

# 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

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

Raymond Fok  
rayfok@cs.washington.edu  
University of Washington  
Seattle, WA, USA

Daniel S. Weld  
dsw@allenai.org  
Allen Institute for AI &  
University of Washington  
Seattle, WA, USA

**Figure 1: Researchers suggest that AI explanations could aid numerous human-AI processes, including decision making, model development, knowledge discovery, and model audit. In this paper, we focus solely on understanding whether explanations are helpful in the context of AI-advised decision making. We claim AI explanations cannot foster appropriate reliance and engender complementary performance in decision making, except in the rare instances in which they efficiently verify the AI's recommendation.**

**ABSTRACT**

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. In contrast to other common desiderata, e.g., interpretability or spelling out the AI's reasoning process, we argue explanations are only useful to the extent that they allow a human decision maker to verify the correctness of the AI's prediction. 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.

**1 INTRODUCTION**

Recent years have seen an explosion of work on explainable AI (XAI), but there have been mixed results on whether explanations actually help humans who are making decisions with AI support. In this decision making context, the role of explanations is to foster appropriate reliance by helping the human understand whether or not the AI's advice should be trusted. Appropriate reliance is desired in order to achieve complementary performance, where the human-AI team performs better than either the human or AI alone [1]. But here we see a confusing montage of results: not only do most papers find explanations don't induce complementary performance more than baseline methods, such as displaying AI accuracy or confidence, but these papers suggest explanations can in fact increase over-reliance, where the human trusts the AI even when it errs. The inconclusive nature of these results raises a huge question for the field of XAI: when are explanations useful?

We focus solely on the process of AI-advised decision making, defined as the following: given an instance of a decision making task, an AI makes a recommendation, and drawing on features of the task, the AI's recommendation, and possibly an explanation for the AI's recommendation, a human decision maker arrives at a final decision (Figure 1). There are many other possible uses for AI explanations [1, 7], including model debugging and auditing, e.g., to help the human understand whether the AI's reasoning will generalize, but our arguments pertain only to decision making.

In this paper, we present a perspective we believe explains the seemingly mixed empirical results found throughout the XAI literature. Furthermore, our proposal is consistent with the way human groups reach consensus on "inductive" tasks [2]. We argue explanations provided by an AI model are helpful in decision making (engender complementary performance [1]) to the extent they allow a decision maker to verify the AI's recommendation. While this theory may appear self-evident,

(1) Most work on XAI has focused instead on creating inherently interpretable models or generating faithful post-hoc explanations of the AI's reasoning process.

(2) Most of these explanations do not support such verification. Explanations which faithfully expose the AI's reasoning process may well be useful for debugging the AI or predicting its ability to generalize, but it does not seem to help human decision makers make judgements on individual task instances. Indeed, most human-subject studies have shown that explanations fail to produce complementary performance in decision making, the sole exceptions are explanations that support answer verification (Table 1).

The rest of this paper is structured as follows. The next section surveys the conflicting results from prior studies on XAI utility. Section 1 details the decision making context, which is our focus in this paper. Section 3 presents our core argument -- explanations must facilitate verification in order to engender complementary performance. Section 5 discusses the concept of appropriate reliance, arguing this term has become overloaded, leading to confusion.

~10,000 words

# Abstracts

# Papers

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 tasks that the use of explanation method, limiting the extent to which a human can compare the objective of complementary performance with that of appropriate reliance, decomposing the latter into the notion of outcome-graded and strategy-graded reliance.

Access Paper:

- Download PDF
- Other Formats

Current browse context: cs.AI

Change to browse by: cs.HC

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

Bookmark

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>

What's complementary performance?..

~150 words

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

Raymond Fok  
rayfok@cs.washington.edu  
University of Washington  
Seattle, WA, USA

Daniel S. Weld  
dsw@allenai.org  
Allen Institute for AI & R  
University of Washington  
Seattle, WA, USA

**ABSTRACT**

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 the AI's prediction. Prior studies find in many decision making contexts that the use of explanation method, limiting the potential benefit of any type of explanation, decomposes the latter into the notion of outcome-graded and strategy-graded reliance.

**1. INTRODUCTION**

Recent years have seen an explosion of work on explainable AI (XAI), but there have been mixed results on whether explanations actually help humans who are making decisions. In this decision making context, the role of appropriate reliance is desired in order to achieve complementary performance, where the human-AI team performs better than either the human or AI alone [1]. But here we see a confusing montage of results: not only do most papers find explanations don't induce complementary performance more than baseline methods, such as displaying AI accuracy or confidence, but these papers suggest explanations can in fact increase over-reliance, where the human-AI team performs worse than either the human or AI alone [2].

We focus solely on the process of AI-advised decision making, defined as the following: given a task, the AI generates a recommendation, and possibly an explanation for the task. A human decision maker arrives at a final decision (Figure 1). There are many other possible uses for explanations [1, 7], including model debugging and auditing, but to help the human understand whether the AI's reasoning will generalize, but our arguments pertain only to decision making.

