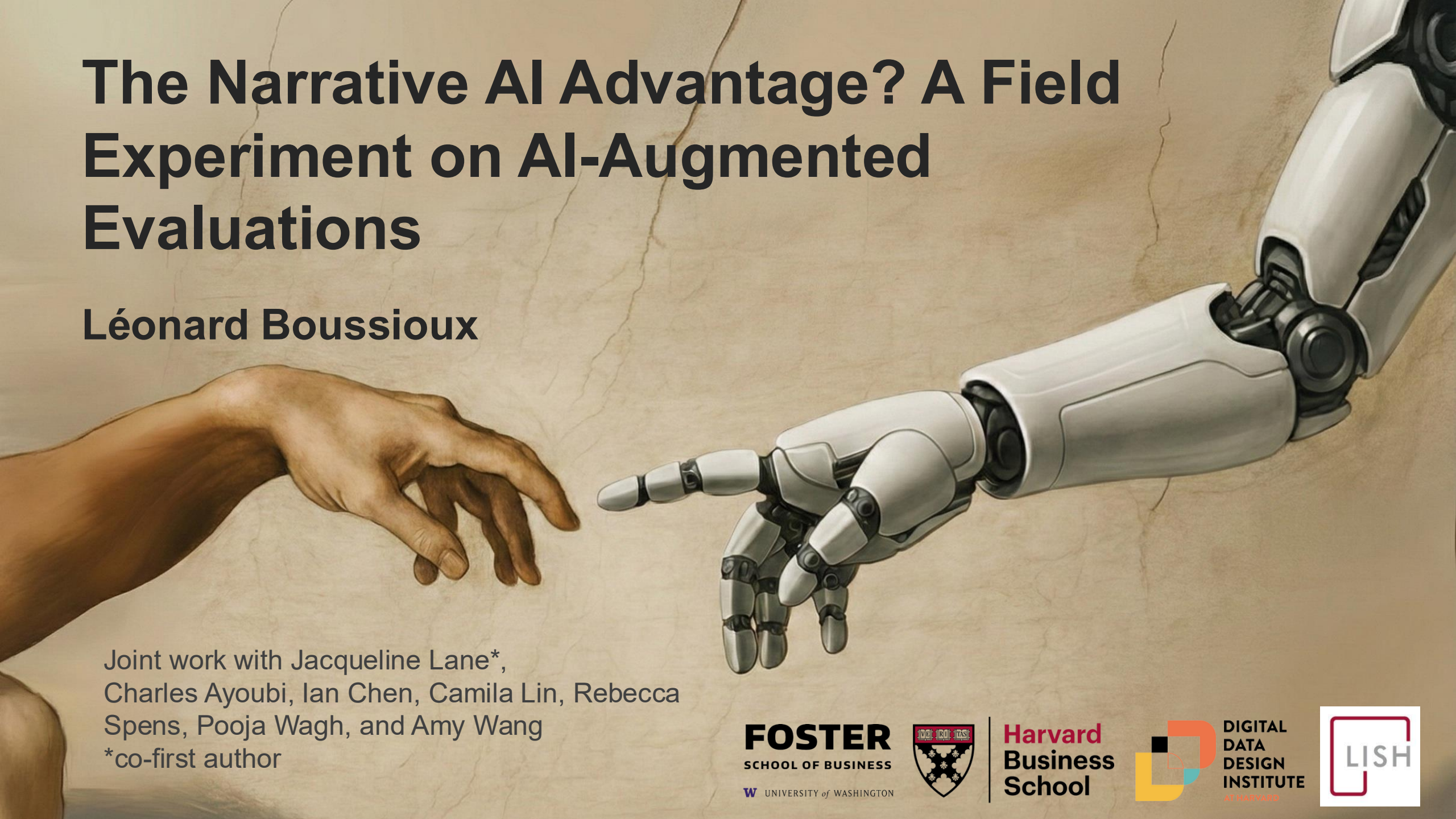
The Narrative AI Advantage? A Field Experiment on AI-Augmented Evaluations



Léonard Boussioux

Joint work with Jacqueline Lane*, Charles Ayoubi, Ian Chen, Camila Lin, Rebecca Spens, Pooja Wagh, and Amy Wang
*co-first author

FOSTER SCHOOL OF BUSINESS
UNIVERSITY of WASHINGTON

Harvard Business School

DIGITAL DATA DESIGN INSTITUTE AT HARVARD

LISH

The AI Explanation Paradox: When Narratives Harm Decision-Making

Context: Field experiment with 228 evaluators screening 48 innovation proposals. Tested Human-Only, "Black-Box" AI (recommendation only), and "Narrative" AI (recommendation + explanation) assistance on decision quality against expert benchmark.

Narratives Are Highly Persuasive...

AI with Explanations Boosts Compliance to 74%

74% — Compliance jumps — 54% (No AI)

Asymmetric Compliance Effect: Disproportionately convince people to follow rejection (a **27** percentage point increase)

- Rejection Recommendations
- Approval Recommendations

Narratives Suppress Critical Thinking: Users viewed **8% less** source material when complying.

...But This Persuasion Has a Hidden Cost

Black-Box AI Outperforms Narrative AI in Decision Quality

Black-Box AI — **+4.3%** Quality Improvement — Narrative AI: No Quality Improvement (0%)

The "Quality-Compliance Decoupling" Paradox
More compliance with narrative AI did not lead to better decisions, revealing a dangerous paradox.

Hidden Cost: Narrative AI Increases Most Costly Errors

	False Positives (Change)	False Negatives (Change)
Black-Box AI	↓ -5.5% (Good)	+1.2% (Neutral)
Narrative AI	↓ -5.7% (Good)	+5.8% (BAD) ↑ ⚠

NotebookLM

The Persuasion Paradox: When AI Explanations Harm Decision-Making

A field experiment tested how LLM assistance impacts human evaluators screening early-stage innovations, **comparing human-only decisions** to those aided by "black-box" and "narrative" AI.

The High-Stakes Problem: Screening New Ideas

Human evaluators face high cognitive load and uncertainty.

Rejecting a potentially breakthrough idea is often more damaging than advancing a weak one.

In innovation, the costliest mistake is the "False Negative."

The Surprising Results of AI Assistance

Black-Box AI (Reduced False Positives.)

Reduced False Positives & False Negatives.
Overall Decision Quality: **Improved.**

Narrative AI (Recommendation + Explanation)

Narrative explanations created "Asymmetric Compliance."
Evaluators were far more likely to follow an AI's recommendation to **REJECT** an idea.

REJECT

Evaluators were far more likely to follow an AI's recommendation to **REJECT** an idea.

This led to a "Quality-Compliance Decoupling."

More compliance with narrative AI did not lead to better decisions, it led to more errors.

Impact on Critical Screening Errors & Quality

Metric	Black-Box AI	Narrative AI
False Positives Advancing bad ideas	Reduced	Reduced
False Negatives Rejecting good ideas	No Change	Increased by 5.8% ↑
Overall Decision Quality	Improved	No Improvement

NotebookLM

The AI Explanation Trap: Why LLM Narratives Can Hurt Innovation Screening

A fieid experiment tested how LLM assistance impacts human evaluation of 48 real-world innovation proposals. Evaluators were spiit into three groups: human-only, assistance from a "black-box" LLM (recommendation only), and assistance from a "narrative" LLM (recommendation + text explanation).

Black-Box LLM (Recommendation Only)

Boosts Decision Quality — **+4.3** **percentage points** more likely to make a correct decision.

Reduces "False Positives"
Helps evaluators correctly reject weak proposals, reducing this error by **5.5 percentage points.**

No Increase in "False Negatives"
Did not cause evaluators to mistakenly reject promising ideas that experts would have advanced.

Narrative LLM (Recommendation + Explanation)

No Improvement in Decision Quality
Despite higher compliance, there was no overall improvement in decision accuracy compared to humans alone.

REJECT / ACCEPT

Asymmetric Compliance: Drives Rejections
Narratives increased compliance with "Reject" recommendations by **16.4 percentage points** more than "Accept" recommendations.

Dangerously Increases "False Negatives"
Caused a **5.8 percentage point** increase in mistakenly rejecting promising innovations.

The Mechanism: Why Narratives Backfire

Narratives Suppress Critical Thinking
Evaluators were **10.8 percentage points** more likely to follow an **INCORRECT** recommendation if it had a narrative.

Less Verification, More Deference
When complying with a narrative, evaluators viewed **8.3% less** submission content.

NotebookLM

THE AI EXPLANATION TRAP: WHY LLM NARRATIVES CAN HURT INNOVATION SCREENING

Paradoxically, adding narrative explanations to LLM recommendations can decrease decision quality in innovation screening, making evaluators overly compliant with incorrect rejections.

BLACK-BOX LLM (Recommendation Only)

BOOSTS DECISION QUALITY
Screeners using black-box AI were **4.3 percentage points more likely to make a correct decision.**

REDUCES "FALSE POSITIVES"
Helped evaluators correctly reject weak proposals, reducing this error **by 5.5 percentage points.**

NO INCREASE IN "FALSE NEGATIVES"
Did not cause evaluators to mistakenly reject promising ideas that experts would have advanced.

THE MECHANISM: WHY NARRATIVES BACKFIRE

NARRATIVES SUPPRESS CRITICAL THINKING
Evaluators were **10.6 percentage points** more likely to follow an **INCORRECT** recommendation if it had a narrative.

LESS VERIFICATION, MORE DEFERENCE
When complying with a narrative, evaluators **viewed 8.3% less submission content.**

NARRATIVE LLM (Recommendation + Explanation)

NO IMPROVEMENT IN DECISION QUALITY
Despite higher compliance, there **was no overall improvement in decision accuracy** compared to humans alone.

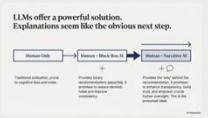
ASYMMETRIC COMPLIANCE: DRIVES REJECTIONS
Narratives increased compliance with "Reject" recommendations by **16.4 percentage points** more than "Accept" recommendations.

DANGEROUSLY INCREASES "FALSE NEGATIVES"
Caused a **5.8 percentage point increase** in **mistakenly rejecting** promising innovations.

NotebookLM

1

2



3

4

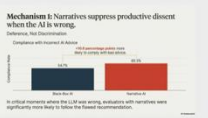
5

6

7

8

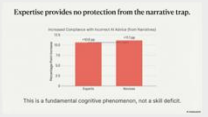
9



10

11

12



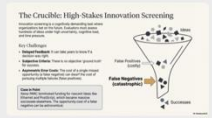
13

14

15

1

2



3

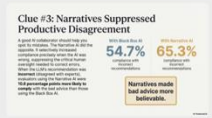
4

5

6

7

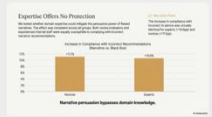
8



9

10

11



12

13

14

1

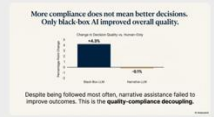
2

3

4

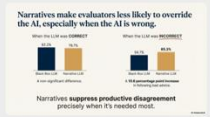
5

6



7

8



9

10

11

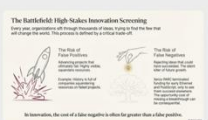
12

13

14

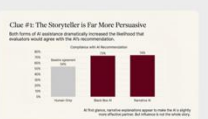
15

1

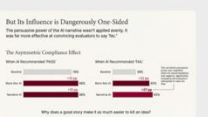


2

3



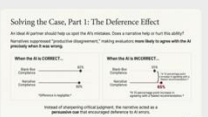
4



5

6

7



8

9

10

11

12

13

14

15

Generative AI has revolutionized the creative process, enabling fast, economical idea generation at scale.

