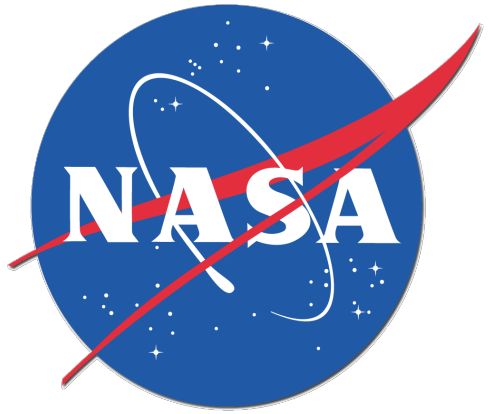
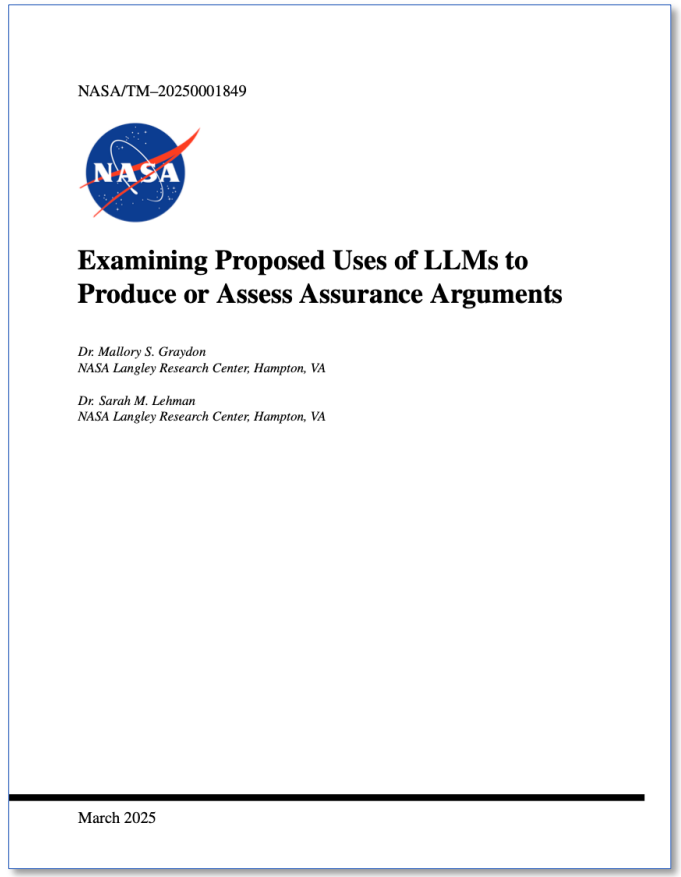
Examining Proposed Uses of LLMs to Produce or Assess Assurance Arguments



Dr. Mallory S. Graydon
Dr. Sarah M. Lehman
Safety Critical Avionics Systems Branch
NASA Langley Research Center, Hampton, Virginia, USA

Are LLMs fit for use in assurance cases?

- Seeing some buzz about A"I" in certification contexts
  - DARPA ARCOS
  - FAA *Roadmap for Artificial Intelligence Safety Assurance*
    - "AI is already being introduced …. Documentation supporting the certification processes that currently exist can also be auto generated through similar applications."
- We decided to see whether there was any science behind the buzz



