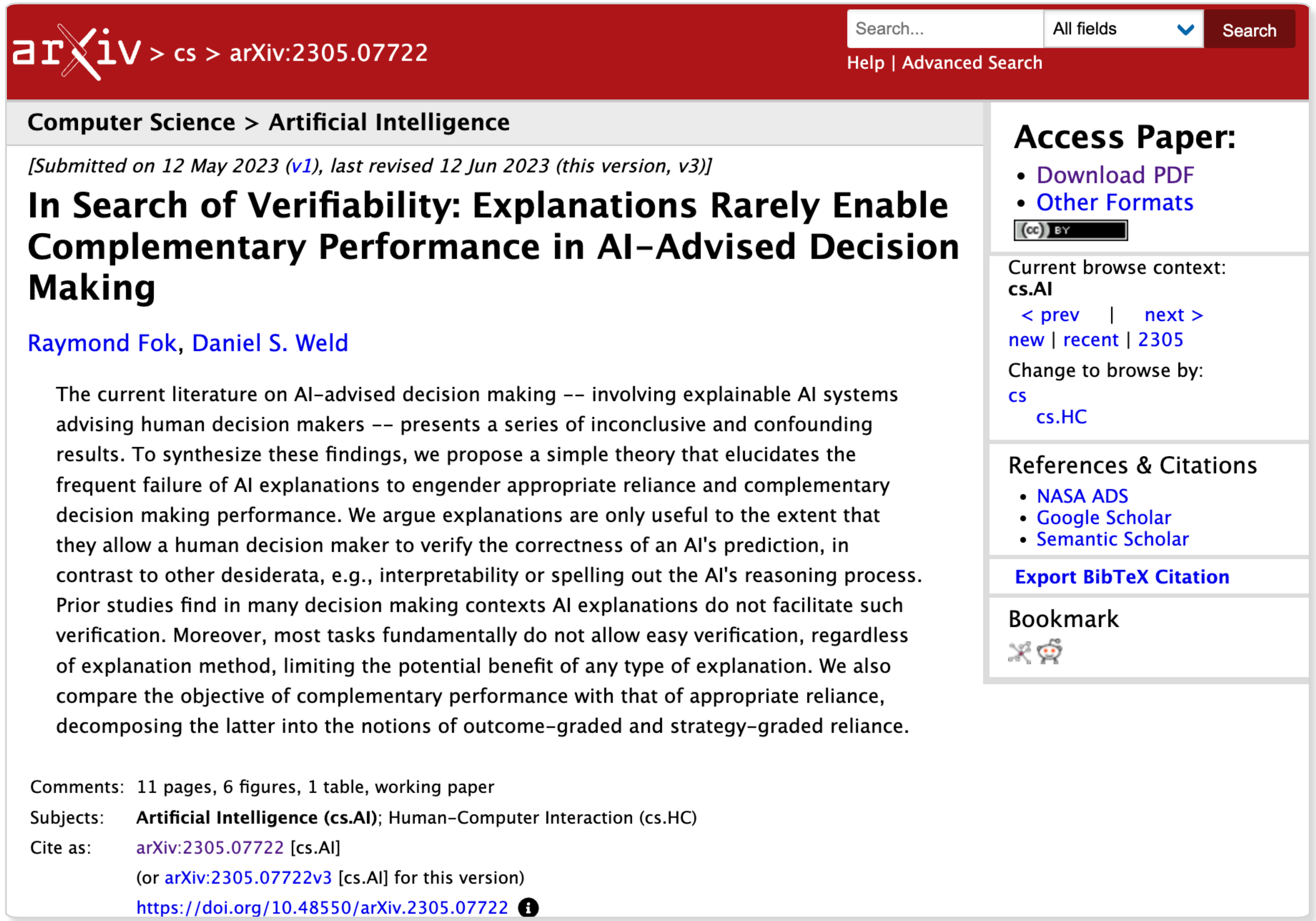
# Expandable Abstracts

*Bridging Scholarly Abstracts and Papers with Recursively Expandable Summaries*

**Raymond Fok** | November 21, 2023

# Abstracts



arXiv > cs > arXiv:2305.07722

Search... All fields Search

Help | Advanced Search

Computer Science > Artificial Intelligence

*[Submitted on 12 May 2023 (v1), last revised 12 Jun 2023 (this version, v3)]*

**In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making**

Raymond Fok, Daniel S. Weld

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

Comments: 11 pages, 6 figures, 1 table, working paper

Subjects: **Artificial Intelligence (cs.AI)**; Human-Computer Interaction (cs.HC)

Cite as: arXiv:2305.07722 [cs.AI]

(or arXiv:2305.07722v3 [cs.AI] for this version)

https://doi.org/10.48550/arXiv.2305.07722

**Access Paper:**

- Download PDF
- Other Formats

Current browse context:
**cs.AI**

< prev | next >

new | recent | 2305

Change to browse by:

cs

cs.HC

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

**Export BibTeX Citation**

Bookmark

~150 words

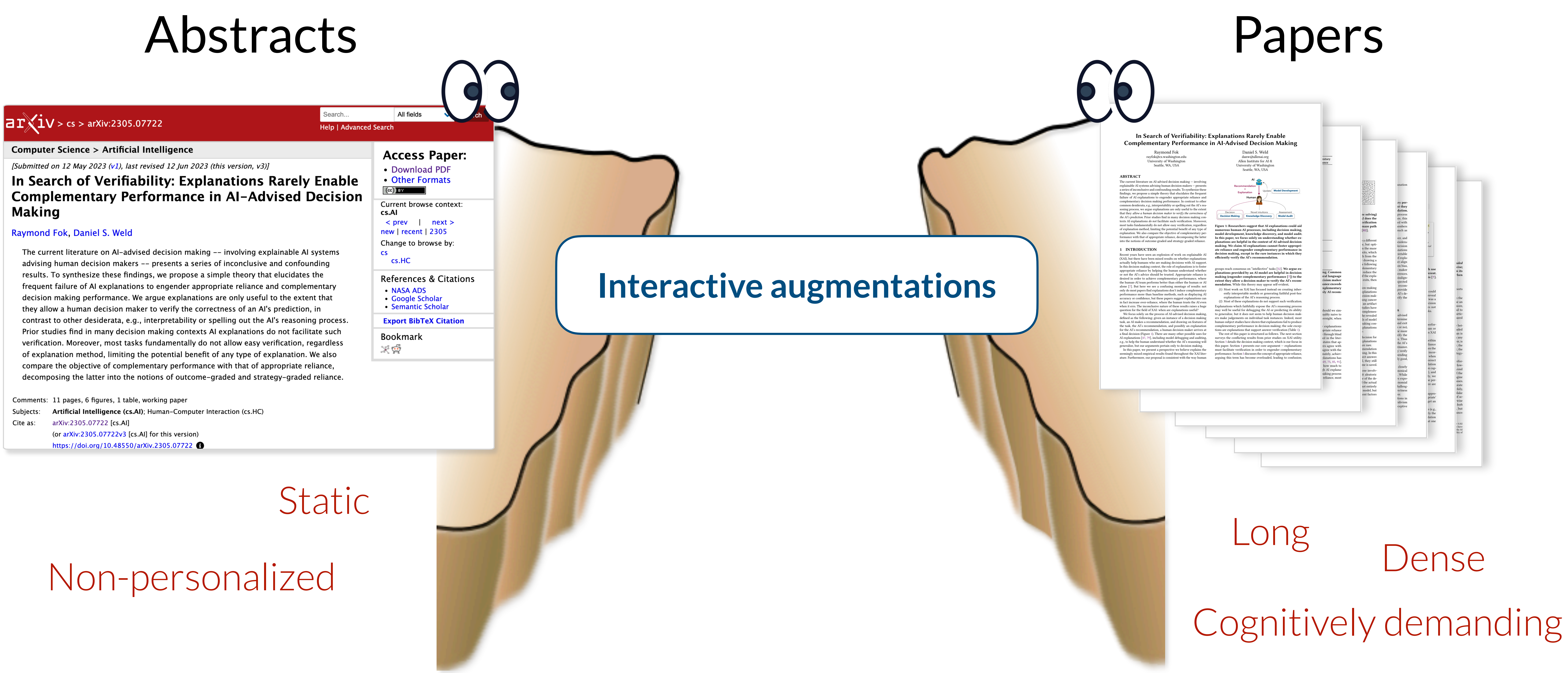
# Papers


~10,000 words

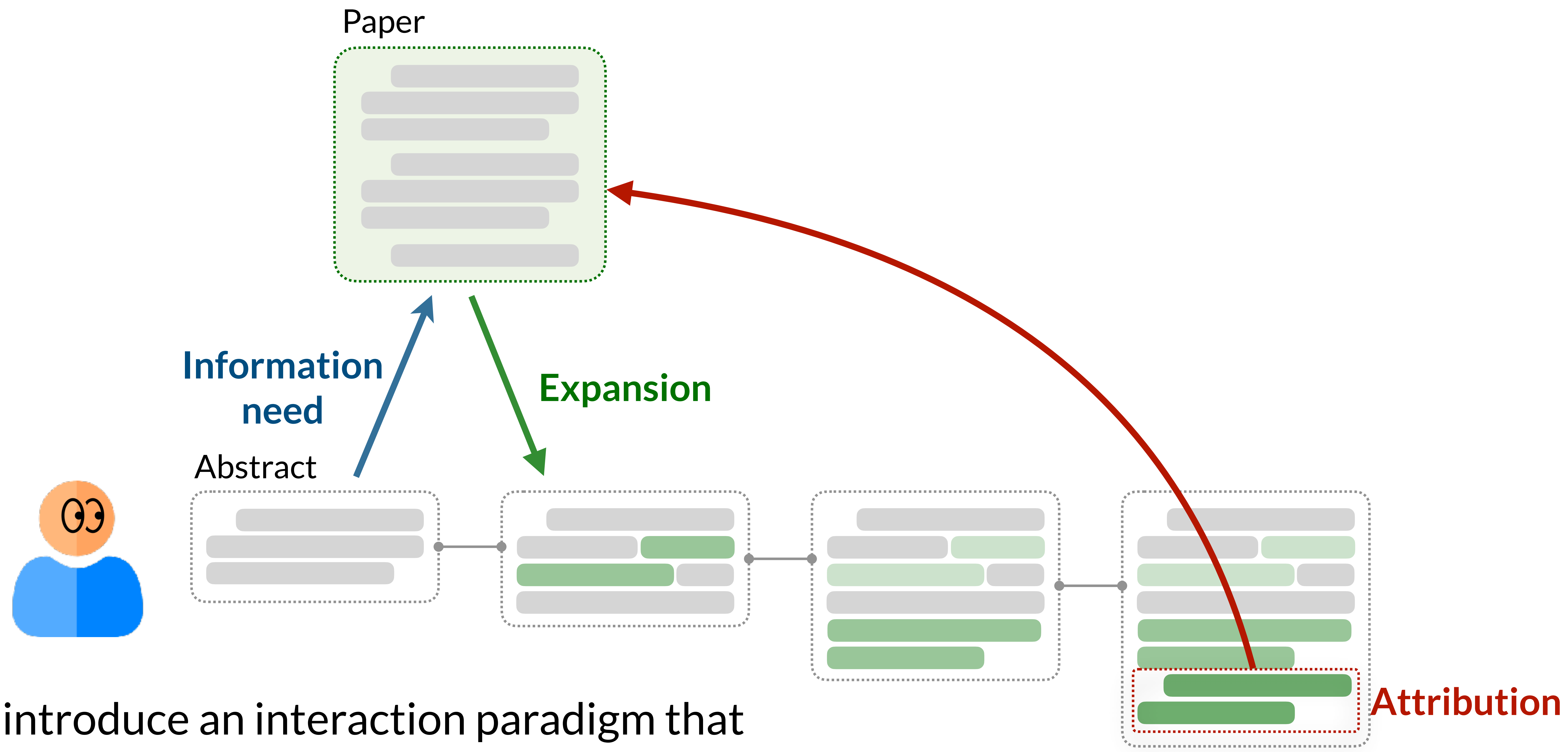
Abstracts | Papers


~150 words | ~10,000 words

# Bridging the information chasm



# Recursive summary expansion




We introduce an interaction paradigm that helps scholars to recursively expand abstracts with information from the full text of papers to address on-demand information needs.

# How can we design a useful expansion interaction?

**Determine what and how to expand**

**Present expansion**

**Provide information provenance**

| Dimension | Alternatives | | | | |
|---|---|---|---|---|---|
| Information needs type | Agnostic | Grounded | Latent | | |
| Information needs source | User-suggested | AI-suggested | Mixed-initiative | | |
| Information needs context | Same doc | Related docs | Open-domain | | |
| Expansion length | Short phrase | One sentence | Several sentences | | |
| Expansion placement | Fluid | Inline | Appended | Popup | Sidebar |
| Expansion delineation | Bold | Italicize | Colorize | Indent | Quote |
| Attribution method | Embedded | Separate | | | |
| Attribution granularity | Phrase | Sentence | Entire expansion | | |
| Attribution length | Phrase | Sentence | Paragraph | Page | |

# What information should an expansion entail?

A formative study with 7 scholars reading paper abstracts highlighted four common information needs, commonly expressed as clarification questions.

| | |
|---|---|
| Instantiation | *"What is an example of…"* |
| Definition | *"What does this mean?"* |
| Motivation | *"Why did they do this?"* |
| Expansion | *"How? Tell me more…"* |

# Selecting expandable entities | AI-initiated

**LLM suggests spans that could be expanded with additional information**

The current liter… …ision making -- involving explainable AI systems advisin… - presents a series of inconclusive and confoun… …ese findings, … theory th… elucidates the frequent failure of AI expl… appropriate reliance and complementary decision… argue explanations are only useful to the exte… that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in m… …ion making contexts AI explanations do not facilitate such ver…ication. …oreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