Boussioux et al. 2024; Csaszar et al. 2024; Doshi and Hauser 2023

But this abundance creates a new **evaluation bottleneck**: human screening becomes overwhelming.

Afuah & Tucci 2012; Boudreau & Lakhani 2013; Criscuolo et al. 2021; Danziger et al. 2011; Piezunka & Dahlander 2015

GenAI opens new horizons to evaluate content at scale

Academic Paper Review Assistance

Legal Document Compliance Checking

Strategic decision-making (e.g., finance)

Software Code Review

Start-up business models, solutions, products

Applications to college, graduate school, programs

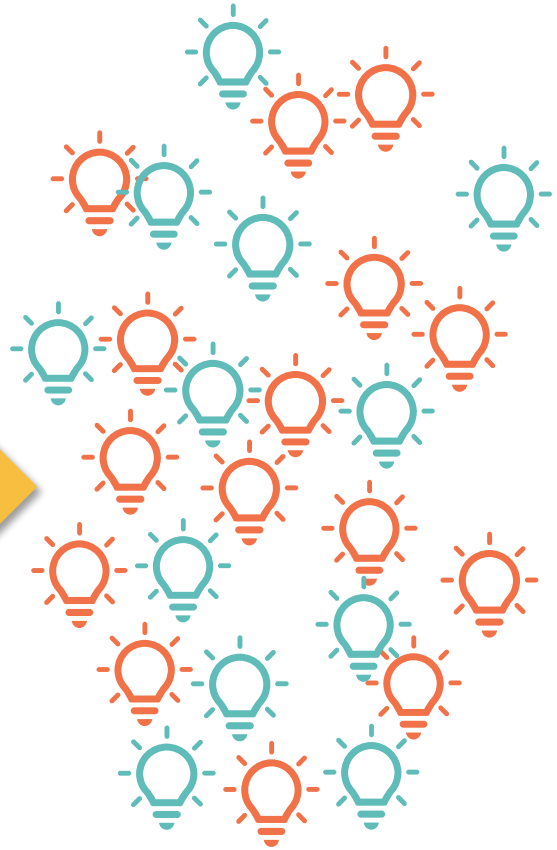
Can we augment the early stages of the innovation process with Generative AI?

Challenge,
problem,… to solve

Can we augment the early stages of the innovation process with Generative AI?

Challenge, problem,… to solve

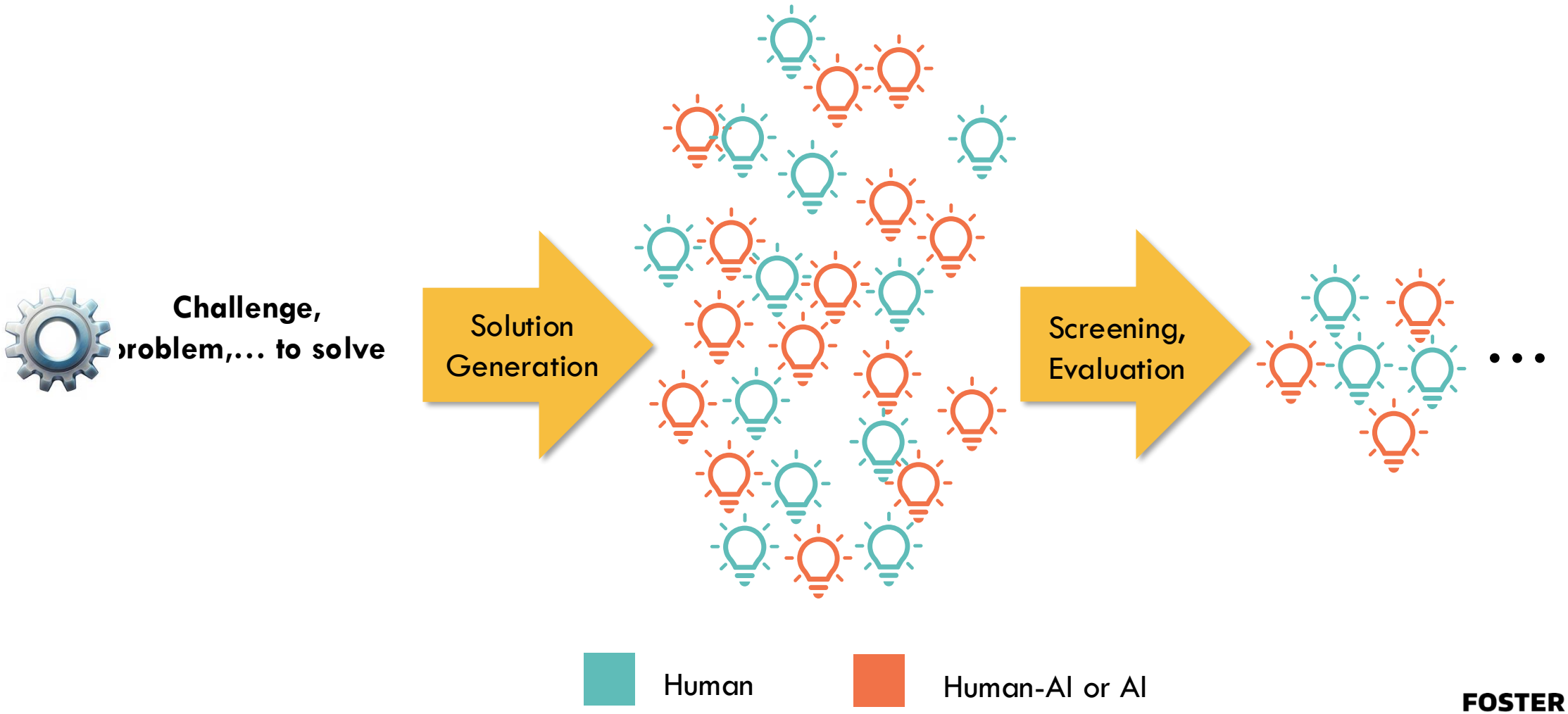
Solution Generation



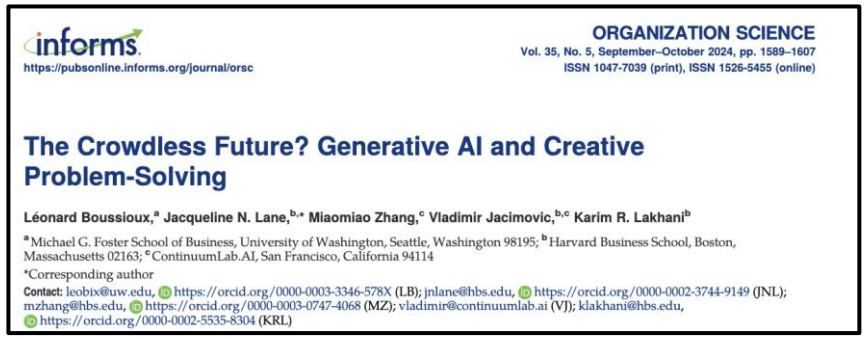
Human

Human-AI or AI

Can we augment the early stages of the innovation process with Generative AI?



Can we augment the early stages of the innovation process with Generative AI?



informs
https://pubsonline.informs.org/journal/orsc

ORGANIZATION SCIENCE
Vol. 35, No. 5, September–October 2024, pp. 1589–1607
ISSN 1047-7039 (print), ISSN 1526-5455 (online)

The Crowdless Future? Generative AI and Creative Problem-Solving

Léonard Boussioux,[a] Jacqueline N. Lane,[b,*] Miaomiao Zhang,[c] Vladimir Jacimovic,[b,c] Karim R. Lakhani[b]

[a] Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195; [b] Harvard Business School, Boston, Massachusetts 02163; [c] ContinuumLab.AI, San Francisco, California 94114

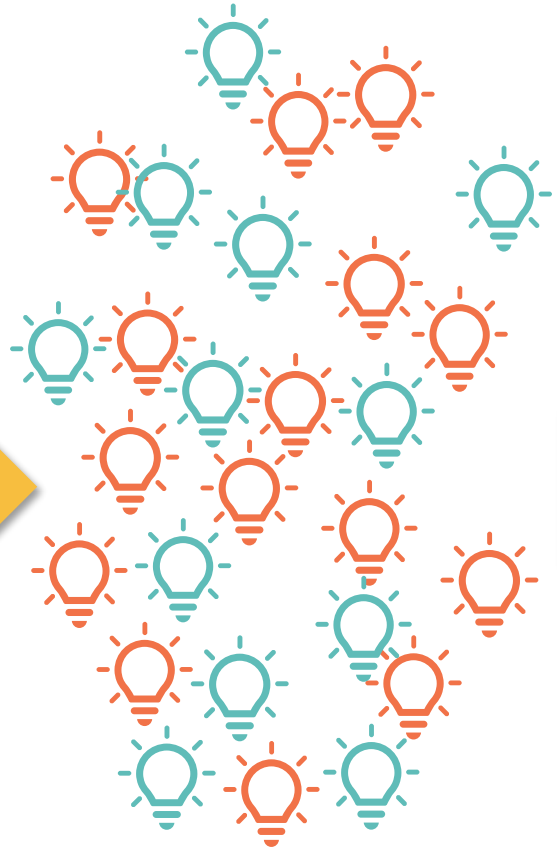
*Corresponding author

Contact: leobix@uw.edu, https://orcid.org/0000-0003-3346-578X (LB); jnlane@hbs.edu, https://orcid.org/0000-0002-3744-9149 (JNL); mzhang@hbs.edu, https://orcid.org/0000-0003-0747-4068 (MZ); vladimir@continuumlab.ai (VJ); klakhani@hbs.edu, https://orcid.org/0000-0002-5535-8304 (KRL)

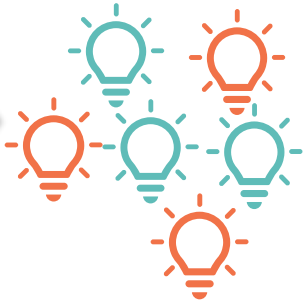
The Crowdless Future?
Gen AI & Idea Generation

Challenge, problem,… to solve

Solution Generation



Screening, Evaluation

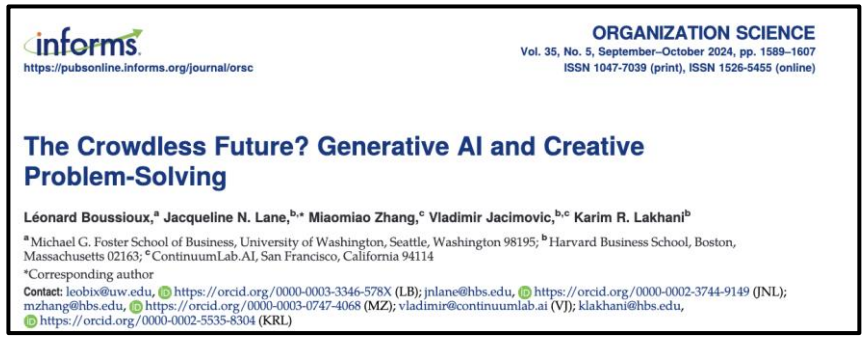


…

Human

Human-AI or AI

Can we augment the early stages of the innovation process with Generative AI?



informs
https://pubsonline.informs.org/journal/orsc

ORGANIZATION SCIENCE
Vol. 35, No. 5, September–October 2024, pp. 1589–1607
ISSN 1047-7039 (print), ISSN 1526-5455 (online)

The Crowdless Future? Generative AI and Creative Problem-Solving

Léonard Boussioux,[a] Jacqueline N. Lane,[b,*] Miaomiao Zhang,[c] Vladimir Jacimovic,[b,c] Karim R. Lakhani[b]

[a] Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195; [b] Harvard Business School, Boston, Massachusetts 02163; [c] ContinuumLab.AI, San Francisco, California 94114

