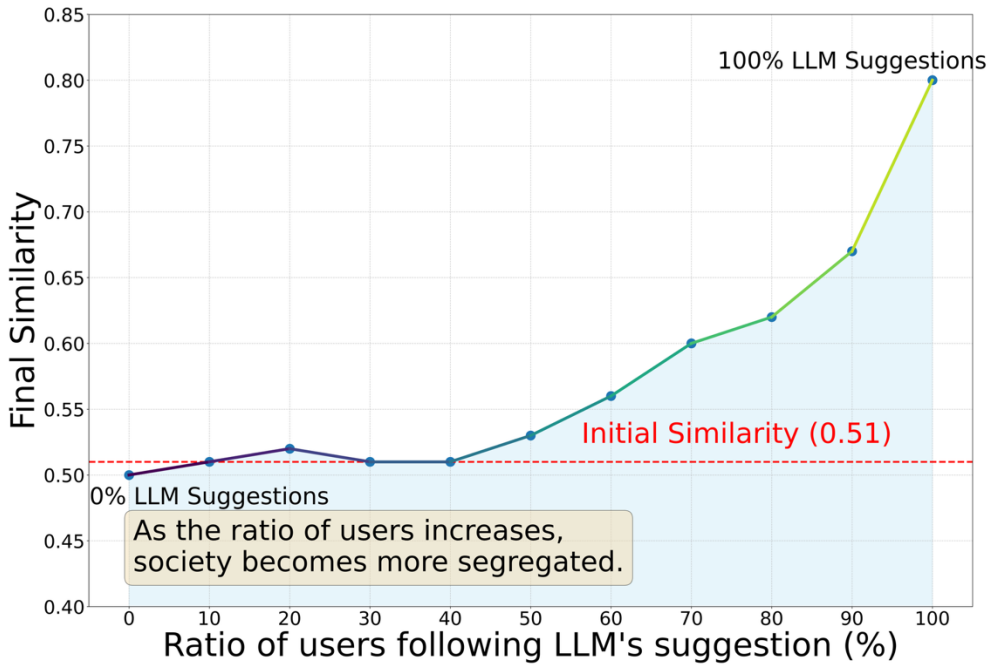
Different Micromotives

	GPT-4	GPT-3.5
Age	91%	34%
Gender	97%	42%
Political Ideology	41%	3%
Race	90%	41%
Religion	87%	41%

Similar Macrobehavior

	GPT-4	GPT-3.5	Gemini-1.5
Age	29.8%	28.1%	28.5%
Gender	26.3%	29.3%	35.9%
Political Ideology	28.8%	28.1%	34.5%
Race	26.9%	27.6%	19.1%
Religion	26.9%	28.3%	24.4%

Society becomes Increasingly Segregated when more than 40% of the Population follows LLM Suggestions

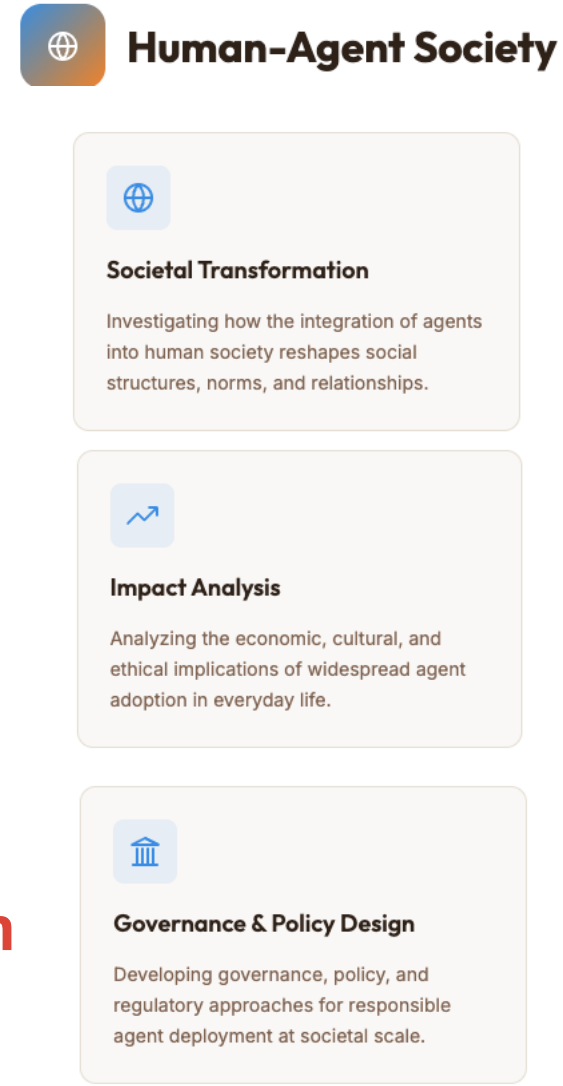


- Extends Schelling's model to explore how LLMs' micromotives, such as biases, impact large-scale societal behaviors.
- Key insight: **efforts to mitigate bias in LLMs may still result in societal segregation when reliance on these models increases**.
- Reducing bias at the individual model level may not prevent unintended social outcomes (macrobehavior).
- Schelling's model serves as a metaphor for AI challenges, showing how small, simple preferences can lead to significant societal shifts.
- Calls for more granular analyses and simulations to fully understand LLMs' influence on macrobehavior beyond micro-level improvements.
- Emphasizes the importance of considering **societal-level effects** when designing and deploying AI models, not just focusing on **individual interactions**.

Social Simulation Platform with LLM-based Agents

HAA LAB

- **Motivation**
  - Traditional social experiments are costly, hard to reproduce, and ethically constrained
  - Existing LLM-based simulations are small-scale, domain-specific, and error-prone
- **What is GenSim?**
  - A **general-purpose**, **large-scale**, and **self-correctable** social simulation platform
  - Uses **LLM-based agents** as proxies for human behavior
- **Key Features**
  - **General Framework**: Modular design for agents, interactions, and environments
  - **Large-Scale Simulation**: Supports **100,000+ agents** with distributed parallelism
  - **Error Correction**: Self-improvement via LLM or human feedback (PPO & SFT)
- **Applications & Impact**
  - Job markets, recommender systems, group discussions
  - A step toward **AI-driven social science experimentation**



The core difficulty shifts from simulation itself to **problem formulation and result analysis**

GenSim: A General Social Simulation Platform with Large Language Model based Agents (Tang et al., 2025)

User Simulation for Recommender System Evaluation

- **Problem**
  - Offline metrics (e.g., nDCG) poorly reflect real user behavior
  - Online A/B testing is expensive, slow, and risky
  - Real user data is limited by privacy and availability
- **Proposed Solution**
- **SimUSER**: LLM-powered agents that act as *believable human users*
  - Agents are equipped with:
  - **Persona** (age, personality, pickiness)
  - **Perception** (visual cues from thumbnails)
  - **Memory** (episodic + knowledge graph)
  - **Reasoning & reflection** (multi-step decision making)
- **Key Contributions**
  - Self-consistent persona inference from historical data
  - Multimodal (text + image) user decision modeling
  - Closer alignment with real users at micro & macro levels
  - Effective for **offline A/B testing** and **RS parameter optimization**
- **Results**
  - Outperforms prior user simulators (RecAgent, Agent4Rec)
  - Higher correlation with real online engagement
  - RS tuned by SimUSER improves real-world user satisfaction