In this paper, we present a perspective we believe explains the seemingly mixed empirical results found throughout the XAI literature. Furthermore, our proposal is consistent with the way human groups reach consensus on "inductive" tasks [3]. We argue explanations provided by an AI model are helpful in decision making (engender complementary performance [1]) to the extent they allow a decision maker to verify the AI's recommendation. While this theory may appear self-evident,

(1) Most work on XAI has focused instead on creating inherently interpretable models or generating faithful post-hoc explanations of the AI's reasoning process.

(2) Most of these explanations do not support such verification. Explanations which faithfully expose the AI's reasoning process may well be useful for debugging the AI or predicting its ability to generalize, but it does not seem to help human decision makers make judgements on individual task instances. Indeed, most human-subject studies have shown that explanations fail to produce complementary performance in decision making, the sole exceptions are explanations that support answer verification (Table 1).

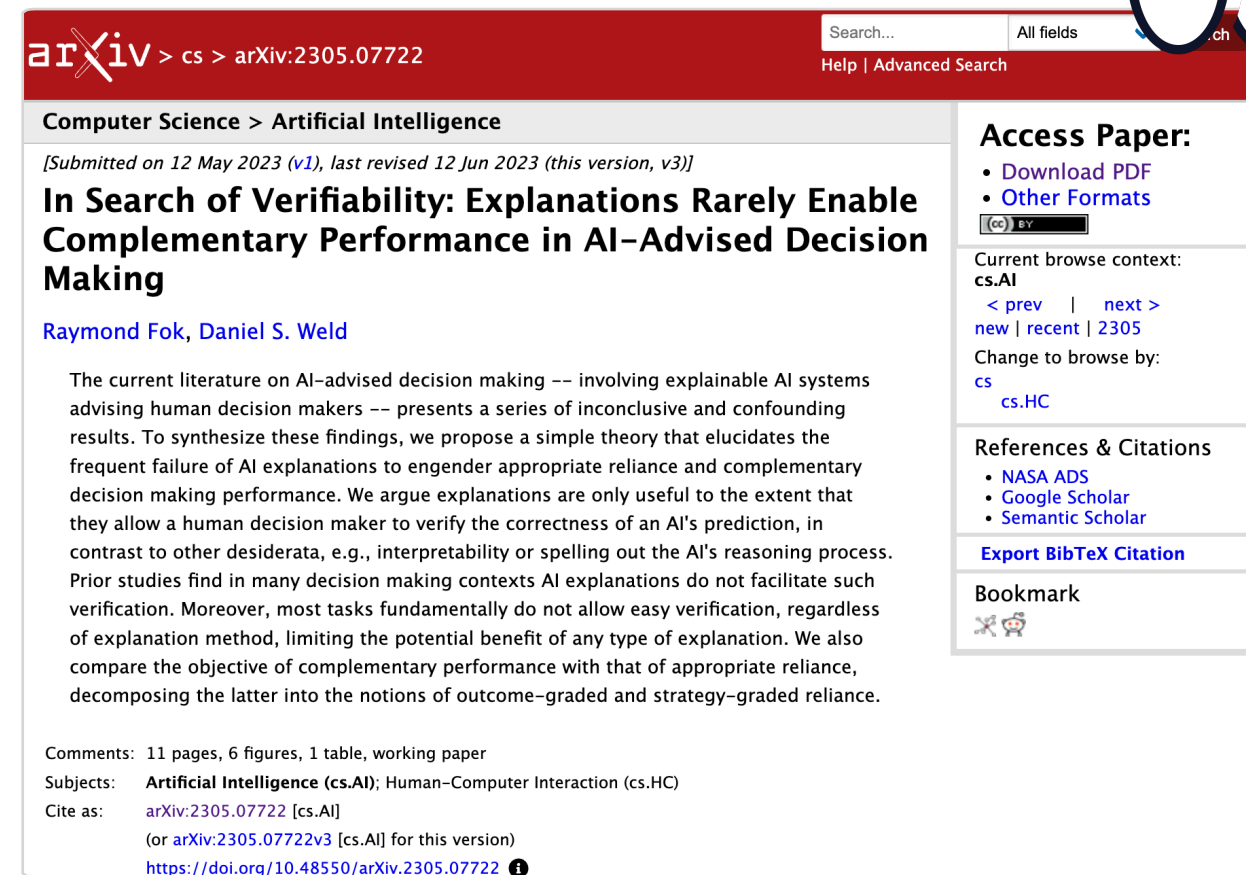
The rest of this paper is structured as follows. The next section surveys the conflicting results from prior studies on XAI utility. Section 1 details the decision making context, which is our focus in this paper. Section 2 presents our core argument -- explanations must facilitate verification in order to engender complementary performance. Section 3 discusses the concept of appropriate reliance, arguing this term has become overloaded, leading to confusion.

~10,000 words

# Bridging the information chasm

## Abstracts

## Papers



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>

Interactive augmentations



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

Raymond Fok, Daniel S. Weld

ABSTRACT

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.