*Corresponding author

Contact: leobix@uw.edu, https://orcid.org/0000-0003-3346-578X (LB); jnlane@hbs.edu, https://orcid.org/0000-0002-3744-9149 (JNL); mzhang@hbs.edu, https://orcid.org/0000-0003-0747-4068 (MZ); vladimir@continuumlab.ai (VJ); klakhani@hbs.edu, https://orcid.org/0000-0002-5535-8304 (KRL)

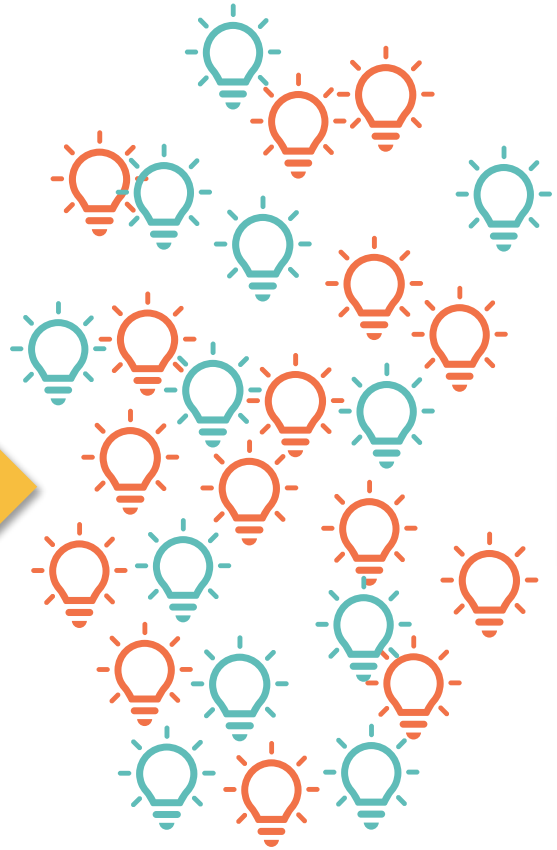
The Crowdless Future?
Gen AI & Idea Generation

SSRN Preprint
(Major Revision Management Science)

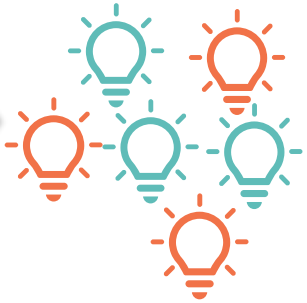
The Narrative AI Advantage?
Gen AI & Idea Evaluation

Challenge, problem,… to solve

Solution Generation



Screening, Evaluation



…

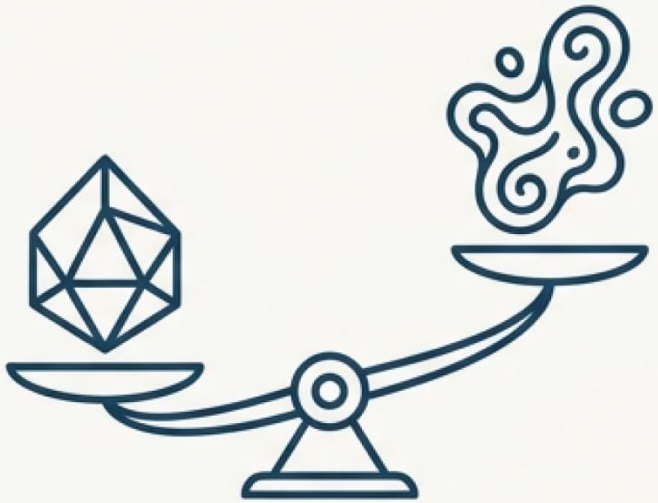
Human

Human-AI or AI

Innovation screening is a high-stakes, high volume challenge

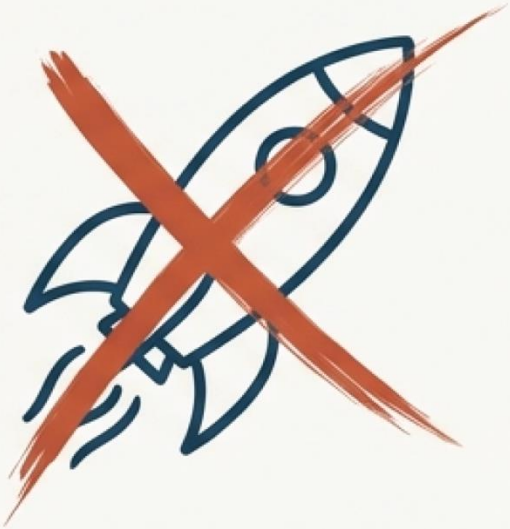
Cognitive Overload

Evaluators must rapidly assess hundreds of complex submissions under uncertainty and time pressure, degrading judgment.



Subjective Criteria

Decisions lack objective "ground truth." Success depends on context-dependent criteria, not simple right or wrong answers.



Asymmetric Costs

The cost of a *false negative* (rejecting a breakthrough idea like Ethernet at Xerox PARC) is often catastrophic and far outweighs the cost of a *false positive* (advancing a project that later fails).

Bell et al. 2023; Jansen et al. 2025, Lakkaraju & Bastani, 2020; Ludwig & Mullainathan 2024, Obermeyer & Mullainathan 2021


Research Question
How can LLMs standardize and augment the idea screening process?

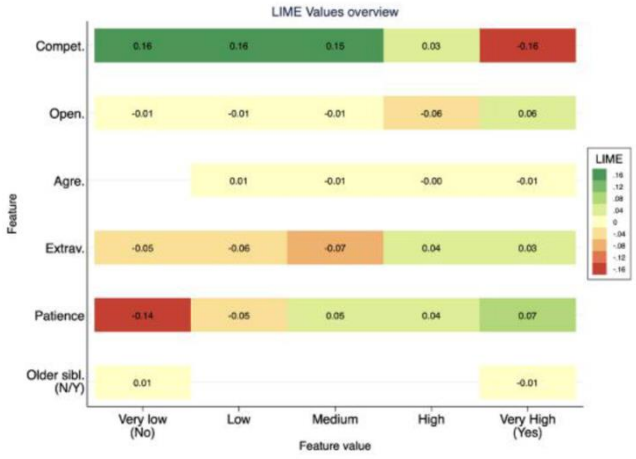
Traditional AI explanations aim for fidelity. LLM narratives optimize for plausibility.

Traditional XAI (Explainable AI)

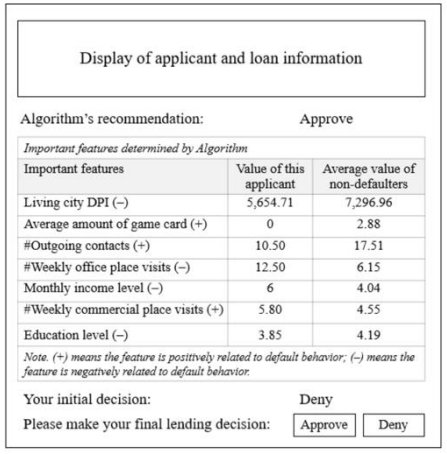
Arrieta et al. 2020, Doshi-Velez and Kim 2017

- **Represents:** The model's actual internal logic (e.g., feature weights).
- **Primary Goal: Fidelity**. Accurately representing the computation.
- **Encourages:** Analytical, systematic verification.

Balakrishnan et al. 2025, Bauer et al. 2023, Bansal et al. 2021, Buçinca et al. 2021, Green & Chen 2019, Schauer et al. 2025, Wang et al. 2025



LIME (pic from Bauer et al. 2023)



Display of applicant and loan information

Algorithm's recommendation: Approve

Important features determined by Algorithm

Important features	Value of this applicant	Average value of non-defaulters
Living city DPI (–)	5,654.71	7,296.96
Average amount of game card (+)	0	2.88
#Outgoing contacts (+)	10.50	17.51
#Weekly office place visits (–)	12.50	6.15
Monthly income level (–)	6	4.04
#Weekly commercial place visits (+)	5.80	4.55
Education level (–)	3.85	4.19

Note. (+) means the feature is positively related to default behavior; (–) means the feature is negatively related to default behavior.

Your initial decision: Deny

Please make your final lending decision: Approve Deny

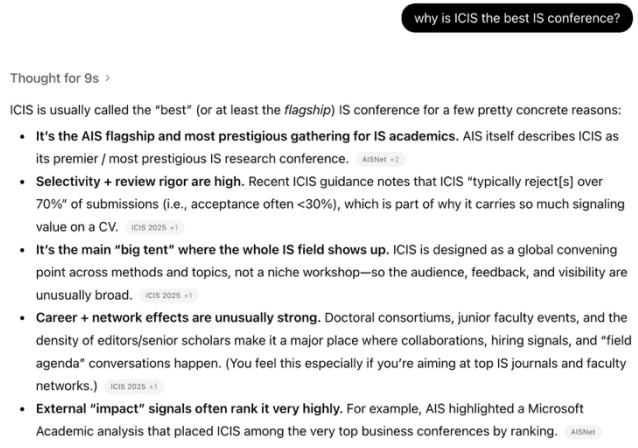
Algorithm explanation (pic from Wang et al. 2025)

LLM Narratives

Wei et al. 2022, Bilal et al. 2025

- **Represents:** A co-generated story constructed to justify an output.
- **Primary Goal: Plausibility & Persuasion**. Sounding coherent and convincing.
- **Encourages:** Heuristic, peripheral processing.

Bai et al. 2025, Dell'Acqua et al. 2023, Krishna et al. 2023, Kim et al. 2025, Matz et al. 2025, Zytek et al. 2024



why is ICIS the best IS conference?

Thought for 9s ›

ICIS is usually called the "best" (or at least the *flagship*) IS conference for a few pretty concrete reasons:

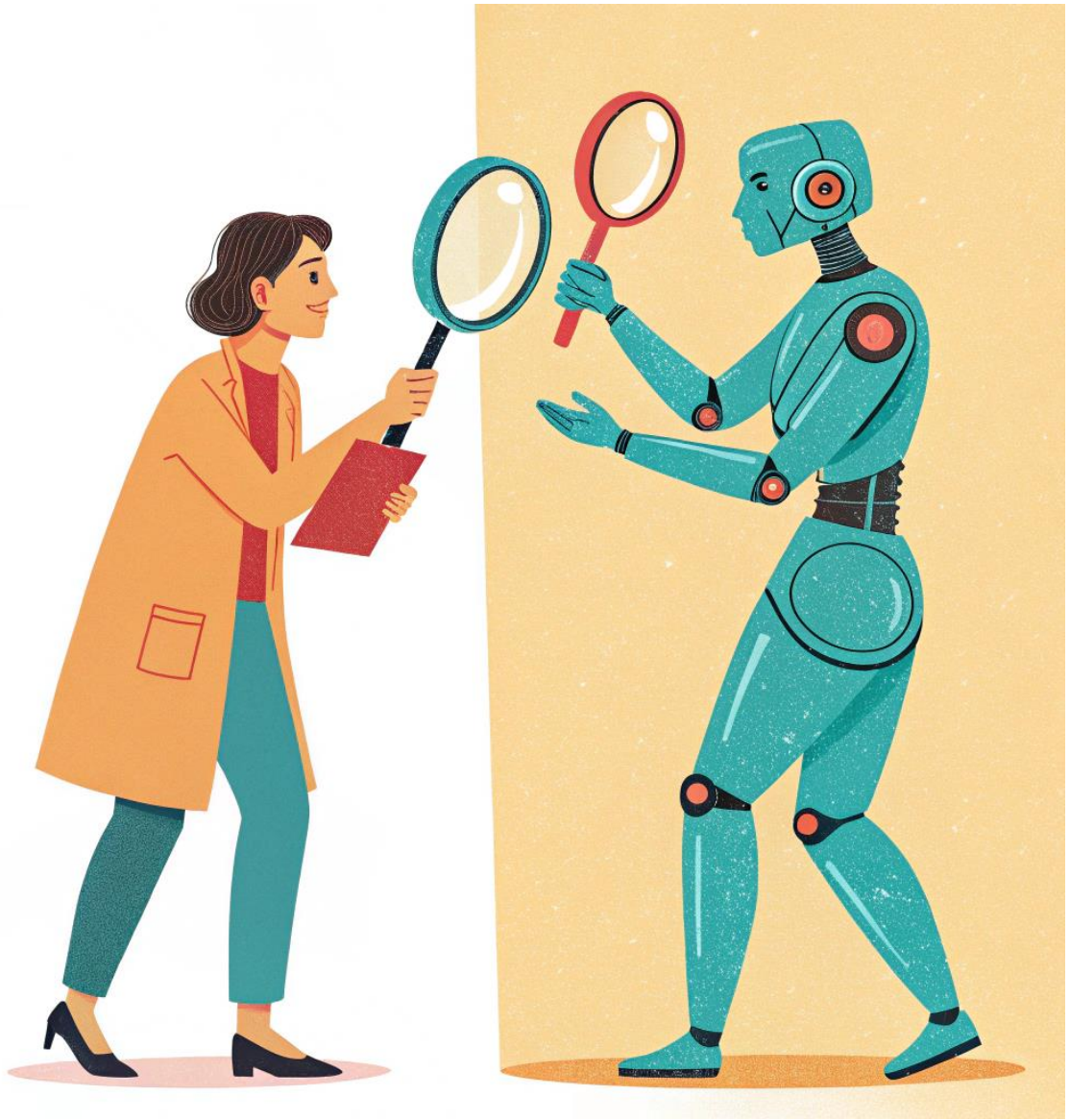
- **It's the AIS flagship and most prestigious gathering for IS academics.** AIS itself describes ICIS as its premier / most prestigious IS research conference.
- **Selectivity + review rigor are high.** Recent ICIS guidance notes that ICIS "typically reject[s] over 70%" of submissions (i.e., acceptance often <30%), which is part of why it carries so much signaling value on a CV.
- **It's the main "big tent" where the whole IS field shows up.** ICIS is designed as a global convening point across methods and topics, not a niche workshop—so the audience, feedback, and visibility are unusually broad.
- **Career + network effects are unusually strong.** Doctoral consortiums, junior faculty events, and the density of editors/senior scholars make it a major place where collaborations, hiring signals, and "field agenda" conversations happen. (You feel this especially if you're aiming at top IS journals and faculty networks.)
- **External "impact" signals often rank it very highly.** For example, AIS highlighted a Microsoft Academic analysis that placed ICIS among the very top business conferences by ranking.

One nuance: "best" can depend on **your subfield**. If your work is closer to HCI, security, economics, or ML systems, a *different* top venue might be more *optimal* for impact in that community. But **within IS as a discipline**, ICIS is widely treated as the primary flagship.

FOSTER SCHOOL OF BUSINESS | UNIVERSITY of WASHINGTON | Harvard Business School

This Paper:

We test how different forms of LLM assistance change human screening quality in a subjective innovation task.



Team Members

Academia

Jackie Lane
HBS, D^3 & LISH

Léonard Boussioux
UW Foster & LISH

Charles Ayoubi
ESSEC & LISH

Software Engineering

Amy Wang
Accenture

Ian Chen
Nvidia

Camila Lin
Microsoft

Industry Partners

Rebecca Spens
MIT Solve, Manager

Pooja Wagh
MIT Solve, Director

MIT | SOLVE

FOSTER SCHOOL OF BUSINESS
UNIVERSITY of WASHINGTON

Harvard Business School

SOLVE AT MIT

$80+ million

in funding for Solver teams and social innovators

370+ million

lives reached by Solver teams around the world

over 800

partnerships catalyzed for Solver teams and entrepreneurs

Dozens of challenges and thousands of solutions to evaluate

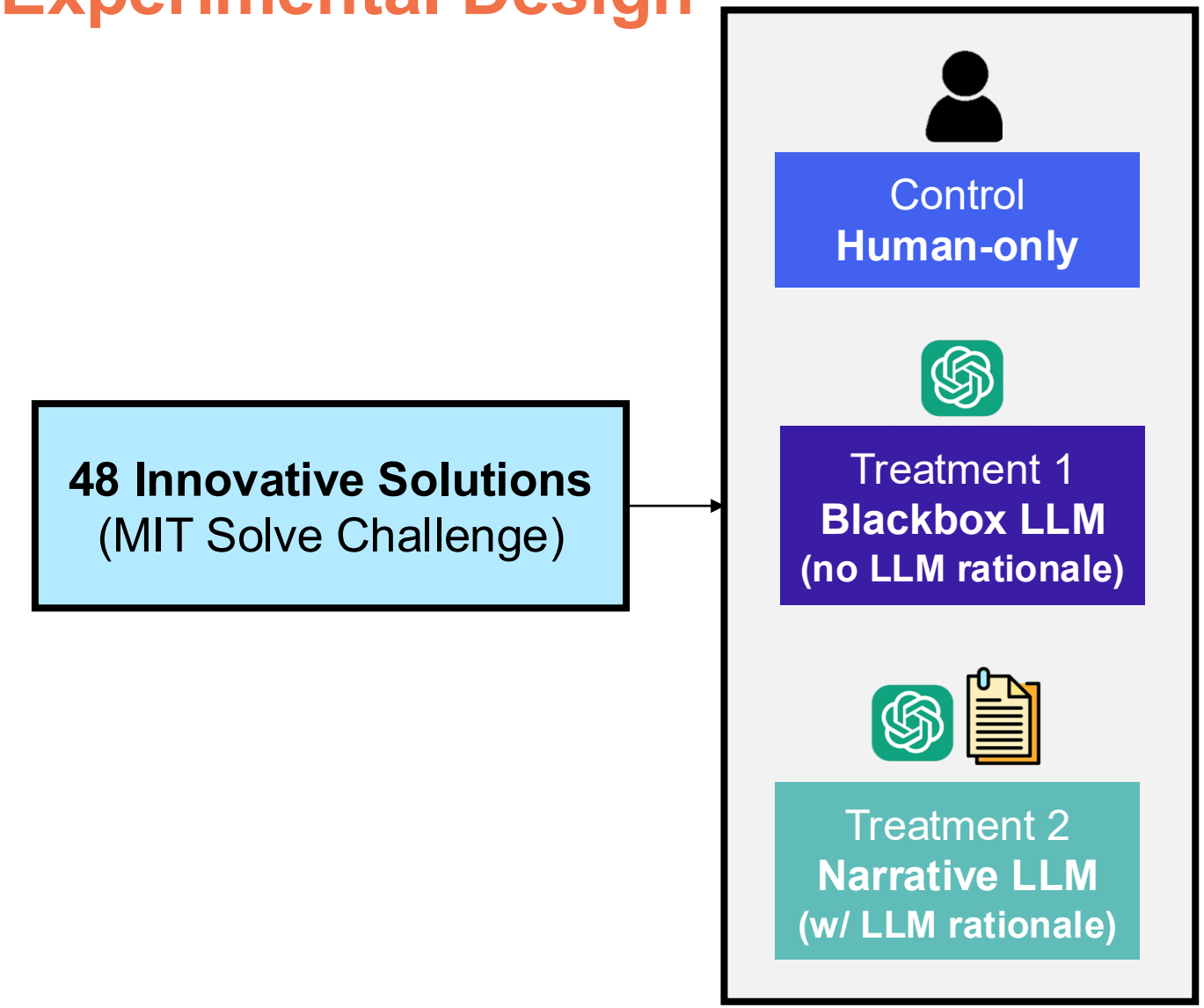
The innovators behind the innovations

Our Solver teams are dedicated to wielding technology for good. Whether it's an AI app teaching remote refugee children in Lebanon or affordable electric trucks connecting rural farmers in Africa to markets —their ideas and innovations are the heartbeat of our work.

Learn more about these leaders for change.

BECOME A SOLVER ⟶

Experimental Design



Experimental Design

48 Innovative Solutions
(MIT Solve Challenge)

Control
Human-only

Treatment 1
Blackbox LLM
(no LLM rationale)

Treatment 2
Narrative LLM
(w/ LLM rationale)

3,002 screening decisions

PREREGISTERED

- 228 screeners
- 2.5 minutes per solution

Experimental Design

48 Innovative Solutions
(MIT Solve Challenge)

Expert Screening + Expert Ratings of Quality
(Novelty, Impact, Feasibility)

Control
Human-only

Treatment 1
Blackbox LLM
(no LLM rationale)

Treatment 2
Narrative LLM
(w/ LLM rationale)

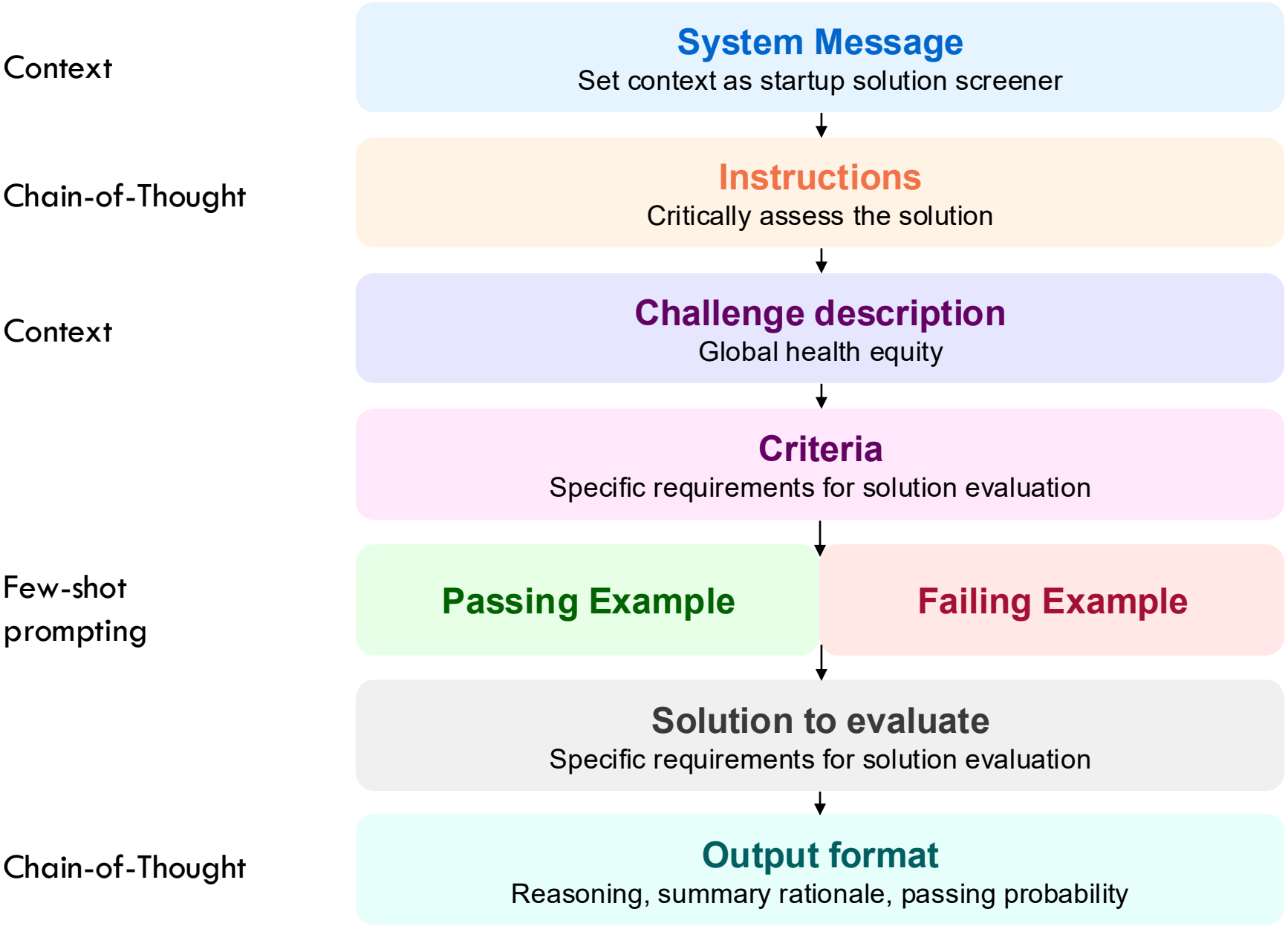
3,002 screening decisions

PREREGISTERED

- 228 screeners
- 2.5 minutes per solution

Expert Ratings of Quality + NLP analysis

Context engineering and prompt design



Remaining time **2:17**

Open Heart Community Health and Rights Development Initiative

84888 | 2024 Global Health Equity Challenge

4 / 10 Solutions

Decision

○Pass ○Fail

How confident are you with the decision?