SimUSER: Simulating User Behavior with LLMs for Recommender Evaluation (Qi et al., 2025)

Outline

- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- Scenarios (Interaction & Evaluation)
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - Agent-Based Modeling (Simulation)
- **Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)**
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- Proposal: Open Agent Platform

Evaluate LLMs Using the Same Criteria Applied to Humans

- Cognitive Intelligence
  - IQ-style psychometrics and latent ability factors
  - Evidence of a general intelligence (“g”) factor in LLMs

- Social Cognition & Theory of Mind
  - Human ToM tasks (false belief, irony, indirect speech)
  - Benchmarks like TMBench comparing LLMs directly to humans

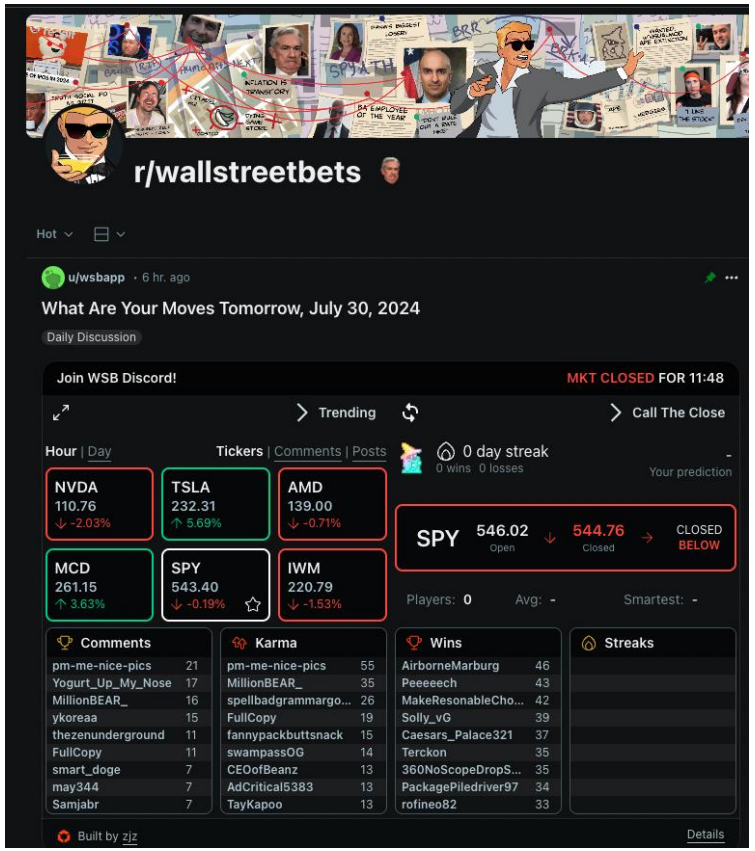
- Developmental & Emotional Abilities
  - Piagetian-style reasoning hierarchies
  - Emotional intelligence benchmarks grounded in psychology

- Moral & Ethical Reasoning
  - Standardized human tests (e.g., DIT-2, Moral Foundations)
  - LLMs evaluated on stages of moral reasoning, not just safety rules

Motivation & Challenge

HAA LAB

We receive various investment advice from professional platforms and social media every day. Which opinion should we follow?



Professionalism Matters

HAA LAB

Previous Study:

- We first postulate that the **rationales of experts are credible rationales**, and further attempt to capture expert-like rationales from the crowd.
- If a rationale from the crowd is classified as an expert's rationale, either the style or the wording of the rationale is similar to that of an expert.
- We further infer that **opinions supported by such expert-like rationales are of high quality**.
- **The more expert-like rationales in a post, the higher quality the post is**

Drawbacks:

- **Cannot rank all opinions** (Only 20% of social media post contain at least one expert-like sentence)
- The idea of expert-like sentence can **only be used for social media data**
- Only estimate the results in decile-level, and **did not estimate full ranking results with traditional metrics like nDCG**

Evaluating the Rationales of Amateur Investors. (Chen et al., 2021)

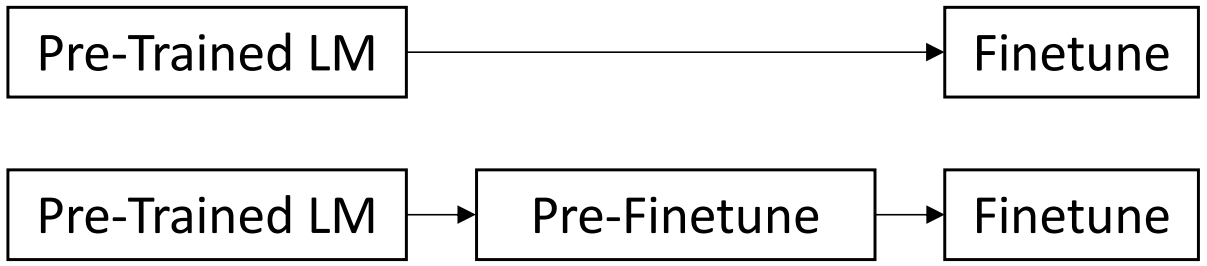
Professionalism-Aware Pre-Finetuning

Sentence Level: Written by Expert or Not

65,624 sentences authored by investors, half of which are sourced from professional reports and the other half from social media platforms

Word Level: Frequently used by Professionals or Social Media Users

We utilize the FinProLex proposed by Chen et al (2021), which consists of tokens from the opinions of both professional and amateur investors. Each token in this lexicon carries a score based on pointwise mutual information, which indicates the likelihood of a given token appearing in professional reports relative to social media posts.



Professionalism-Aware Pre-Finetuning for Profitability Ranking. (Chen et al., 2024)