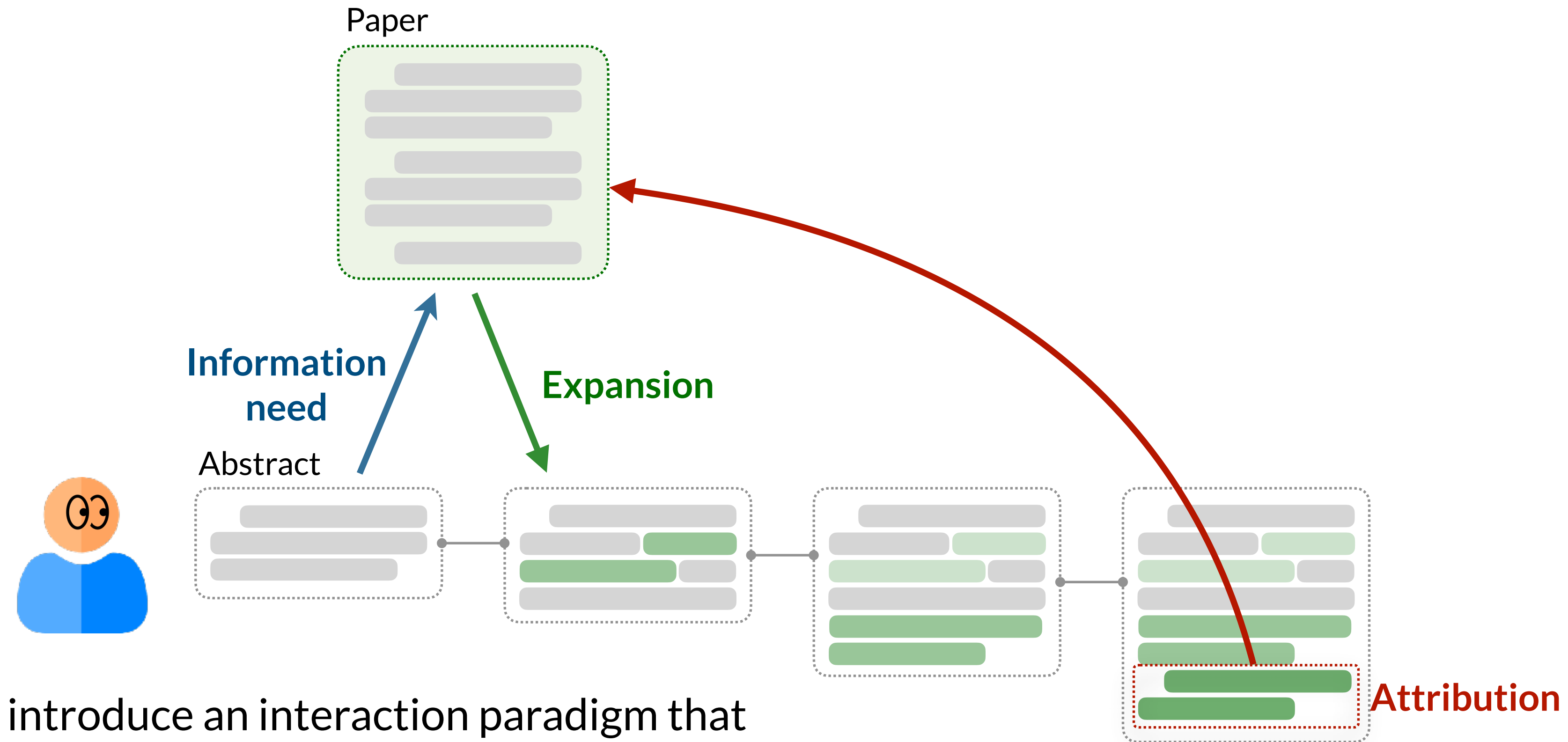
1 INTRODUCTION

Recent years have seen an explosion of work on explainable AI (XAI), but there has been mixed results on whether explanations actually help human decision makers with AI support. In the decision-making context, the role of explanations is to foster appropriate reliance by helping the human understand why the AI made a certain recommendation. Appropriate reliance is defined as when the human appropriately performs, when the human will perform better than otherwise the human or AI performance alone. In this paper, we propose a simple theory that elucidates the frequent failure of AI explanations to engender appropriate reliance. 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.

Static  
Non-personalized

Long  
Dense  
Cognitively demanding

# 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 literature on AI decision making -- involving explainable AI systems advising and confounding theory that elucidates the frequent failure of AI explanations -- presents a series of inconclusive findings. *How do they define appropriate reliance?* *Tell me more about this..* appropriate reliance and complementary decision-making. *Why?* 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.



# 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 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.




*What does this mean?*

**User selects any arbitrary span to expand**

# Reducing expansion effort via one-click expansion actions

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, easy verification, regardless of explanation benefit of any type of explanation. We propose complementary performance with that of appropriate reliance, decomposing the latter into the notions of outcome-graded and strategy-graded reliance.

What is the difference between outcome and strategy graded reliance?

 Define  Expand  Why

Definition & Instantiation

Expansion

Motivation

AI-generated question, inferring user intent based on context

# 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 complementary decision making with that of appropriate reliance, decomposing the latter into 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.

How is strategy-graded reliance calculated?



Define



Expand



Why

# 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”.

Tell me more about “strategy-graded reliance”.