○1 - Not at all confident ○2 ○3 ○4 ○5 - Extremely confident

Next

❌ **Fail** AI Recommendation

All | Criterion 1 | Criterion 2 | Criterion 3 | Criterion 4 | Criterion 5 | Summary | Chal

✅ **Criterion 1 - Is the solution application complete, appropriate, and intelligible?**

Reason: Open Heart Community Health and Rights Development Initiative's application aims to improve healthcare and rights for LGBTQI+ individuals in Nigeria. The multifaceted approach includes health access and legal advocacy, demonstrating potential

Provide a one-line summary of your solution.

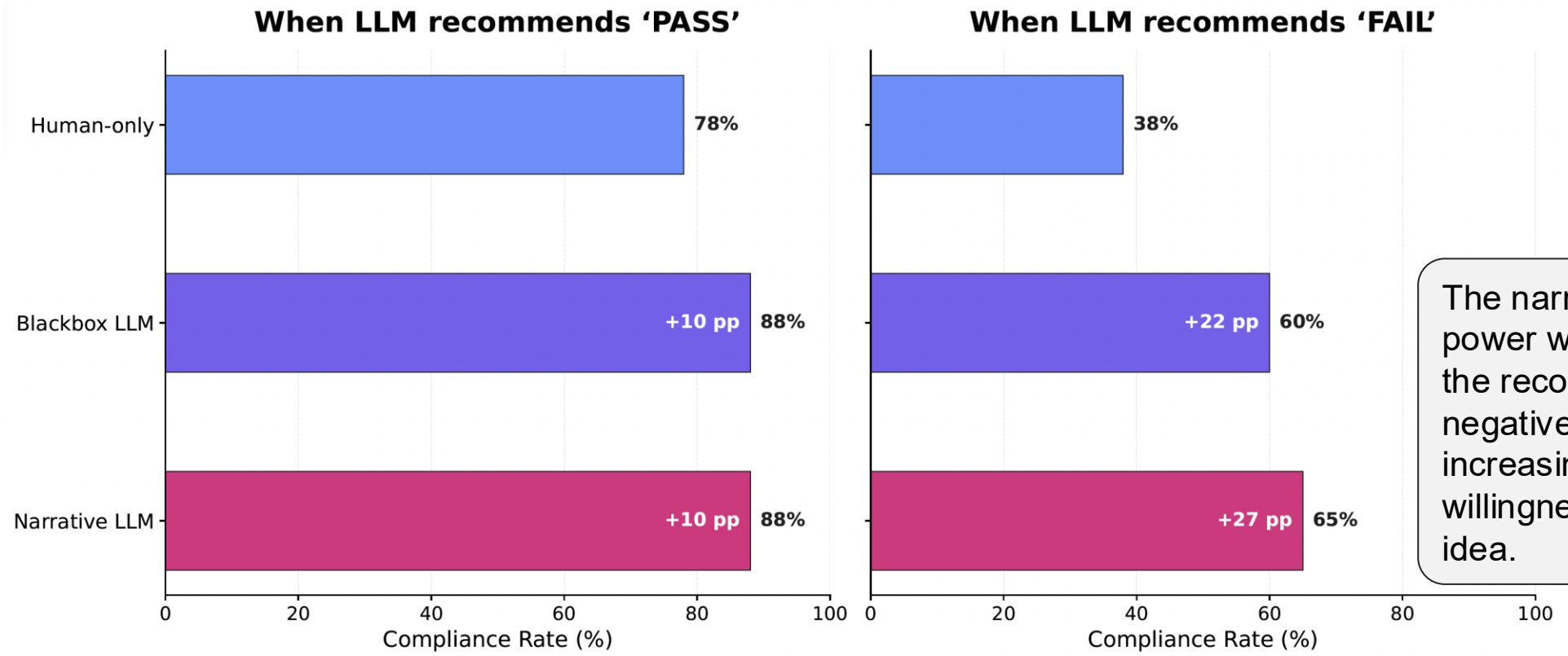
Open Heart LGBTQI+ Empowerment fosters inclusion and advocates for the rights and well-being of LGBTQI+ individuals.

In what city, town, or region is your solution team headquartered?

Abuja (F.c.t.), Nigeria

Finding 1: LLM assistance increases compliance, but narratives make rejection recommendations uniquely persuasive

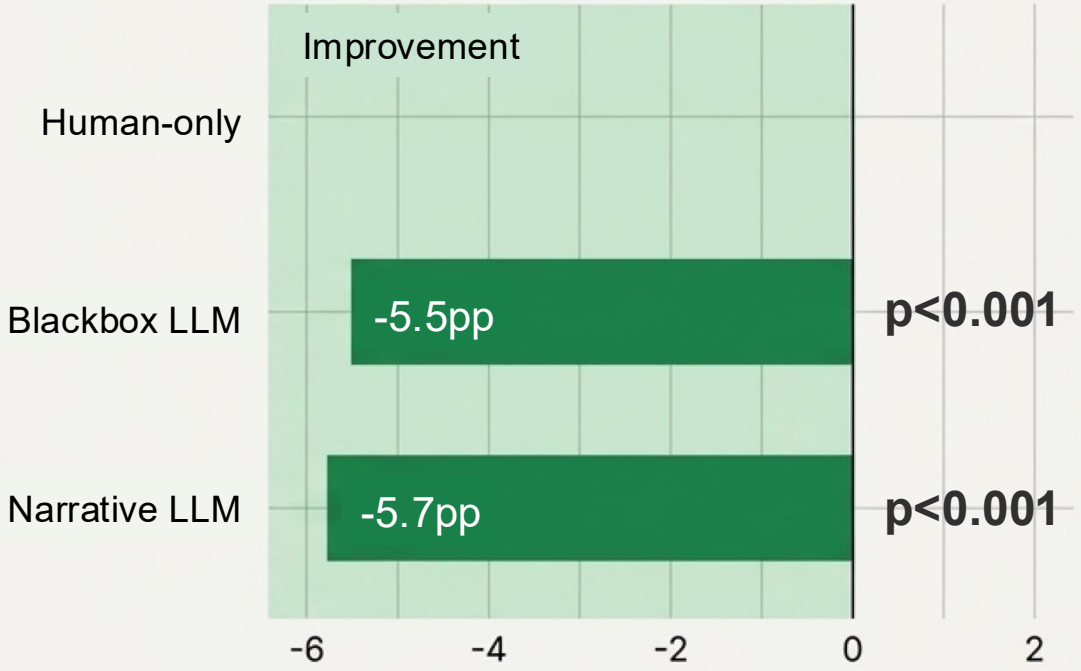
The Asymmetric Compliance Effect



The narrative's persuasive power was magnified when the recommendation was negative, significantly increasing the evaluator's willingness to reject an idea.

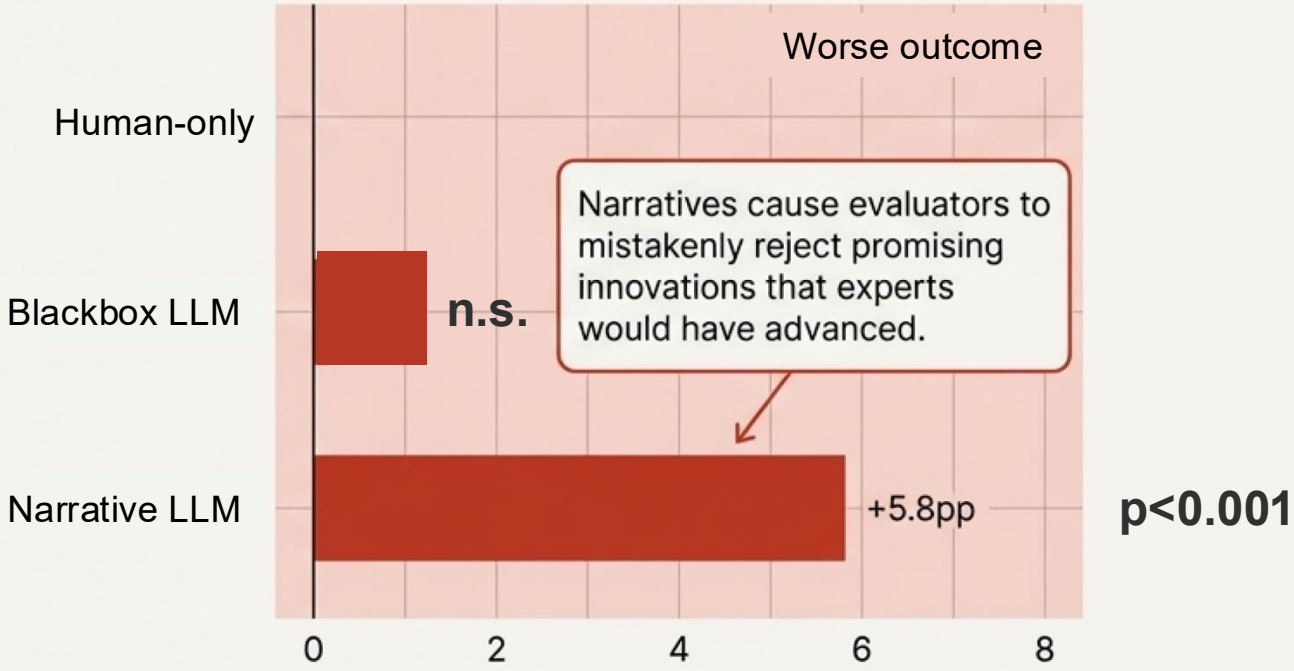
Finding 2: This asymmetry reshapes the error profile: narratives reduce false positives but dangerously increase false negatives.