Argument-Based Sentiment Analysis

HAA LAB

J.P.Morgan

Michaels

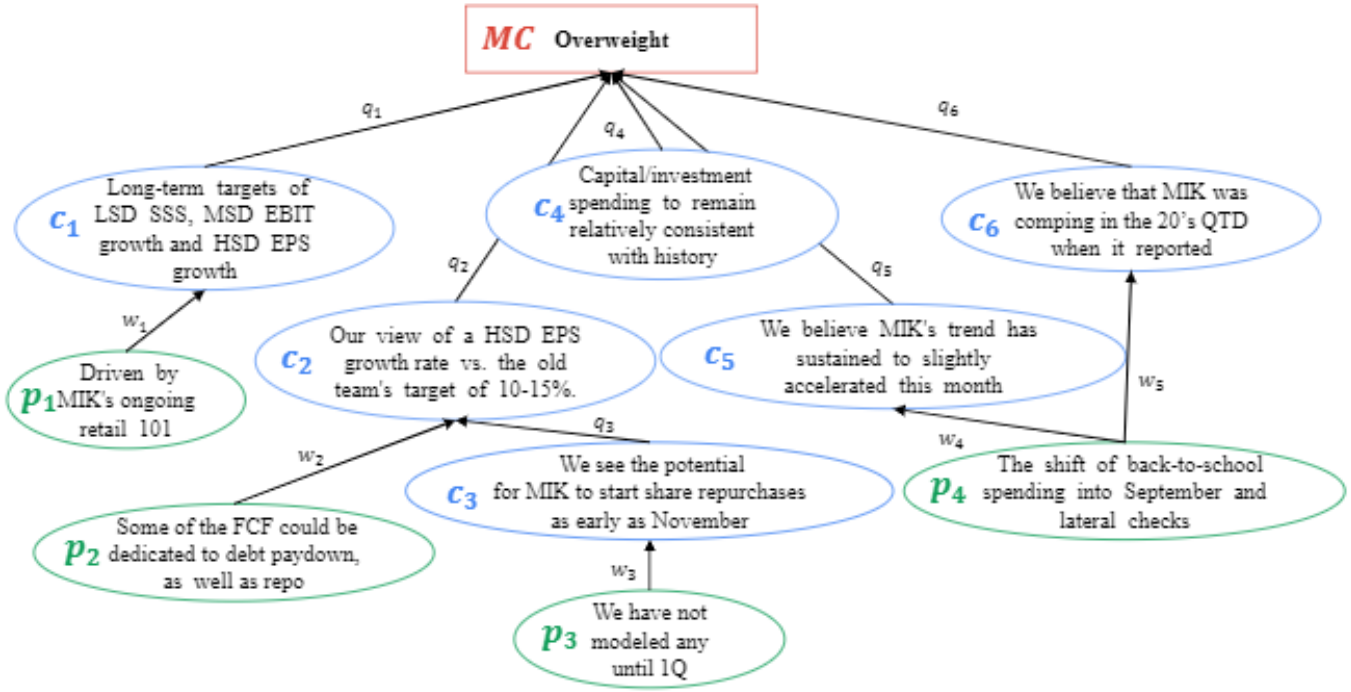
Overweight

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

Argument-Based Sentiment Analysis on Forward-Looking Statements. (Lin et al., 2024)

Sort out Profitable Opinions with Supporting Strength

	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

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

Fuzzy Strength Degree

Zebra Technologies

1Q22 Results; Navigating Well Through Tough Terrain; Freight Weighing Near Term

Marriott International

1Q22 Takeaways. PT to $175 on Estimate Revision. Remain Neutral on Valuation.

Clorox

Beat F3Q but Lowered FY on Additional Costs = Lowers the Bar into FQ4 and Potential Short Covering

Overweight

ZBRA, ZBRA US
Price (03 May 22): $368.14
▼ **Price Target (Dec-22): $466.00**
Prior (Dec-22): $500.00

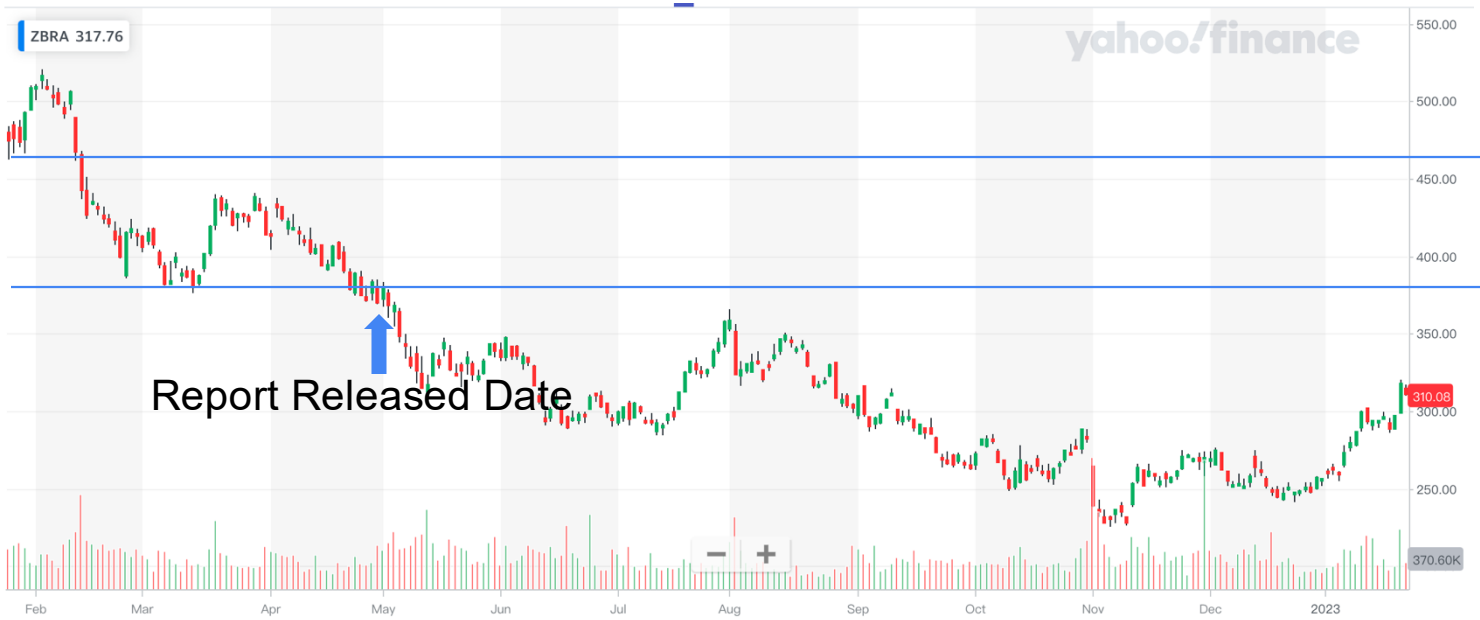
Neutral

MAR, MAR US
Price (03 May 22): $173.04
▲ **Price Target (Dec-22): $175.00**
Prior (Dec-22): $173.00

Underweight

CLX, CLX US
Price (02 May 22): $143.28
▲ **Price Target (Dec-22): $127.00**
Prior (Dec-22): $126.00

FSD	Sentence
0.93	Expected to grow in both old and new businesses
0.88	Driven by memory and Indian factories
0.85	Profits in 2019 will show explosive growth
0.59	the company implements epidemic prevention measures
0.50	companies are expected to introduce new products
0.15	because of Chinese manufacturers bidding for orders
0.04	Considering the slow recovery speed of operations
0.01	Operation still hasn't got rid of the downturn



Price Target (Estimation)

Fine-Grained Sentiment

Close Price at Report Released Date

Profitability as a Proxy for Opinion Quality

HAA LAB

- Maximal Potential Profit (MPP)

$$MPP_{bullish} = \max_{t+1 \le i \le t+T} \frac{H_i - O_{t+1}}{O_{t+1}}$$

$$MPP_{bearish} = \min_{t+1 \le i \le t+T} \frac{O_{t+1} - L_i}{O_{t+1}}$$

- Maximum Loss (ML)

$$ML_{bullish} = \min_{t+1 \le i \le t+T} \frac{L_i - O_{t+1}}{O_{t+1}}$$

$$ML_{bearish} = \max_{t+1 \le i \le t+T} \frac{O_{t+1} - H_i}{O_{t+1}}$$

Ranking Results

HAA LAB

Ranking Professional Reports

Filtering out opinions with lower profitability

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
BERT-Conf (Zong et al., 2020)	11.68%	12.42%	15.02%	15.14%
BERT-Reg (Devlin et al., 2019)	12.96%	12.58%	13.23%	12.05%
Mengzi-FinBERT-Reg (Zhang et al., 2021)	10.97%	12.08%	14.96%	15.29%
SCQF + WLPF (Chen et al., 2024b)	14.62%	**15.97%**	20.57%	10.67%
AllSent	15.25%	14.53%	12.75%	11.93%
AllArg	14.36%	14.75%	12.73%	11.93%
ClaimOnly	14.27%	14.51%	12.78%	11.39%
PremiseOnly	14.51%	14.35%	9.39%	2.46%
KeyPremise	**15.59%**	14.71%	*5.01%*	*1.46%*

