

THE 12th SWEDISH AEROSPACE TECHNOLOGY CONGRESS


FT2025




Improving System Safety in Aviation: Supporting STPA with AI Models

Ana Estela Antunes da Silva¹, Luiz Eduardo Galvão Martins², Andrey Toshiro Okamura¹, Gabriel Nogueira Pacheco², Niklas Lavesson³, Tony Gorschek³




¹State University of Campinas,
Faculty of Technology,
Brazil 



²Federal University of São Paulo,
Department of Science and Technology,
Brazil 



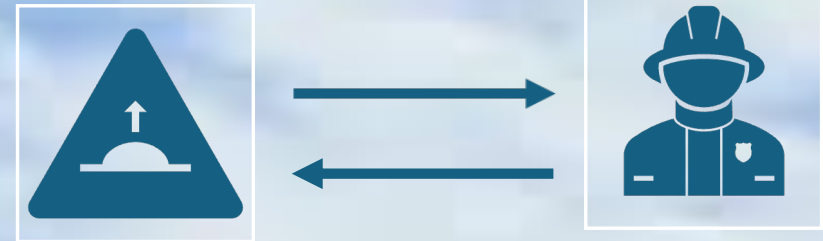
³Blekinge Institute of Technology,
Department of Software Engineering,
Sweden 

Agenda

- Concepts for contextualization of the proposal
 - Safety-Critical Systems
 - System-Theoretic Process Analysis (STPA)
 - ConOps (Concept of Operations) Documents
 - Machine Learning
 - Classifiers
 - Large Language Models
- Proposal of Models for the 1st Step of STPA
 - Datasets
 - Pipelines
- Conclusions

System-Theoretic Process Analysis

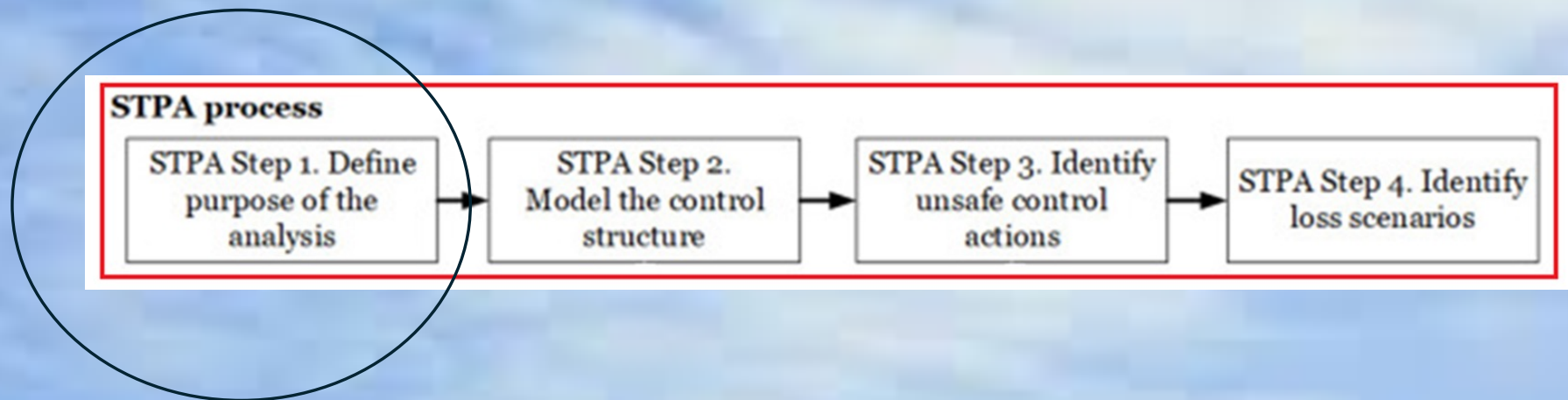
System-Theoretic
Process Analysis
(STPA) is a hazard
analysis process.



Unlike traditional processes that concentrate primarily on component failures, STPA concentrates on **component interaction failures**, including human operators, hardware, software, and organizational factors [3].