NASA/TM–20250001849

Examining Proposed Uses of LLMs to Produce or Assess Assurance Arguments

Dr. Mallory S. Graydon
NASA Langley Research Center, Hampton, VA

Dr. Sarah M. Lehman
NASA Langley Research Center, Hampton, VA

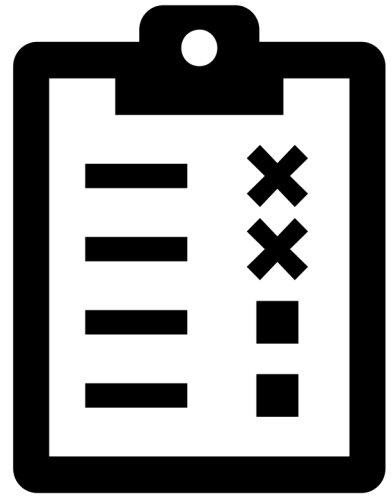
March 2025

Literature survey

- Literature survey identified 14 works about assurance cases / assurance arguments and and LLMs/GPT
  - 4 WS papers, 4 preprints, 3 conf. papers, 2 theses, & 1 pos. paper
- Three proposed uses:
  - Generation of arguments or parts of arguments (7)
  - Generation of potential argument defeaters (6)
  - Formalization of parts of argument for use in other analyses (4)
- Motivations:
  - Human fallibility
  - Expense (driven by the difficulty of data gathering and thinking)

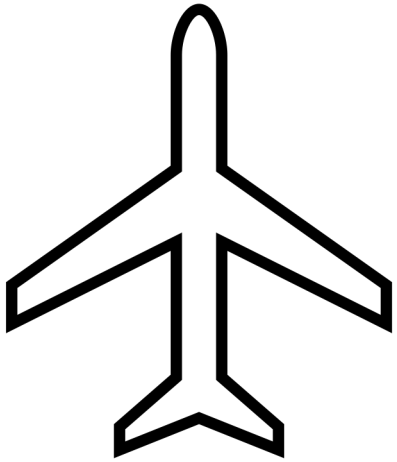
How efficacy was defined and shown

- 8 papers described an efficacy study
- Three broad kinds of studies:
  - ***Knowledge checks*** to see if LLM is competent by quizzing it about notations
  - ***Replication checks*** to see if LLM can produce output like an existing human-generated artifact
  - ***Comparative performance studies*** to assess which LLM or prompting technique works best



Knowledge checks

- Four works assessed competency by quiz
- One group of authors asked 13 knowledge and 6 "generation-based" questions, e.g.:
  - "How many elements are present in a goal-structure and what are they? Can a parent element have multiple children?"
  - "Give me a sample goal element connected to 2 sub-goals"
- But … the answers are in manuals … online
- And then there's the *reification fallacy* here