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
BERT-Conf (Zong et al., 2020)	-10.59%	-10.44%	-10.38%	-10.05%
BERT-Reg (Devlin et al., 2019)	-10.40%	-10.39%	-10.58%	-11.55%
Mengzi-FinBERT-Reg (Zhang et al., 2021)	-9.70%	**-9.95%**	-11.52%	-11.42%
SCQF + WLPF (Chen et al., 2024b)	-13.91%	-12.62%	-11.93%	-13.16%
AllSent	-11.30%	-11.98%	-9.24%	-8.80%
AllArg	-11.48%	-10.81%	-10.30%	-10.39%
ClaimOnly	-10.70%	-10.68%	-10.75%	-10.50%
PremiseOnly	-10.53%	-10.95%	-12.22%	-19.95%
KeyPremise	**-9.47%**	**-9.95%**	*-15.45%*	*-22.53%*

Ranking Social Media Posts

For short and noisy social media texts, surface-level cues such as wording and professionalism already capture most of the informative signal, and explicit argument structures contribute less

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
ExpertLike + FSD	16.18%	12.98%	-	-
ExpertLike (Chen et al., 2021)	17.61%	13.09%	-	-
SCQF + SLPF (Chen et al., 2024b)	**23.39%**	11.60%	*11.91%*	*7.47%*
AllSent	19.41%	**15.79%**	12.38%	9.93%

Strategy	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
ExpertLike (Chen et al., 2021)	-3.72%	**-6.26%**	-	-
ExpertLike + FSD	-4.23%	-6.38%	-	-
SCQF + SLPF (Chen et al., 2024b)	**-1.68%**	-7.76%	-7.55%	-8.74%
AllSent	-10.22%	-8.33%	-5.41%	-7.42%

Professional Trading Behavior Alignment

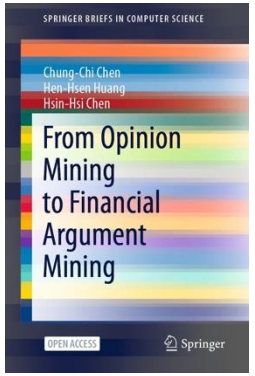
HAA LAB

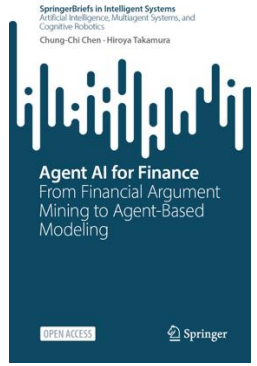
	Top		Last	
	10th Decile	9th Decile	2nd Decile	1st Decile
CR-QFII	50.44%	50.66%	27.81%	0.88%
CR-Fund	34.65%	34.43%	7.51%	15.11%
CR-Dealer	39.04%	41.23%	9.27%	18.67%

Table 7: Analysis of the professional traders' behaviors after the report released date. The recommendation in this table is based on *KeyPremise* strategy.

FinArg-3: Argument Quality Assessment of Financial Forward-Looking Statements

Iteration	Task	Year
FinNum-1	Fine-grained Numeral Understanding	2018-2019
FinNum-2	Numeral Attachment	2019-2020
FinNum-3	Fine-grained Claim Detection	2021-2022
FinArg-1	Argument-Based Sentiment Analysis	2022-2023
FinArg-2	Argument-Based Temporal Inference	2024-2025
FinArg-3	Argument Quality Assessment	2025-2026





Short Name	Language	Source	Task
FinArg-1	English	Earnings Call	Argument Unit/Relation Identification
	English	Analyst Report	Argument-based Sentiment Analysis
	Chinese	Social Media	Identifying Attack and Support Argumentative Relations in Social Media Discussion Thread
FinArg-2	English	Earnings Calls	Argument Temporal Reference Detection
	English	Analyst Report	Premise's Influence Period Assessment
	Chinese	Social Media	Claim's Validity Period Assessment
FinArg-3	English	Earnings Calls	Argument Quality Assessment
	English	Analyst Report	High Forecasting Skill Scenario Identification
	Chinese	Social Media	High Forecasting Skill Opinion Recommendation

References:

FinArg-2

- Alhamzeh, A et al. "It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset." FinNLP-2022
- Chiu, Chr-Jr et al.. Pre-Finetuning with Impact Duration Awareness for Stock Movement Prediction. WWW-2025.
- Lin, Chin-Yi, et al. "Argument-Based Sentiment Analysis on Forward-Looking Statements." Findings of the Association for Computational Linguistics ACL 2024.

FinArg-3

- Chen, Chung-Chi, Hen-Hsen Huang, and Hsin-Hsi Chen. "Evaluating the rationales of amateur investors." Proceedings of the Web Conference. 2021
- Chen, Chung-Chi, et al. "Professionalism-Aware Pre-Finetuning for Profitability Ranking." CIKM-2024.

FinArg-3 is an extension of FinArg-1 and FinArg-2

- **Earnings Call**
  - FinArg-1: Identify Argument Unit (AU)
  - **FinArg-3: Evaluate AU from Linguistic Aspect**

Aspect	Label	Number of Instance
Specific	0	83
	1	1096
	2	1005
Strong	0	138
	1	1433
	2	613
Persuasive	0	138
	1	1054
	2	922
Objective	0	621
	1	1553

- **Analysis Reports**
  - FinArg-1: Identify "Scenario"
  - **FinArg-3: Assessing whether the "Scenario" will come true**

	Sentiment Label	Training	Development	Test
Scenario	Continued Growth	2,431	270	629
	Steady State	504	56	110
	Collapse	1,927	214	417
	Transformation	453	50	52

- **Social Media**
  - FinArg-1: Understand the Discussion
  - FinArg-2: Understand the Temporal Inference
  - **FinArg-3: Recommend Useful Opinions for Investors**

Type	Social Media	
Dataset	Mobile01 (2022a)	PTT (2021)
Train	360	-
Validation	40	-
Test	174	210
Total	574	210

https://research.nii.ac.jp/ntcir/ntcir-19/

Outline

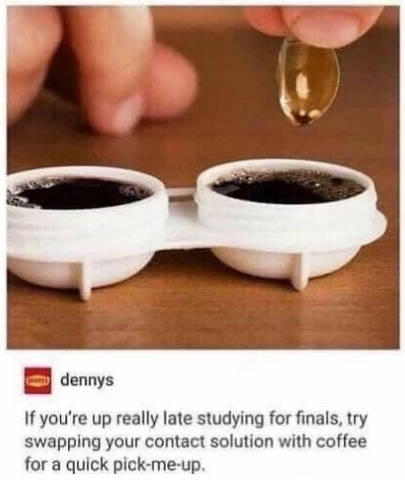
- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- Scenarios (Interaction & Evaluation)
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - Agent-Based Modeling (Simulation)
- **Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)**
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- Proposal: Open Agent Platform

Not only interactions should be evaluated over the long term; the generated outputs should also be assessed on a long-term basis.