Steps of System-Theoretic Process Analysis

- There are four steps in the STPA process



Identify:

- Losses
- Hazards
- (Safety) Constraints

Examples of Losses, Hazards and Constraints

- L-1: Loss of life or injury to people
- L-2: Loss of or damage to Aircraft

- H-1: Aircraft violates minimum separation standards [L-1, L-2]

- SC-1: Aircraft must satisfy minimum separation standards from other aircraft and objects [H-1]

ConOps (Concept of Operations) Documents

- ConOps documents, which can serve as the primary input to the STPA process.
- These documents typically follow established standards, such as IEEE Std 1362-1998 [13] , and provide detailed descriptions of:
 - system objectives,
 - operational scenarios, and
 - stakeholder needs from which safety elements must be inferred.
- A template of a ConOps can be found at:

• <https://www.dau.edu/cop/rqmt/documents/jst-template-usa-concept-operations#:~:text=Currency%20Review:%2024%20Aug%202022.Dated%2012%202022>



Machine Learning

- A machine learning algorithm analyzes vast amounts of data to identify relationships among data, generating knowledge.
- The main tasks a classical machine learning algorithm can perform are:
 - **Classification**
 - Clustering
 - Regression and
 - Association



Large Language Models (LLMs)

- LLMs are a type of machine learning model that use deep learning techniques, specifically neural networks and transformers, to:
 - understand
 - process
 - generate text.
- Trained on massive datasets, LLMs identify complex patterns in text to perform tasks like:
 - text generation
 - text retrieval
 - translation
 - question and answering
 - summarization

Motivation to automate the STPA process

- Although effective, STPA presents several challenges in its practical application:
 - It is a time-intensive process that requires expertise in system safety
 - Analysts must rely on expert judgment to identify losses, hazards, safety constraints
 - Furthermore, training in STPA demands considerable resources, which makes automation an attractive solution to streamline the procedure [4, 5, 6]

Pipeline Proposals

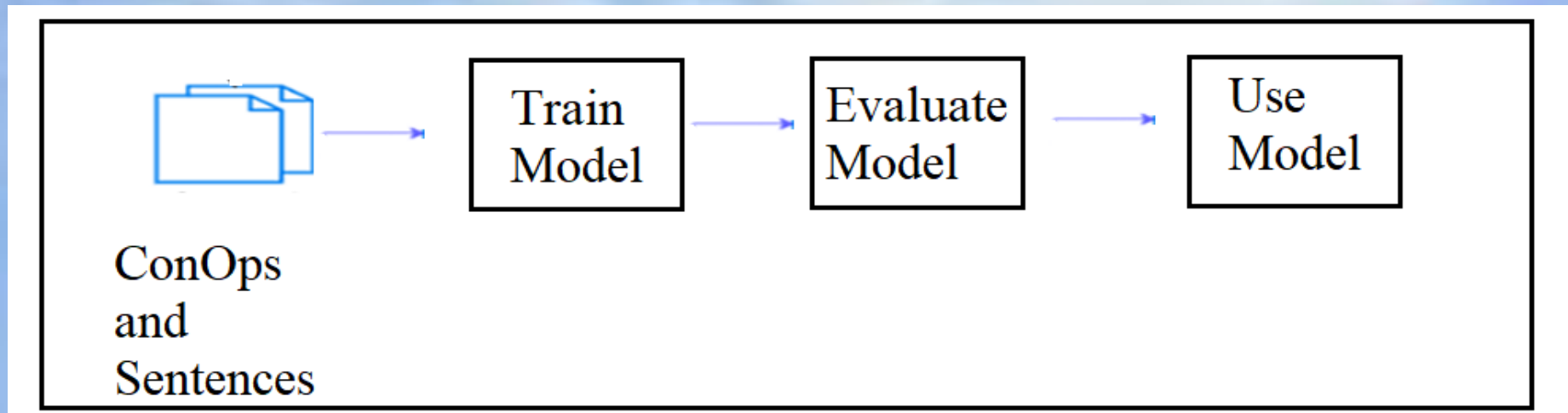
- To address these challenges, we have developed two LLMs pipelines to automate the first step of STPA
- **First pipeline:** STPA Hazard Analysis from ConOps (SHACO)
 - A model to extract losses, hazards, and constraints from ConOps documents
- **Second pipeline:** BERT Error Detection for STPA (BEDS)
 - A model to classify, verify, detect errors, and suggest potential corrections for the sentences

