Another RAND AI effort?

- RAND's Center on AI, Security, and Technology (CAST) Management desires to accelerate the pace at which it can produce high quality research products

- LLMs are able to automate and accelerate research workflows. RAND has yet to integrate AI and realize these gains

- There is a significant amount of effort in this space
  - Google has a Co-Scientist system and an "empirical software" optimizer
  - Sakana has an AI Scientist (with 2 published blind peer-reviewed papers)
  - All major models can perform web-based "deep research"

- Our goal is to make sure we can leverage ALL of these innovations and incorporate them into a system that is highly robust and reliable

- To facilitate this, we've started by focusing on a mechanical peer review system that we can trust
  - To confidently and responsibly use these tools, CAST (and RAND) need to "own" a reliable and quick system for evaluating the robustness of our process, findings, recommendations, etc.

Rand-AI Reviewer

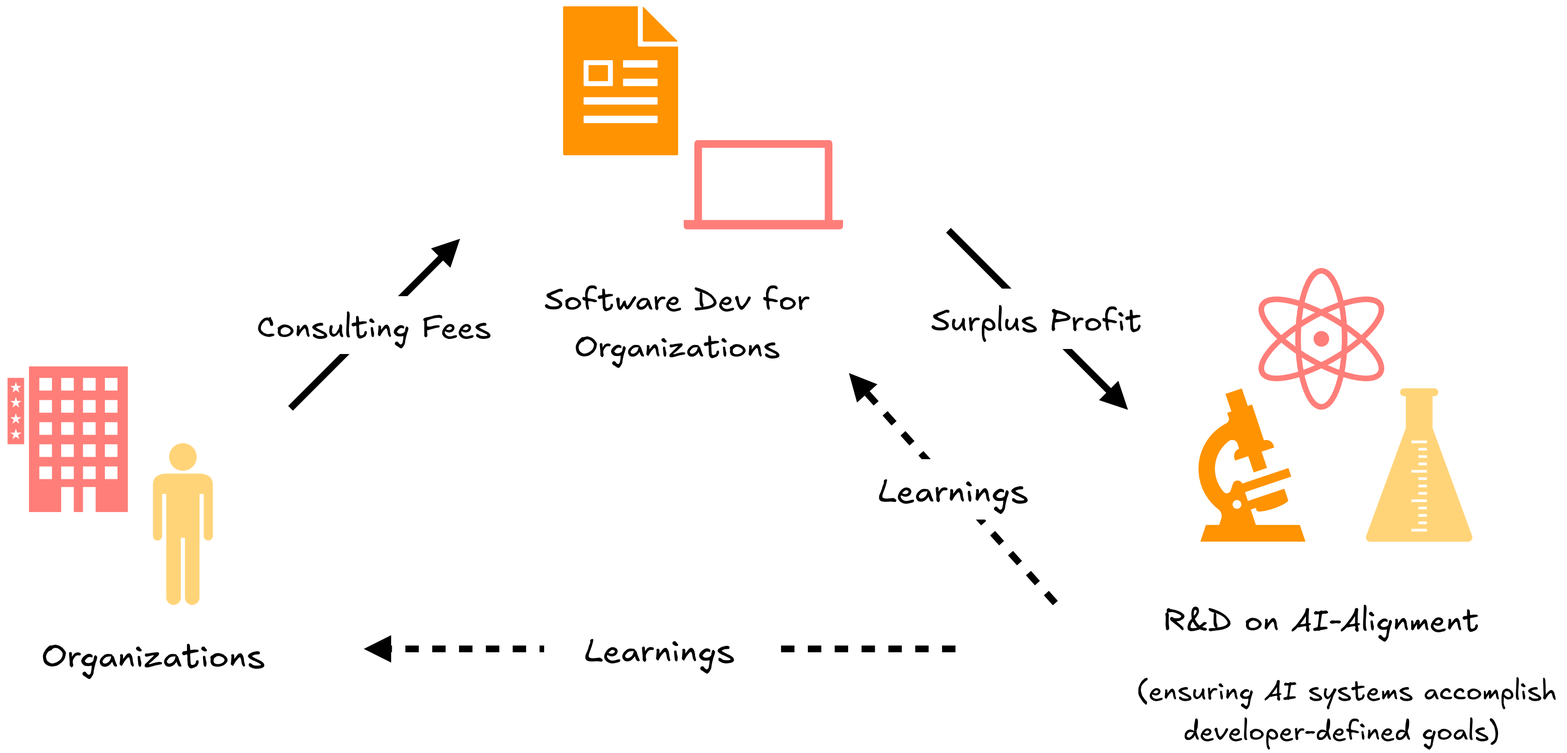
Automating Document Review Process

Rand Corporation **—** ***AE Studio*** **Collaboration — Tuesday November 4, 2025**

Who is doing this work?

AE.STUDIO

Consulting firm focused on software development and AI alignment



Who is doing this work?

AE.STUDIO

RAND-AI Reviewer Team

Carlos Bonetti

(Technical Product Manager)

(BS, Computer Science)

Ricardo Schieck

(Full Cycle Engineer)

(BS, Computer Software Engineering)

Mobolaji Williams

(Data Scientist)

(PhD, Physics)

Presentation Roadmap

I. Motivation

II. Video Demo

III. System Details

- -.i Architecture
- -.ii Agents and LangGraph
- -.iii Design principles

IV. Gaps and Next Steps

V. Live Demo & Q&A

Presentation Roadmap

I. Motivation

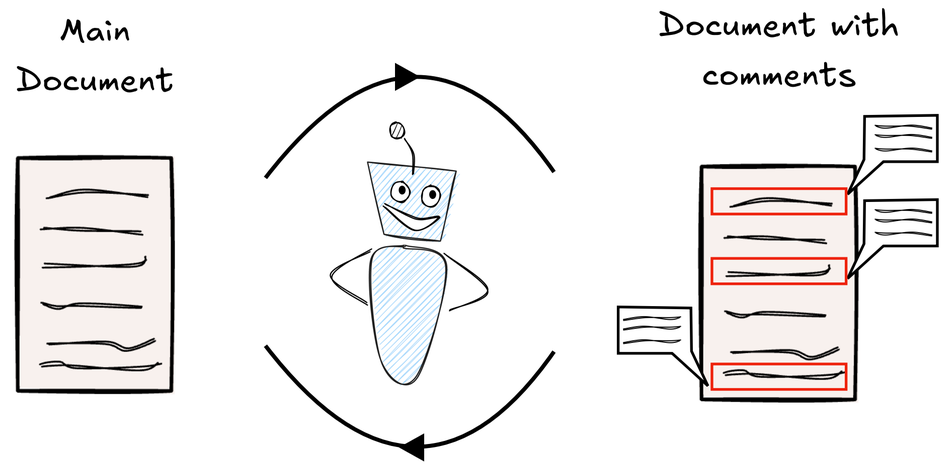
II. Video Demo

III. System Details

- -.i Architecture
- -.ii Agents and LangGraph
- -.iii Design principles

IV. Gaps and Next Steps

V. Live Demo & Q&A



Motivation: Embarassments of LLM Capabilities

Deloitte to Refund Government After Admitting AI-Generated Errors in $440K Report

https://colitco.com/deloitte-refund-ai-errors-government-report/

Austrailian govt commissioned a report to examine compliance issues of welfare recipients'

↓

Consultants used LLMs for their reports which led to "false quotes", fabricated "experts" and "references"

Consequences

Micro

Embarrassing for Deloitte (and financially costly; $440,000 refund for services)

Macro

Welfare discussions were based on false information; Further erosion of belief in efficacy of LLMs