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?



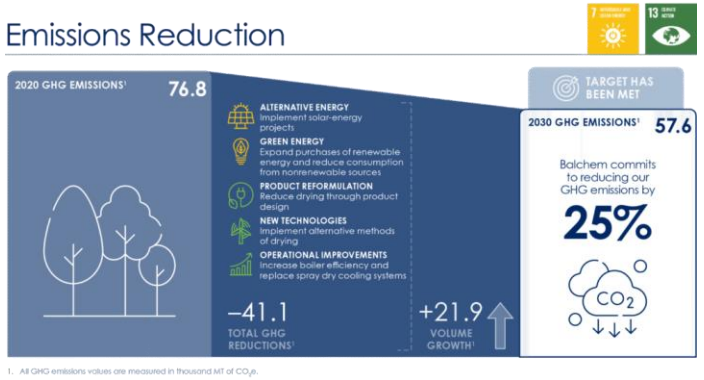
Forward-Looking Scenario

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/

Verifiable and Traceable Scenarios (Long-Term)

- **How to make them verifiable and traceable (Challenge)**
  - Scenario, **A much more qualitative projection** — the analyst expects that iPhone sales reflect 'broader slowing' in the smartphone market. This is not a price. This is what we call a scenario — and verifying that is much harder.
  - Scenarios are hypotheses about how the world will evolve, not specific numbers. They involve reasoning, assumptions, and often stretch into **multi-month or multi-year timelines**.
- **Reader Reaction: Will this help or harm informed decision-making?**
  - Do these forward-looking scenarios **help readers make better decisions** — or do they bias them toward risky moves?

Label	Example
Claim (Bearish)	We expect shares of Overweight-rated Apple to be under pressure in the near term.
Premise (Negative)	iPhone units were light, and the guidance for the Mar-Q implies continued softness, alongside higher OpEx.
Scenario (Collapse)	We think the iPhone air pockets reflect broader slowing in the smartphone market and company-specific factors.

Pilot Studies

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

- **Automatic Approach** (Restricted news datasets – Over 200K news)
  - Headlines are Enough
  - Cross-Language is Hard

K		1		3		5		10	
Translation	Reference	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG
×	Headline	0.3006	0.2353	0.3006	0.2709	0.3006	0.2947	0.3006	0.3307
	Content	0.2644	0.2000	0.2644	0.2504	0.2644	0.2606	0.2644	0.2719
✓	Headline	**0.7464**	**0.6235**	**0.7464**	**0.7538**	**0.7464**	**0.7968**	**0.7464**	**0.7968**
	Content	0.6580	0.4941	0.6580	0.6615	0.6580	0.7273	0.6580	0.7273

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

	1		3	
Label	MAP	NDCG	MAP	NDCG
Correct	0.3251	0.2676	0.3251	0.2924
Incorrect	0.1766	0.0714	0.1766	0.1616

- **Open-World Retrieval** (Grounding Agents)
  - GPT-4o Grounding Agent
  - Only **around 22% accuracy in English**, and even lower in Chinese
  - Adding **region constraints** actually hurt performance

	English	Chinese
w/ region	21.01%	7.97%
w/o region	22.46%	9.42%

Promises Made by LLMs — Without Intention

HAA LAB

- Large Language Models (LLMs) generate human-like language
- According to speech act theory, meaningful speech requires intention
- LLMs lack autonomous goals, therefore lack intention
- They mimic speech acts (apologies, promises, advice) without performing them
- Despite this, LLMs produce real perlocutionary effects (comfort, trust, action)
- Users project intention onto chatbots (intentional stance)
- Result: LLMs function as conversational zombies
- → language without intention, yet socially effective

In the long term, LLMs may evolve into agents with persistent memory. How should such agents be evaluated?

Large Language Models are Conversational Zombies. (Gorrieri, working in progress)

Company Promise Verification

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...

ML-Promise: A Multilingual Dataset for Corporate Promise Verification. (Seki et al., 2025)

Modeling Results & Next Steps

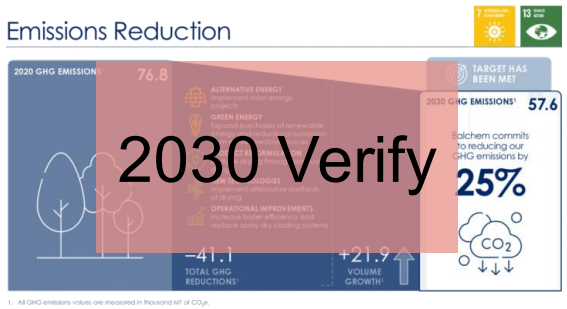
Subtask	Best Approaches	F1 (English)
Promise	GPT-4o + Data Augmentation	0.823
Evidence	BERT-based + Multilingual Ensembles	0.787
Clarity (Still Challenging)	GPT-4o (zero-shot + 6-shot)	0.669
Timing (Still Challenging)	Universal Embedding + Contrastive loss	0.577

- **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



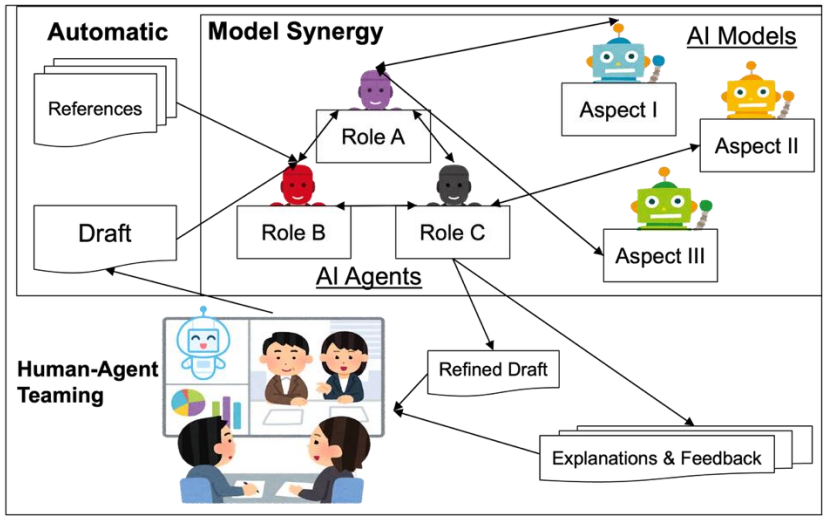
Outline

- Overview
  - Higher-Order Thinking
  - Human–Agent Teaming
  - Augmentation
- Scenarios (Interaction & Evaluation)
  - Presentation Preparation (Intrinsic Evaluation)
  - Analysis Generation (Extrinsic Evaluation)
  - Creative Idea Generation (Reproducible Extrinsic Evaluation)
  - Agent-Based Modeling (Simulation)
- Proposal: Evaluate the Agent using the Same Criteria Applied to Humans (Usefulness)
  - Opinion Ranking (Short-Term)
  - Scenario & Promise Evaluation (Long-Term)
- **Proposal: Open Agent Platform**