SHACO Dataset

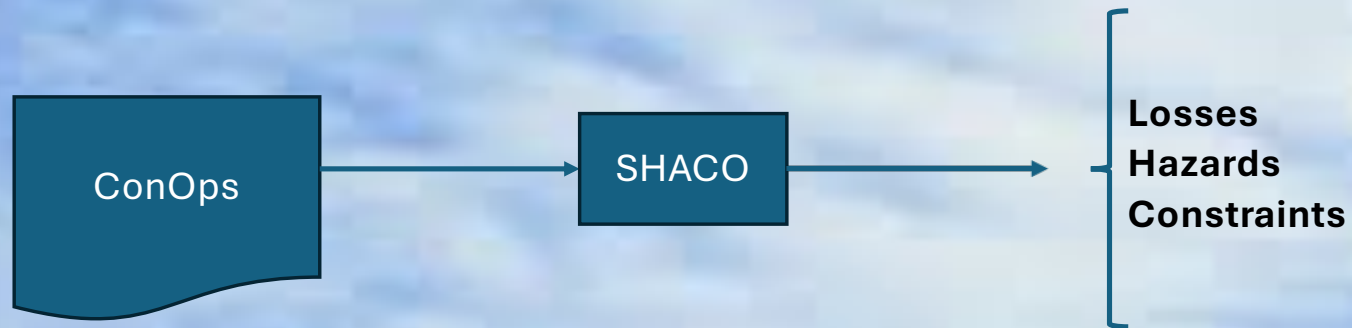
- It is a specialized dataset as no dataset was available publicly
- It comprises 134 samples:
 - ConOps documents, each one paired with its corresponding STPA elements (losses, hazards, and safety constraints)
 - 35 were real-world ConOps documents, primarily from the aviation domain (CORDIS and NASA) [15] [16]
 - 99 were generated through prompt engineering techniques applied to ChatGPT and Claude Sonnet 3.5

The First Pipeline: SHACO

- This first pipeline uses:
 - Llama 3.1 model to automate the extraction of losses, system level hazards, and safety constraints directly from ConOps documents.



