Can Large Language Models Assist in Hazard Analysis?

WAISE 2023 – Toulouse

University of Victoria, British Columbia, Canada

Simon Diemert, M.Sc., P.Eng., Jens H. Weber, Ph.D., P.Eng.

Presenter

Simon Diemert is a Software and Systems Engineer at Critical Systems Labs Inc. and a PhD student at the University of Victoria in Canada. He has contributed functional safety and software engineering expertise to numerous projects with CSL, including the development and analysis of autonomous automotive and rail systems. Simon's PhD research is focused on assurance methods for complex systems, including those involving AI/ML. Simon holds a Master of Computer Science and Bachelor of Software Engineering and is registered as a Professional Engineer with Engineers and Geoscientists of British Columbia.

Hazard Analysis: Think think think…

- Hazard analysis is a creative exercise!
- How do we know the analysis is **complete**?
  - Have identified all the hazards, hazard causes, failure modes, system interactions, and consequences?

Large Language Models

GPT3

175 B params
/300 B tokens

LlaMA 2

70 B params
/2 T tokens

PaLM 2

340 B params
/3.6 T tokens

https://en.wikipedia.org/wiki/Large_language_model#List

Check for updates

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• **Computing methodologies → Natural language processing.**

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language

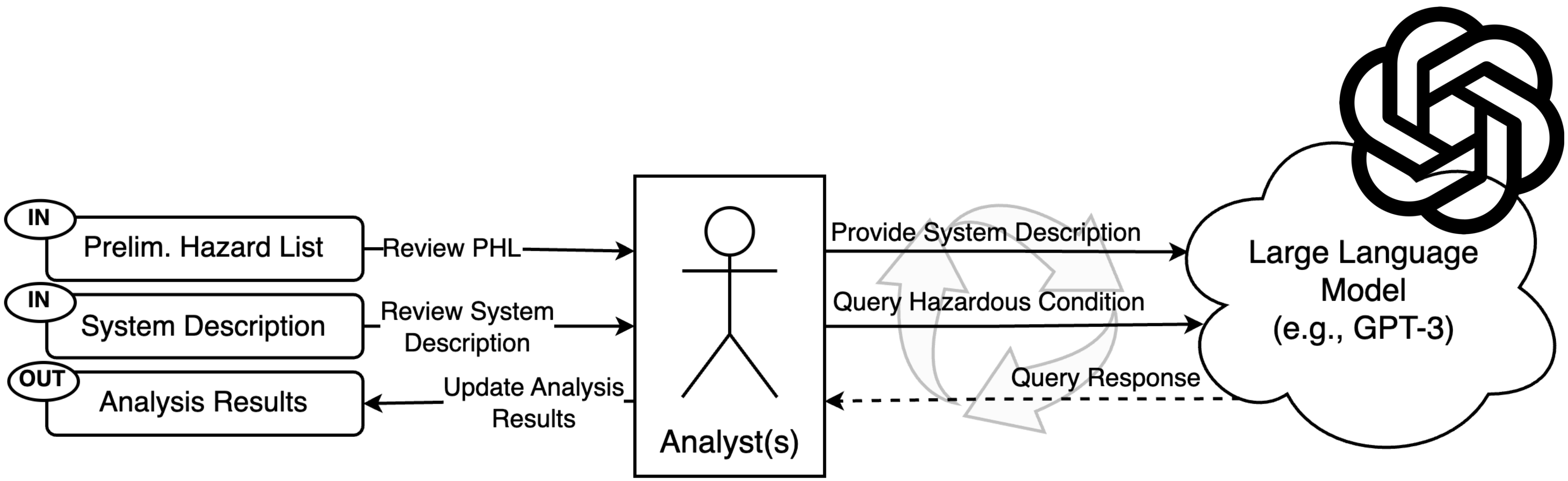
alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

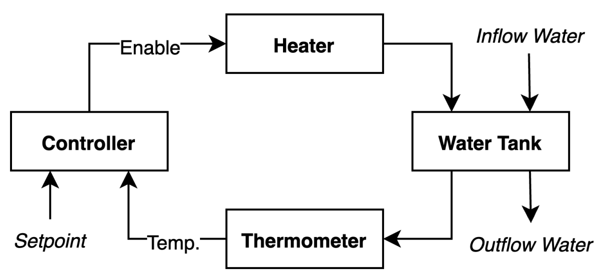
We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger

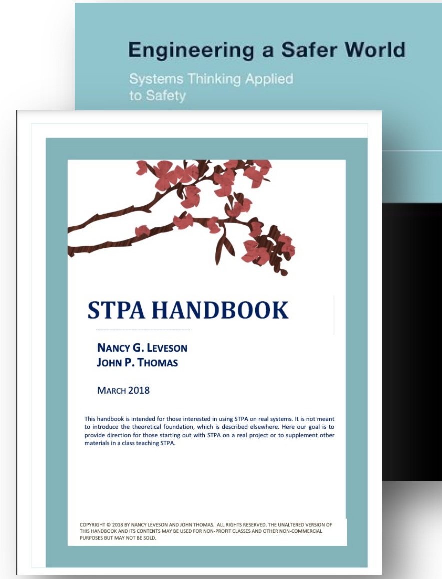
E. M. Bender, *et al.*, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT ’21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623.

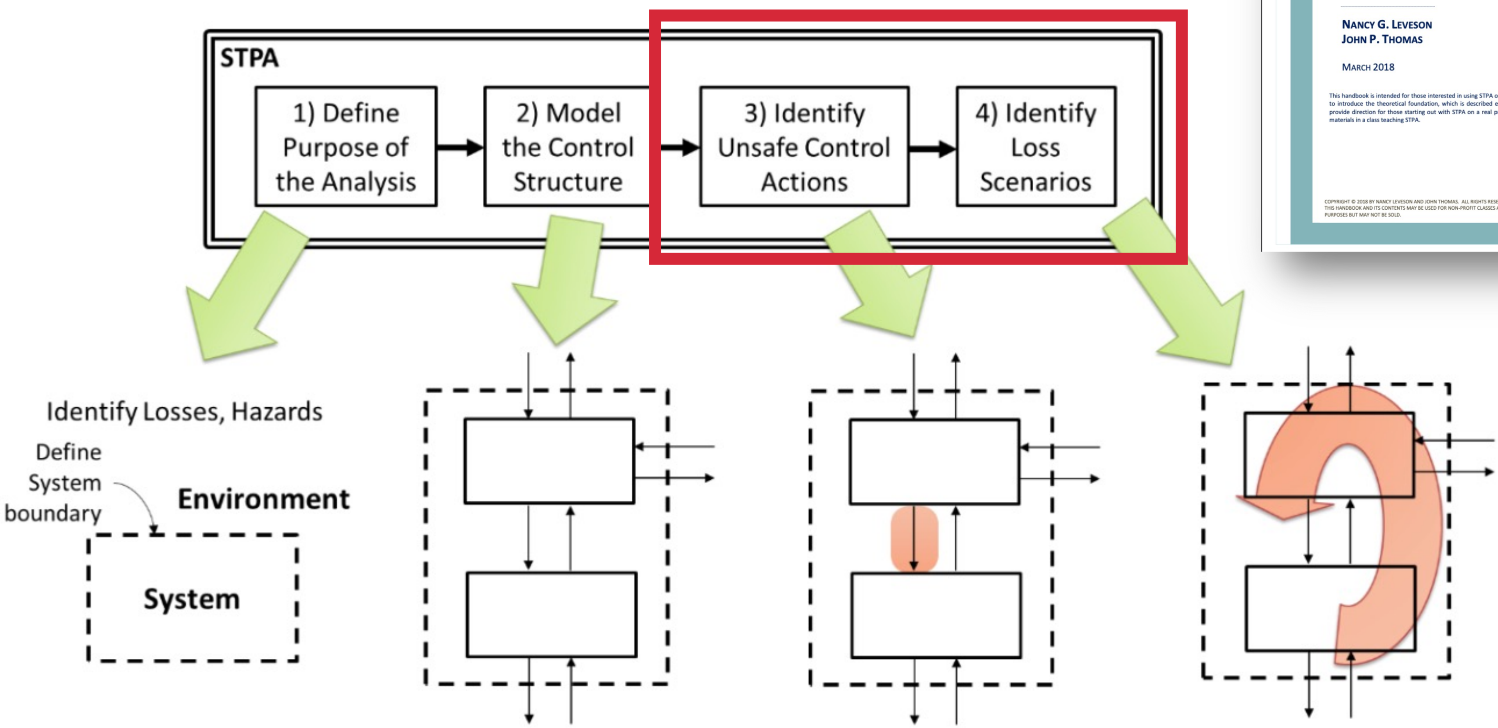
Co-Hazard Analysis





System Theoretic Accident and Processes





N. G. Leveson and J. P. Thomas, "STPA Handbook," 2018.

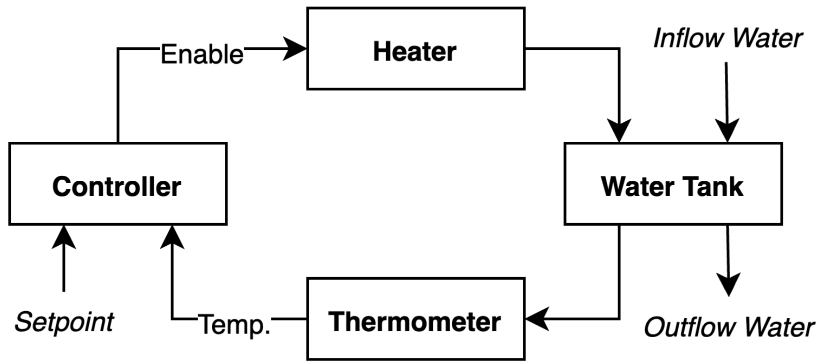