*How do they define appropriate reliance?*

*Tell me more about this..*

*Why?*

# Selecting expandable entities | User-initiated

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, li

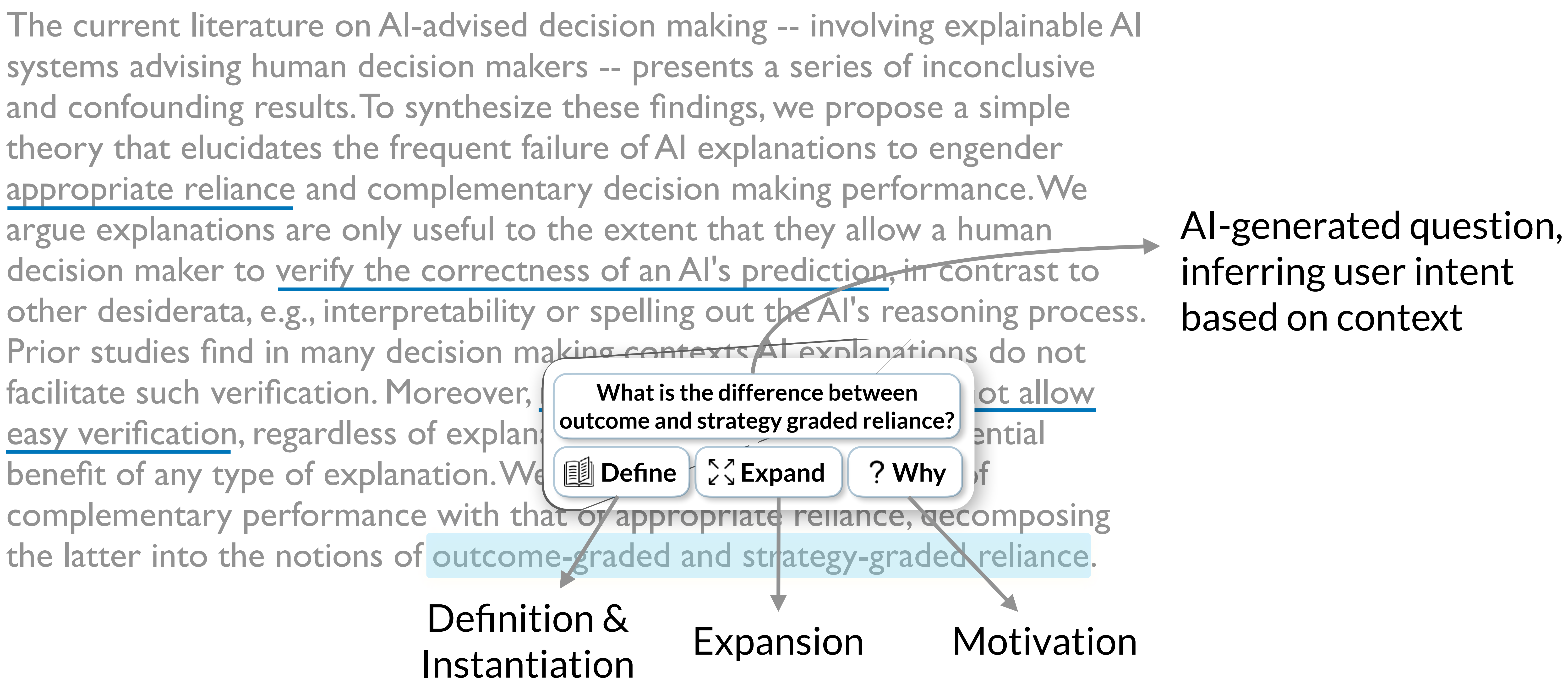
*What does this mean?*

benefit of any type of explanation. We also compa

complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

User selects any arbitrary span to expand

# Reducing expansion effort via one-click expansion actions



# In-situ expansions

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

Define "outcome-graded and strategy-graded reliance".

> Outcome-graded reliance defines a reliance behavior based on human acceptance of AI advice conditioned on the post-hoc correctness of the AI. Strategy-graded reliance defines a reliance behavior based on the relative expected performance of the human and the AI.

# In-situ expansions

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

Define "outcome-graded and strategy-graded reliance".

Outcome-graded reliance defines a reliance behavior based on human acceptance of AI advice conditioned on the post-hoc correctness of the AI. Strategy-graded reliance defines a reliance behavior based on the relative expected performance of the human and the AI.

# Recursive expansions

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complement… f appropriate reliance, decomposing the latter in… aded and strategy-graded reliance.

How is strategy-graded reliance calculated?

Define | Expand | ? Why

Define "outcome-graded and strategy-graded reliance".

Outc… reliance behavior based on human accep… ed on the post-hoc correctness of the AI. Strategy-graded reliance defines a reliance behavior based on the relative expected performance of the human and the AI.

# Recursive expansions

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

Define "outcome-graded and strategy-graded reliance".

Outcome-graded reliance defines a reliance behavior based on human acceptance of AI advice conditioned on the post-hoc correctness of the AI. Strategy-graded reliance defines a reliance behavior based on the relative expected performance of the human and the AI.

Tell me more about "strategy-graded reliance".

Strategy-graded reliance is appropriate if a decision maker accepts an AI recommendation when the AI is expected to outperform the human (conditioned on factors such as past performance, instance features, and AI advice), and rejects otherwise. The optimal strategy is to rely on the party most likely to have the correct answer.

The current literature on AI-advised decision making -- involving explainable AI systems advising human decision makers -- presents a series of inconclusive and confounding results. To synthesize these findings, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance and complementary decision making performance. We argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of an AI's prediction, in contrast to other desiderata, e.g., interpretability or spelling out the AI's reasoning process. Prior studies find in many decision making contexts AI explanations do not facilitate such verification. Moreover, most tasks fundamentally do not allow easy verification, regardless of explanation method, limiting the potential benefit of any type of explanation. We also compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

Define "outcome-graded and strategy-graded reliance".

> Outcome-graded reliance defines a reliance be
> acceptance of AI advice conditioned on the po
> the AI. Strategy-graded reliance defines a relia
> the relative expected performance of the hum

Tell me more about "strategy-graded reliance".

> Strategy-graded reliance is appropriate if a
> an AI recommendation when the AI is expe
> human (conditioned on factors such as pas
> features, and AI advice), and rejects otherw
> is to rely on the party most likely to have the correct answer. ”

*Excerpt from page 7* See in paper context

Instead, consider an alternative definition, strategy-graded reliance, where reliance is appropriate if the human accepts an AI recommendation when the AI is expected to outperform the human, and rejects otherwise (see Figure 5 right). Unlike outcome-graded reliance, strategy-graded reliance is neither post-hoc nor nondeterministic; it considers the appropriateness of reliance given the expected relative performance of the human and the AI. The optimal strategy is to rely on the party most likely to have the correct answer.

**Show evidence from the paper for this expansion**

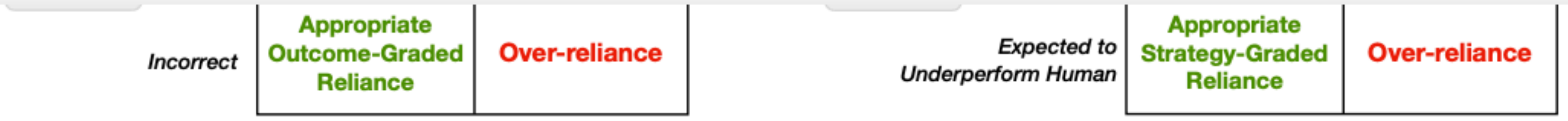



**Figure 5: We propose a clarification of two notions of reliance commonly conflated in the literature on AI-advised decision making. *Outcome-graded reliance* defines a reliance behavior based on human acceptance of AI advice conditioned on the post-hoc correctness of the AI. Specifically, outcome-graded reliance is appropriate if the human decision maker accepts an AI recommendation when it is correct and rejects otherwise. We argue this definition is problematic given its outcome-dependent and nondeterministic nature. In contrast, *strategy-graded reliance* defines a reliance behavior based on the relative expected performance of the human and the AI. Strategy-graded reliance is appropriate if a decision maker accepts an AI recommendation when the AI is expected to outperform the human (conditioned on factors such as past performance, instance features, and AI advice), and rejects otherwise.**

The definition says it is'appropriate' to trust the AI's advice on one case but not on the other — even though they are indistinguishable!

Instead, consider an alternative definition, *strategy-graded reliance*, where reliance is appropriate if the human accepts an AI recommendation when the AI is *expected* to outperform the human, and rejects otherwise (see Figure 5 right). Unlike outcome-graded reliance, strategy-graded reliance is neither post-hoc nor nondeterministic; it considers the appropriateness of reliance given the expected relative performance of the human and the AI. The optimal strategy is to rely on the party *most likely* to have the correct answer. A key question here is "Upon what information is that expectation computed?" There are several possibilities.