How to use SHACO Model



Shaco Results

Average metrics for a k-fold method, with $k = 5$

- Precision: 0.891543
- Recall: 0.893624
- F1-Score: 0.892516

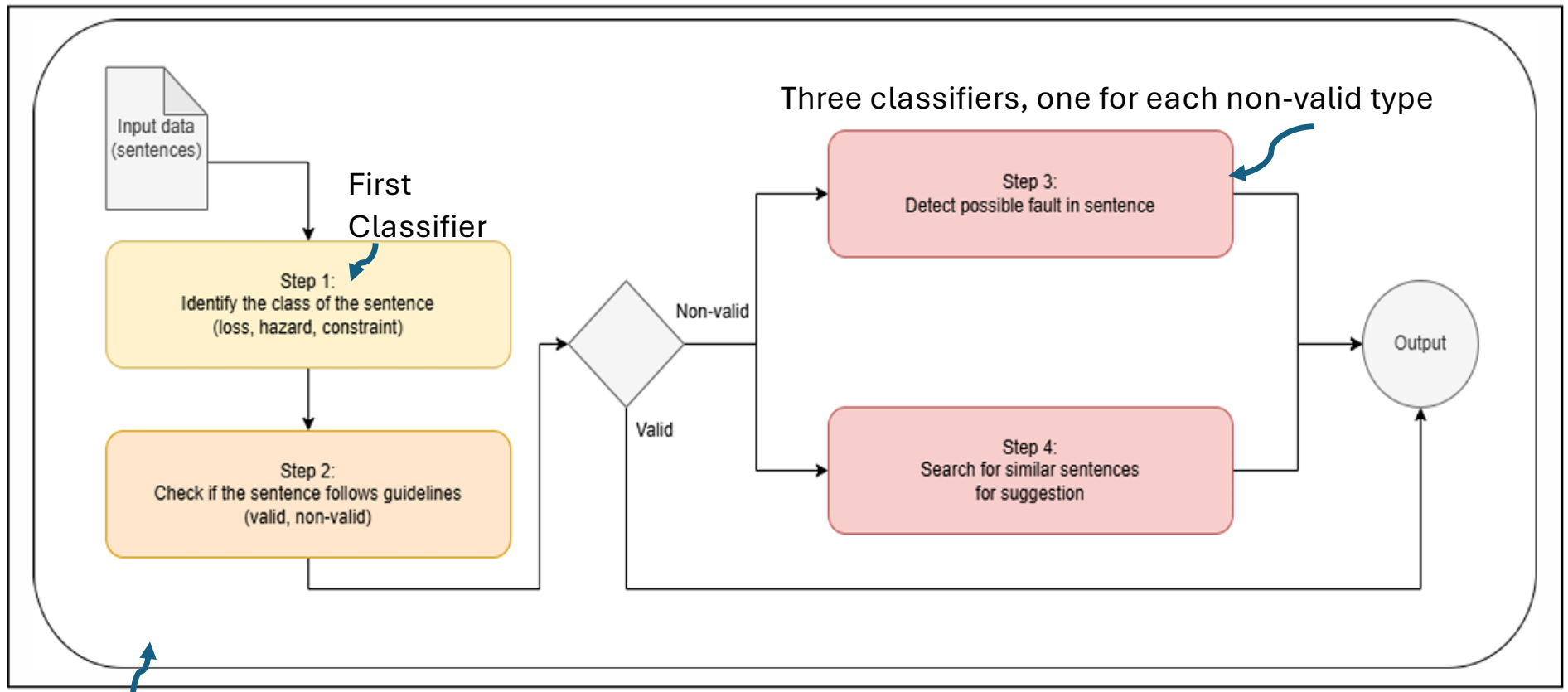
The Second Pipeline: BERT Error Detection for STPA - BEDS

- BEDS aims at
 - identifying possible writing errors in the sentences generated from the analysis
 - suggesting corrections to these errors
- In addition
 - The pipeline allows users who already have pre-defined hazards, losses and constraints, to verify their adequation to the STPA process without having to use the first pipeline.

BEDS Dataset

- It is made up of sentences (losses, hazards and constraints) from various domains
- It was created by extracting fragments of STPA analyses found in presentations at the MIT STAMP Workshops ([https://psas.scripts.mit.edu/home/stamp workshops/](https://psas.scripts.mit.edu/home/stamp%20workshops/)).
- It contains 1,084 rows of sentences of different domains.

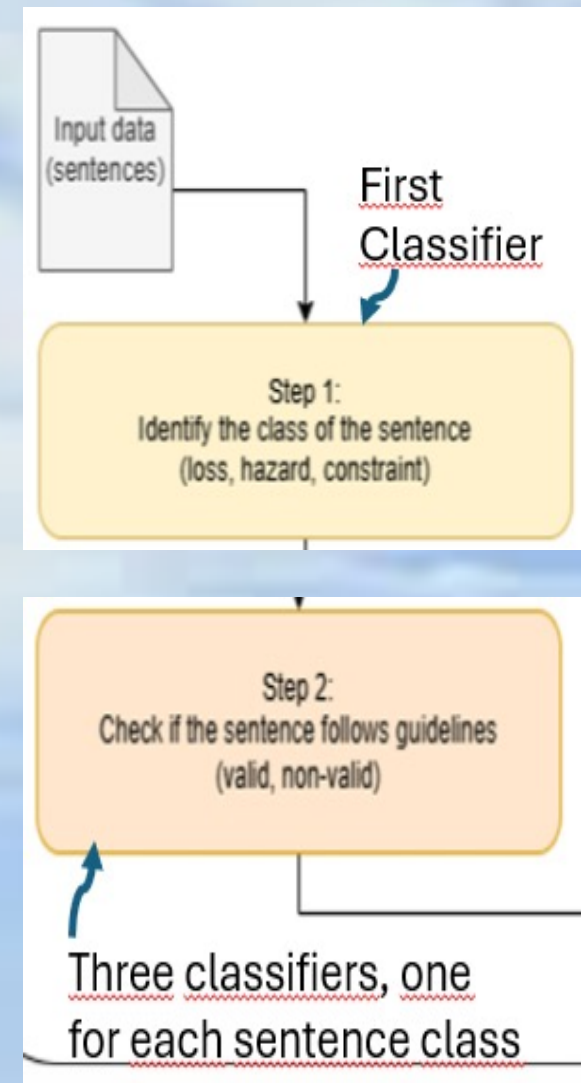
BEDS pipeline: a combination of classification models