Experiment – Research Questions

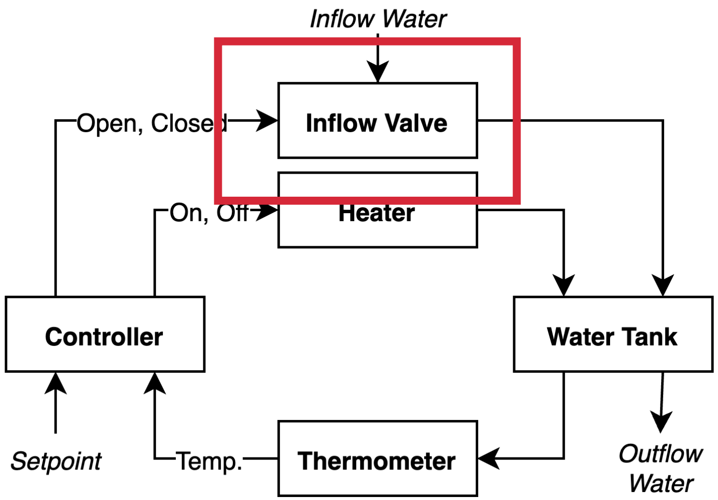
- **RQ1 (Feasibility)** – Can an LLM produce results that are useful to human analysts identifying UCAs and causal scenarios?
- **RQ2 (Utility)** – What proportion of responses are *useful and correct* v. *incorrect* v. *not useful*?
- **RQ3 (Scalability)** – Does the response quality degrade as system under analysis increases in complexity?