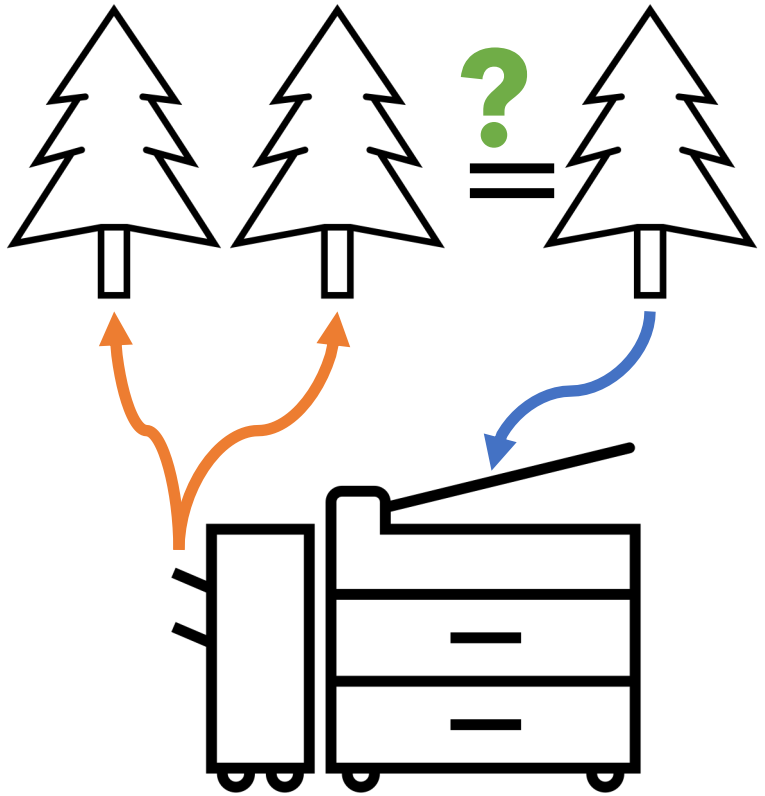


Ceci n'est pas un avion

With apologies to René Magritte

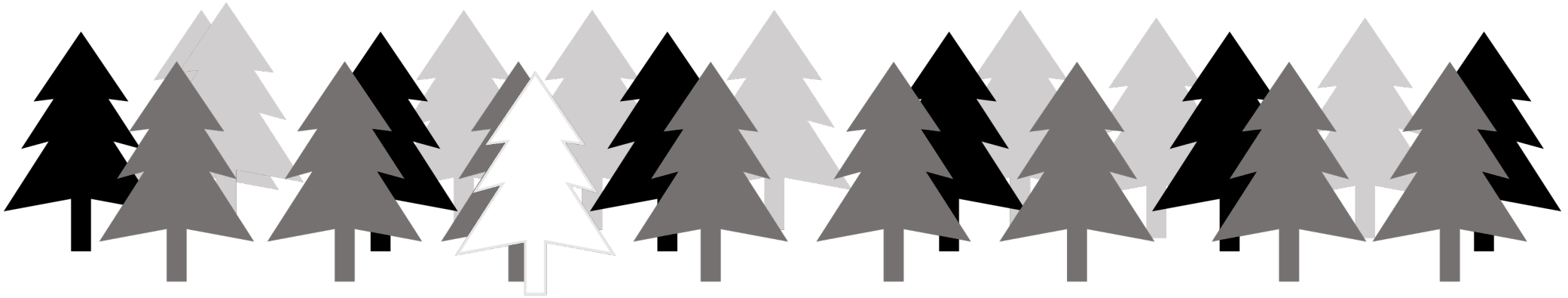
Replication checks

- 6 works tested whether LLMs could replicate human-generated argument fragments or defeaters
  - E.g., one group asked an LLM to find defeaters for 58 propositions in an argument about the Large Hadron Collider
    - Human assessors rated "complete match, partial match, or no match"
    - The LLM "completely identified more than half of all defeaters"
  - Others found the LLM didn't produce the "right number" of defeaters (which is???)



With apologies to Alvy Ray Smith ...

A Safety Argument is **Not** a Big Tree,
A Safety Argument is **Not** a Big Tree,
A Safety Argument is **Not** a Big Tree!
(And Evidence is **Not** an Artifact)



Threats to validity of replication checks

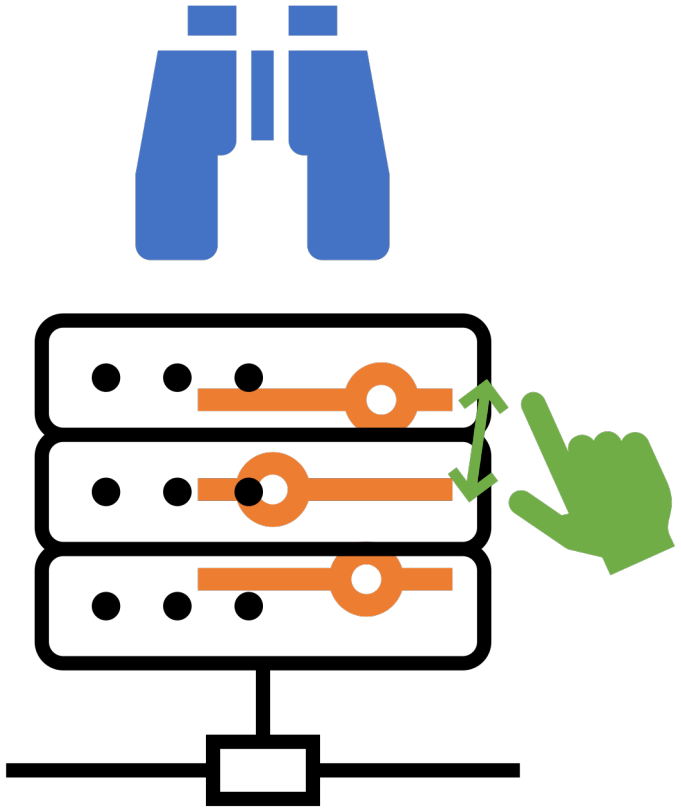
- 6 works tested whether LLMs could replicate human-generated argument fragments or defeaters … but:
  - The LLM might just be reproducing examples in its training data
    - The examples came from papers that predate the LLMs' training cut-off date
  - Are these good examples to replicate?
    - These "safety cases" were small examples from academic papers
    - Real safety assurance needs a lot that isn't mentioned in typical academic toy examples: see ARP4754B, ARP4761A, DO-178C, DO-254, etc.
- Bottom line: *there is no evidence that LLMs have an ability that generalizes to full safety arguments for any domain*