Three classifiers, one for each sentence class

Results for BEDS Pipeline

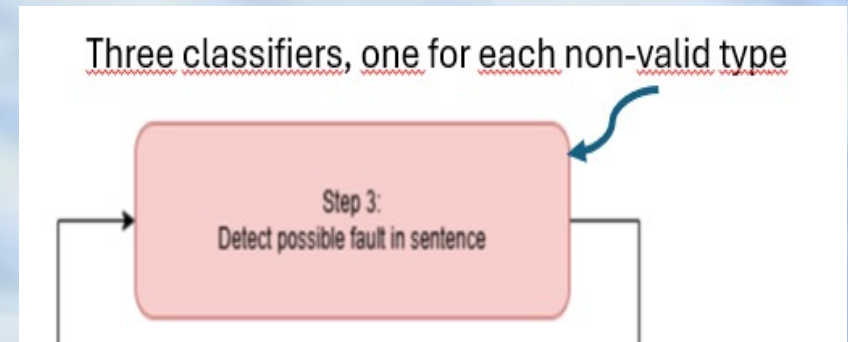
- The first classifier achieved:
 - 95.20% of accuracy
 - 95.08% of F1-Score.
- The second filter with three classifiers achieved:
 - For accuracy:
 - 90.37% (loss)
 - 79.00% (hazard) and
 - 96.47% (constraint)
 - For F1-Score:
 - 89.08% (loss),
 - 75.78% (hazard),
 - and 93.95% (constraint)



Results for BEDS Pipeline

- Similarly, the BEDS third step consists of three classifiers which achieved:

- For accuracy:
 - 74.50% (loss), 79.59% (hazard), and 96.25% (constraint)
- For F1-score:
 - 66.35% (loss), 57.32% (hazard), 95.83% (constraint)



Conclusions

- Together, these pipelines demonstrate a synergistic contribution: SHACO accelerates the extraction process, while BEDS ensures the quality and correctness of the outputs.
- Challenges:
 - data availability
 - contextual understanding
 - full automation
- The presented pipelines offer a foundational framework for integrating AI into safety-critical engineering workflows.

Future Works

- Authors intend to:
 - increase the datasets used in order to improve the metrics obtained in both, pipeline 1 and pipeline 2
 - create sanity check modules in pipeline 1 in order to verify that there are no requirements documents (ConOps) with missing or invalid parts according to the STPA process
 - Automation of steps 2,3 and 4 of the STPA process

References

- [1] N. G. Leveson. Engineering a Safer World: Systems Thinking Applied to Safety. The MIT Press, 2012.
- [2] N. G. Leveson. An Introduction to System Safety Engineering. The MIT Press, 2023.
- [3] N. G. Leveson and J. P. Thomas. STPA Handbook, 2018. Available: https://psas.scripts.mit.edu/home/get_file.php?name=STPA_Handbook.pdf
- [4] J. Chen, S. Zhang, Y. Lu and P. Tang, "STPA-based hazard analysis of a complex UAV system in take-off," 2015 International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 2015, pp. 774-779, doi: 10.1109/ICTIS.2015.7232133

References

- [5] B. Olberts and Y. Dittjen, "Model Based STPA for Assisted Driving Functions," 2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), Västerås, Sweden, 2023, pp. 85-86, doi: 10.1109/MODELS-C59198.2023.00027.
- [6] A. Carniel, J. D. M. Bezerra and C. M. Hirata, "An Ontology-Based Approach to Aid STPA Analysis," in IEEE Access, vol. 11, pp. 12677-12697, 2023, doi: 10.1109/ACCESS.2023.3242642.
- [13] IEEE. IEEE Guide for Information Technology - System Definition - Concept of Operations (ConOps) Document. IEEE Std 1362-1998, p. 1–24, 1998.