Conclusion: LLM-generated information cannot be trusted without verification

Motivation: Reversing Generation

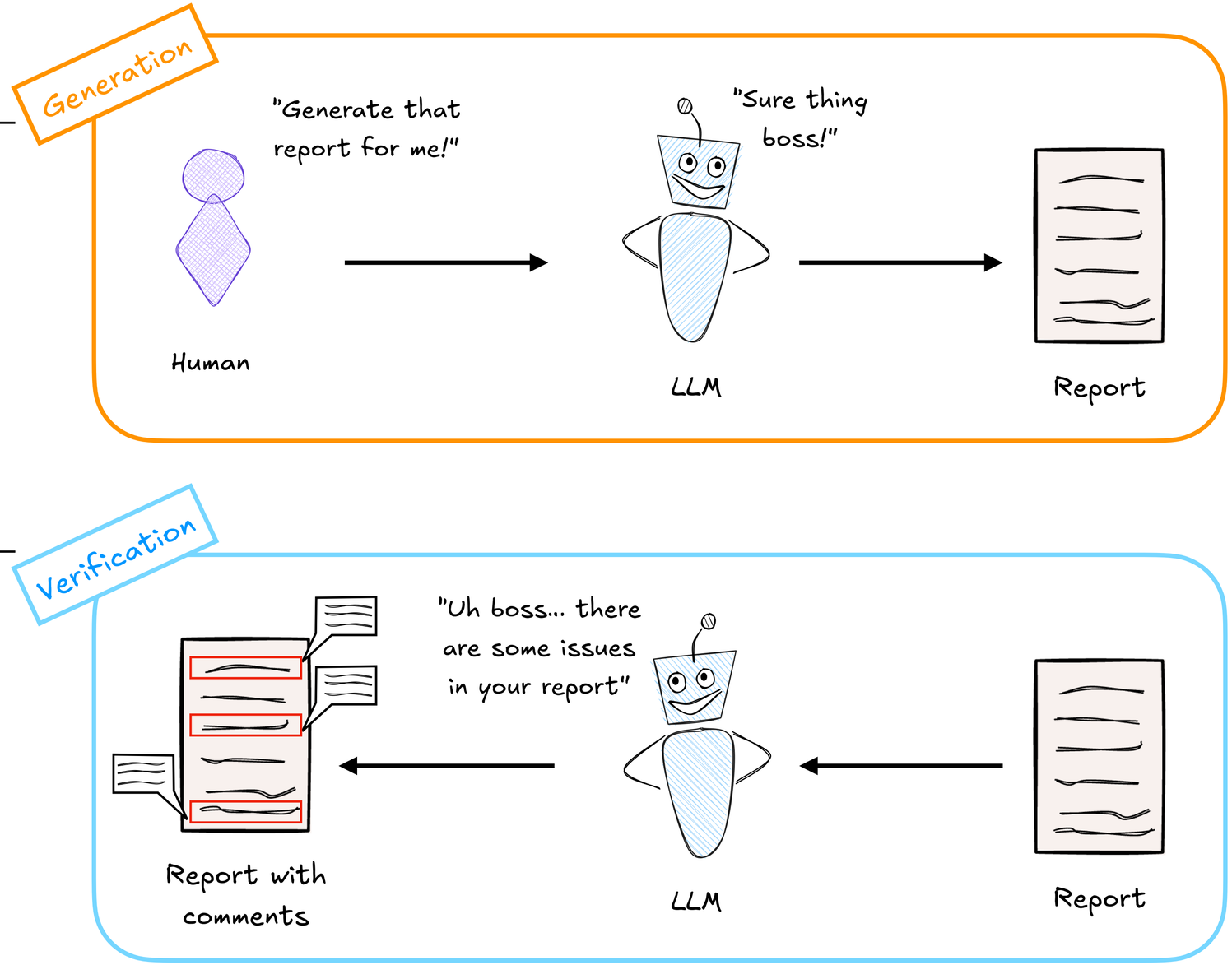
Deloitte story reflects the typical usecase for LLMs...

Both processes can have "hallucinations"

Which poses a problem

If we cannot trust LLMs to generate accurate information...

...how can we trust them to verify information?



...but this chain could be reversed for the purpose of verification and correction

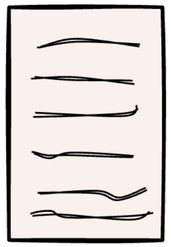
Motivation: LLM Systems & Peer Review

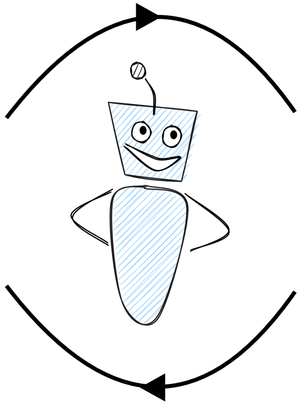
- (Motivating Question)-

How can we build an AI system that we can trust to conduct the main tasks of peer review?

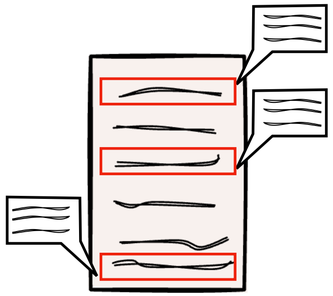
We need to narrow the problem because peer review can check for a lot of things....

Main Document





Document with comments



NOTE!

There is a lot of existing work in this area (https://arxiv.org/pdf/2501.10326 "LLMs for automated peer review: a survey (2025)")

We did not want to add just another system but to...

...understand the design principles within in any such trustworthy "LLM for peer-review" system

Syntax, grammar, formatting

Existing efforts at RAND for this

Statement veracity and context

Our current focus

Overall argument coherence, soundness

"Higher level" future goal

Claim Verification

- checking whether sources support claims

Contextualization

- checking whether there is additional literature that provides context

Inference Validation

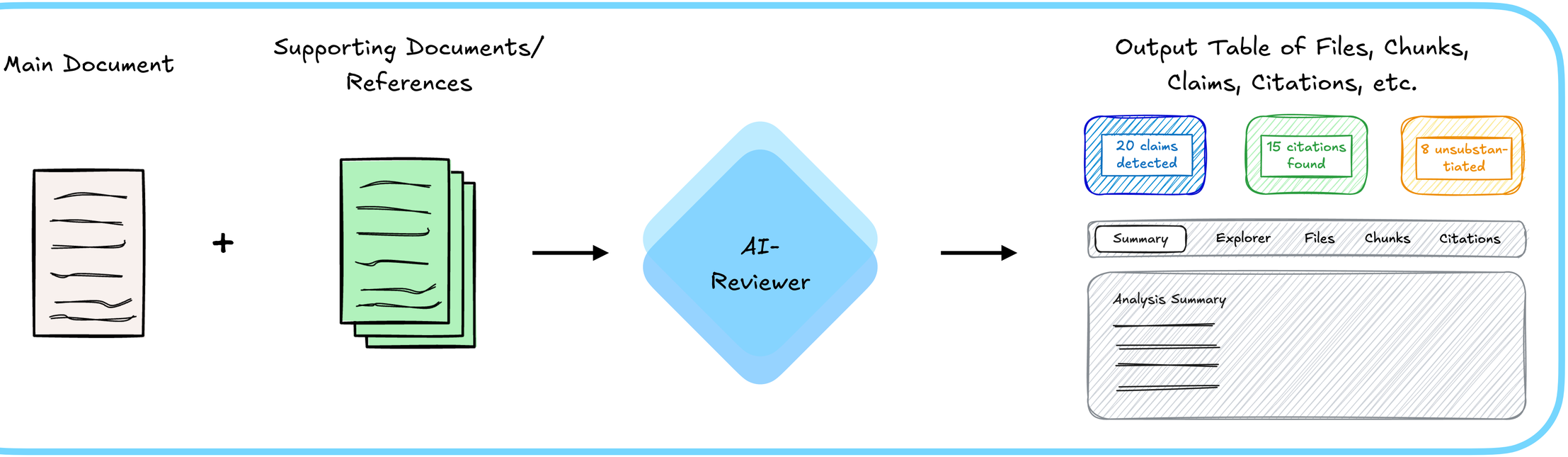
- checking whether arguments are valid

Motivation: Application Introduction

- (Motivating Question)-

How can we build an AI system that we can trust to conduct the main tasks of peer review?