- **Past performance:** If past experience shows the AI is more likely to be correct than the human, it might be appropriate to defer to the AI even without information about this particular decision instance. Note this policy cannot produce complementary performance.
- Previous characteristics + **instance features:** Conditioning on the current instance (i.e., specific details of the task at hand) can lead to complementary performance. For example, if a driver knew her auto-drive car was less prone to accidents when on the freeway, she might confidently take her hands off the wheel in that situation — even if she knew that she was the better driver on winding country roads. When automated, this type of conditioning resembles the human-AI delegation workflow discussed at the end of Section 6.2.
- Previous characteristics + **the AI's recommendation:** Conditioning on the AI's recommendation allows the human to adopt a policy of the form "I know the AI is conservative and very unlikely to err with a false positive, so I will accept positive recommendations and only scrutinize instances when the AI offers a negative recommendation."
- Previous characteristics + **the AI's explanation:** In this paper, we have argued this condition rarely improves upon the previous strategy, and only when the explanation supports verification.

In contrast to complementary performance, which refers to the *team's* measured performance, both notions of reliance define an attribute of the *human's behavior* relative to the AI. We believe the strategy-graded definition of reliance is the better objective. To illustrate the shortcomings of outcome-graded reliance, consider a decision making task in which the human is historically 60% accurate, while the AI is 99.999% accurate. On any given instance of the task, if the human is uncertain of the answer, is it appropriate to rely on the AI's recommendation? Intuitively, the answer seems a clear 'yes'. But if the human later discovers the AI was wrong, the outcome-graded definition says "Inappropriate," while the strategy-graded definition matches intuition and says "Appropriate."

Outcome-graded reliance is similar to complementary performance in the sense that both qualities can only be measured post hoc. However, there are subtle differences between these notions, beyond the fact that one measures a pattern of human behavior and the other the performance of a human-AI team. To elaborate, consider a three-way classification problem, where widgets must be graded A, B, or C. Suppose Clare and Dave are both 80% accurate at the task while the AI is only 10% accurate. Luckily, the AI outputs verifiable explanations, so both Clare and Dave can perfectly tell when the AI is correct. Suppose Clare follows the policy of accepting the AI's recommendation when it is correct, and otherwise choosing randomly. Dave also accepts the AI's recommendation when it is correct, but solves the problem himself when it is not. According to the definition, *both Clare and Dave have perfect outcome-graded reliance*, but their strategies lead to very different expected team performance: 55% for Clare and 82% for Dave.

Given the limitations of the outcome-graded definition of appropriate reliance, we suggest researchers focus on the strategy-graded, or eschew the term 'appropriate reliance' altogether. We argue overall performance is a better objective when evaluating a

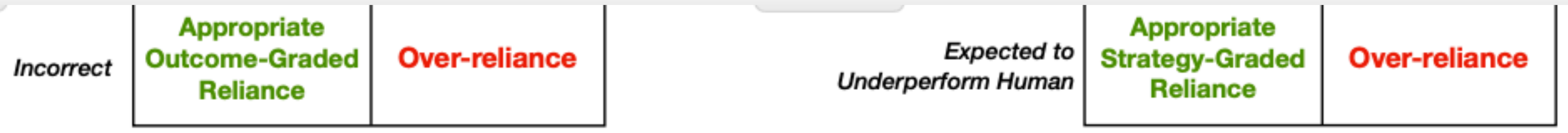



**Figure 5: We propose a clarification of two notions of reliance commonly conflated in the literature on AI-advised decision making. *Outcome-graded reliance* defines a reliance behavior based on human acceptance of AI advice conditioned on the post-hoc correctness of the AI. Specifically, outcome-graded reliance is appropriate if the human decision maker accepts an AI recommendation when it is correct and rejects otherwise. We argue this definition is problematic given its outcome-dependent and nondeterministic nature. In contrast, *strategy-graded reliance* defines a reliance behavior based on the relative expected performance of the human and the AI. Strategy-graded reliance is appropriate if a decision maker accepts an AI recommendation when the AI is expected to outperform the human (conditioned on factors such as past performance, instance features, and AI advice), and rejects otherwise.**

The definition says it is'appropriate' to trust the AI's advice on one case but not on the other — even though they are indistinguishable!

Instead, consider an alternative definition, *strategy-graded reliance*, where reliance is appropriate if the human accepts an AI recommendation when the AI is *expected* to outperform the human, and rejects otherwise (see Figure 5 right). Unlike outcome-graded reliance, strategy-graded reliance is neither post-hoc nor nondeterministic; it considers the appropriateness of reliance given the expected relative performance of the human and the AI. The optimal strategy is to rely on the party *most likely* to have the correct answer. A key question here is "Upon what information is that expectation computed?" There are several possibilities.

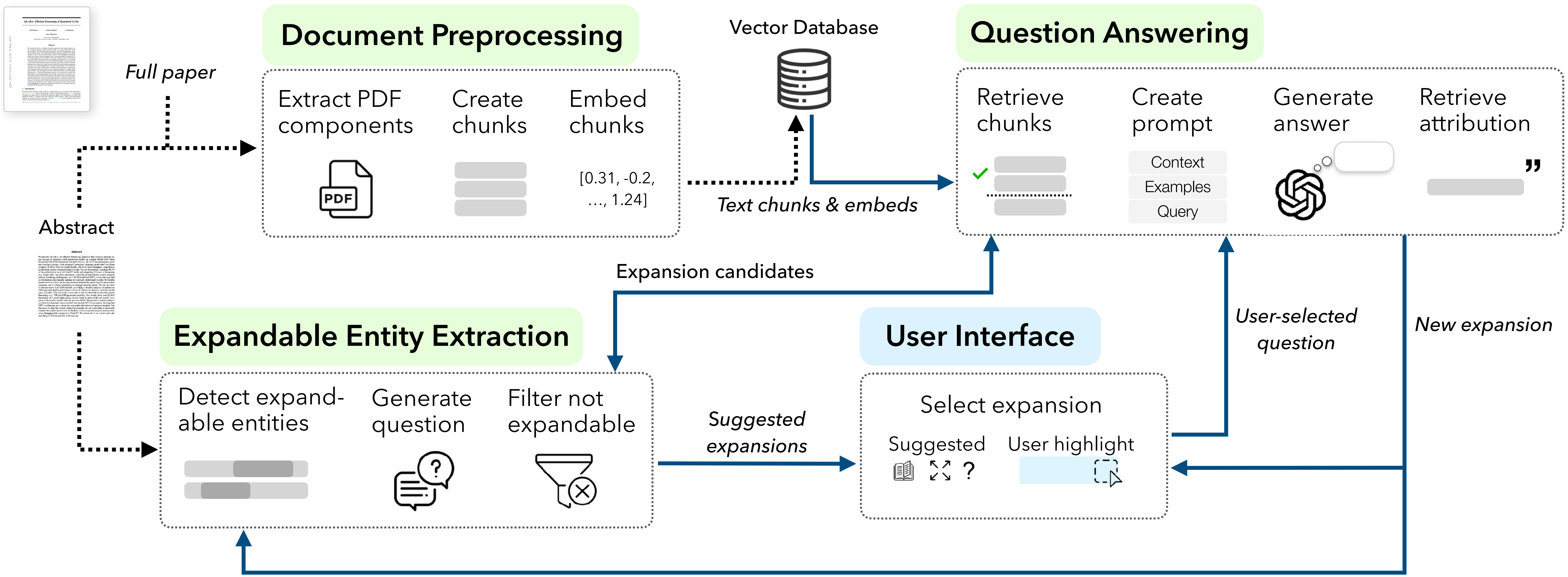
- **Past performance:** If past experience shows the AI is more likely to be correct than the human, it might be appropriate to defer to the AI even without information about this particular decision instance. Note this policy cannot produce complementary performance.
- Previous characteristics + **instance features:** Conditioning on the current instance (i.e., specific details of the task at hand) can lead to complementary performance. For example, if a driver knew her auto-drive car was less prone to accidents when on the freeway, she might confidently take her hands off the wheel in that situation — even if she knew that she was the better driver on winding country roads. When automated, this type of conditioning resembles the human-AI delegation workflow discussed at the end of Section 6.2.
- Previous characteristics + **the AI's recommendation:** Conditioning on the AI's recommendation allows the human to adopt a policy of the form "I know the AI is conservative and very unlikely to err with a false positive, so I will accept positive recommendations and only scrutinize instances when the AI offers a negative recommendation."
- Previous characteristics + **the AI's explanation:** In this paper, we have argued this condition rarely improves upon the previous strategy, and only when the explanation supports verification.