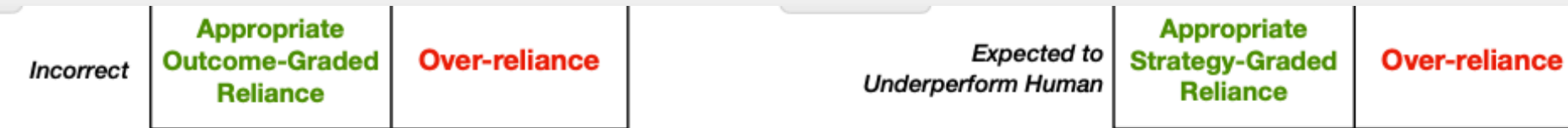
Outcome-graded reliance defines a reliance based on the acceptance of AI advice conditioned on the performance of the AI. Strategy-graded reliance defines a reliance based on the relative expected performance of the human and the AI. Strategy-graded reliance is appropriate if a human accepts an AI recommendation when the AI is expected to outperform the human (conditioned on factors such as past performance, features, and AI advice), and rejects otherwise. The optimal strategy 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

Def  
st





**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

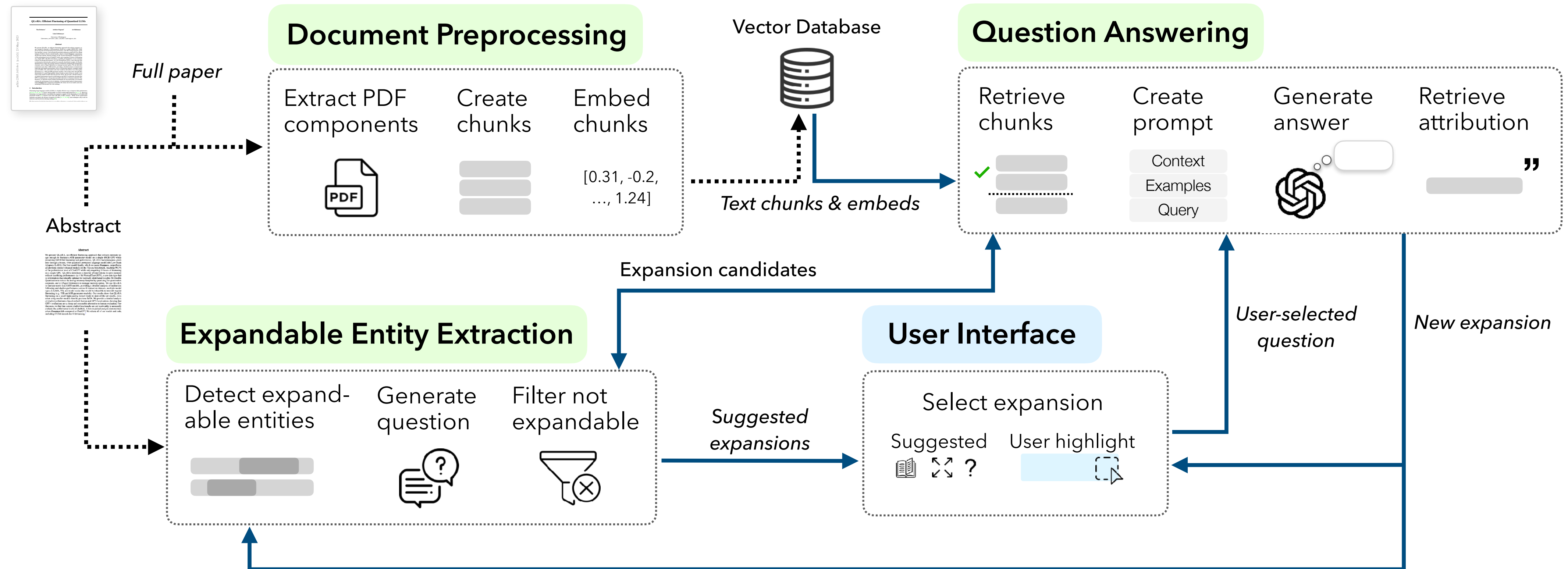
previous strategy, and only when the explanation supports verification.

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.”