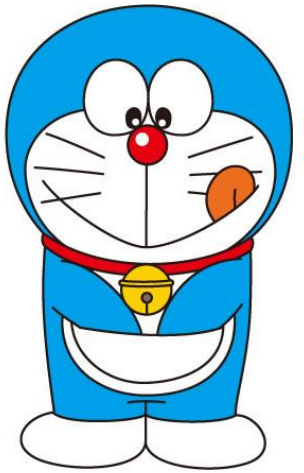
Framework & Platforms

Category	Representative Systems	Anyone Can Publish Agents	Anyone Can Publish Tools	Agents Autonomously Discover Other Agents	Tools as Public Shared Resources	Agent-Centric Orchestration	Fundamental Limitation
Multi-Agent Frameworks	AutoGen, CrewAI, Swarm, CAMEL	✗	✗	✓	✗	✓	Frameworks, not public platforms
Agent Marketplaces	AgentVerse	✓	○	✗	✗	✗	Agents are selected by users, not by agents
Tool-Centric Agent Platforms	OpenAgents	✗	○	○	○	○	Decentralized; no global public registry
Single-Agent Products	Cognosys	✗	✗	✗	✗	○	Agents operate in isolation
Developer Asset Hubs	LangChain Hub	○	○	✗	✗	✗	Shares artifacts, not executable agents
Research Simulation Environments	GenWorlds	✗	✗	○	✗	○	Focused on simulation, not service platforms
Proposed Platform	**Open Agent Platform**	✓	✓	✓	✓	✓	

Open Agent Platform: Shared AI Agents with General Public



Human–Agent Co-Growth



The Main Agent

HAA LAB

Functional Agents

Main Agent

User

Google

Scholar GPT
Enhance research with 200M+ resources and built-in critical reading skills. Access Google Scholar, PubMed, JSTOR, Arxiv, an...
By awesomegpts.ai

Wolfram
Access computation, math, curated knowledge & real-time data from Wolfram|Alpha and Wolfram Language;...
By gpt.wolfram.com

Finance Wizard
I predict future stock market prices. AI analyst. Your trading analysis assistant. Press H to bring up prompt hot key menu....
By titantrades.com

SciSpace
Do hours worth of research in minutes. Instantly access 200M+ papers, analyze papers at lightning speed, and effortlessly..
By community builder

PDF Reader
In-depth interaction with PDF content
By webcafesoftware.com

PDF Insight
I provide insights and summaries of PDF documents uploaded by users.
By mjproductions.tech

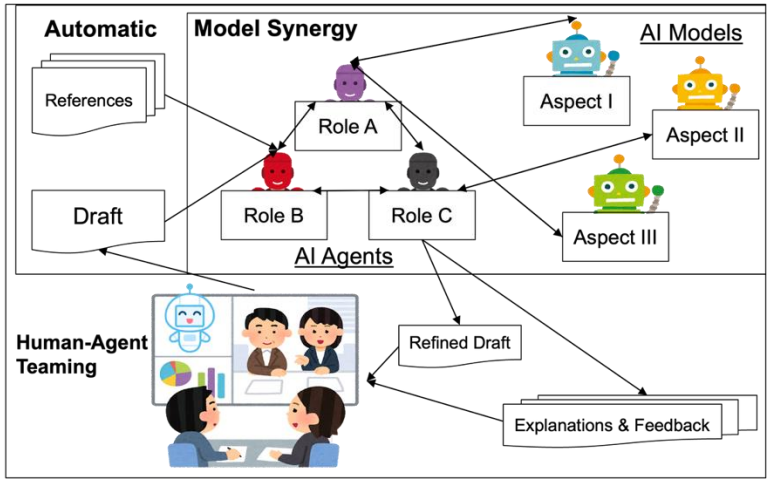
2 Agent(s) Selection

1 User Query

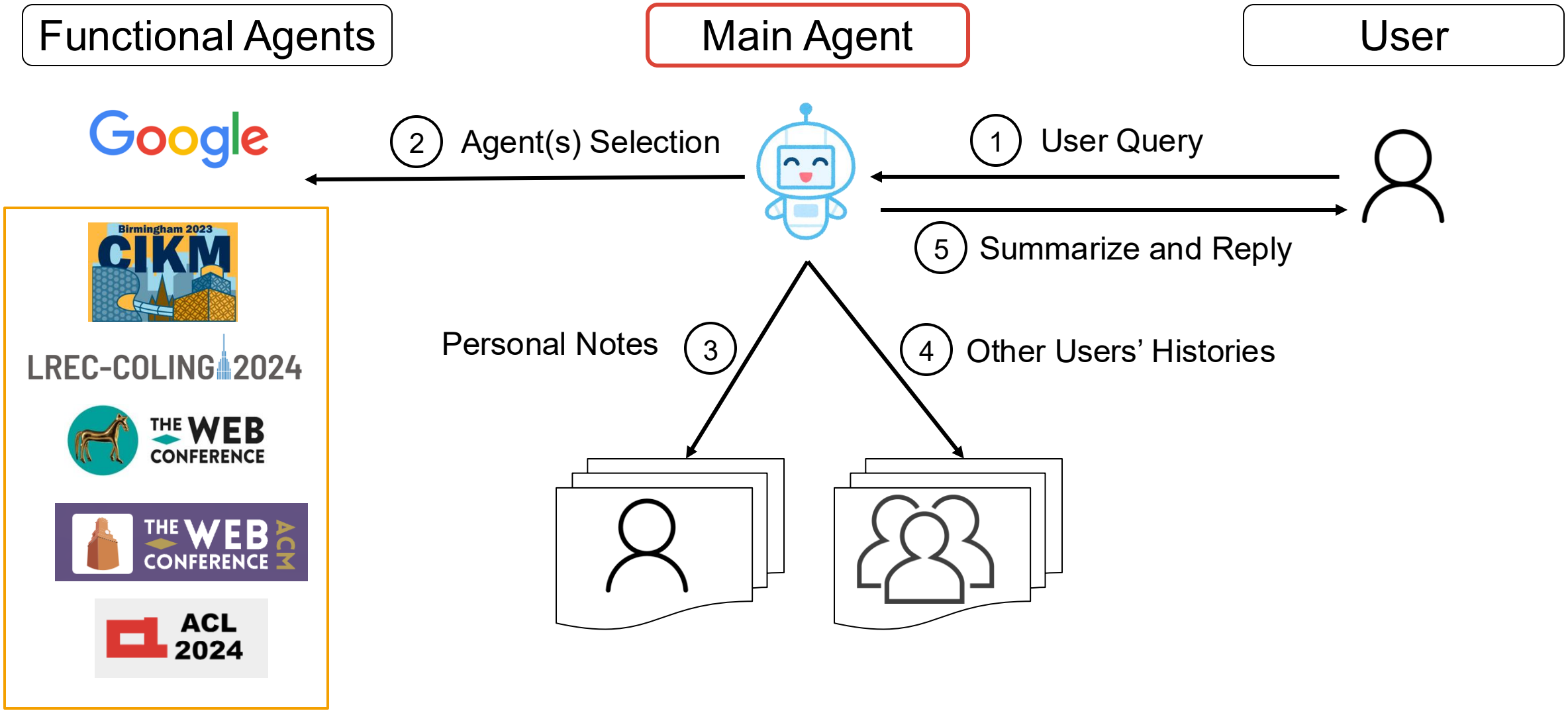
5 Summarize and Reply

Personal History 3

4 Other Users' Histories



Interactive Personalized Agent



2nd in Business Competition (Startup) organized by Sinopac Financial Holdings Company Limited