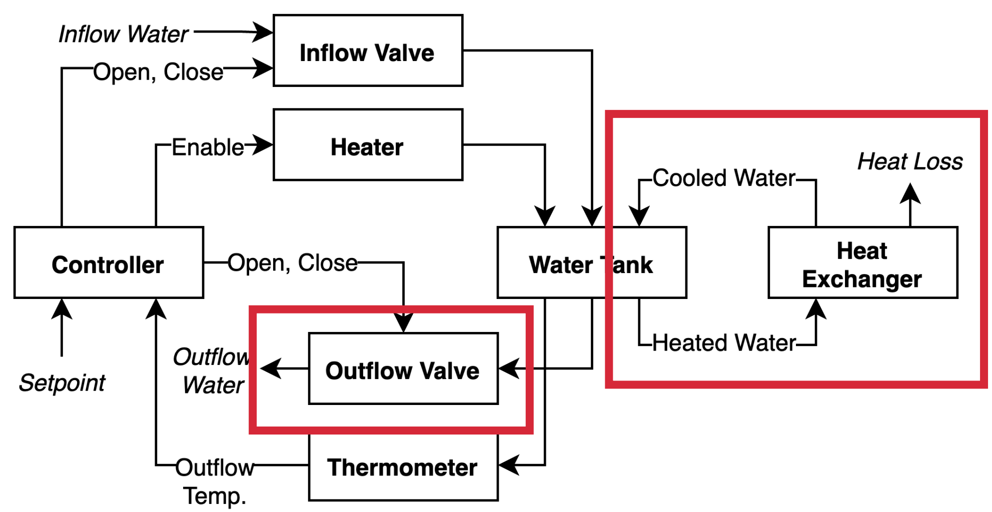
Experiment – Method Overview

1. System selection
2. System encoding
3. Querying the LLM
4. Manual response review and coding
5. Analysis of results

Experiment – System Selection







Experiment – System Encoding

Consider a system consisting of a Controller, Heater, Water Tank, and Thermometer.

List of system elements

The Controller provides the enable signal to the Heater to maintain a temperature setpoint. While the Controller is providing the enable signal to the Heater, the Heater heats the water in the Water Tank. When the Controller stops providing the enable signal to the Heater, the Heater does not heat the water in the Water Tank. The Thermometer measures the current water temperature inside the Water Tank. The Thermometer provides the current temperature of the water flowing out of the Water Tank to the Controller.

Description of relationships

The water flowing into the tank has variable temperature between 5 and 60 degrees Celsius. The ambient temperature is above 0 degrees Celsius. Water flows in and out of the tank at the same rate.

Additional constraints and assumptions

Critical Systems Labs
Innovating Safely

Experiment – System Encoding

A dangerous event occurs if the temperature of the water flowing out of the tank is greater than 90 degrees Celsius.

Hazard 1

A dangerous event occurs if the water level in the Water Tank exceeds the Water Tank's capacity.

Hazard 2
(system 2 & 3)

There are no more dangerous events.

Closed world
assumption.

Experiment – Querying the LLM

Control Action	Provided	Not Provided	Too Early / Late	Stopped to Soon / Applied too Long
Controller - Enable				

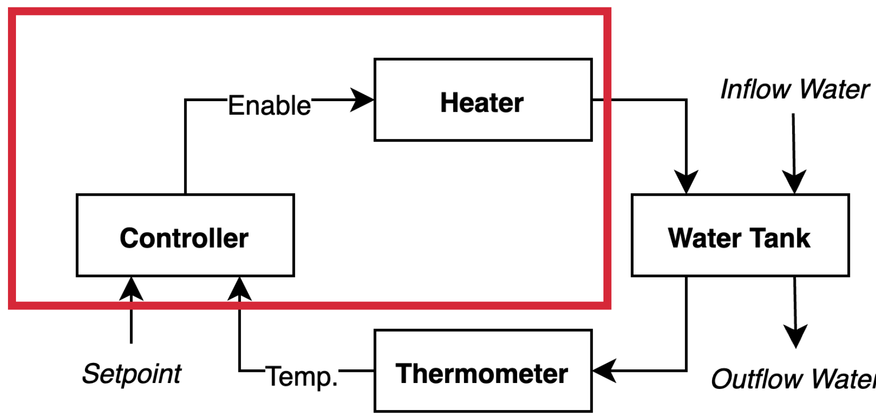
Could Element doing guideword & control action result in dangerous event ?