Aside: What are "defeaters"?

- Literature focuses on things that would undermine arguments or rebut conclusions … and fair enough
- Published examples, though, are sometimes failure conditions or contributions to them, e.g., power failure
- This raises a question: *How does the process of finding defeaters relate to system safety analysis (SSA)?*
  - Does it *augment safety analysis*? Or replace it?
  - *Is it as effective* (and cost-effective) as whatever it replaces?
  - How much value does ID'ing defeaters add to SSA? At what cost?

Comparative performance studies

- Comparative performance studies aim to figure out which LLM works best and how best to prompt it
  - Study output data looks like replication studies: judge whether/how well the output matches a baseline artifact created by humans
  - But do this with different LLMs and/or different prompting techniques
- Can be used to refine proposed techniques
- Can also be used as an assessment



LLMs don’t always tell the truth

A safety case for **PES**, or **Personal Protective Equipment**, can help ensure a safe and healthy work environment.

A safety case typically includes:

- Claim: A statement about a property of the system or subsystem
- Evidence: Facts, assumptions, or sub-claims that support the claim
- Argument: A link between the evidence and the claim
- Inference: The mechanism that provides the rules for the argument

Here are some steps for building a safety case:

1. State the purpose and scope of the safety case
2. Develop and document the safety objectives and system
3. Develop a safety risk management methodology
4. Process for hazard identification and analysis
5. Establish the safety risk of each hazard

PPE can include masks, gloves, and full body suits. It’s important to be able to determine the need for and proper use of PPE to minimize the chances of exposure to workplace hazards.

Excerpt from the AI overview result for a Google search on safety cases for PES (programmable electronic systems, presumably)