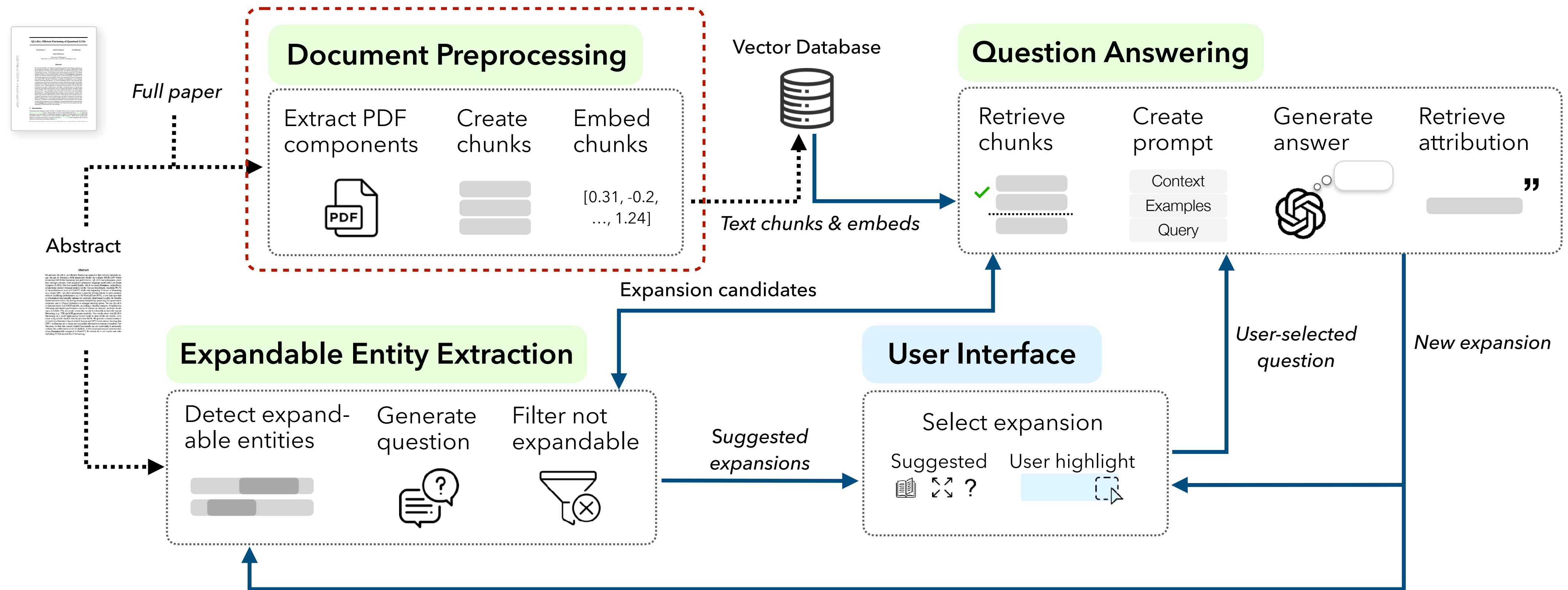
Outcome-graded reliance is similar to complementary performance in the sense that both qualities can only be measured post hoc. However, there are subtle differences between these notions, beyond the fact that one measures a pattern of human behavior and the other the performance of a human-AI team. 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