In contrast to complementary performance, which refers to the *team's* measured performance, both notions of reliance define an attribute of the *human's behavior* relative to the AI. We believe the strategy-graded definition of reliance is the better objective. To illustrate the shortcomings of outcome-graded reliance, consider a decision making task in which the human is historically 60% accurate, while the AI is 99.999% accurate. On any given instance of the task, if the human is uncertain of the answer, is it appropriate to rely on the AI's recommendation? Intuitively, the answer seems a clear 'yes'. But if the human later discovers the AI was wrong, the outcome-graded definition says "Inappropriate," while the strategy-graded definition matches intuition and says "Appropriate."

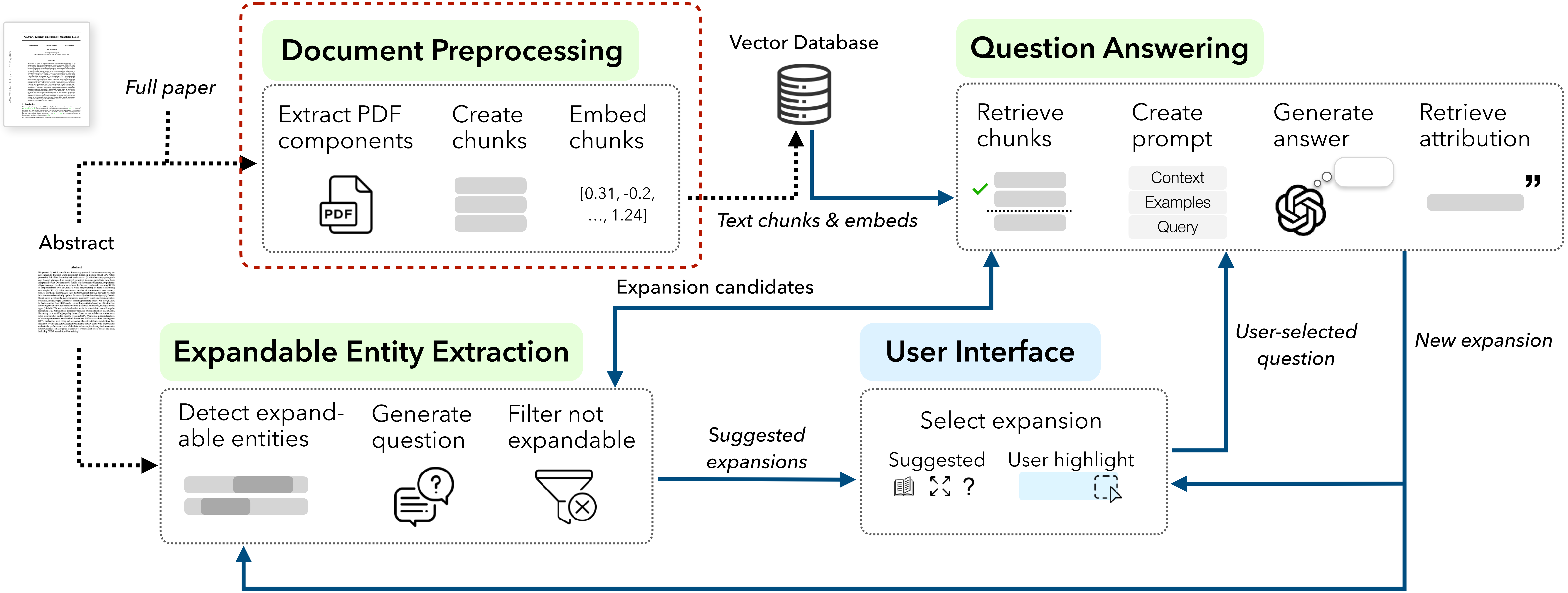
Outcome-graded reliance is similar to complementary performance in the sense that both qualities can only be measured post hoc. However, there are subtle differences between these notions, beyond the fact that one measures a pattern of human behavior and the other the performance of a human-AI team. To elaborate, consider a three-way classification problem, where widgets must be graded A, B, or C. Suppose Clare and Dave are both 80% accurate at the task while the AI is only 10% accurate. Luckily, the AI outputs verifiable explanations, so both Clare and Dave can perfectly tell when the AI is correct. Suppose Clare follows the policy of accepting the AI's recommendation when it is correct, and otherwise choosing randomly. Dave also accepts the AI's recommendation when it is correct, but solves the problem himself when it is not. According to the definition, *both Clare and Dave have perfect outcome-graded reliance*, but their strategies lead to very different expected team performance: 55% for Clare and 82% for Dave.

Given the limitations of the outcome-graded definition of appropriate reliance, we suggest researchers focus on the strategy-graded, or eschew the term 'appropriate reliance' altogether. We argue overall performance is a better objective when evaluating a