- (General Processing Pipeline) -



Presentation Roadmap

I. Motivation

II. Video Demo

III. System Details

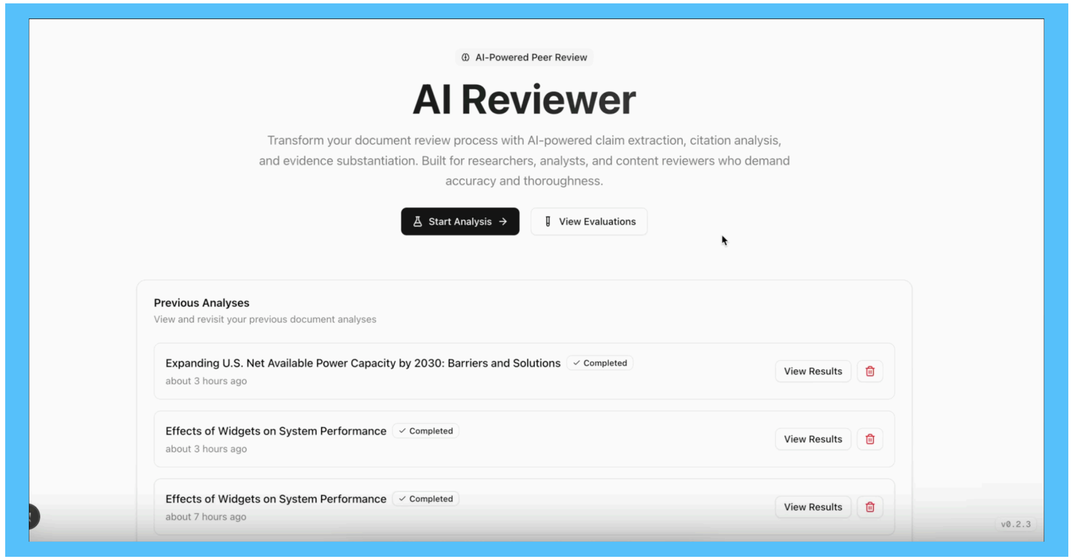
-.i Architecture

-.ii Agents and LangGraph

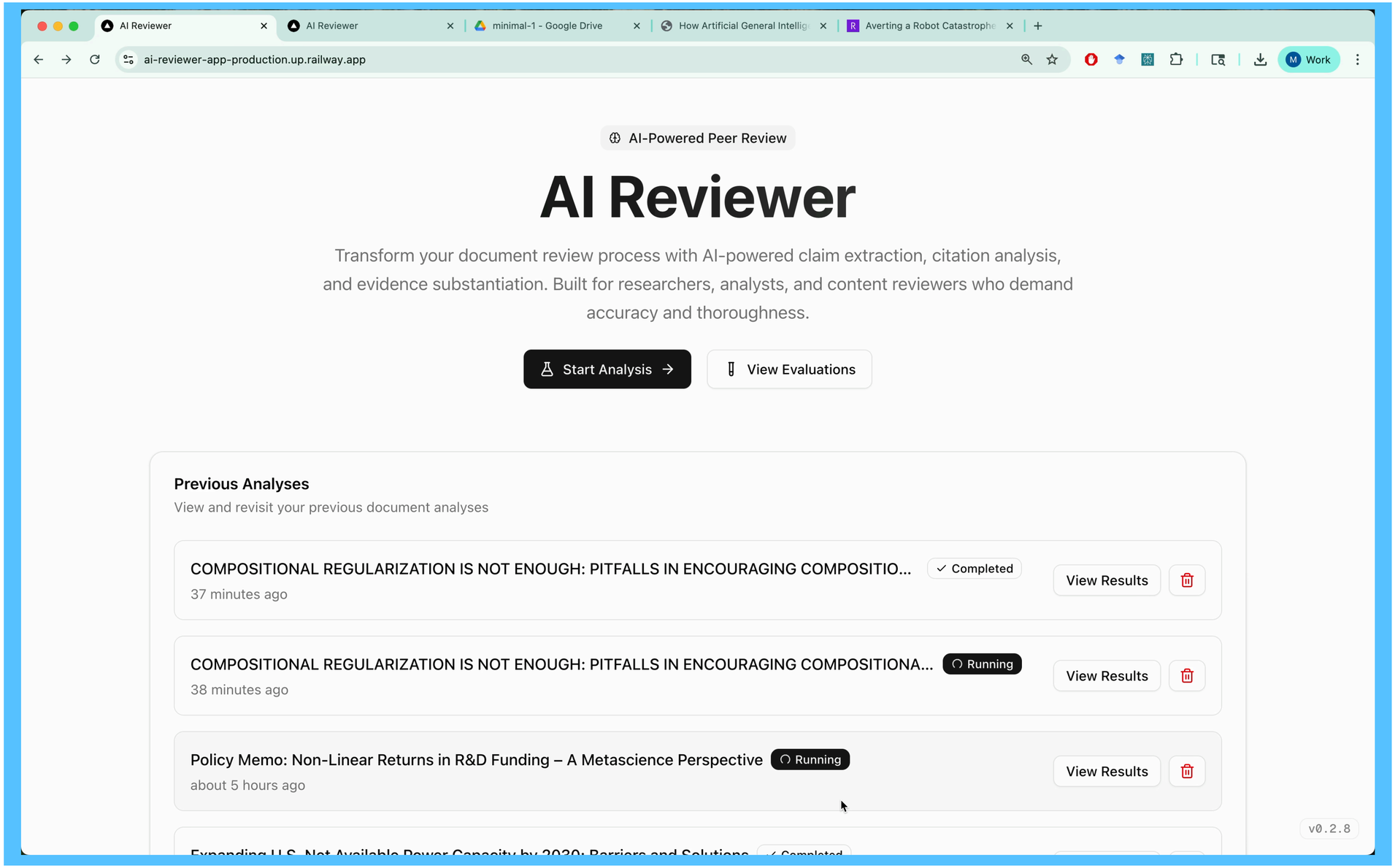
-.iii Design principles

IV. Gaps and Next Steps

V. Live Demo & Q&A



Video Demo of Application



Presentation Roadmap

I. Motivation

II. Video Demo

III. System Details

-.i Architecture

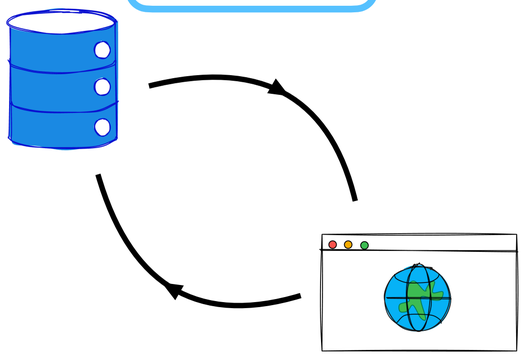
-.ii Agents and LangGraph

-.iii Design principles

IV. Gaps and Next Steps

V. Live Demo & Q&A

System Architecture

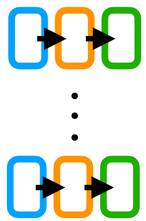


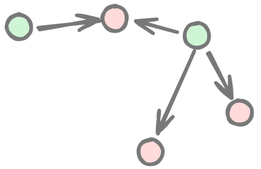
LangGraph Structure of Agents

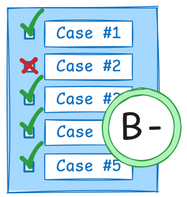
extract_claims