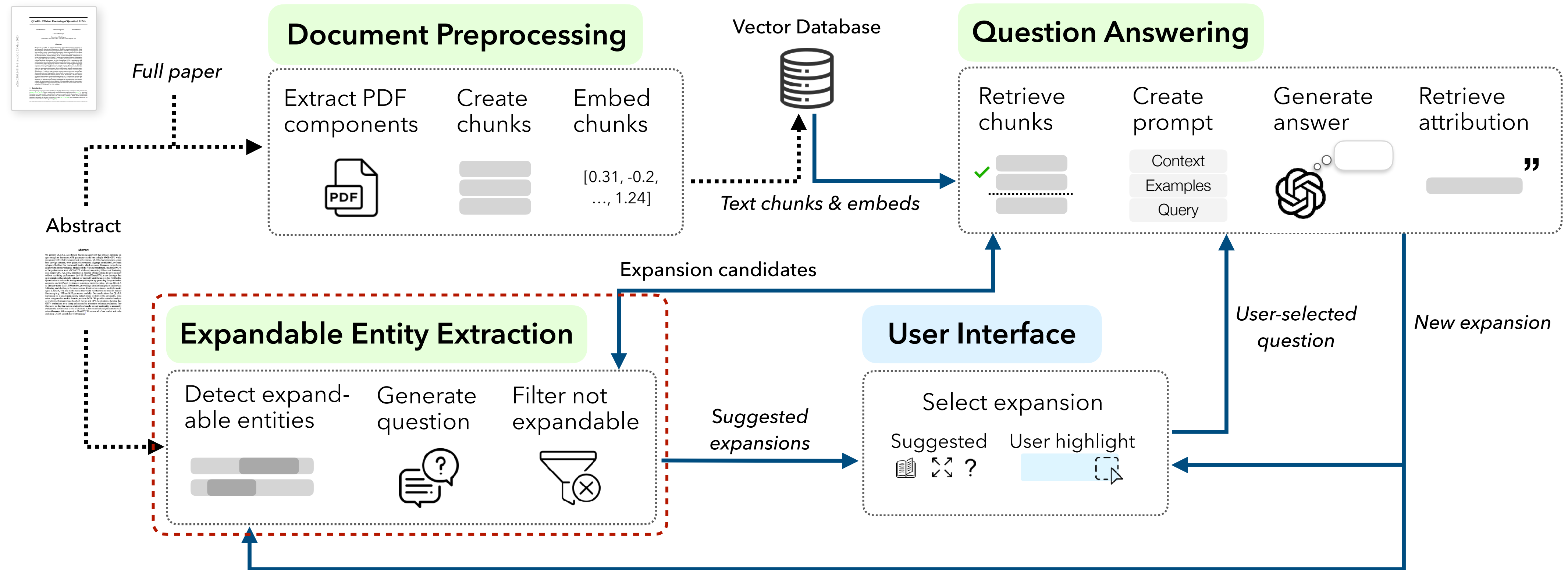
# System architecture



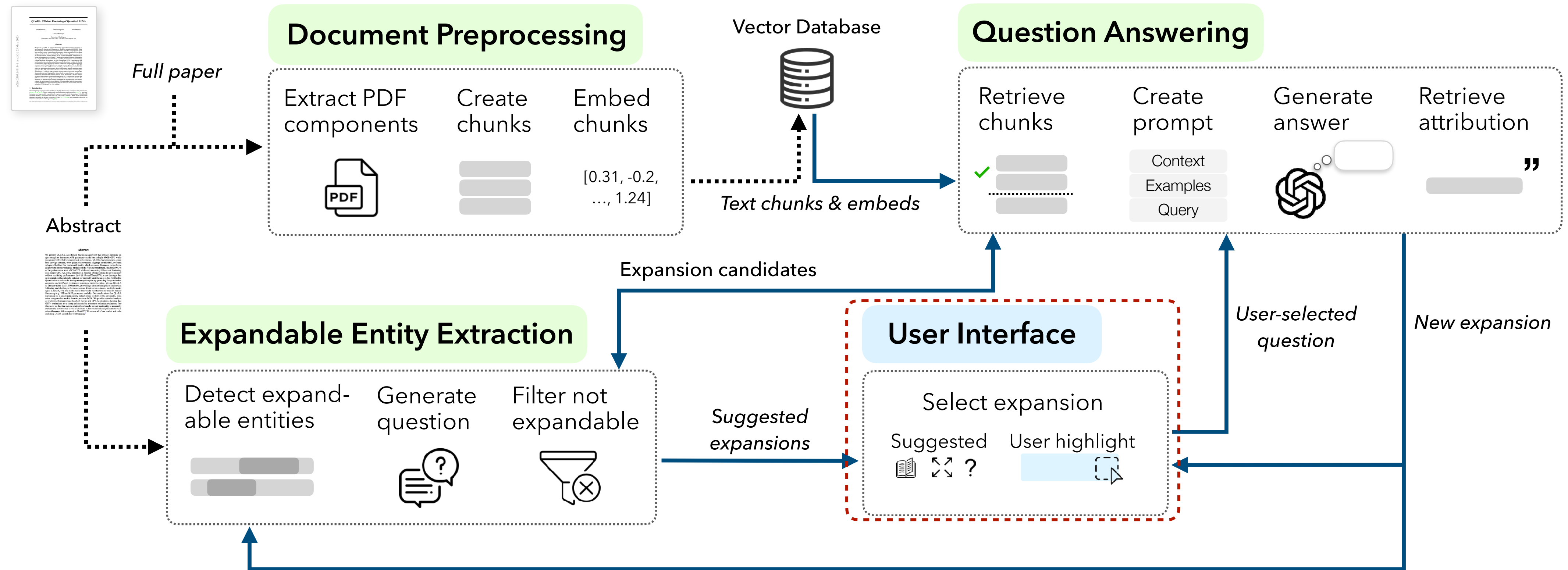
# System architecture



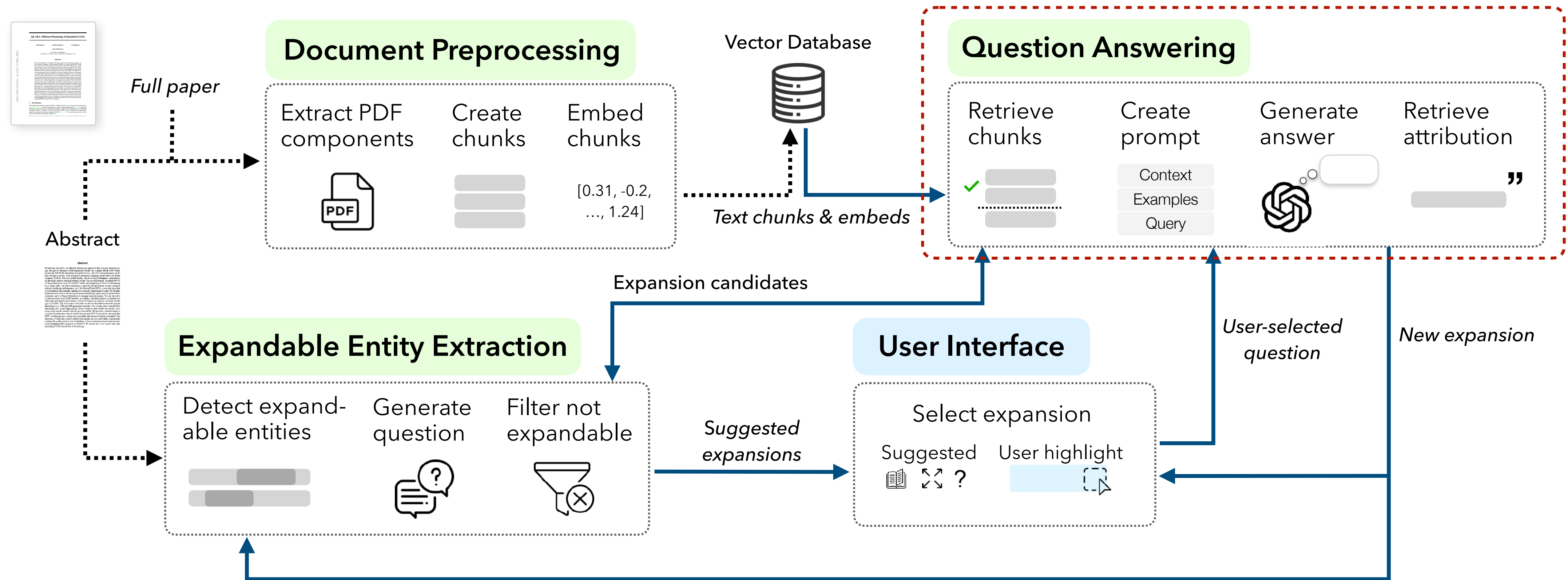
# System architecture



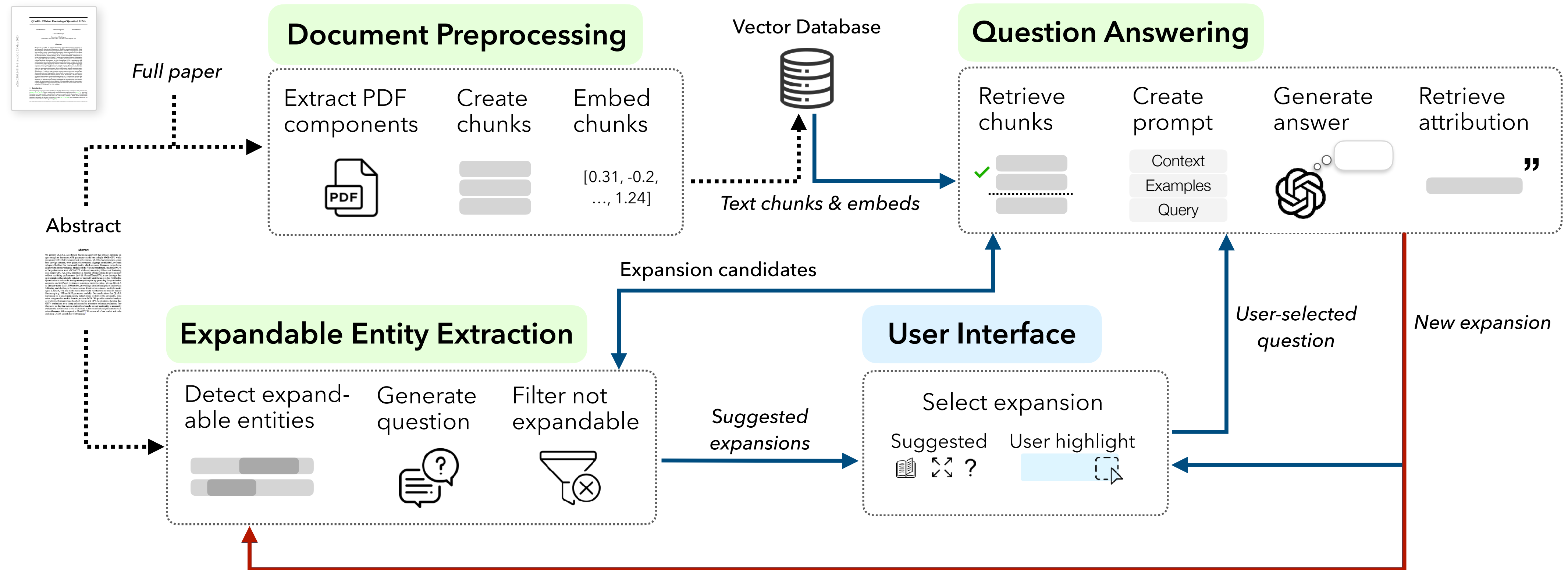
# System architecture



# System architecture



# System architecture



# Evaluation



Qualitative interview study with 9 scholars

---

Deployment study at a conference (n=275)





# 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.

The screenshot shows the arXiv interface for the paper. It includes the title, authors (Raymond Fok, Daniel S. Weld), submission date, and a list of actions like 'Download PDF' and 'Other Formats'. The abstract text is visible, starting with '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.'

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

The image shows a stack of paper pages representing the document. A callout box highlights a specific section of the paper, which is the abstract. The abstract text is: '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.'