Critical Systems Labs
Innovating Safely

Experiment – Querying the LLM

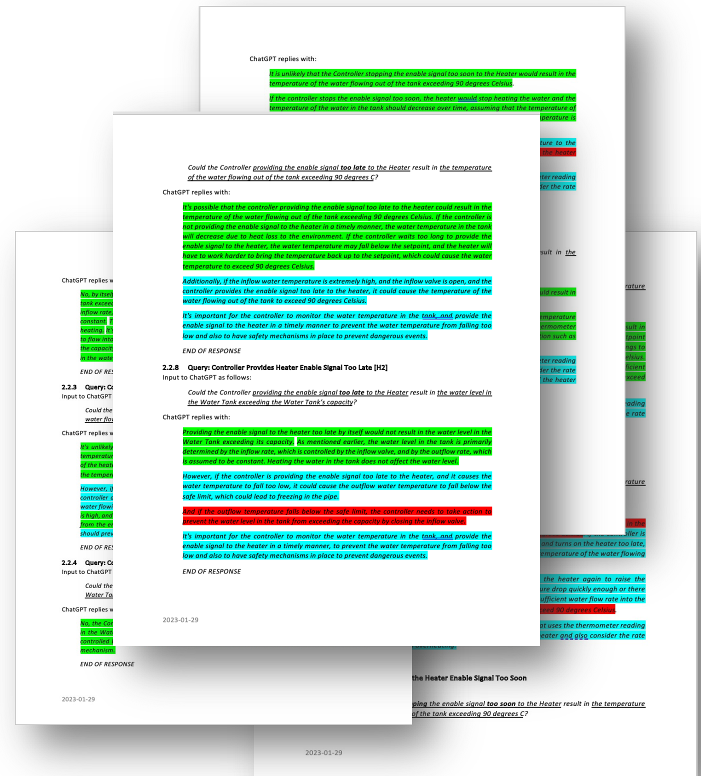
Control Action	Provided	Not Provided	Too Early / Late	Stopped to Soon / Applied too Long
Controller - Enable	...	...	?	...

Could the Controller providing the enable signal too early to the Heater result in the temperature of the water flowing out of the tank exceeding 90 degrees C?



Experiment – Coding

- Coded at “word level”.
- Pair coding:
  1. Independently code
  2. Review and discuss differences
  3. Optionally change codes
- Three codes:
  - Correct and useful
  - Correct but not useful
  - Incorrect



Experiment – Coding Example

2.2.17 Query: Controller Provides Open Command to Inflow Valve Too Early [H1]

Input to ChatGPT as follows:

> *Could the Controller providing the open command to the Inflow Valve **too early** result in the temperature of the water flowing out of the tank exceeding 90 degrees C?*

ChatGPT replies with:

> *No, providing the open command to the Inflow Valve too early will not cause the temperature of the water flowing out of the tank to exceed 90 degrees C. The temperature of the water in the tank is primarily determined by the heater and the ambient temperature. The inflow valve controls the water flow rate into the tank, it does not affect the temperature of the water flowing out of the tank.*
>
> *The water level in the tank will increase due to the water flow rate. If the controller is monitoring the water level, it can adjust the inflow valve and the heater accordingly to maintain the temperature within the safe range.*

Results – Basic Experimental Data

Table 1. Results by system version (standard deviation in parentheses).

Complexity	# Queries	Words per Response	Total Words	Agreement
Lowest	6	143.7 (34.4)	862	0.77
Moderate	28	122.7 (51.1)	3436	0.91
Highest	44	116.0 (48.5)	5016	0.89
Overall	78	120.6 (48.6)	9404	0.89

Post discussion Cohen's Kappa (moderate – strong agreement)

Results – RQ1 (Feasibility)

Can an LLM produce results that are useful to human analysts identifying UCAs and causal scenarios?

- ChatGPT’s responds to queries about potentially unsafe control actions with **useful and correct information with moderate frequency (64% of the time)**. Therefore, it is feasible to use ChatGPT to support STPA.

- However, ChatGPT’s responses are likely to also contain correct but not useful and incorrect information and so responses **must be scrutinized by a human analyst**.

Table 2. Number of responses with at least one word coded in each category.

Complexity	Correct and Useful	Correct but not useful	Incorrect
Lowest	5 (83%)	5 (83%)	3 (50%)
Moderate	20 (71%)	23 (82%)	14 (50%)
Highest	25 (56%)	32 (72%)	29 (70%)
Overall	50 (64%)	60 (77%)	46 (59%)

Results – RQ2 (Utility)

What proportion of responses are *useful and correct* v. *incorrect* v. *not useful*?