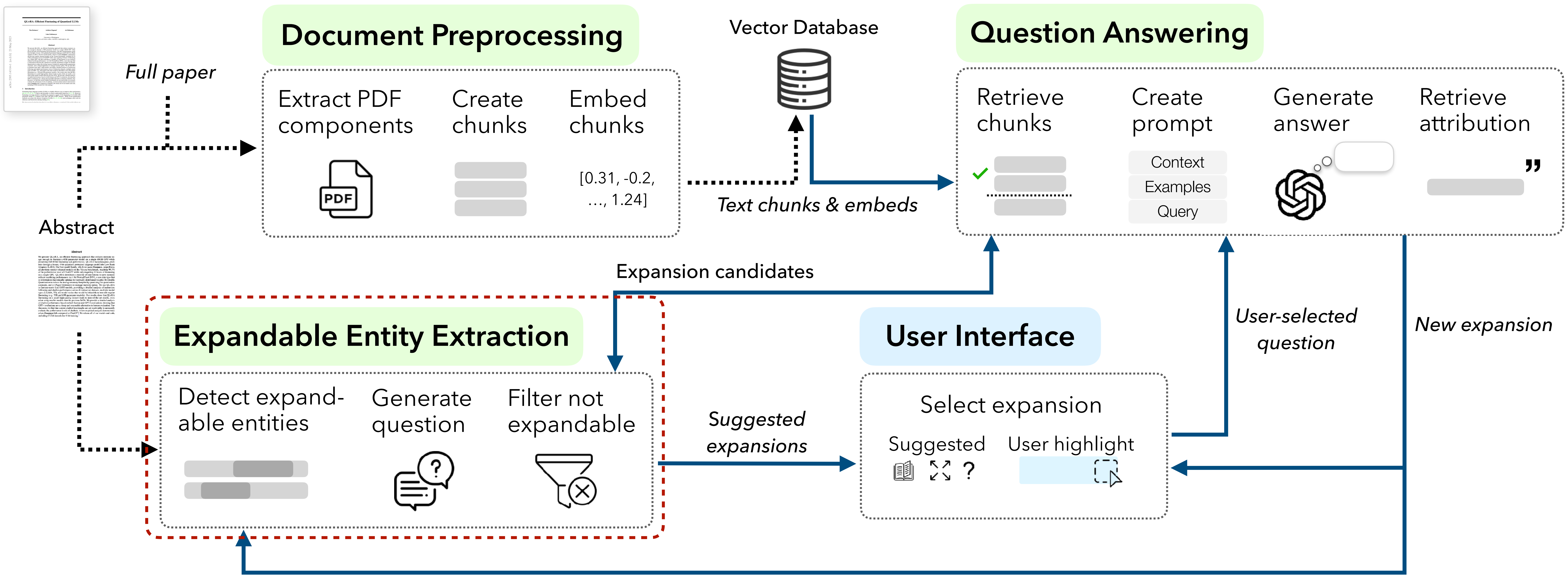
# System architecture



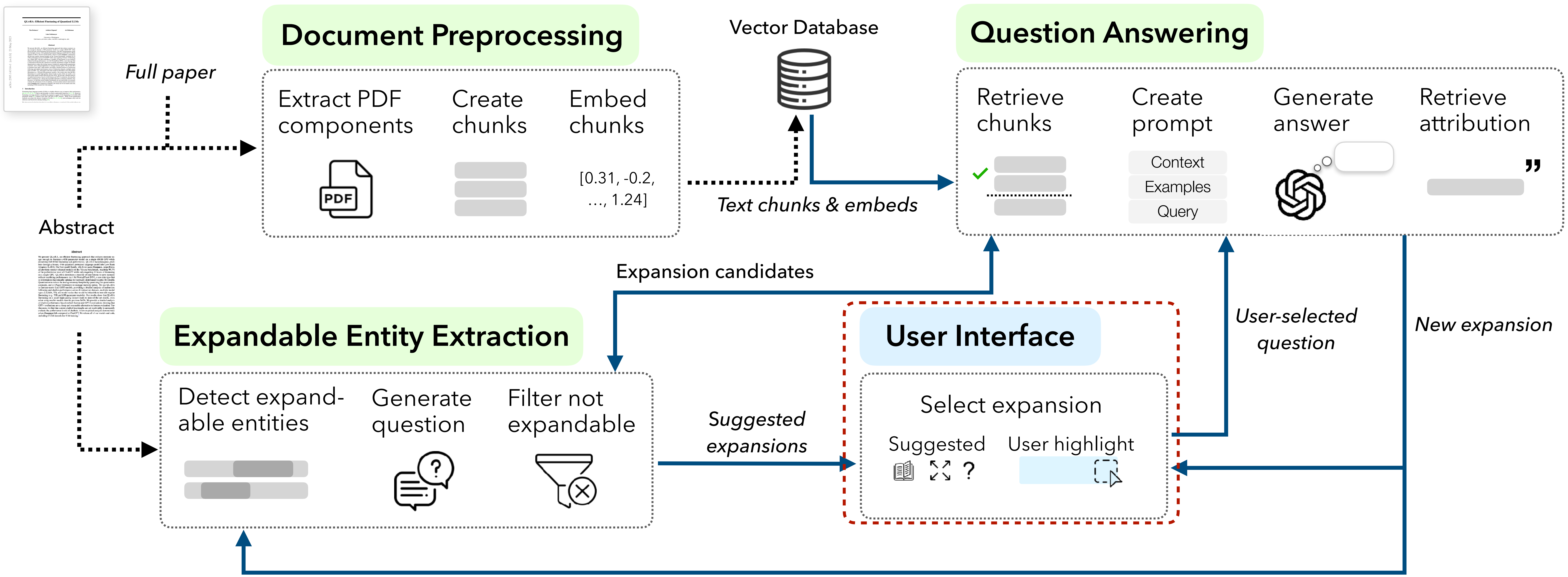
# System architecture



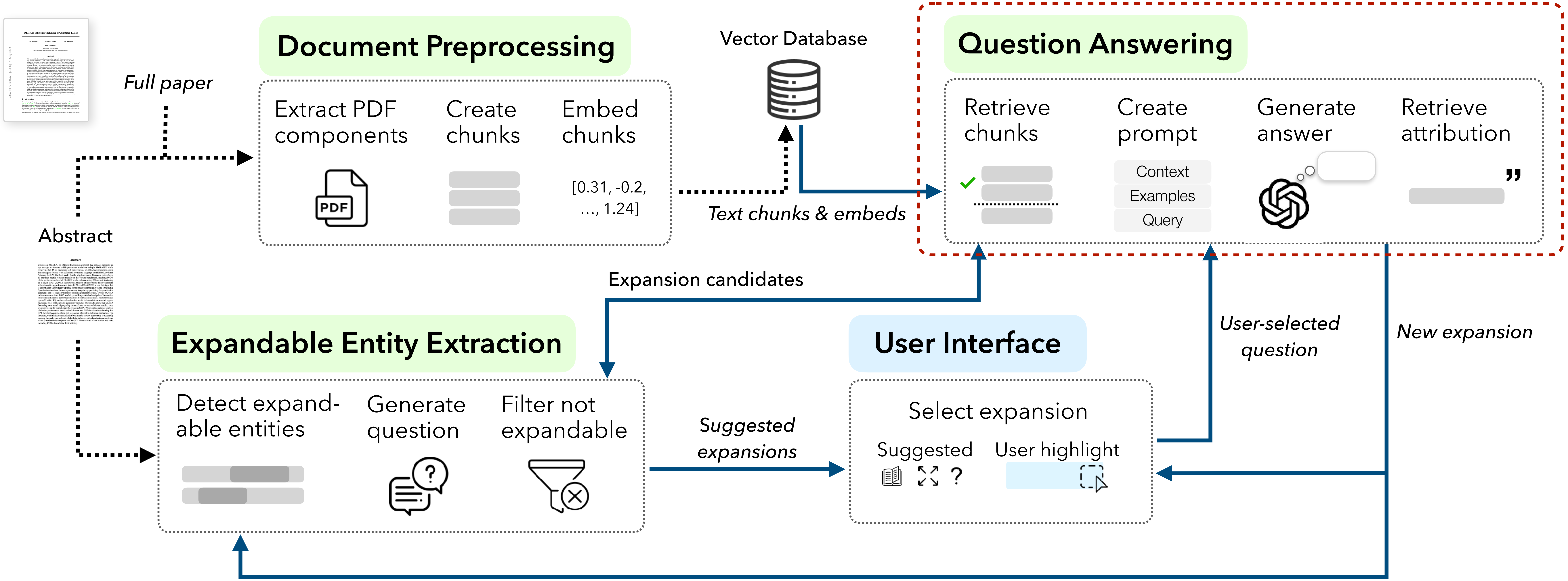
# System architecture



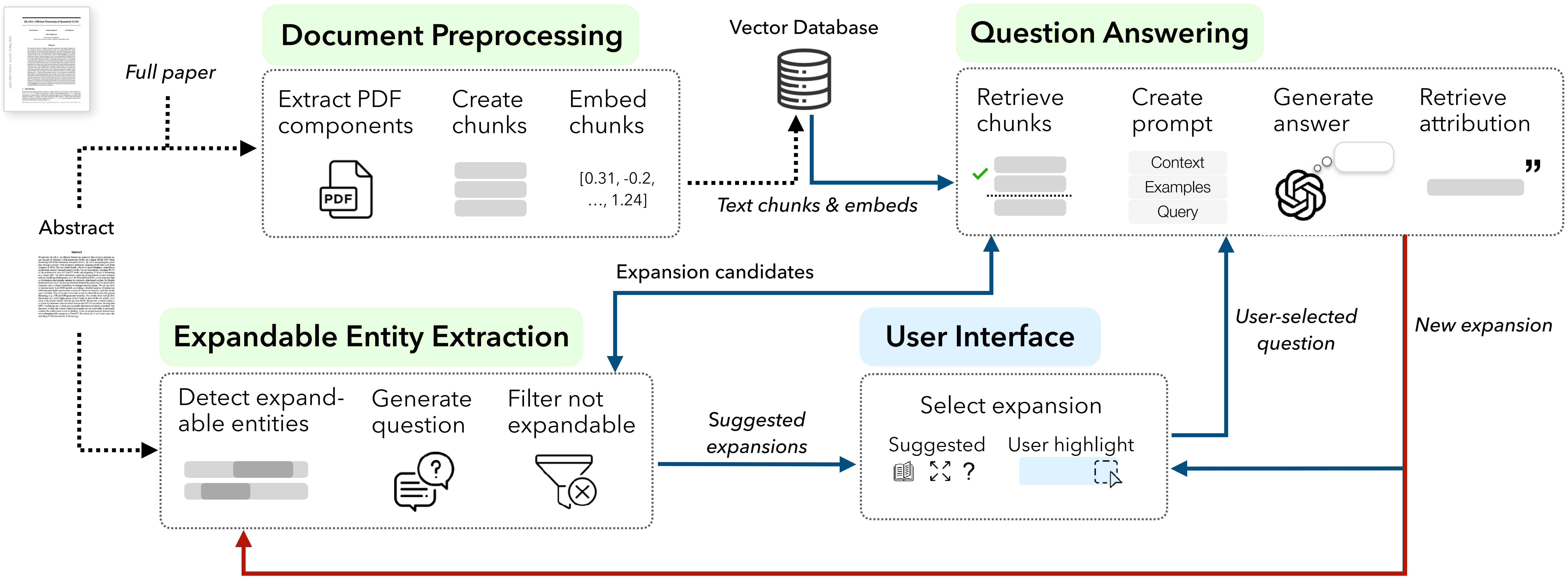
# System architecture



# System architecture



# System architecture



# Evaluation

Qualitative interview study with 9 scholars

Deployment study at a conference (n=275)

VLDB2023 vancouver

# Interview study

What are the benefits and disadvantages of using expandable abstracts to triage scientific papers?

3 - 5 seed papers relevant to their current research interests

↓

20 - 25 other papers via S2 recommendations API

↓

Exploration (~30min) over the list of abstracts

## Findings | Overall utility

**Participants liked how the expansions allowed them to surface details from the paper using simple interactions with the abstracts rather than manually searching for them over the full papers.**

Abstracts have a common structure and served as a jumping-off point to pull in information from different sections when needed.