# Findings

## How did participants use each of the features?

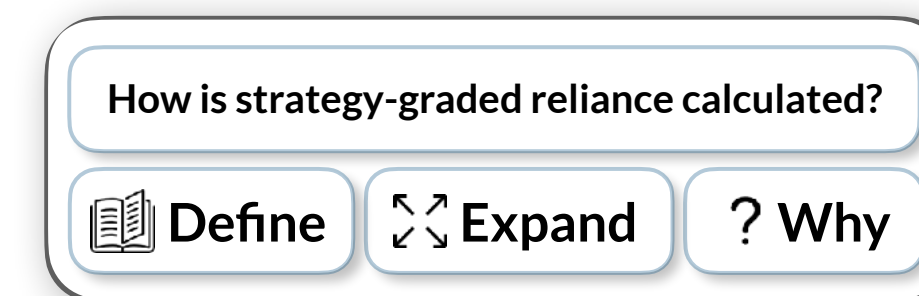
### Participants

...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

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



# 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.



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

# Findings

## How did scholars use each of the features?

### Scholars

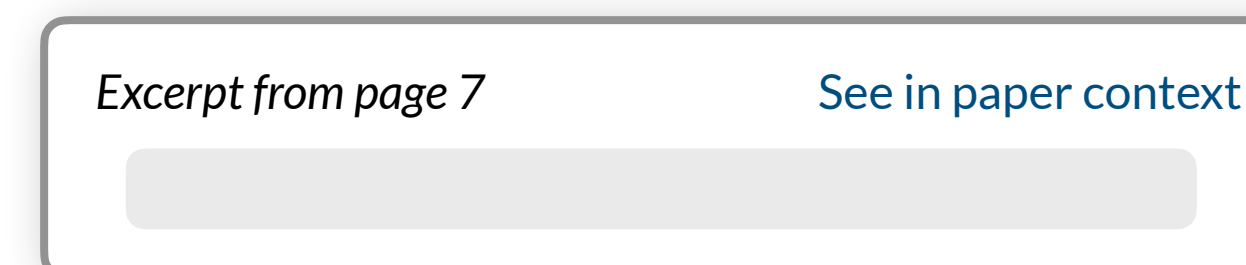
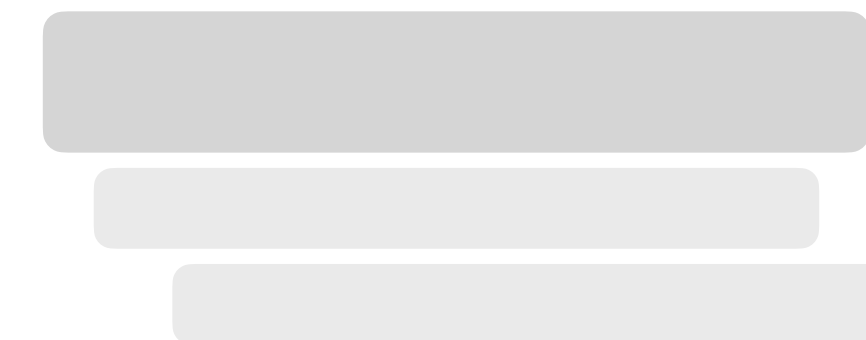
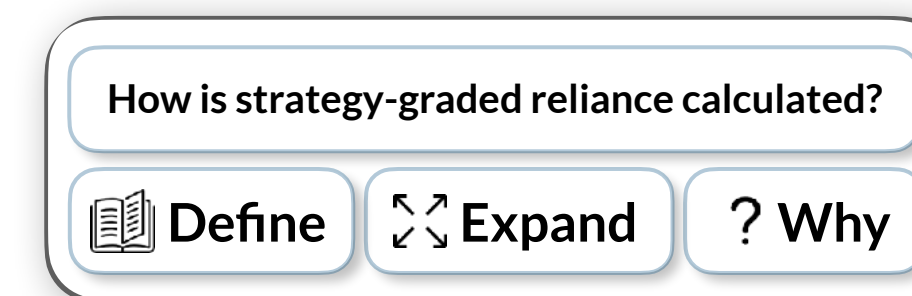
...more often selected an AI-suggested expandable entity (80%) rather than highlighting their own (20%)

...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

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



Different usage patterns likely due to different levels of engagement.

# Risks of expandable abstracts

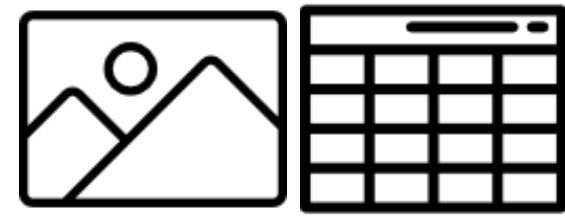
Augmented intelligence for scholars may harm pedagogical and self-learning processes, especially for novice scholars.

“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

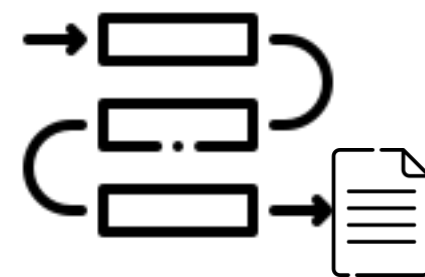
# Improving expandable abstracts



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

...outperforms all previous openly released models on the Vicuna benchmark, performance level of ChatGPT while only requiring 24 hours GPU. QLoRA introduces a number of innovations to save performance: (a) 4-bit NormalFloat (NF4), a new data type that is theoretically optimal for normally distributed weights (b) double quantization to reduce the average memory footprint by quantizing the quantization constants, and (c) paged optimizers to manage memory spikes.

Define "double quantization"

How does double quantization reduce..

Define Expand ? Why

What is double quantization?..

Double quantization is the process of quantizing the quantization constants to reduce the memory footprint without degrading performance. The paper's experiments show that QLoRA finetuning on a small high-quality dataset leads to state-of-the-art results, even when using smaller models than the previous SoTA...

We use QLoRA to finetune more than 1,000 models, providing a detailed analysis of instruction following and chatbot performance across 8 instruction datasets, multiple model types (LLaMA, T5), and model scales that would be infeasible to run with regular finetuning (e.g. 33B and 65B parameter models). Our results show that QLoRA finetuning on a small high-quality dataset leads to state-of-the-art results, even when using smaller models than the previous SoTA...