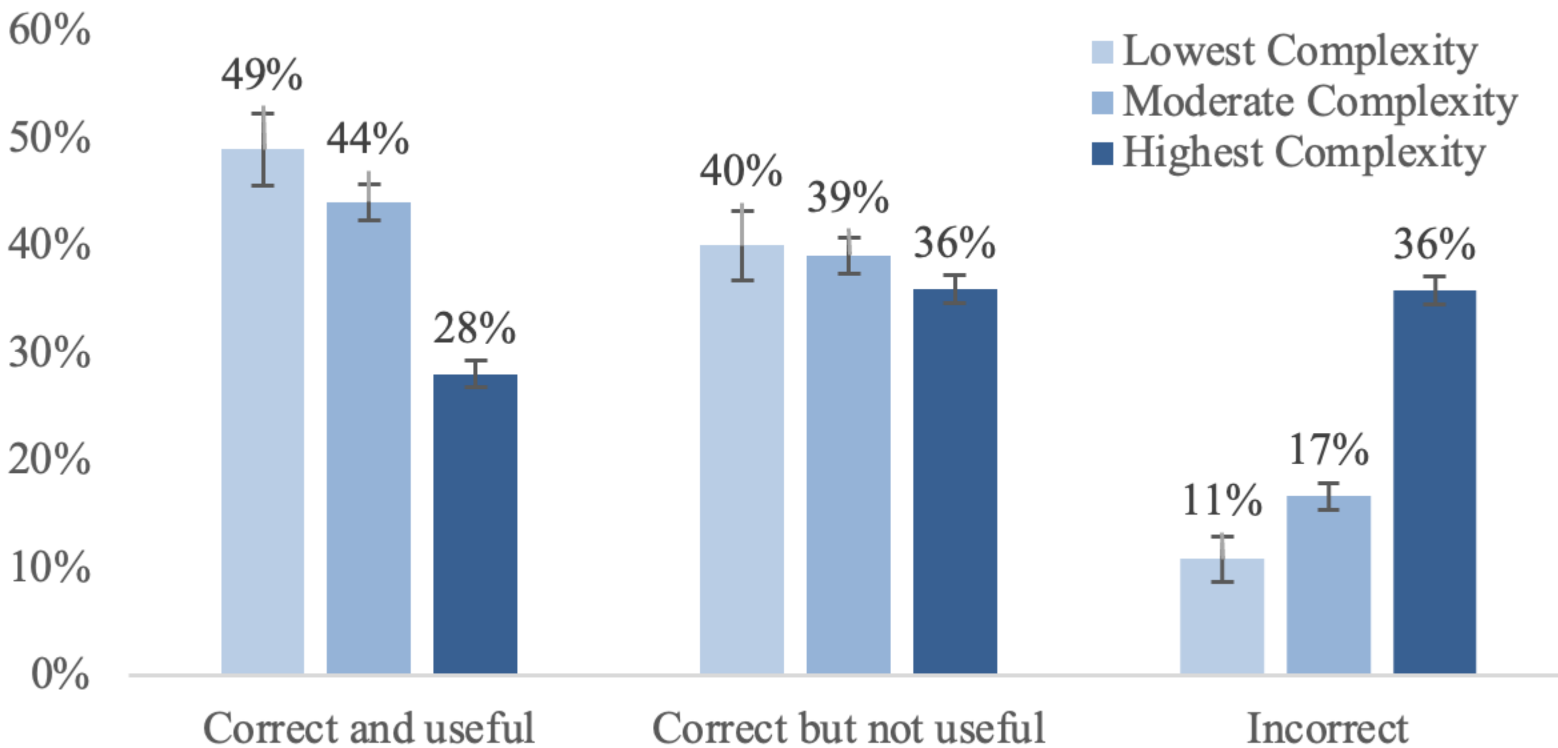


Fig. 4. Proportion of codes assigned by system complexity (95% CI shown).

Results – RQ2 (Utility)

What proportion of responses are *useful and correct* v. *incorrect* v. *not useful*?

- When used for CoHA with STPA, between **one quarter to one half of the content** in ChatGPT's responses is correct and useful information.
- Human analysts performing CoHA with ChatGPT will have to **sift through responses** to find it.
- ChatGPT's responses were **moderately useful**.

Results – RQ3 (Scalability)

Does the response quality degrade as the system under analysis increases in complexity?