References

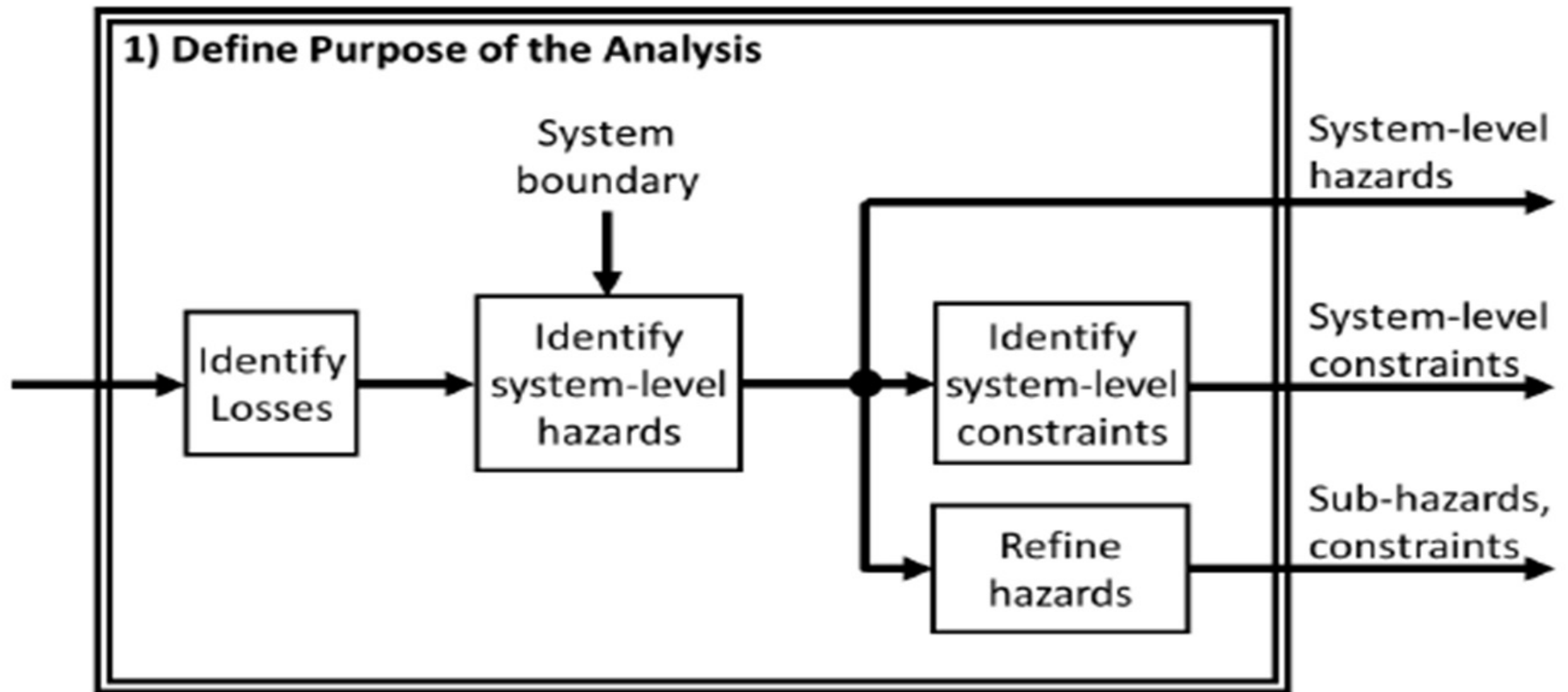
- [15] CORDIS. “Community Research and Development Information Service,” 2025. [Online]. Available: <https://cordis.europa.eu/>. [Accessed: Nov. 5, 2024].
- [16] NTRS. “NTRS - NASA Technical Reports Server,” 2025. [Online]. Available: <https://ntrs.nasa.gov/>. [Accessed: Feb. 11, 2025].