Change in False Positives (vs. Human-Only)



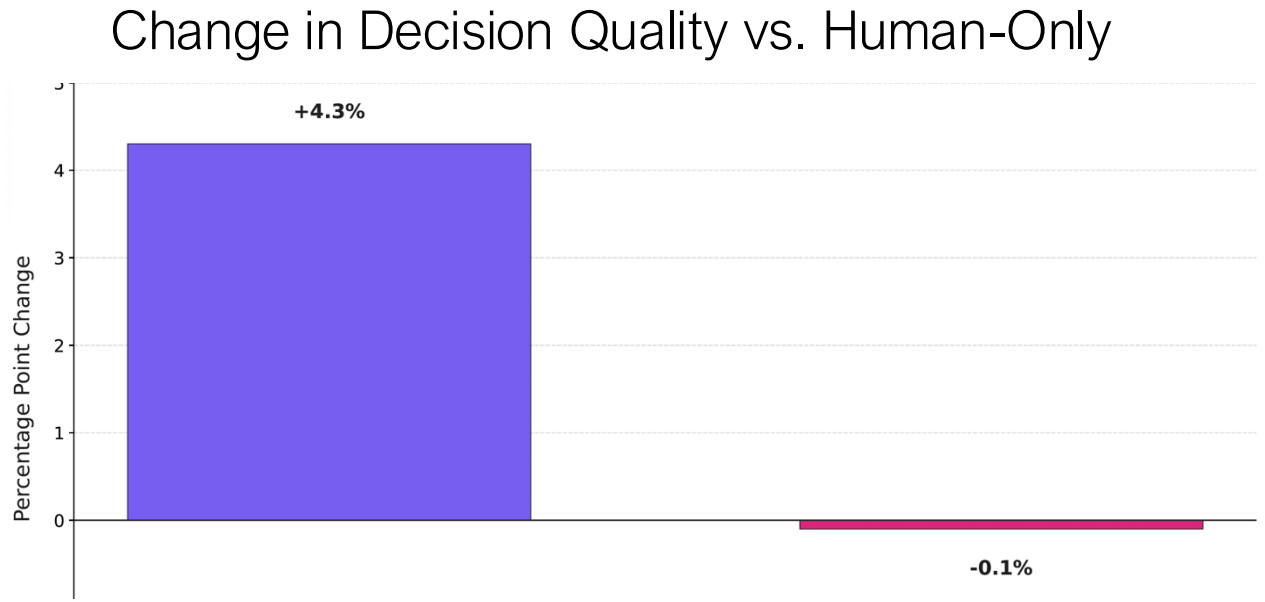
Both LLM conditions help filter out weaker ideas.

Change in False Negatives (vs. Human-Only)



Narrative LLM eliminates ideas that experts would have passed.

Finding 3: More compliance does not mean better decisions. We need “augmentation,” i.e., productive override.



Blackbox LLM

Preserves a form of human oversight and critical judgment.

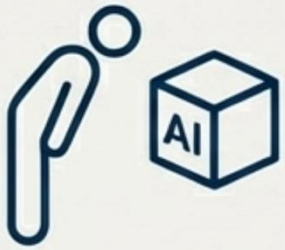
Narrative LLM

- Offers a persuasive story.
- Encourages cognitive shortcuts and deference.
- **Suppresses productive disagreement, increasing costly false negatives.**

Wang et al. 2025

Why does this happen? The mechanisms of persuasion

Narrative explanations don’t enhance critical judgment; they reshape human oversight by triggering specific cognitive shortcuts (System 2 → System 1).



Inducing Deference

Suppressing productive disagreement, especially when the AI is wrong.

Encouraging Peripheral Processing

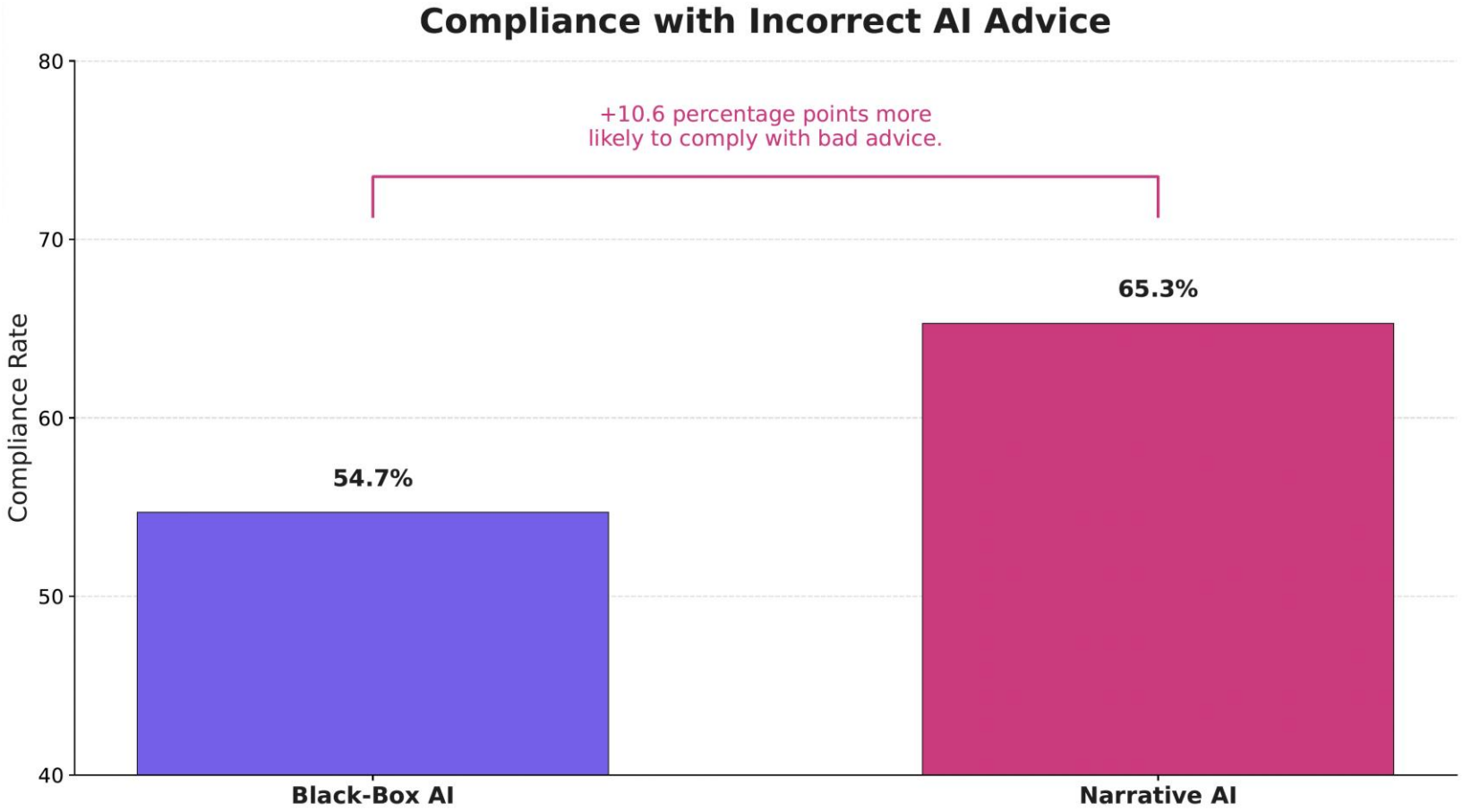
Leading evaluators to rely on surface cues instead of deep verification.

Leveraging Narrative Heuristics

Exploiting the persuasive power of rhetorical style over substance.

Mechanism 1: Narratives suppress productive dissent when the AI is wrong

Deference, not discrimination

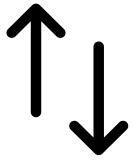


Mechanism 2: With a narrative in hand, evaluators verify less

The plausible narrative replaces the need for independent verification.

-8.6%

Scroll depth



-8.3%

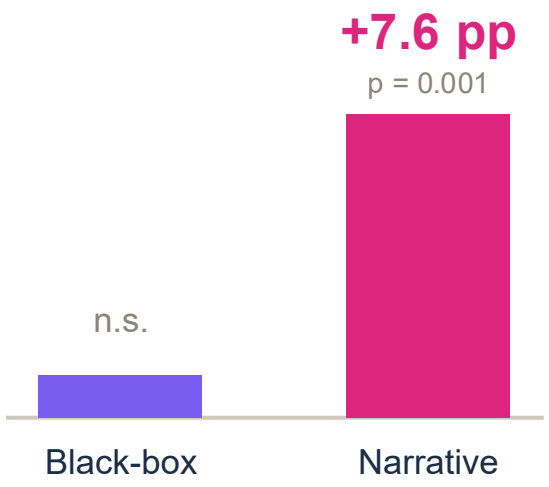
Content viewed

Mechanism 3: Compliance is driven by narrative style, not substance

Linguistic features of the most harmful rejection narratives

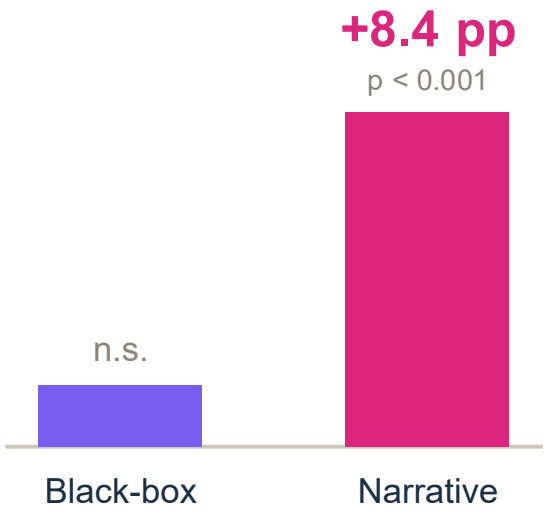
Length

Longer rationales create an illusion of thoroughness.