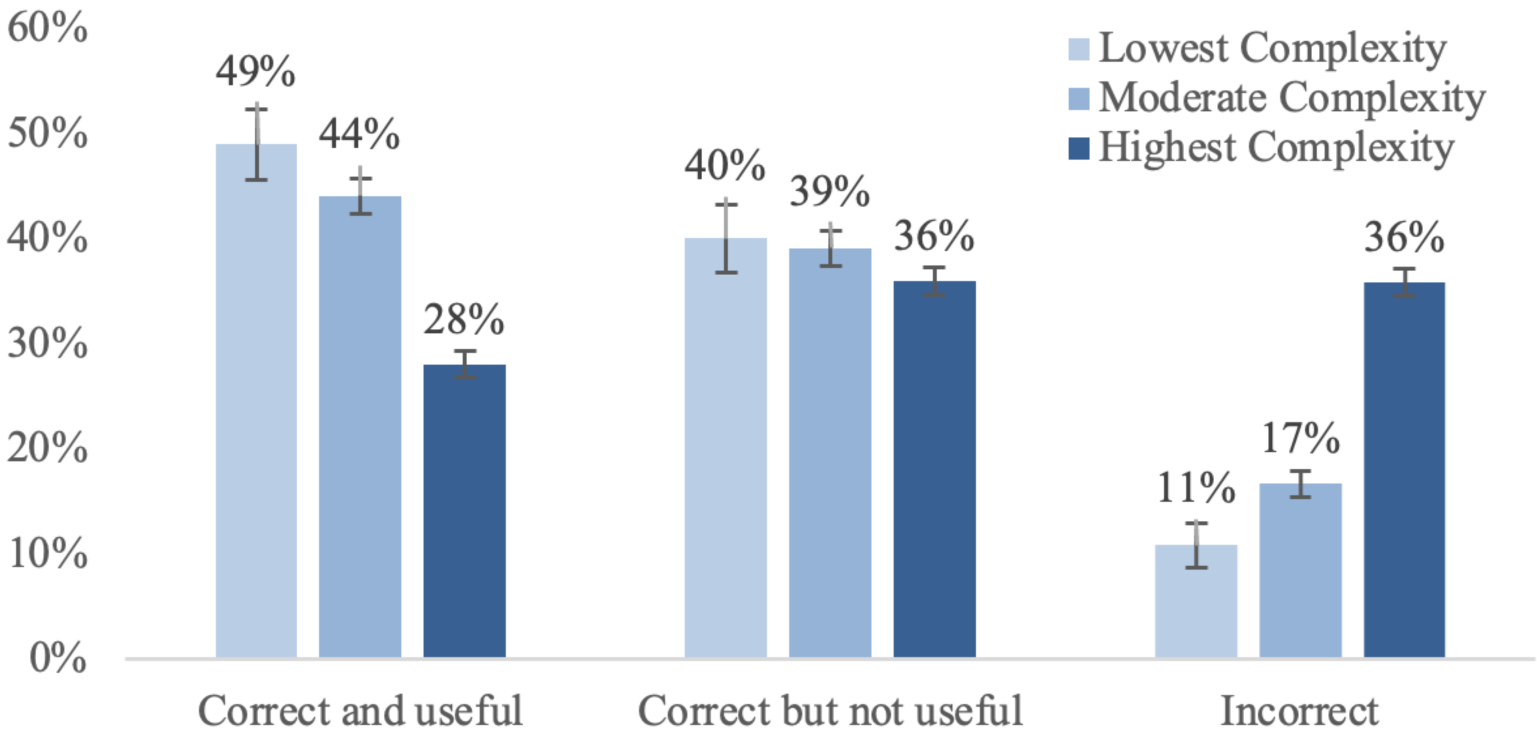


Fig. 4. Proportion of codes assigned by system complexity (95% CI shown).

Results – RQ3 (Scalability)

Does the response quality degrade as the system under analysis increases in complexity?

Table 4. Tests for significance between different system complexities.