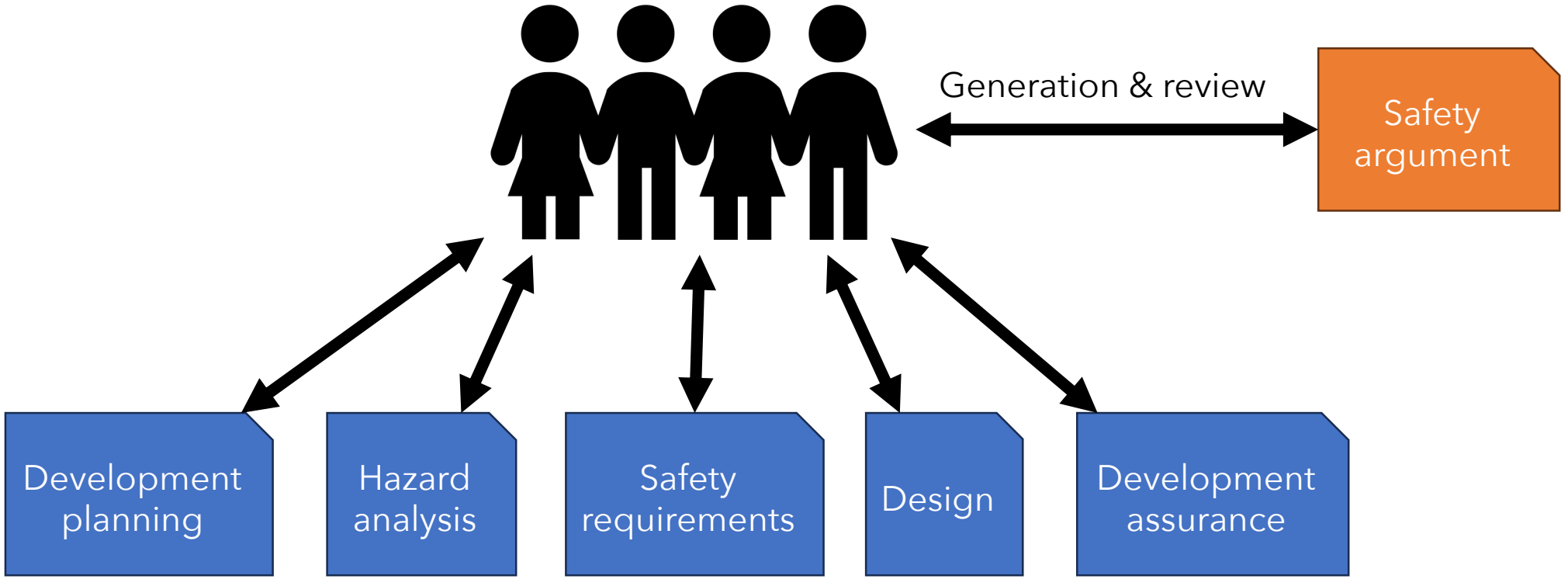
LLMs as producers of Frankfurtian BS

- In 1986, Harry Frankfurt wrote *On bullshit*
  - BS is not a lie: it is *speech/text produced without regard for the truth*
- In 2024, Hicks and Slater asserted that *ChatGPT is bullshit* in the Frankfurtian sense
  - LLMs "have been plagued by persistent inaccuracy in their output; these are often called 'AI hallucinations.' **We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit*.**"