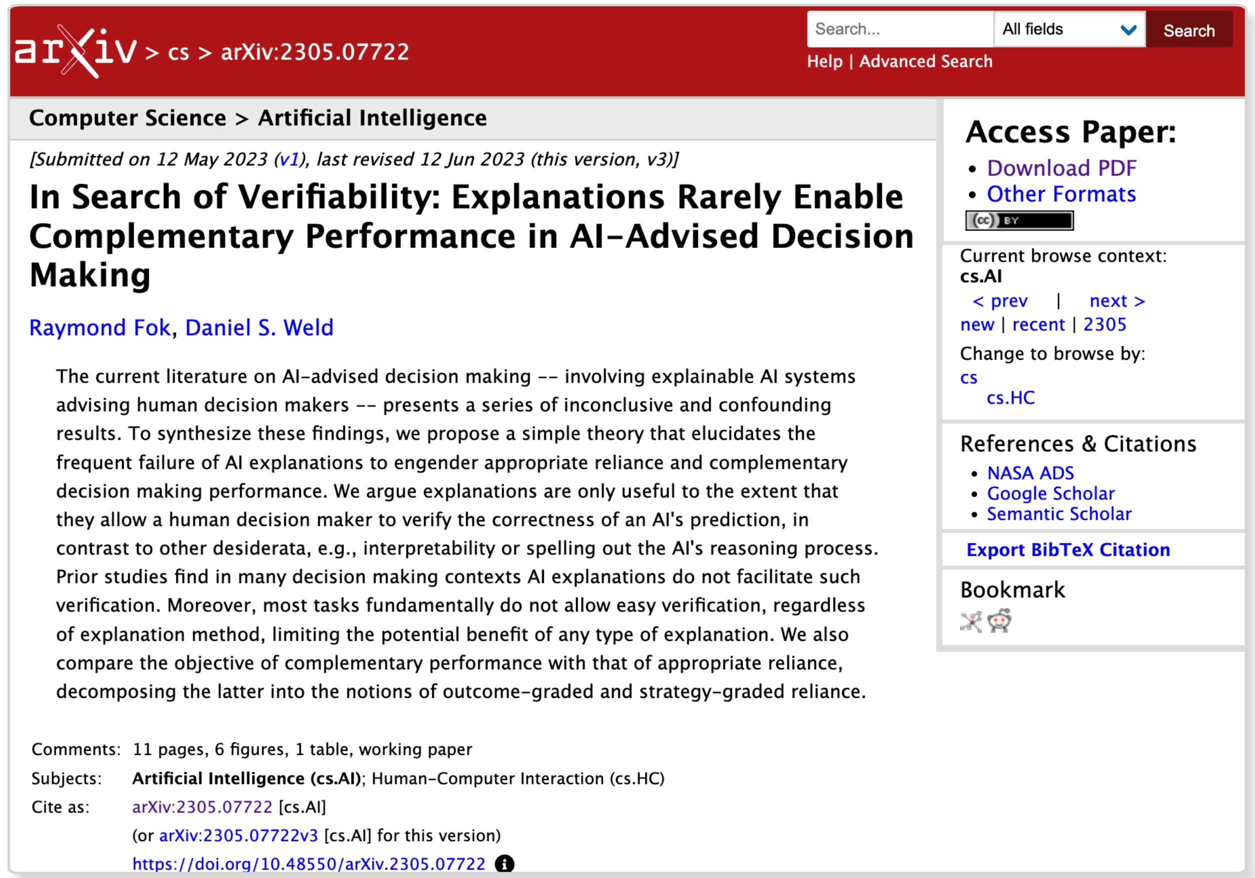
**Current LLMs achieve more than passable performance on both 1) answering user's questions and 2) inferring a user's information need based on context.**

Participants were surprised at the quality of the generated expansions. Everything "looked" factual, and seemed to "extract meaning" from the paper rather than just summarize.

Participants mentioned how the AI-generated questions "seemed to almost read my mind when I click on something or highlight something."

# Findings | Multiple layers of affordances

**The concert of mixed-initiative interactions satisfied the majority of user's information needs while reading an abstract.**




AI-suggested expandable entities

User manually highlighted entities

Static expansion actions (Define, Expand, Why)

AI-generated question

Use attribution to jump into the paper

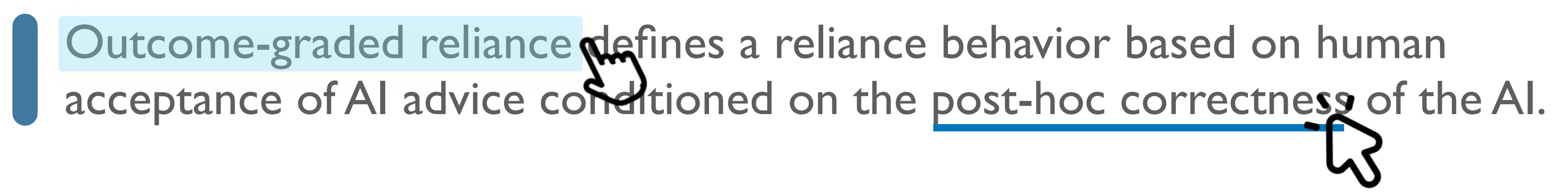
Free-form QA

# Findings

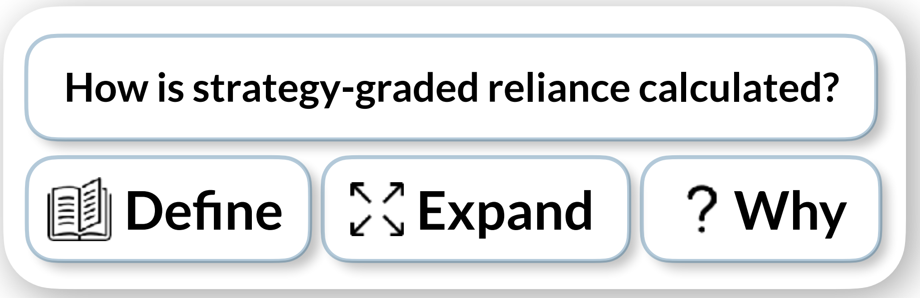
## How did participants use each of the features?