categorize_claims

Design Principles

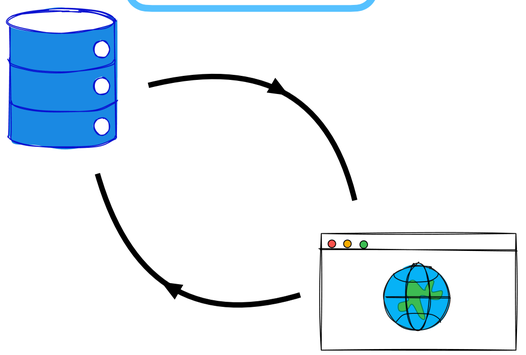


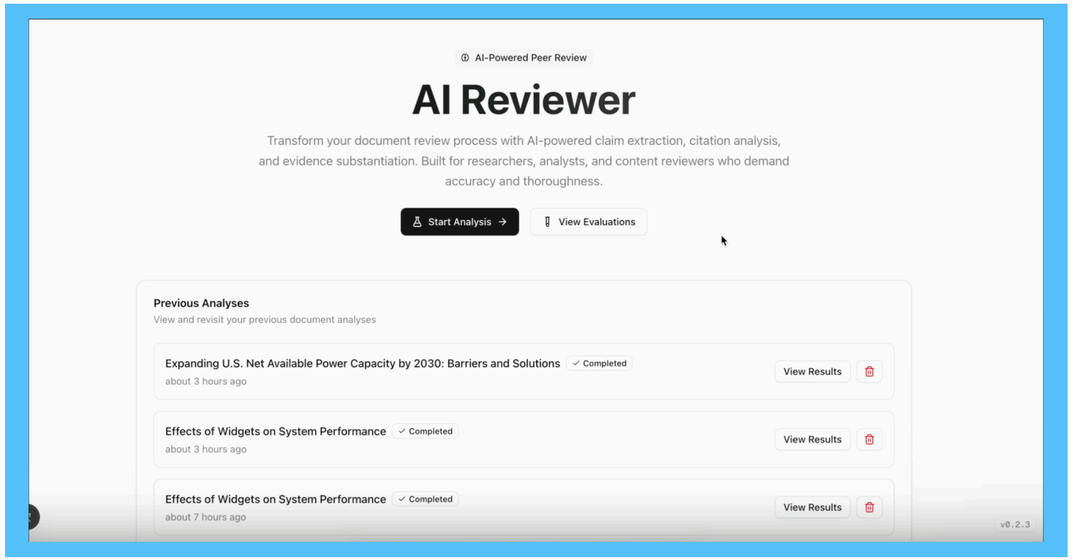




System Details: Architecture

System Architecture



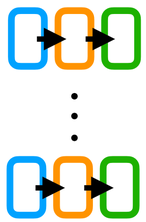


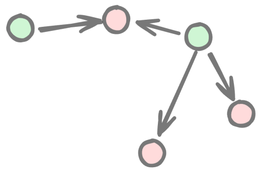
LangGraph Structure of Agents

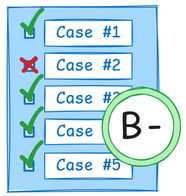
extract_claims

categorize_claims

Design Principles

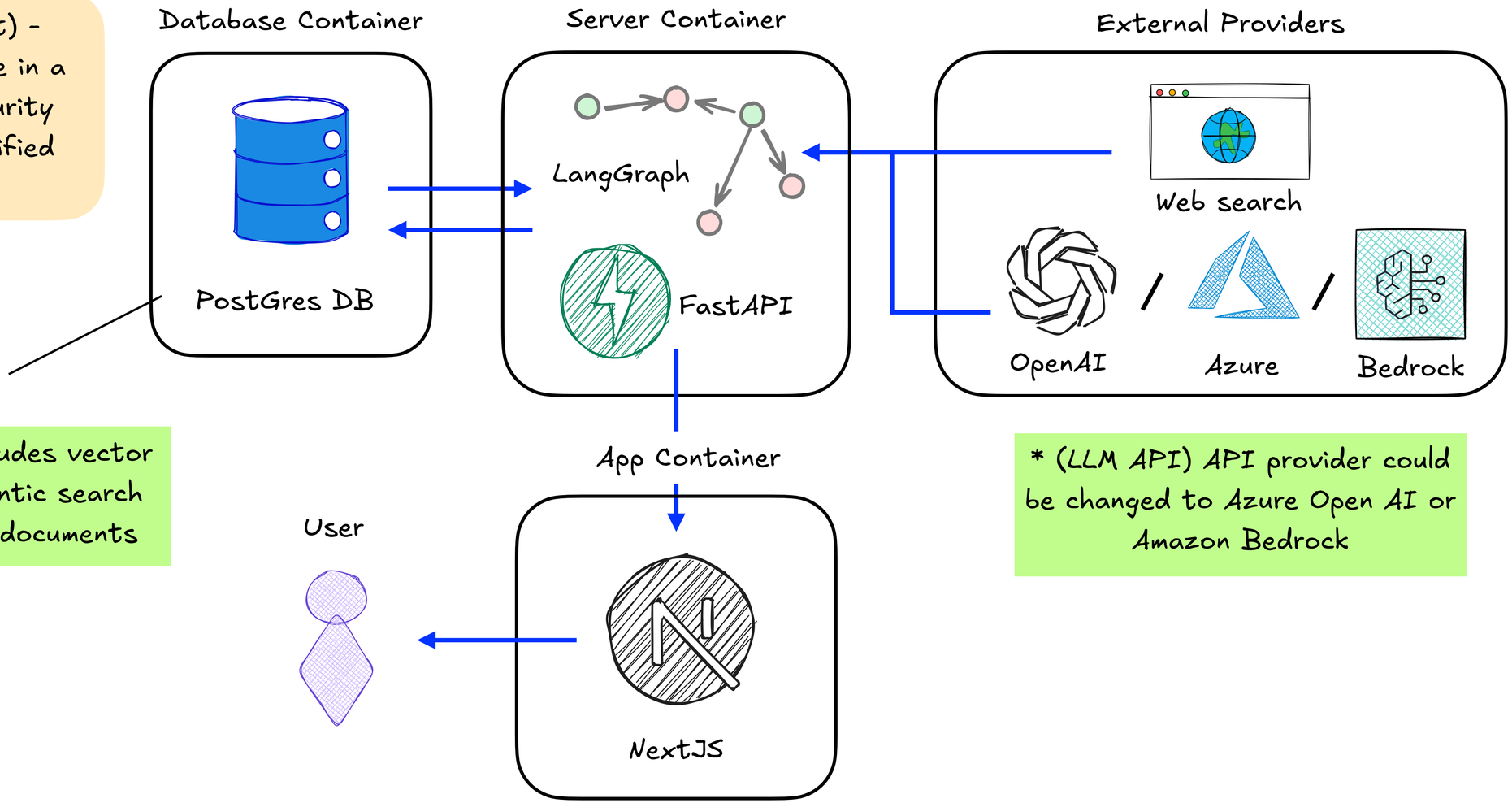






System Details: Architecture

- (Architecture Constraint) - System should be deployable in a variety of information security contexts (GovCloud, Classified cloud, etc.)

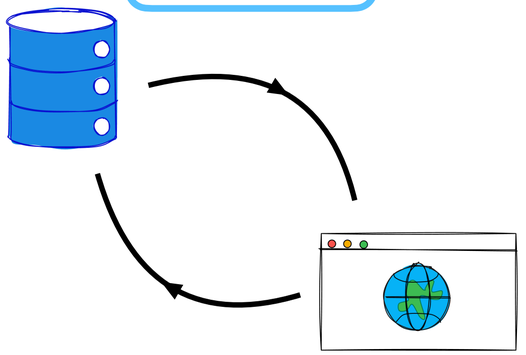


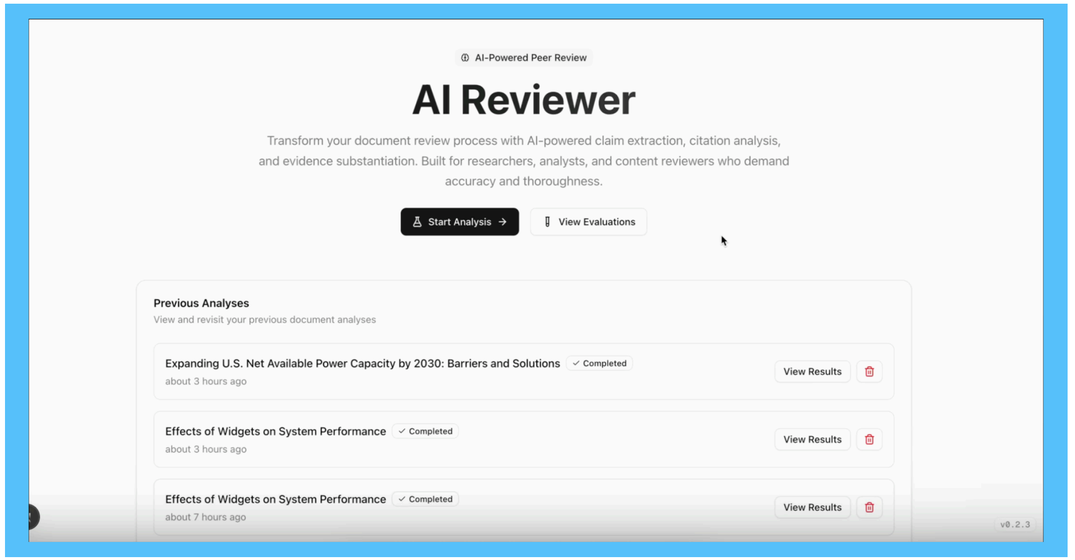
* (Vector DB) Includes vector database for semantic search across supporting documents