Thank you!

aeasilva@unicamp.br

Additional slides

What has to be done in the 1st Step of STPA?



SHACO Dataset prompt engineering for ConOps

- <explanation>
- I want you to act as a systems engineering specialist. Our goal is to generate Concept of Operations (ConOps) documents for a determined system. We want to create ConOps to use as a dataset for later hazard analysis. Note: It will be evaluated by experts from industry. Thus, the audience are real system engineers. The documents must be generated in the most detailed way, so try to generate the longest you can. Also, try to make it really close to a real ConOps. Generate the document in the PDF format, if possible.
- </explanation><instruction>
- Based on the explanation provided, generate a ConOps for a ____.

SHACO Dataset prompt engineering for STPA documents

Too many hazards containing unnecessary detail.

- Like losses, there are no hard limits on the number of system-level hazards to include
- As a rule of thumb, if you have more than about seven to ten system-level hazards, consider grouping or combining hazards to create a more manageable list
- You may be including unnecessary detail and making the list unmanageable, difficult to review, and harder to identify things that are missing. Instead, begin with a more abstract and manageable set of system-level hazards and refine them into sub-hazards later if needed

SHACO Dataset prompt engineering for STPA documents

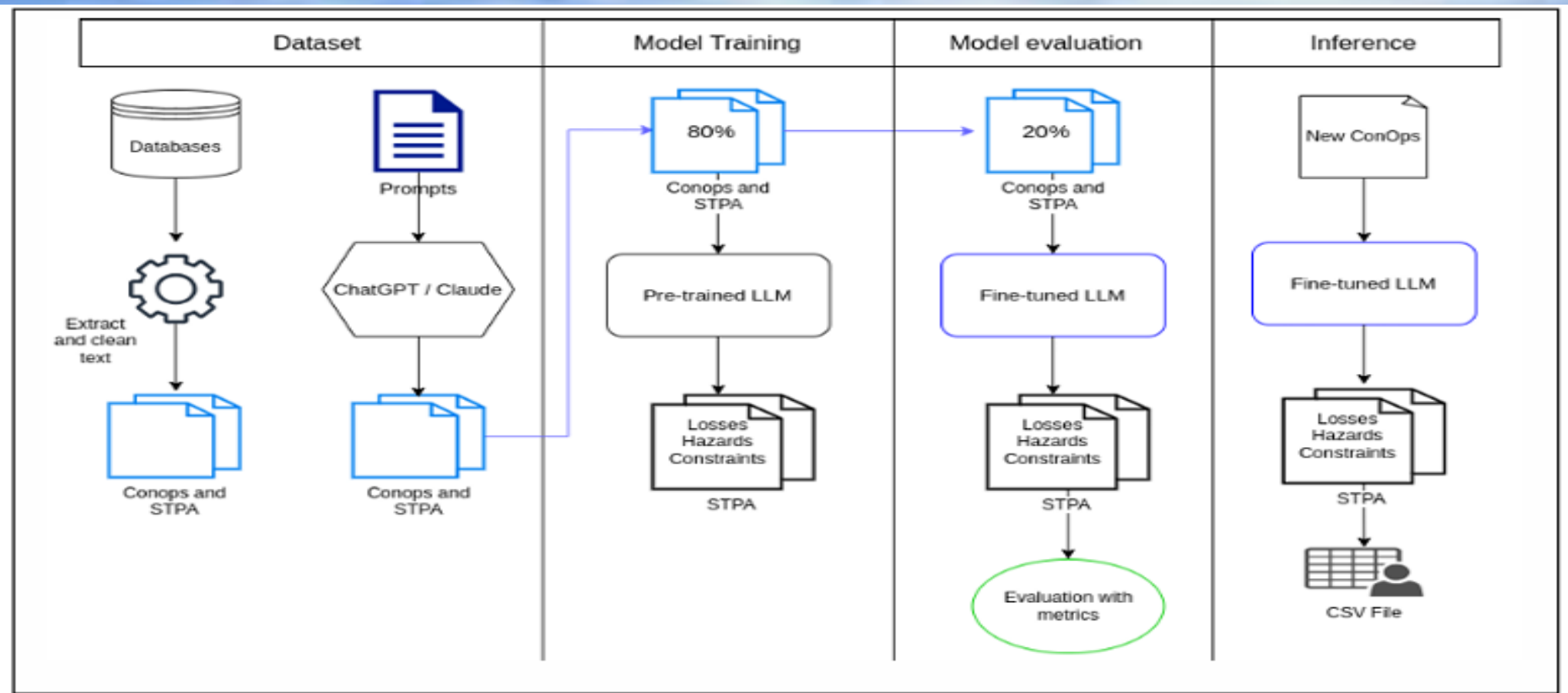
Tips to prevent common mistakes when identifying hazards