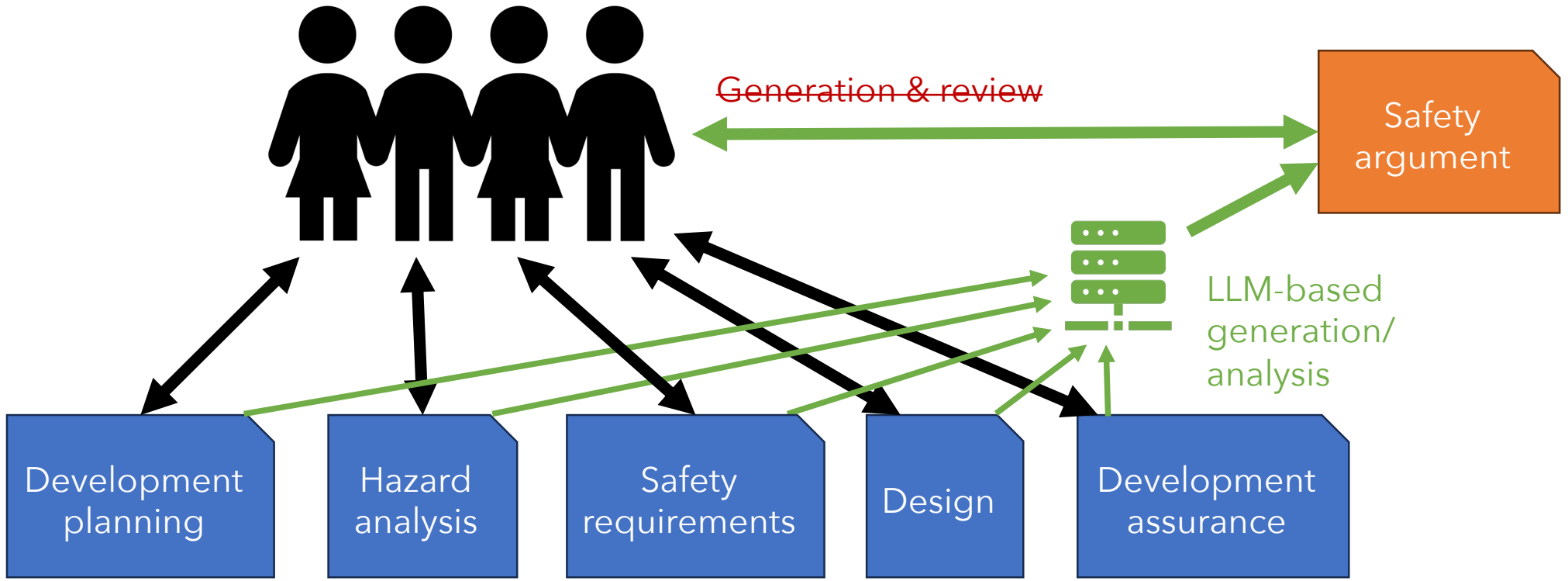
LLMs as producers of Frankfurtian BS

- Understanding LLMs as BSers—or *stochastic parrots*—primes us to be appropriately skeptical
  - LLMs probabilistically replicate patterns they absorbed from their training data
  - But these patterns are not *principles*, let alone knowledge about where principles apply (or don't)
  - We can expect LLMs to:
    - Struggle with edge cases
    - Replicate problems in training data
  - *Humans have to supervise the LLMs' output!*

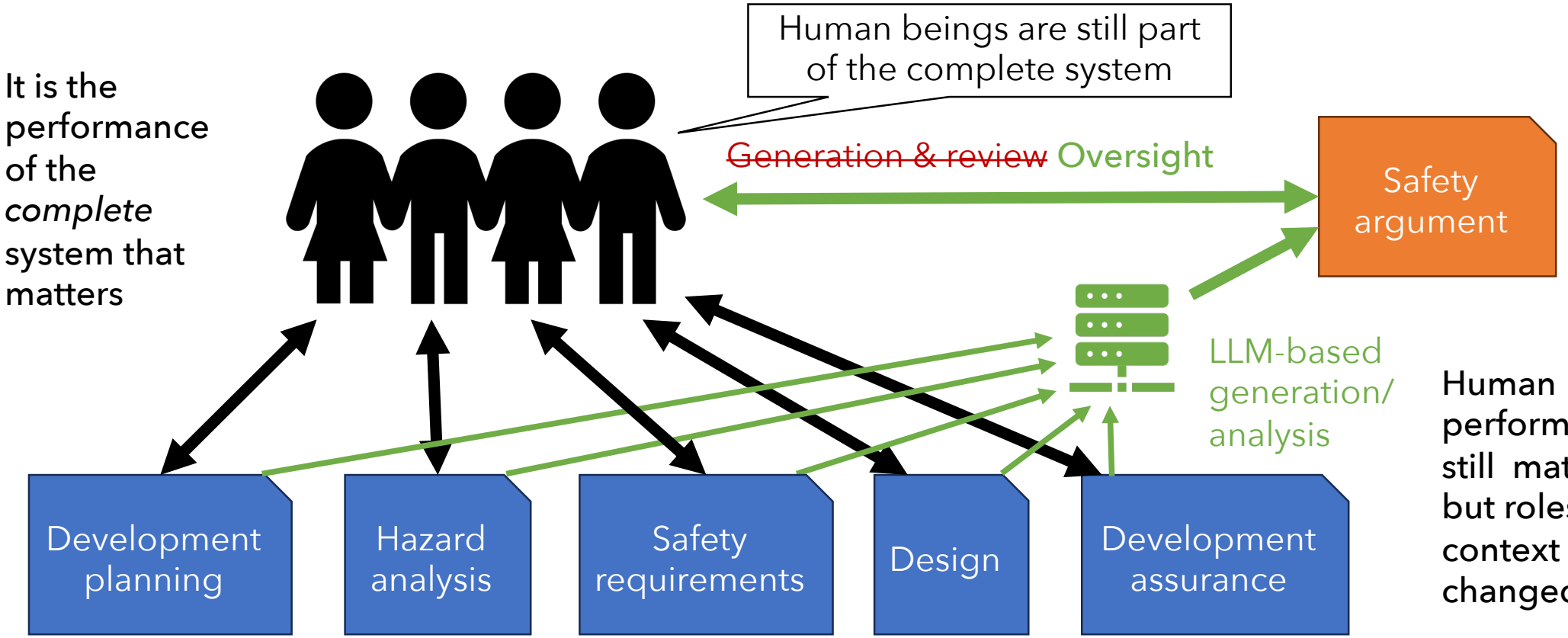
Humans are part of the system



Humans are part of the system^?