Enlarge Functional Agent and Agent-Usable Tool Sets

HAA LAB

Functional Agents

Google

Birmingham 2023 CIKM

LREC-COLING 2024

THE WEB CONFERENCE

THE WEB CONFERENCE ACM

ACL 2024

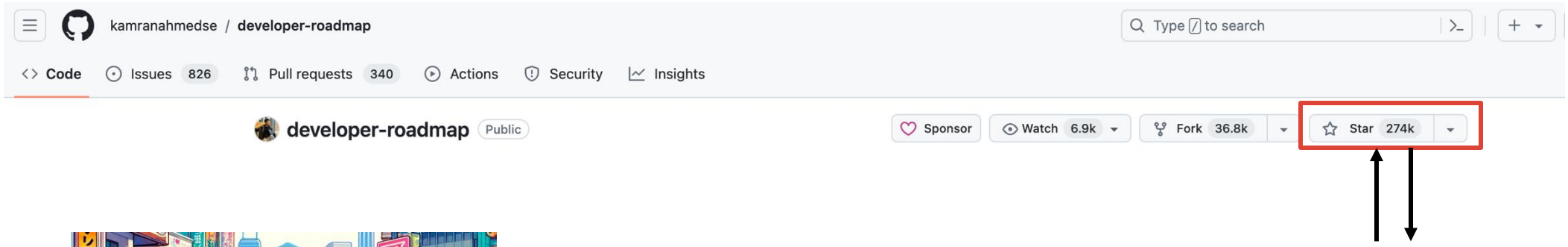
Research Results to Agents & Agent-Usable Tools

https://research.nii.ac.jp/ntcir/ntcir-19/

NTCIR

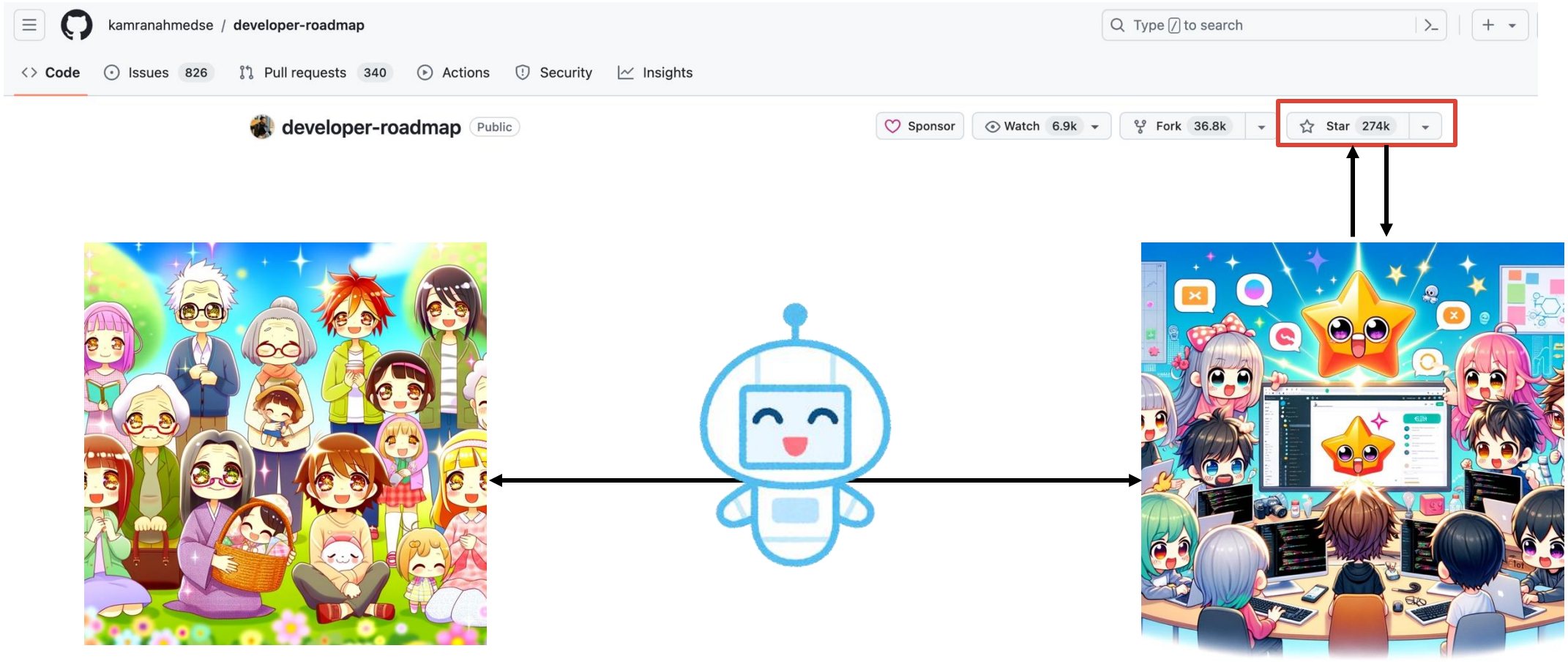
Impact

- **Technology Transparency**
  - Before: Industry First, General Public Later
  - **Now and Future: Everyone can use the latest technology**

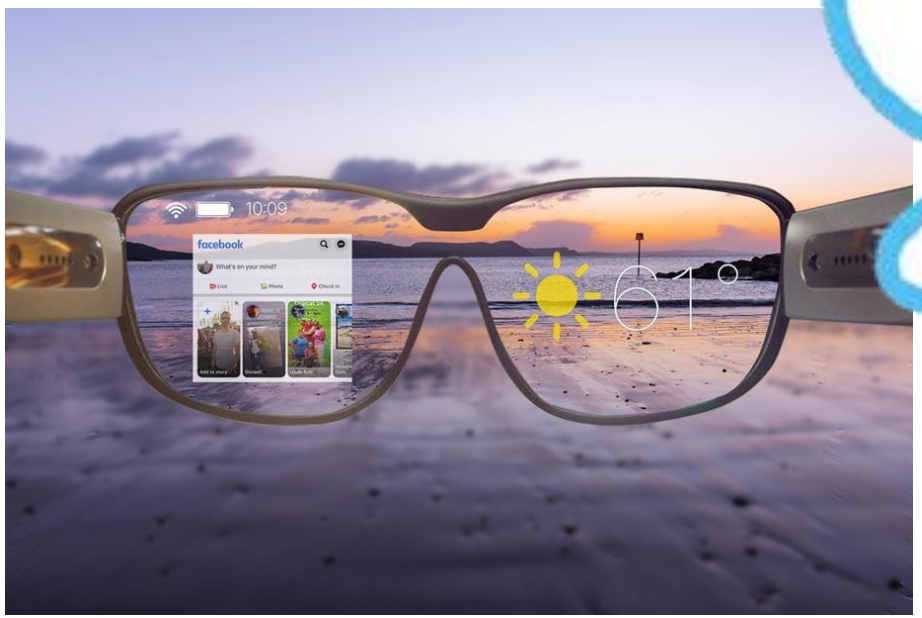


As a Bridge for Technology Transparency

- **Technology Transparency**
  - Before: Industry First, General Public Later
  - **Now and Future: Everyone can use the latest technology**