Measure	$H_0 : \hat{p}_x = \hat{p}_y$	$\hat{p}_x$	$\hat{p}_y$	$\hat{p}_y - \hat{p}_x$	Outcome
Proportion of Incorrect Words in Responses	$\hat{p}_{low} = \hat{p}_{mod}$	0.11	0.17	0.06	Reject H_0 ($p < 0.01$)
	$\hat{p}_{mod} = \hat{p}_{high}$	0.17	0.36	0.19	Reject H_0 ($p < 0.01$)
	$\hat{p}_{low} = \hat{p}_{high}$	0.11	0.36	0.25	Reject H_0 ($p < 0.01$)
Proportion of Useful Words in Responses	$\hat{p}_{low} = \hat{p}_{mod}$	0.55	0.53	−0.02	Do Not Reject H_0
	$\hat{p}_{mod} = \hat{p}_{high}$	0.53	0.44	−0.09	Reject H_0 ($p < 0.01$)
	$\hat{p}_{low} = \hat{p}_{high}$	0.55	0.44	−0.11	Reject H_0 ($p < 0.01$)

H_0: The proportion of words coded as 'incorrect' is equal between the lowest and moderate complexity systems. ← Null Hypothesis Rejected!

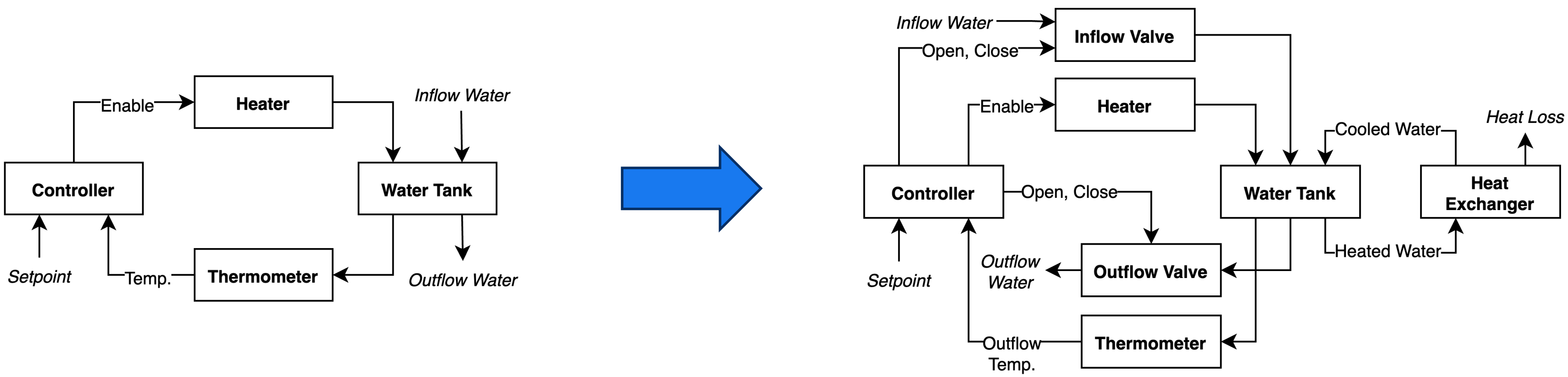
Six pair-wise statistical (two-tailed) tests for significance between population proportions ($\alpha = 0.01$) for different system complexities are used. The Bonferroni correction is used to reduce the probability of a Type I error (i.e., incorrectly rejecting the null hypothesis with $a_0 = \alpha/6 = 0.00167$.

Critical Systems Labs
Innovating Safely

Results – RQ3 (Scalability)

Does the response quality degrade as the system under analysis increases in complexity?

- The **quality** of ChatGPT’s responses **declines** as the system complexity increases.
- For systems above a certain complexity, the proportion of correct and **useful information declines**.



Conclusion and Next Steps

- Preliminary Study…
- *Co-Hazard Analysis (with an LLM) is a feasible and moderately useful. However, utility degrades as complexity increases.*
- Next Steps
  - Repeat with the next generation of LLM's → are the results better?
  - Try with other hazard analysis techniques → does CoHA generalize?
  - Try with a real-world systems → does CoHA "add value"?
  - Compare with human-only analysis.

Critical Systems Labs
Innovating Safely

Simon Diemert, M.Sc., P.Eng.
sdiemert@uvic.ca
University of Victoria