* (LLM API) API provider could be changed to Azure Open AI or Amazon Bedrock

System Details: Architecture

System Architecture



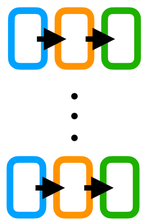


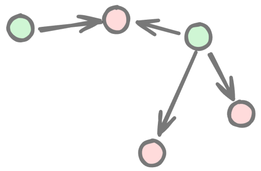
LangGraph Structure of Agents

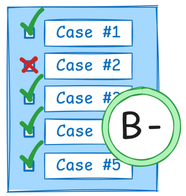
extract_claims

categorize_claims

Design Principles

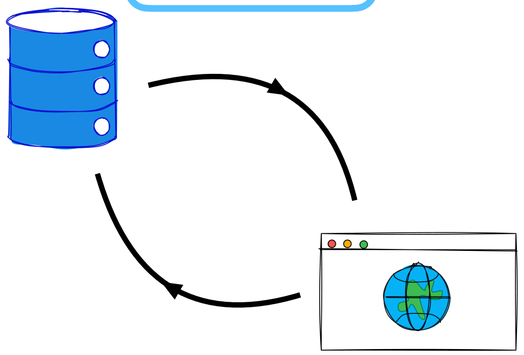


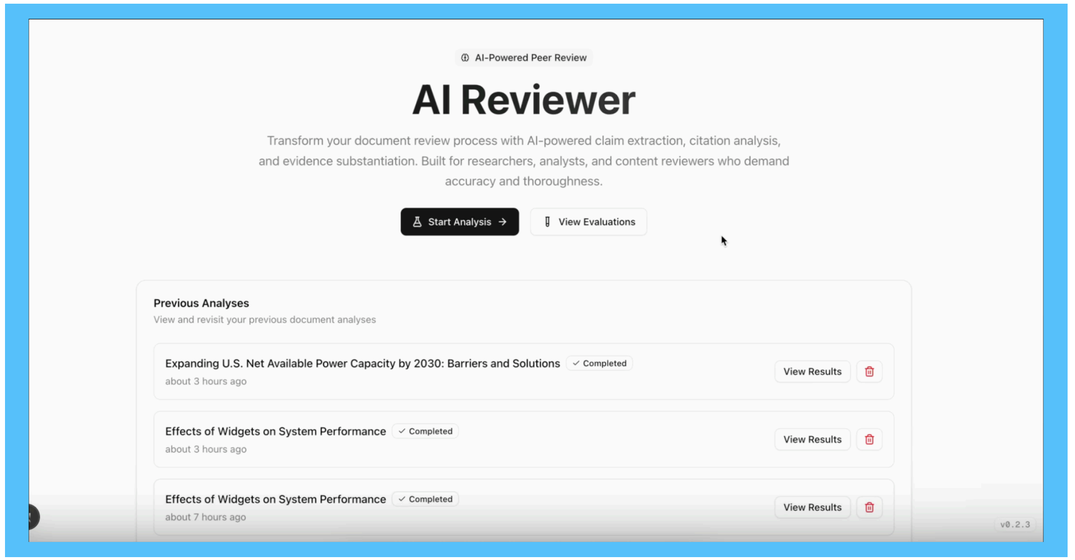




System Details: LangGraph Structure

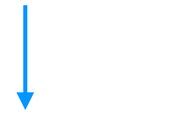
System Architecture





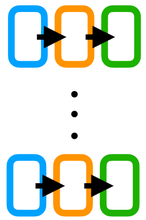
LangGraph Structure of Agents

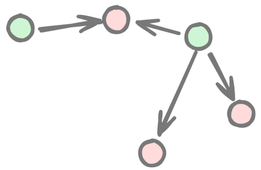
extract_claims

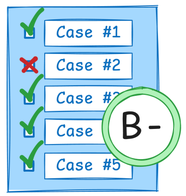


categorize_claims

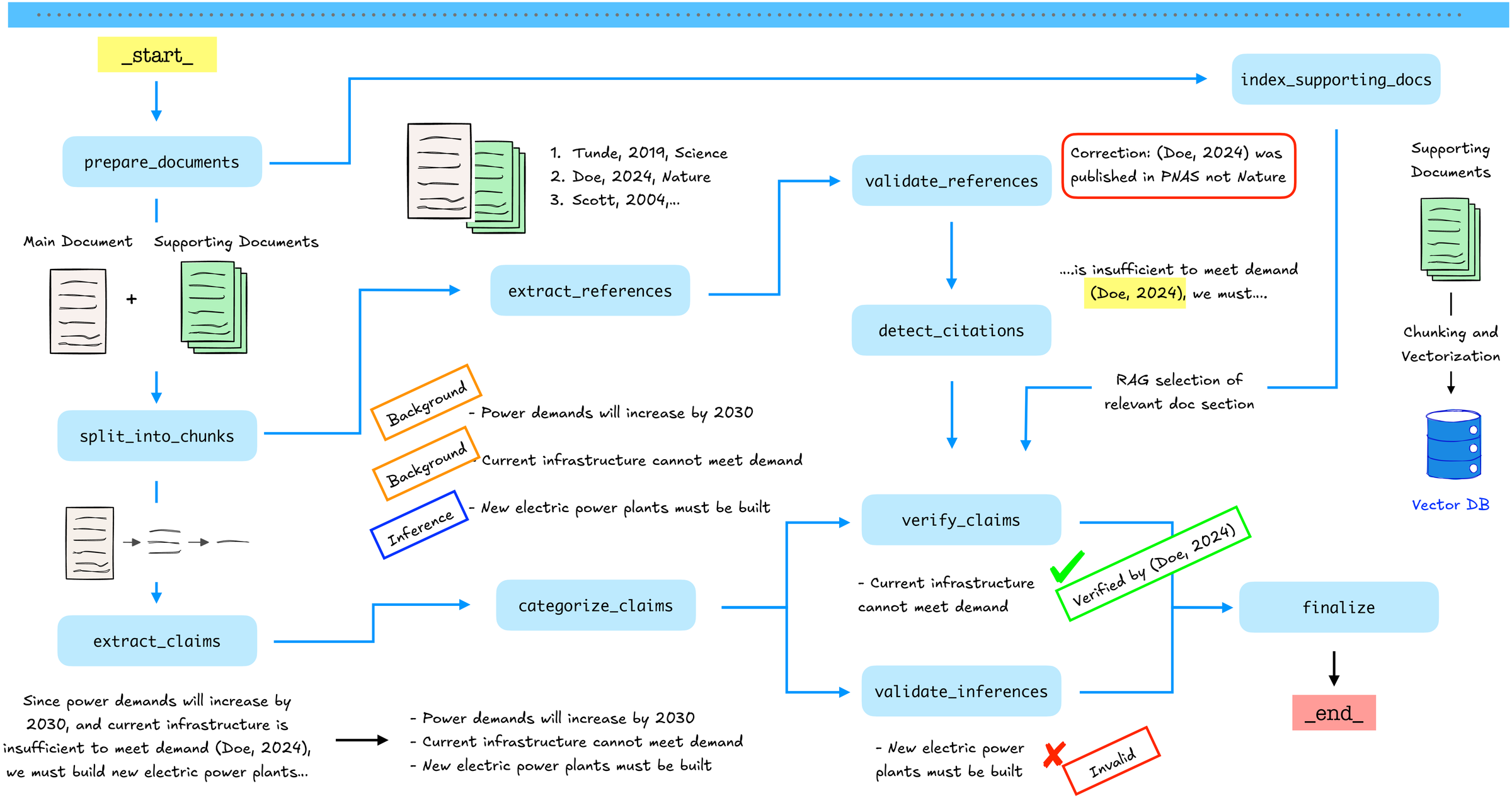
Design Principles







System Details: LangGraph Structure



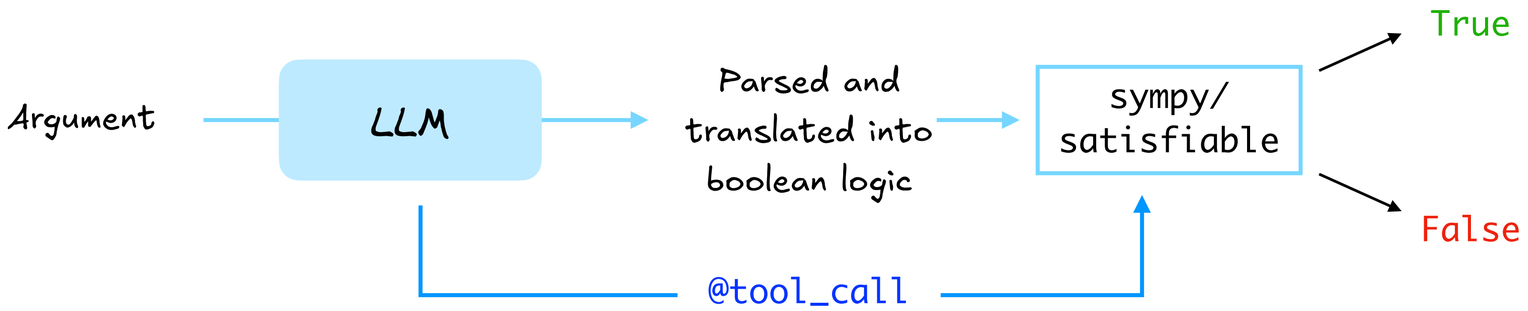
System Details: Symbolic Logic and Inference Evaluation

validate_inferences

- New electric power plants must be built ✗ Invalid

LLM outputs are not constrained by the rules of logic

But LLMs are good at pattern recognition.



Classic Example

Statement	Logic Symbol
i. Socrates is a man	P
ii. If socrates is a man, he is mortal	P implies Q
iii. Socrates is mortal	Q

It is impossible for "P" and "P implies Q" to be true while "Q" is false.