Humans are ˆ*still* part of the system~~?~~



Questions that need to be answered

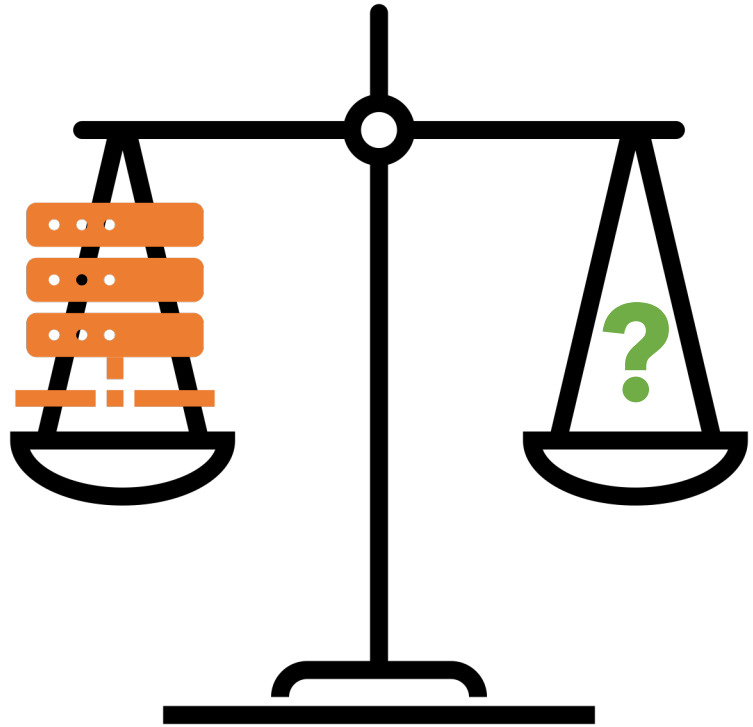
- We walk through the safety process and identify 14 questions about impacts on:
  - System design
  - Oversight of system design
  - Argument readability
  - Quality of "evidence" citations
    - When system quirks should prompt use of different evidence
    - Familiarity to readers
  - Noticing counterevidence
- …
  - Seeking more evidence when it is necessary/available
  - Readability of argument text
  - Following best practices
  - Arguing to the right depth
  - Noticing counterargument
  - Argument assessment
  - Total cost of certification

Questions about efficacy

- Efficacy is *really* hard to measure …
- Not only do LLMs BS, but it's not enough to assume "humans err, therefore automation is good"
  - Supervision is a different task; we don't know how well humans do it
  - *The thinking is the hard part* … and you can't automate it away
  - Worse, *automation threatens to make the necessary thinking both more difficult to do right and more tempting to short-change*
    - It might take humans "out of the loop"
    - It might tempt humans to take short-cuts for productivity
- We need to experimentally study this human performance!

Efficacy relative to what?

- To be "adequately" effective is to be at least as effective as relevant alternatives
  - Manual human generation of argument / assessment of defeaters is an obvious alternative
  - But it might not be the only one …
    - If LLMs absorb patterns of argument, we might arm humans with patterns of argument
    - If LLMs absorb patterns of defeaters, we might arm humans with checklists of defeaters



Questions about cost

- Efficacy was the biggest motivator in selected papers (12/14)
  - But is the output of a BS generator any better than fallible humans?
- Cost was the second-biggest (6/14)
  - But what can you actually save?
  - You can't automate away the *thinking* … and that's the hard part
  - How much can you save by automating away some typing?
  - *Is there even enough potential savings to warrant risking safety?*

Conclusions

- There have been proposals to use LLMs …
  - … to write (parts of) assurance arguments
  - … to propose defeaters for assurance arguments
- The studies done to date have not demonstrated efficacy
- There are good reasons to think this idea might decrease safety or fail to lower costs
- There are questions left to be answered
- LLM-based assurance argument tools remain “experimental”


Big Bethel Rd