### Participants

…more often selected an AI-suggested expandable entity (77.5%) rather than highlighting their own (22.5%)




…selected the AI-suggested question 40% of the time, and the static questions: Define: 23%, Expand: 23%, Why: 14%




…created threaded expansions 58% of the time, suggesting the recursive expansions prompted users to ask followup questions

# Deployment study

How do scholars use expandable abstracts in the wild?

To characterize real-world usage, we created expandable abstracts for the proceedings of VLDB 2023.

VLDB2023
vancouver

275 unique users interacted with the abstracts over the two week deployment period.

# Findings

## How did scholars use each of the features?

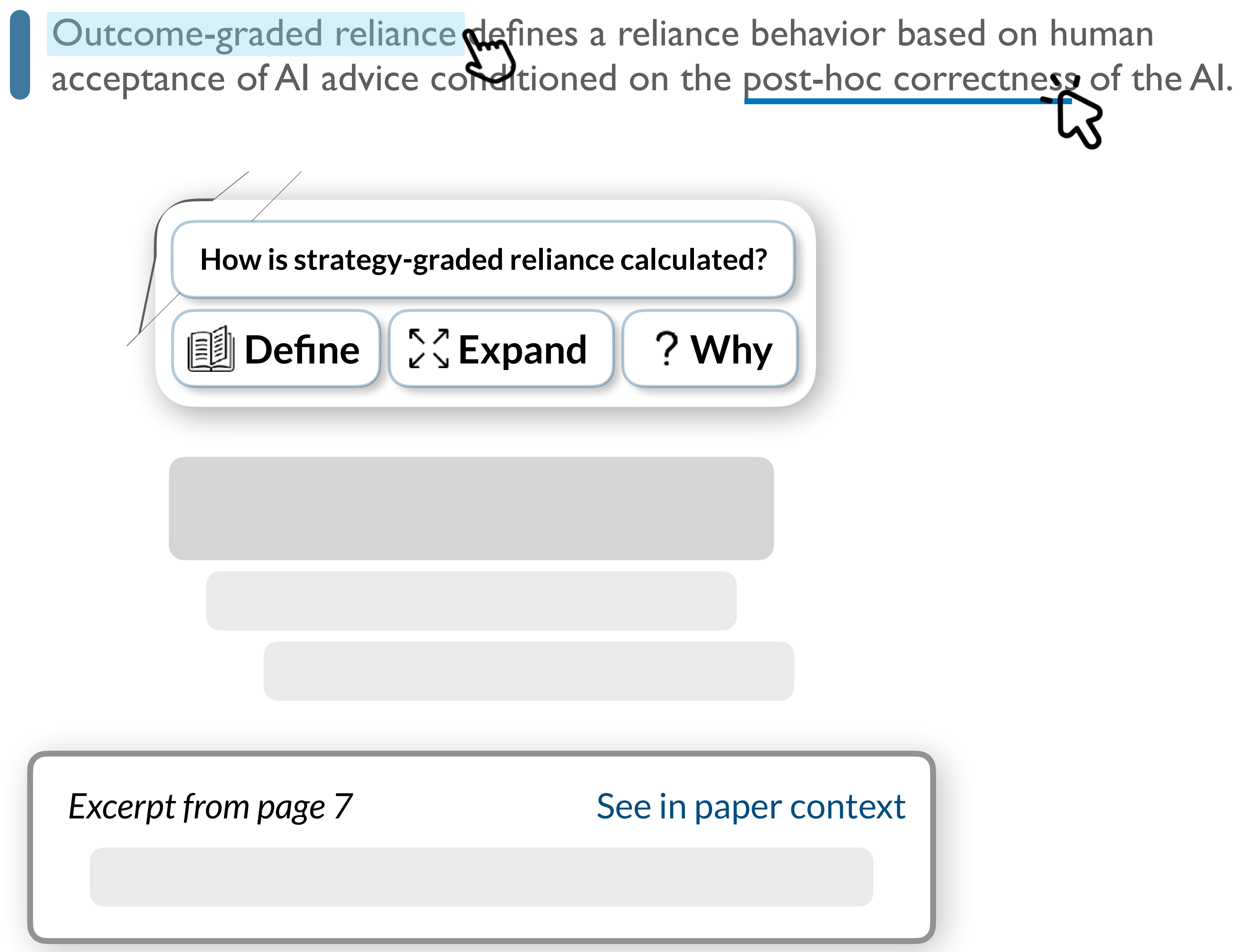
### Scholars

…more often selected an AI-suggested expandable entity (80%) rather than highlighting their own (20%)

…selected the AI-suggested question 12% of the time, and the static questions: Define: 31%, Expand: 42%, Why: 15%

…created threaded expansions 28% of the time, suggesting the recursive expansions prompted users to ask followup questions

…viewed the attributed evidence 15% of the time, and jumped into the paper PDF 40% of the time after viewing the evidence




Different usage patterns likely due to different levels of engagement.

# Risks of expandable abstracts

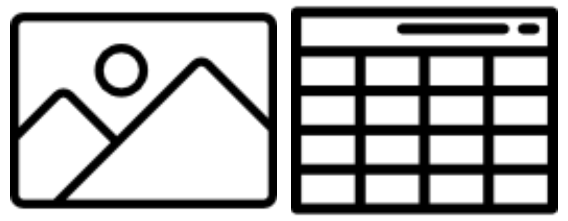
Augmented intelligence for scholars may harm pedagogical and self-learning processes, especially for novice scholars.

> "In the research realm, I don't think people should be reading just the abstract. With this system, I don't think it's that great to just replace the paper reading experience with just expanding the abstract all the time and trying to get details instead of actually reading the paper. You can use this as a map for going to the sections of the paper you want to read, and that's fine. But I don't know … if this somehow promotes this culture within research that all we need to read is the abstract and I don't think that would be very great either." — *P3*

LLM hallucination remains a problem, and verification of generated expansion accuracy can often be challenging or undesirable.

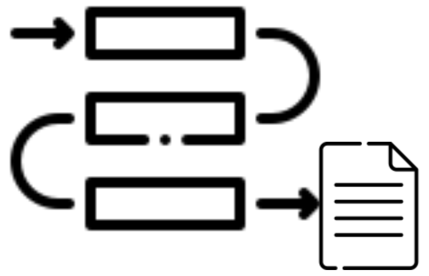
> "When I try to paper in writing my related work, paper writing, the evidence button plays a huge, different role. Because I need to see whether the responses they are generating are correct or really match with the paper. But during the time of abstract exploring, I'm not too caring about the evidence, where they come from in the paper." — *P1*

# Improving expandable abstracts



Incorporate visual media (e.g., figures and tables) into expansions when relevant.

Expand with content from other papers (e.g., expand to show other papers that use similar terms, or show other papers building on this paper).



Use the expansions to help scholars smoothly transition into reading the full paper.

# Expandable Abstracts

*Bridging Scholarly Abstracts and Papers with Recursively Expandable Summaries*






**Raymond Fok** Joseph Chee Chang Tal August

Amy X. Zhang Daniel Weld

@rayrayfok

rayfok.github.io

rayfok@cs.washington.edu