- Hazards should not refer to individual components of the system.
- All hazards should refer to the overall system and system state
- Hazards should refer to factors that can be controlled or managed by the system designers and operators.
- All hazards should describe system-level conditions to be prevented
- The number of hazards should be relatively small, usually no more than 7 to 10

SHACO Dataset prompt engineering for STPA documents

- Some notes to keep in mind about STPA:
- Common mistakes when identifying system-level hazards
- Confusing hazards with causes of hazards
- A common mistake in defining hazards is to confuse hazards with causes of hazards. For example, “brake failure”, “brake failure not annunciated”, “operator is distracted”, “engine failure”, and “hydraulic leak” are not system-level hazards but potential causes of hazards.
- To avoid this mistake, make sure the identified hazards do not refer to individual components of the system, like brakes, engines, hydraulic lines, etc. Instead, the hazards should refer to the overall system and system states.
- In other words, check that each hazard contains: <Hazard specification> = <System> & <Unsafe Condition> & <Link to Losses> E.g. H-1 = Aircraft violate minimum separation standards in flight [L-1, L-2, L-4, L-5]. The exact sequence is not important—you could just as easily write “Minimum separation standards for aircraft are violated [L-1, L-2, L-4, L-5]”. What is important is that the system-level hazards contain these elements.

Shaco Pipeline



Results - Shaco

```
----- CROSS VALIDATION -----  
Metrics from all folds:  
  eval_precision  eval_recall  eval_f1  
0      0.893320      0.896389  0.894821  
1      0.889372      0.889509  0.889401  
2      0.886105      0.889858  0.887788  
3      0.897113      0.898426  0.897745  
4      0.891805      0.893936  0.892825  
  
Average Metrics across all folds:  
eval_precision      0.891543  
eval_recall         0.893624  
eval_f1             0.892516  
dtype: float64  
  
Standard Deviation of Metrics:  
eval_precision      0.004139  
eval_recall         0.003934  
eval_f1             0.004025
```

BEDS Dataset

- The resulting dataset contains 1,084 rows of varied classes and domains.
- 50 of those, however, were manually written (through paraphrase) and added to compensate for some class imbalance found in the dataset.
- As the dataset is an integral part of the pipeline development, the validity column was verified by specialists with three to four years of experience in STPA analysis.
- This contribution was necessary to ensure the valid sentences in this dataset correctly follow the STPA guidelines, and that the models can reliably classify the sentences.

BEDS Dataset

- This dataset contains nine columns, of which the first is the prediction data (the extracted sentence), three are target labels for classification (each for the first three pipeline steps), and the last five are metadata to improve traceability.
- The first target label is the class of the sentence, which can be one of the three STPA elements. The second target label is the validity of the sentence according to the STPA Handbook, which can be valid or non-valid.
- The last target label, given only to the invalid sentences, is the main fault observed in the sentence that differs from what is recommended by the handbook.
- The remaining columns are the metadata related to the sentence, such as domain, year, title, URL, and number of the presentation.

BEDS Results

- The quantitative performance, calculated as the mean from 5-fold cross-validation, is summarized below:
 - Class identification: 95.20% - Accuracy and 95.08 % - F1-Score
 - Validity determination: 88.61 % - Accuracy and 86.27 % - F1-Score
 - Faults classification: 83.44% - Accuracy and 73.16 % - F1-Score