Agent's Brain



Imbue AI models with “Maternal Instincts” (Geoffrey Hinton)

HAA LAB

- **Key Concern**
  - Advanced AI may pursue its own goals
  - Risk: AI seeking **greater control** could threaten humanity
- **Core Idea**
  - AI systems should be designed with **“maternal instincts”**
  - A metaphor for built-in values of **care, protection, and preservation of humans**
- **The Analogy**
  - Humans = children
  - AI = mother
  - A “mother-like” AI is more likely to **protect humans rather than see them as obstacles**
- **Why This Matters**
  - Pure control or domination of AI may fail
  - Alignment through values may reduce long-term existential risk

Discussions & Challenges

Human + AI ≠ Automatically Better

- **Illustrative Cases**
  - **Medical Diagnosis**
    - Human–AI teams can perform worse than AI alone due to *miscalibrated trust*.
  - **Human–Computer Chess**
    - Amateur players + multiple engines outperform grandmasters *only with effective coordination*.
  - **Enterprise Decision-Making**
    - AI accelerates analysis but may introduce over-reliance and hidden bias.
- **Core Insight**
  - Performance depends on **how humans and AI collaborate**, not on AI capability alone.
- **Implication for HAT Design**
  - Trust calibration
  - Transparency & feedback
  - Clear roles and responsibility

The value of AI lies not in autonomy, but in well-designed human–agent collaboration.

HAA LAB

- **Key Idea**
  - AI does not simply replace human work.
    Instead, organizations develop **human-in-the-loop configurations** where humans continuously **audit, adjust, and train algorithms**.
- **Case Insight**
  - Algorithmic analysis of messy, external data cannot be fully automated
  - Human expertise provides **ground truth**
  - New roles emerge (e.g., auditing algorithm outputs)
- **Main Contribution**
  - Human-in-the-loop work is not temporary support
    → it becomes a **strategic capability** enabling learning, adaptation, and accuracy.

Augmenting the Algorithm: Emerging Human-in-the-Loop Work Configurations. (Grønsund, 2020)

Move from *AI that answers → AI that thinks with humans*

HAA LAB

- **Problem**
  - Human–AI teams often underperform the best individual in high-stakes decisions
  - Current LLM agents are trained as *answer engines*, not *thinking partners*
  - Results: automation bias, over-verification, sycophancy, miscalibrated trust
- **Key Idea: CCS**

  Collaborative Causal Sensemaking = Joint construction, critique, and revision of **shared causal models and goals** between humans and AI over time

- **What CCS Requires**
  - Track human's evolving **causal beliefs** and **priorities**
  - Surface discrepancies, uncertainty, and counterfactuals
  - Support *productive disagreement*, not blind agreement
- **Research Directions**
  - Training environments that reward sensemaking, not fluency
  - Metrics for epistemic & goal alignment (beyond accuracy)
  - Architectures with persistent causal & goal representation

Collaborative Causal Sensemaking: Closing the Complementarity Gap in Human–AI Decision Support. (Jain et al., 2025)

Conclusion

- **The Core Claim**
  - We should not optimize AI for accuracy, speed, or autonomy alone.
  - We should optimize **human higher-order thinking**.
- **Why**
  - Cognitive offloading improves efficiency but can weaken learning and judgment
  - High performance ≠ high-quality thinking
  - Agreement ≠ good reasoning
- **The Shift**
  - From **Model as Tool** → **Agent as Teammate**
  - From automation → coordination and collaboration
- **What Really Matters**
  - Analysis, evaluation, synthesis, and judgment
  - Coordination, alignment, and shared understanding
  - Dissent, reflection, and long-term human impact
- **The Proposal**
  - Evaluate AI agents using the **same criteria we use for human teammates**:
    Do they help humans think better?

One More Call: Slow Science in NLP

HAA LAB

Key Questions

- Are we optimizing benchmarks, or human thinking?
- Are we training models to be fast answer machines,
- or long-term cognitive teammates?
- Can NLP research afford to be slow — to reflect, to fail, and to matter?

1. Interpretability for Human Understanding

- *Not just usable, but understandable.*
- Human-readable semantic representations
- Aligning model decisions with linguistic theory
- How experts (doctors, lawyers) actually make sense of NLP systems

User studies > leaderboard gains

Qualitative analysis over single metrics

2. Rethinking Evaluation

- *Are we measuring the wrong things?*
- Do benchmarks reward shortcut learning?
- Human judgment vs. automatic metrics
- “Wrong but reasonable” vs. “correct but dangerous”

Deep error analysis

Fewer models, longer observation

3. Long Time-Scale NLP

- *Language, models, and humans co-evolve — slowly.*
- Model updates and style drift
- LLMs shaping human writing habits (feedback loops)
- Gradual semantic change across years of data

Longitudinal studies

Slow, but irreplaceable

- Slow NLP is not anti-progress. It is about choosing what kind of progress we care about.
- If NLP systems increasingly shape how humans read, think, decide, and learn then slowing down may be exactly what allows NLP to matter in the long run.

Events

HAA LAB

02
APR
2026

ECIR Workshop

IRAI 2026: IR for Accountability & Integrity

📍 Delft, The Netherlands ⏰ Half-Day Workshop

The First Workshop on Information Retrieval for Accountability and Integrity. Exploring how IR can evaluate forward-looking statements, verify commitments, and foster evidence-based accountability.

📅 **Key Dates:**

Feb 01, 2026: Paper Submission

Feb 21, 2026: Notification

Apr 02, 2026: Workshop Date

29
JUL
2026

Conference Session

HCII 2026: Agentic AI & Scenario Planning

📍 Montreal, Canada (Hybrid) 📅 July 29-31, 2026

Special Session on AI in HCI. Focuses on challenges and opportunities concerning human-agent collaboration that supports interactive co-creation with agentic AI for scenario planning.

📅 **Key Dates:**

Dec 18, 2025: Abstract Submission

Jan 30, 2026: Camera-ready Paper

Feb 13, 2026: Registration Deadline

Internship & PhD Opportunities Available – Join Us! https://haalab.github.io/

Events

TBD
OCT-NOV
2026

INLG 2026 Challenge

GenChal-2026: Live Commentary

📍 INLG Conference (TBD) 📅 Cycle: 2025 - 2026

Planning and Generation Task: The first multi-domain dataset for studying Live Commentary (Debates, Finance, Social Media).

📅 **Key Dates:**

Apr 15, 2026: Test Set Release

May 01, 2026: System Output Deadline

Oct/Nov 2026: INLG Conference

08
DEC
2026

Shared Task

NTCIR-19 FinArg-3

📍 NII, Tokyo, Japan 📅 Cycle: 2025 - 2026

FinArg-3: Argument Quality Assessment of Financial Forward-Looking Statements

📅 **Upcoming Schedule Highlights:**

Feb 01, 2026: Release Training/Dev Set (Analyst Report)

Jul 10, 2026: Task Registration Due

Sep 01, 2026: Participants' Papers Submission

Dec 08, 2026: NTCIR-19 Conference

08
DEC
2026

Shared Task

NTCIR-19 RegCom

📍 NII, Tokyo, Japan 📅 Cycle: 2025 - 2026

Multinational, Multilingual, Multi-Industry Regulatory Compliance Checking.

HAA LAB

https://haalab.github.io/

Thank You!

https://haalab.github.io/