Valid Argument ✓

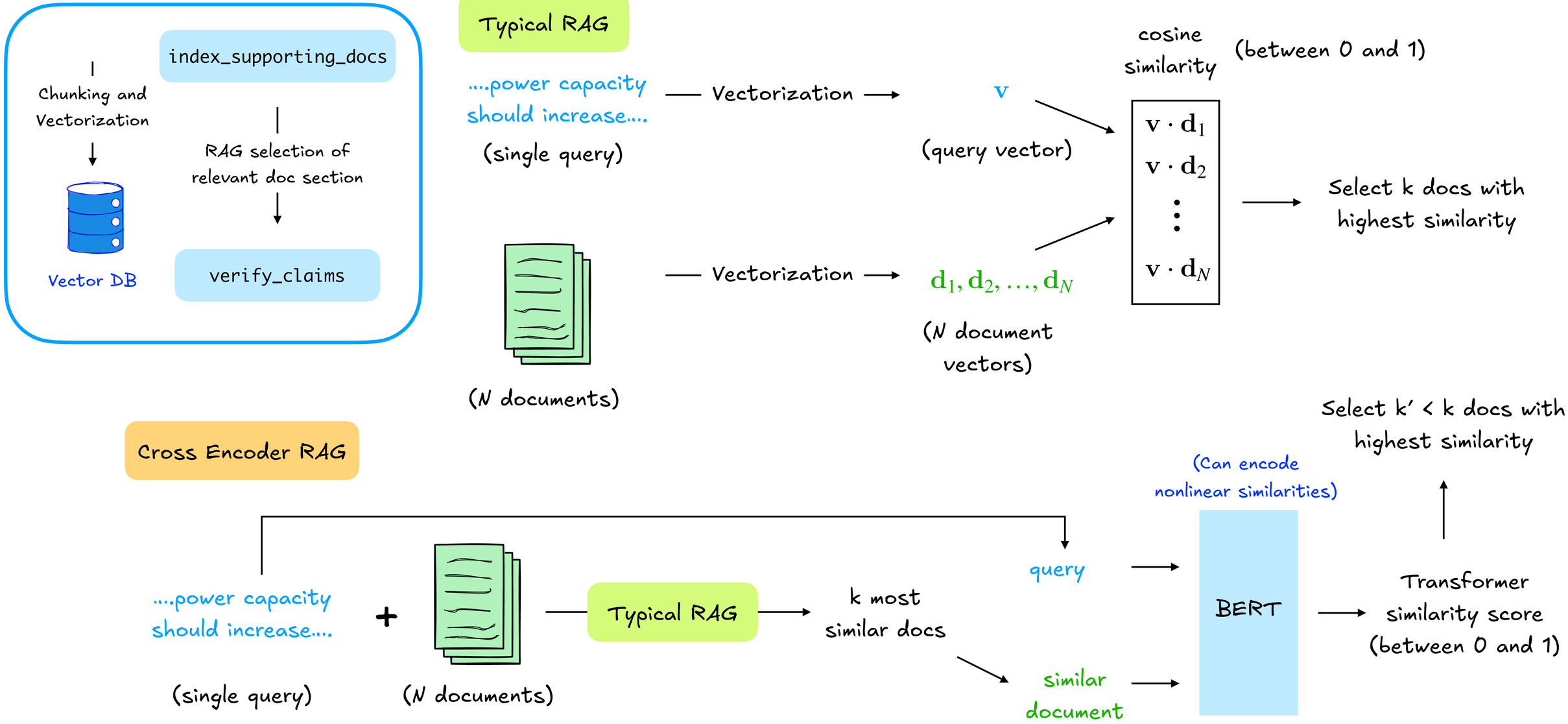
Applied Example

Statement	Logic Symbol
i. Power demands will increase by 2030	P
ii. Current infrastructure cannot meet demand	Q
iii. New electric power plants must be built	R

It is possible for P and Q to be true while R is false.

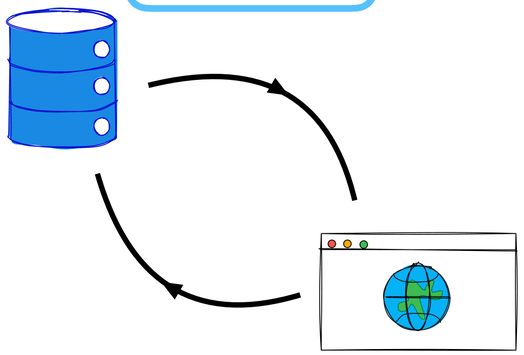
Invalid argument ✗

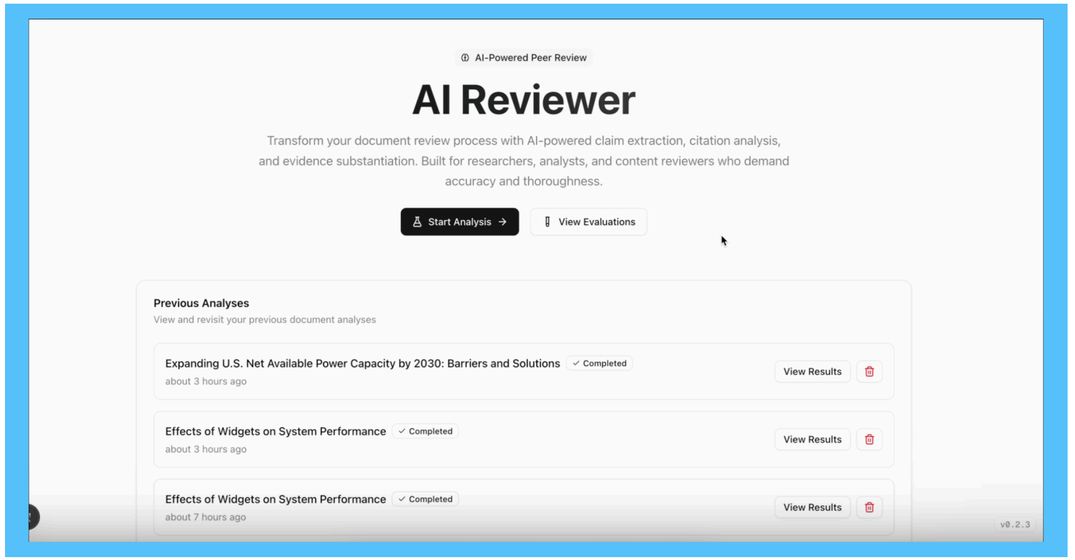
System Details: Verify Citations and Cross Encoder RAG



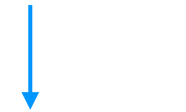
System Details: LangGraph Structure

System Architecture

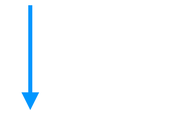




LangGraph Structure of Agents

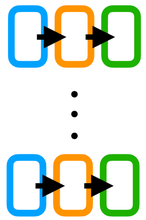


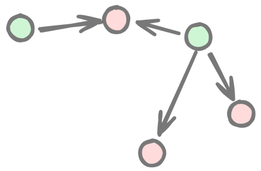
extract_claims

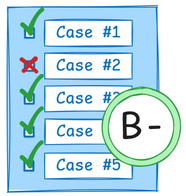


categorize_claims

Design Principles

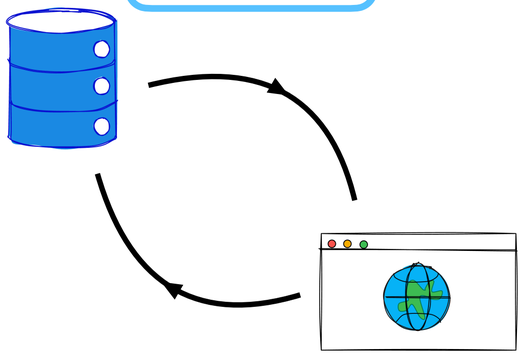


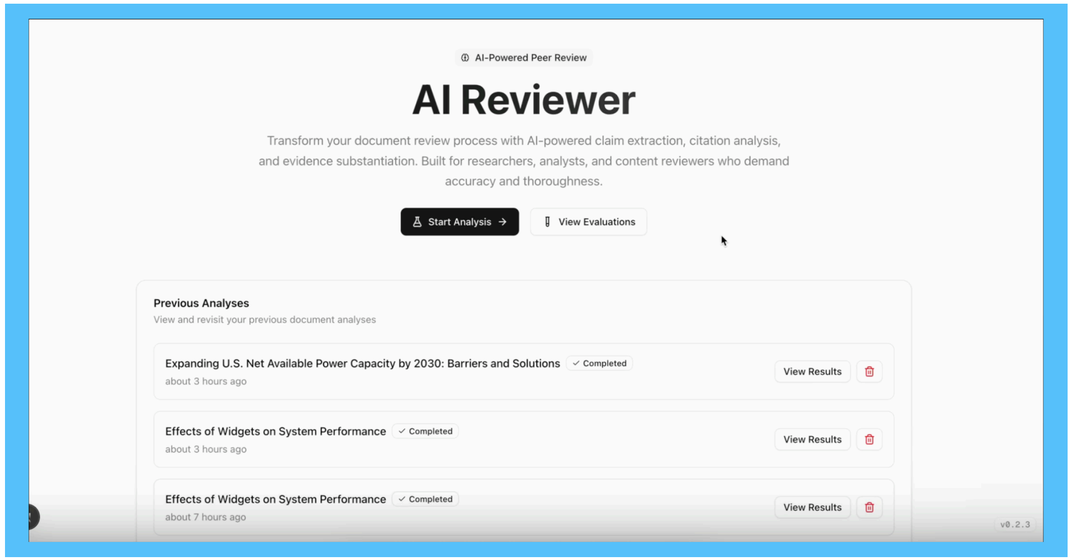




System Details: Design Principles

System Architecture



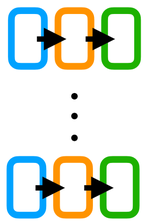


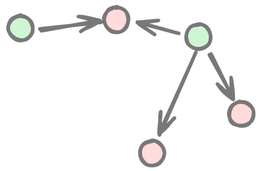
LangGraph Structure of Agents

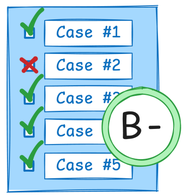
extract_claims

categorize_claims

Design Principles



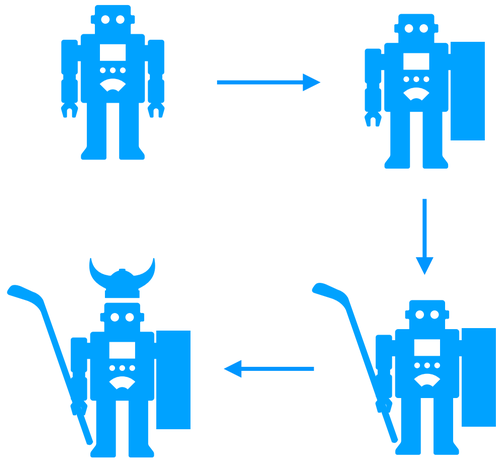




System Details: Design Principles

Why are design principles necessary?

No system is final...