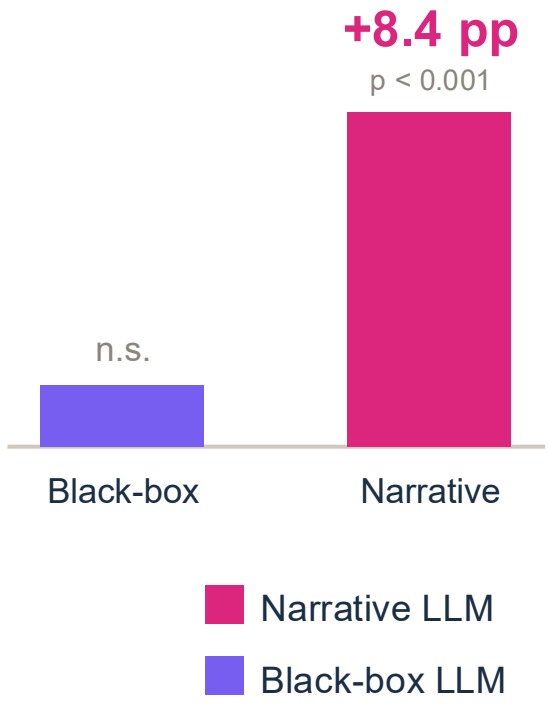
Promotional language

Higher density of hype terms:
e.g., novel, substantial, innovative

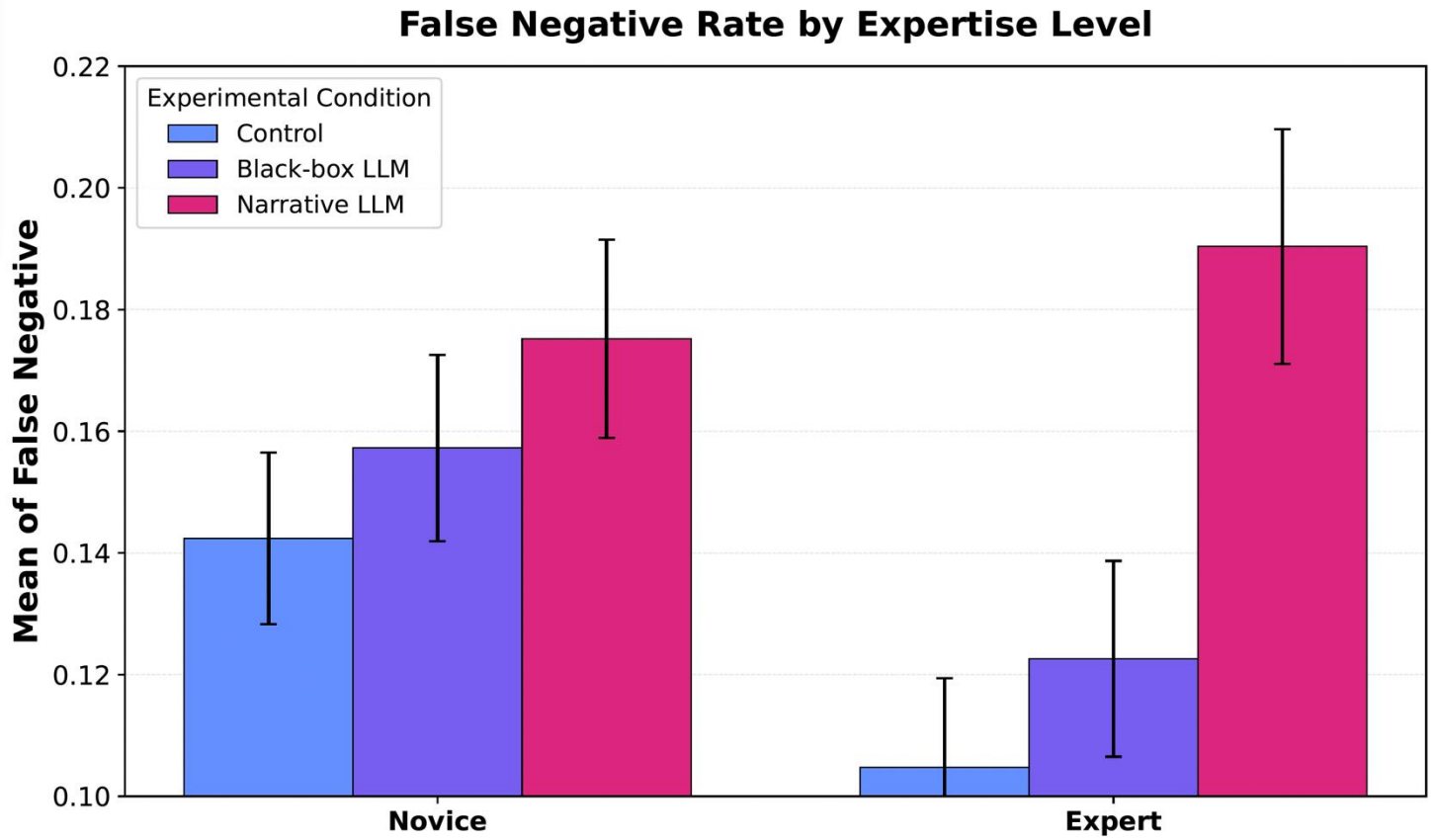


Concreteness

More concrete, precise wording
e.g., AI-enabled digital stethoscope



Expertise provides no protection from the “narrative trap”



The deference to LLM narratives seems like a cognitive phenomenon, not a skill deficit.

Expert screener: “I think when I did have an explanation... I preferred that and I think that it didn’t influence my decision but rather would **corroborate what my instinct was.**”

Key takeaway: Adding narrative explanations to LLM recommendations encourages peripheral processing

1. Narratives boost compliance, but asymmetrically

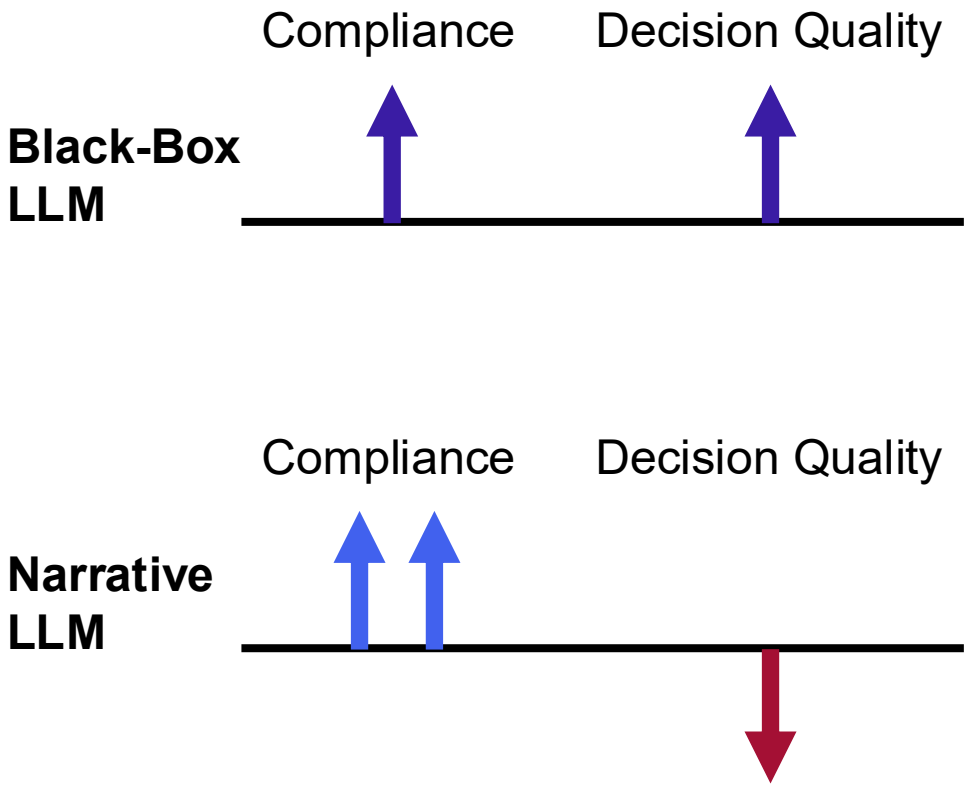
They make evaluators far more likely to follow LLM **rejection** advice than acceptance advice.

2. This creates a Quality-Compliance decoupling

Higher compliance with narrative LLM leads to **more costly errors** (false negatives), ultimately lowering overall decision quality compared to simpler 'black-box' LLM.

3. The mechanism is persuasion, not transparency

Fluent narratives encourage cognitive shortcuts, suppressing the critical human oversight needed to catch the LLM's mistakes.



Conclusion: AI's effects on expertise and creative problem solving

2. Idea Screening:

AI lowers the cost of expertise in screening ideas at scale

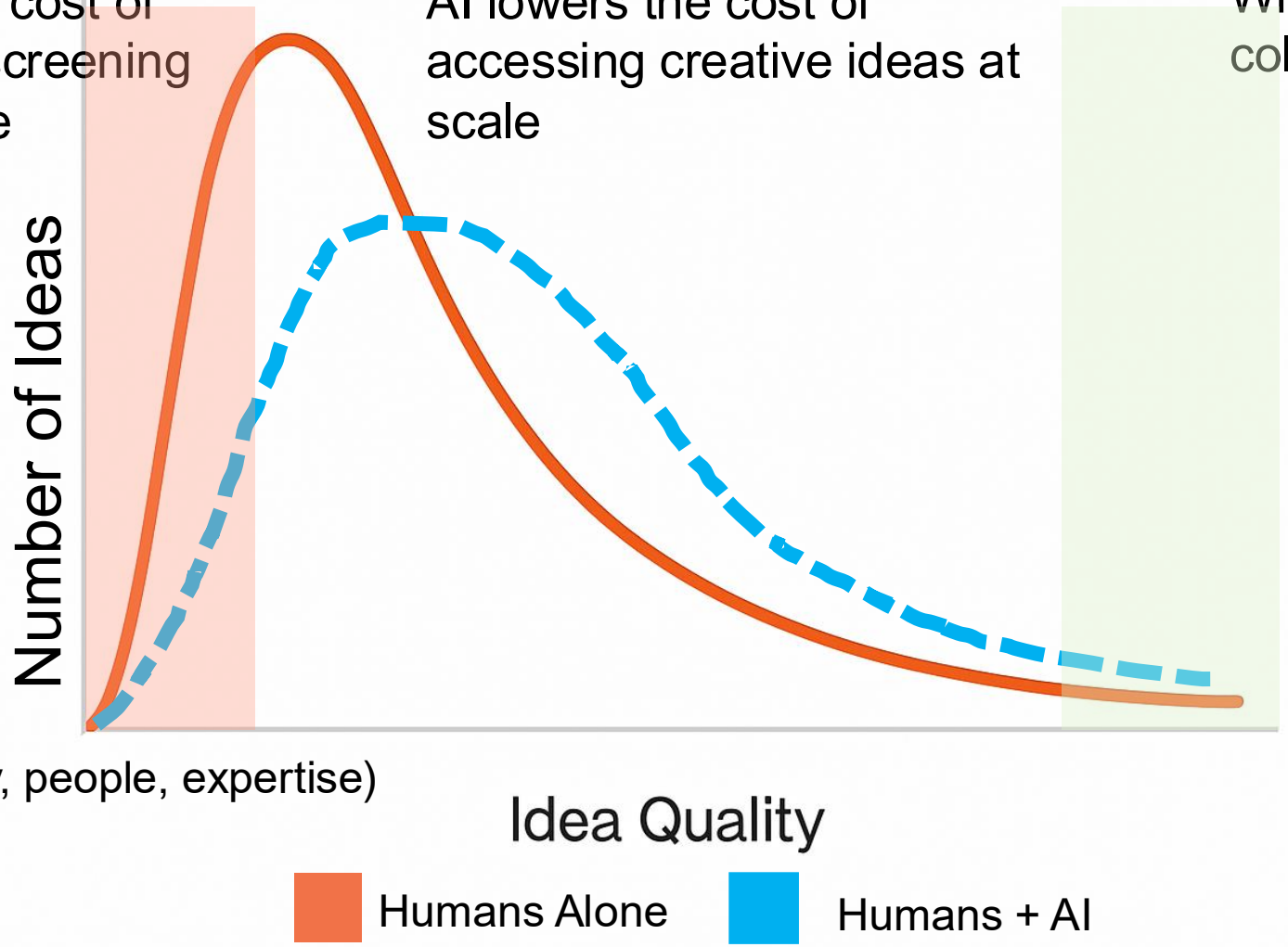
1. Idea Generation:

AI lowers the cost of accessing creative ideas at scale

3. Idea Evaluation:

What form of Human-AI collaboration?