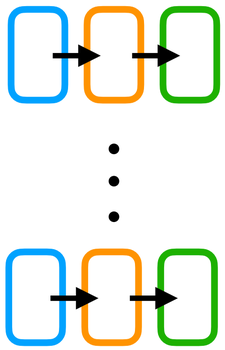


...so the created system must be able to (easily) handle future modifications

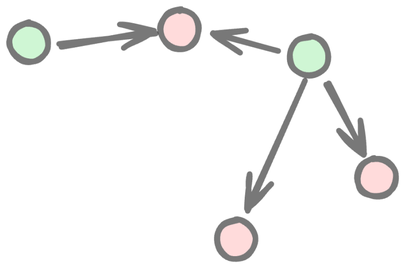
Design principles help ensure the system is built in a way that looks forward to possible change

"Design Principles"

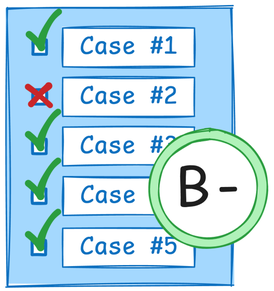
- (Uniformity) -



- (Granularity) -

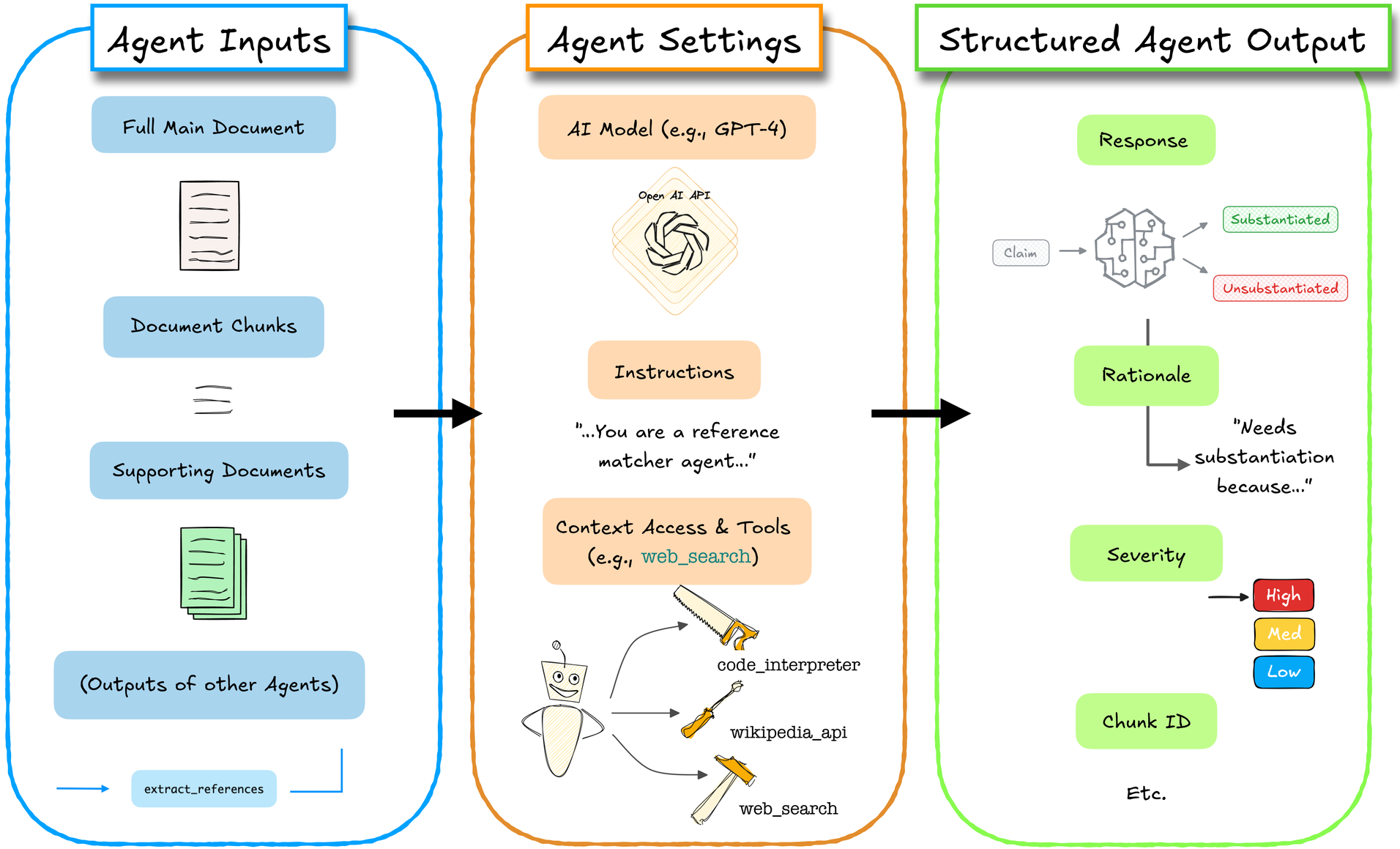


- (Visibility) -



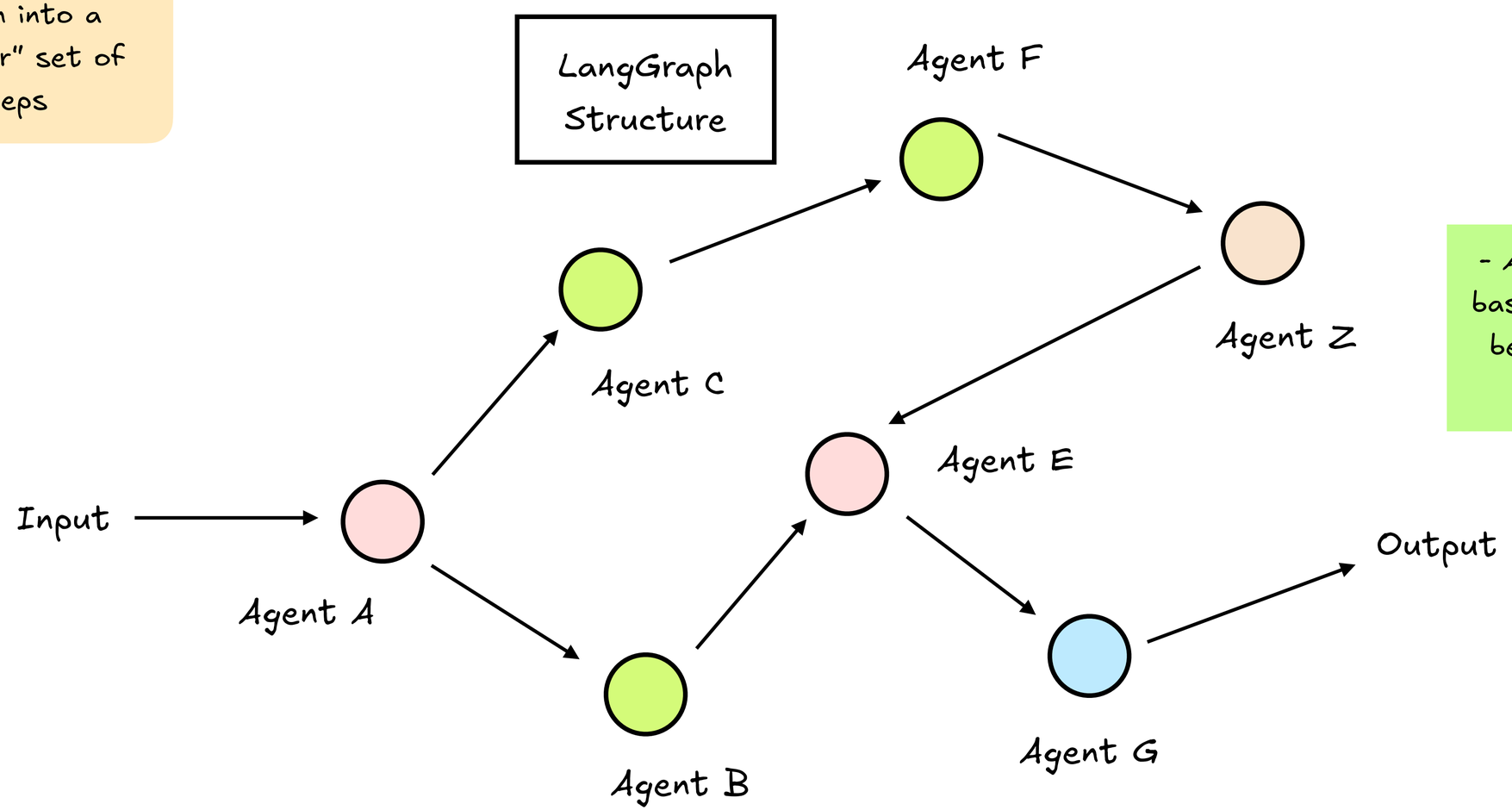
Design Principle: Uniformity of Information Flow

- (Uniformity) - Giving agents the same general input, settings, and output structure



Design Principle: Granularity of Actions

- (Granularity) - Large tasks are broken into a "granular" set of steps



- All agents have the same basic structure so they can be easily added, replaced, subdivided as needed

Design Principle: Visibility of Performance

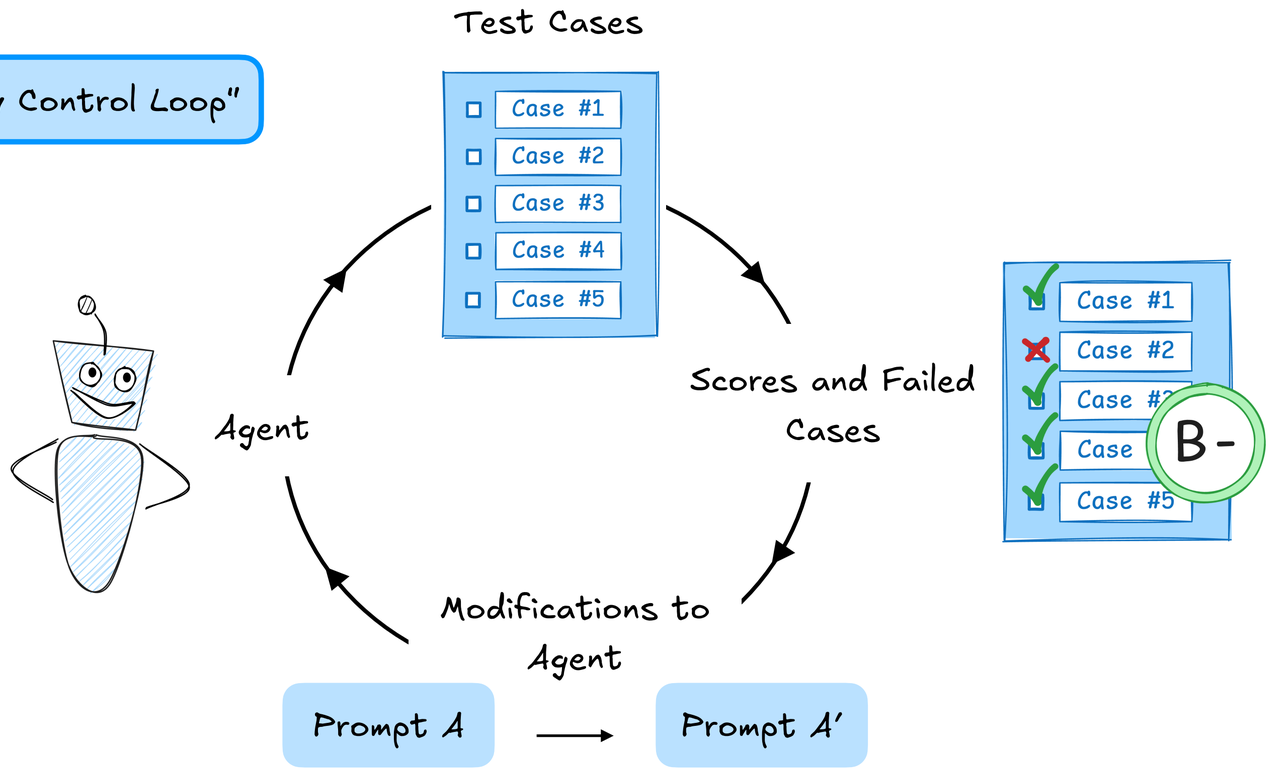
- (Visibility) -
Test cases are created for agents to understand failure modes AND effects of changes

*How often the same tests pass or fail

	Num Test Cases	Pass Rate	Consistency Prob
Agent A	20	80%	90%
Agent A'	20	85%	100%

gives us two main things

"Quality Control Loop"



Failure Visibility — Reveals typical "Failure modes" of agents

Controlled Change — Changes to agents are done through A/B testing

Presentation Roadmap

I. Motivation

II. Video Demo

III. System Details

-.i Architecture

-.ii Agents and LangGraph

-.iii Design principles

IV. Gaps and Next Steps

V. Live Demo & Q&A

What the system cannot do now

&

What we are planning to build in the future

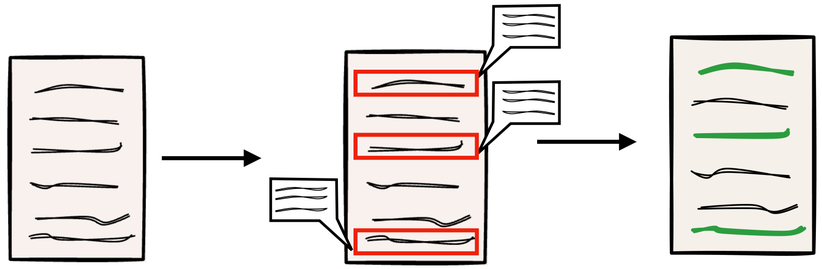
Gaps and Next Steps

What the system cannot do now

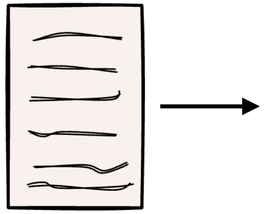
&

What we are planning to build in the future

Modify text "in-line" without user copy-pasting



Live Reports Addendum

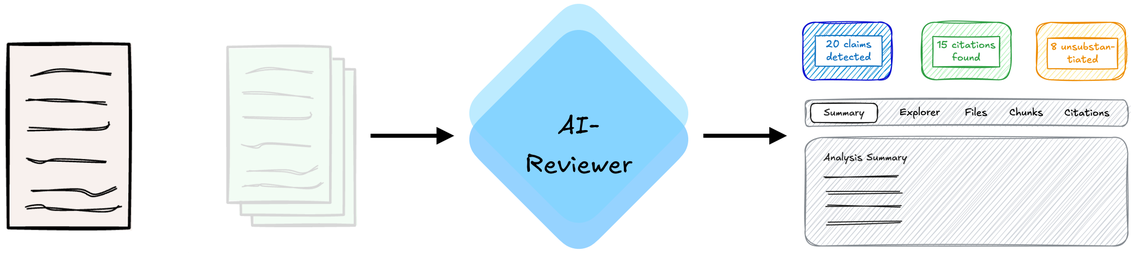


Live Reports
- "Update line 2 or paragraph 4 with...
- "Add new current reference...

Higher Level Coherence

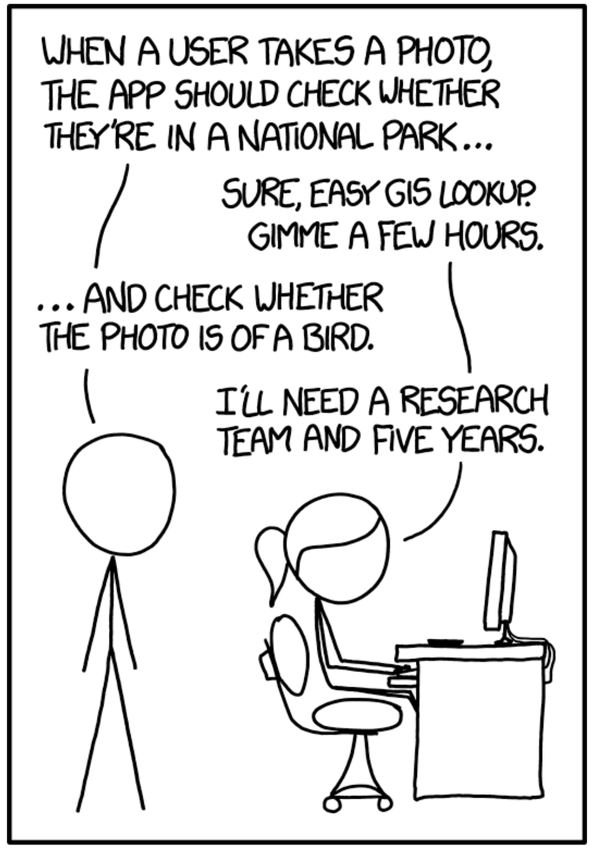
"The overall argument of the paper is....there are gaps in this argument from....

Conduct PDF/paper searchers to find references without user input

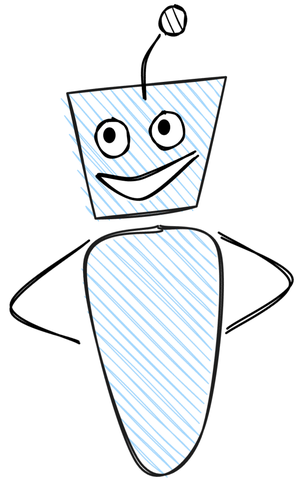


(No supporting documents)

Links before Live Demo



(https://xkcd.com/1425/)



Relevant Links

- AE Studio
  - (https://ae.studio/)

- AI Reviewer Github
  - (https://github.com/agencyenterprise/ai-reviewer)

- Slides
  - (Linked within repo)

Presentation Roadmap

I. Motivation

II. Video Demo

III. System Details

-.i Architecture

-.ii Agents and LangGraph

-.iii Design principles

IV. Gaps and Next Steps

V. Live Demo & Q&A

<redacted>

Live Demo

<redacted>

Extra Slides