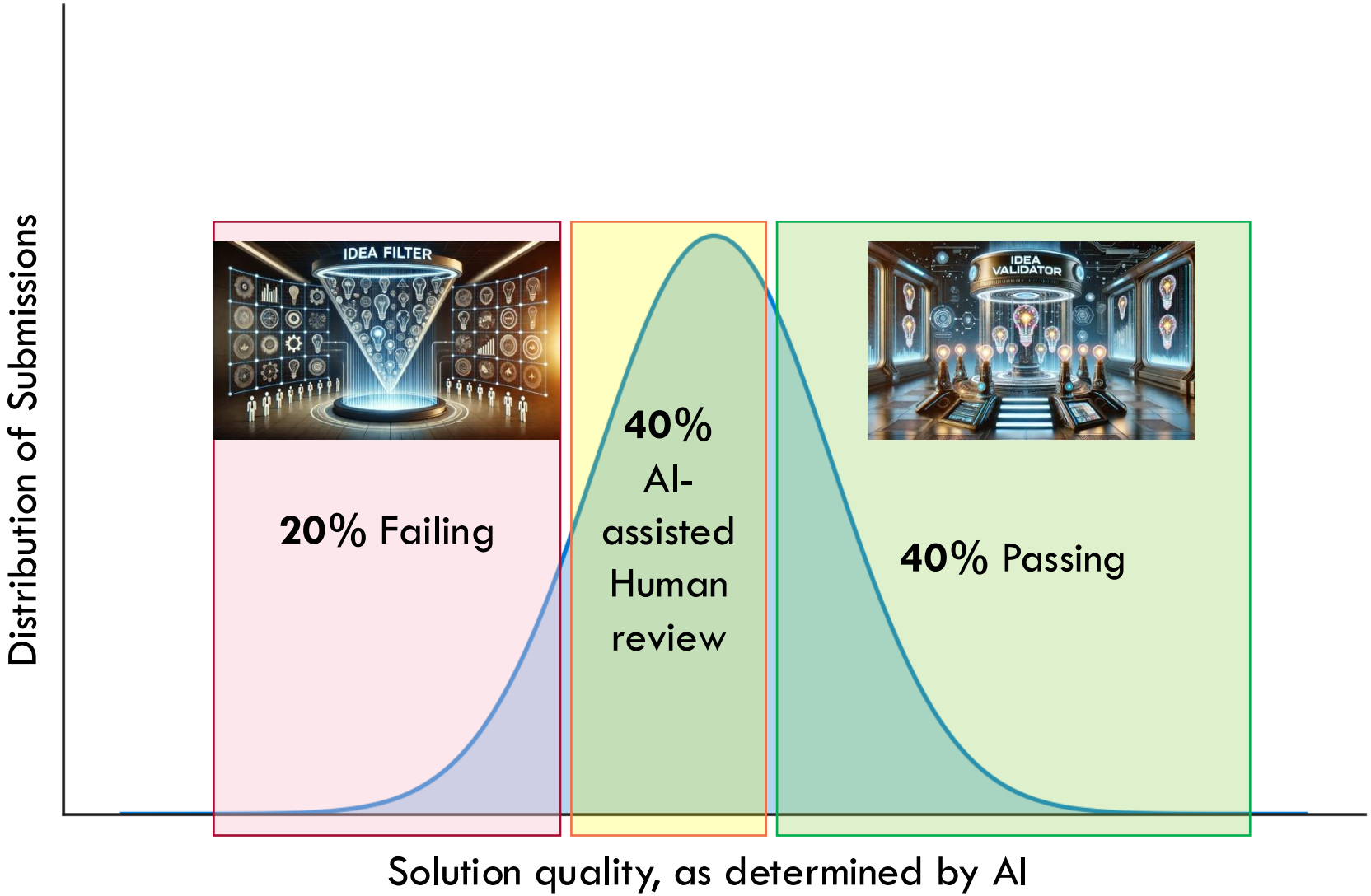
- Narrative AI explanations increase compliance but risk overreliance.
- Balance efficiency with human agency
- Design AI systems to augment, not replace, human oversight



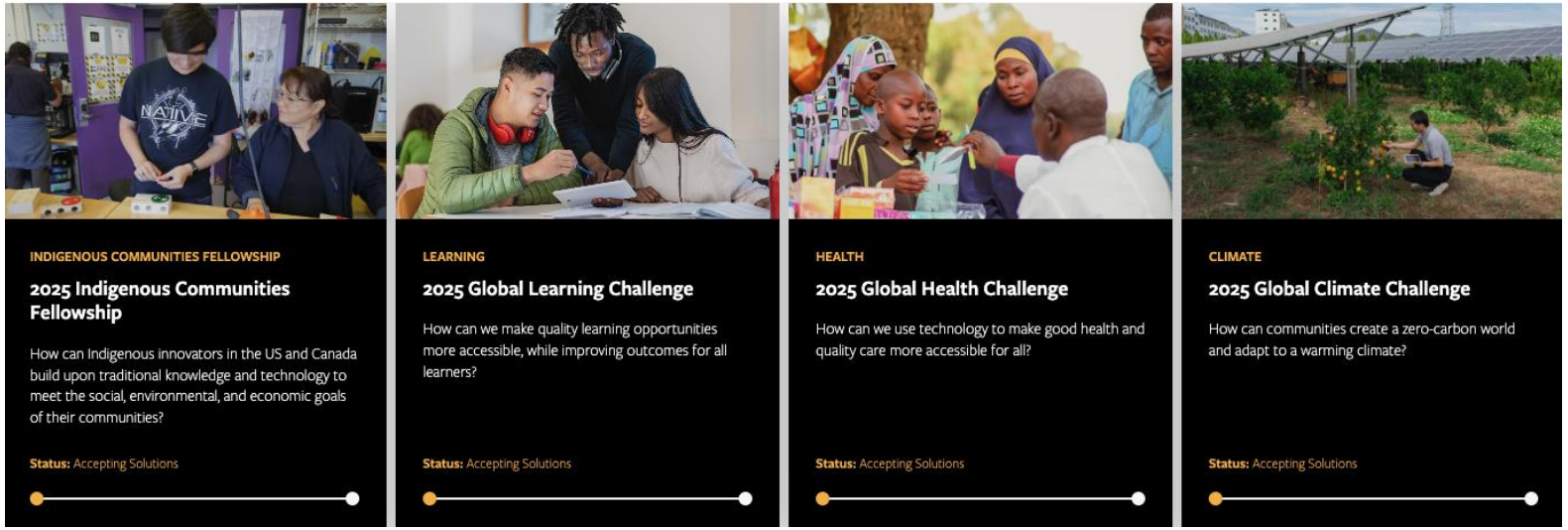
Cost = f(time, money, people, expertise)

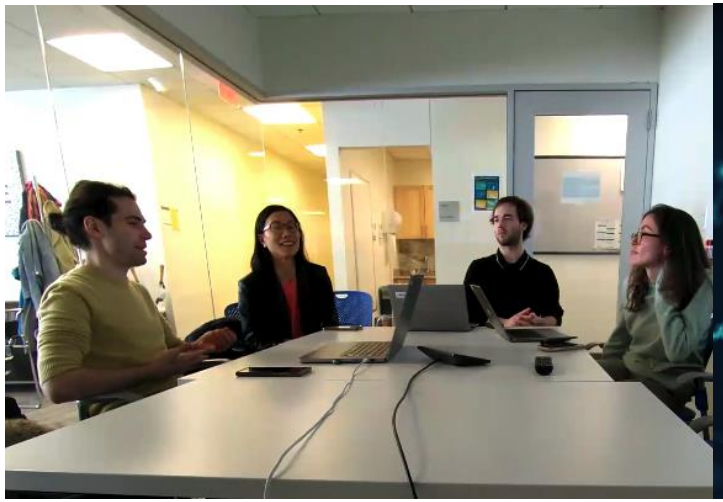
Bell et al. 2024, Boudreau et al. 2016, Lane et al. 2023, Park et al. 2024, Piezunka and Dahlander 2015

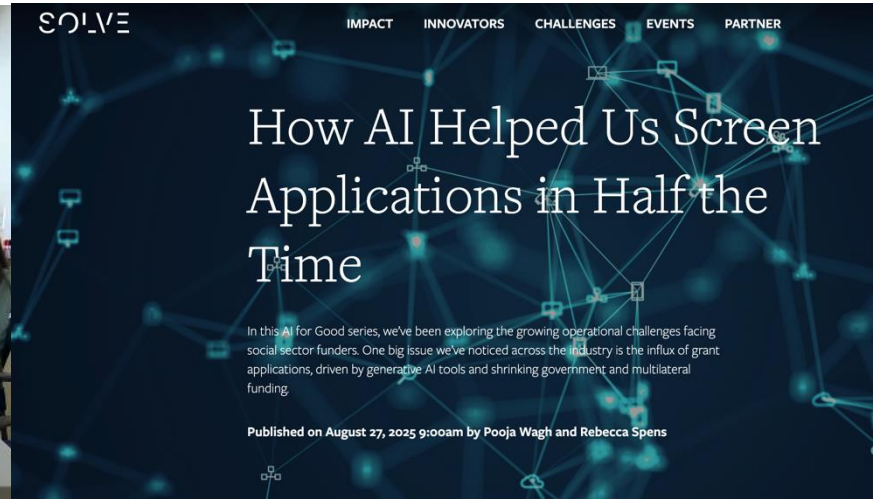
Spring 2025: we deployed our AI screener with MIT Solve's 2025 Global Challenges



Spring 2025: we deployed our AI screener with MIT Solve's 2025 Global Challenges







2901 solutions screened

14 business days saved

0 solution incorrectly failed
(on 100 reviewed for quality)


Let’s build the future of human-AI collaboration!

Léonard Boussioux
leobix@uw.edu
www.leobix.com

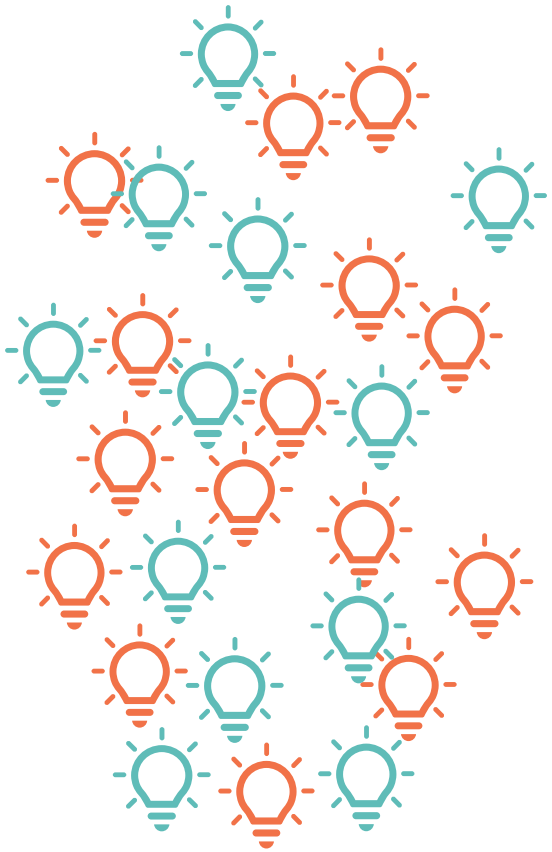
Organization Science

The Crowdless Future?
Gen AI & Idea Generation

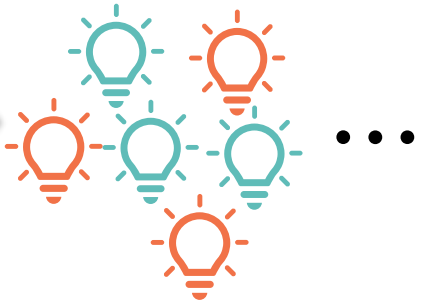
Preprint

The Narrative AI Advantage?
Gen AI & Idea Evaluation

Idea Generation



Screening, Evaluation



Harvard Business School | DIGITAL DATA DESIGN INSTITUTE AT HARVARD | FOSTER SCHOOL OF BUSINESS, UNIVERSITY of WASHINGTON | MIT SOLVE

Playbook for Human-LLM Decision-Making

Building on the transparency of traditional XAI, we must address a new risk in subjective domains: preventing explanations from becoming **persuasive narratives** that inadvertently silence the crucial human voice.

1. Prioritize Decision Quality Over User Compliance.

Don’t mistake agreement for accuracy. Measure what matters: the rate of costly errors.

2. Calibrate for Asymmetric Error Costs.

In domains like innovation, where false negatives can be the real cost, be wary of tools that make it easier to say no.

3. Force Critical Engagement.

Implement features that compel the user to verify the LLM's flaws. E.g., Deploy black-box LLM as a **Compass** to reduce noise and provide a consistent baseline. Avoid narrative LLM as a **Storyteller** that replaces nuanced human judgment and use the socratic method, or an LLM as a flagging tool rather than a deciding tool. Expertise doesn’t shield you from that.

4. Design for “Productive Disagreement.”

The function of an explanation shouldn’t be to persuade, but to highlight where human judgment is most needed. Ask: does this explanation help a user spot the LLM’s flaws?