



# Cognition and Computation in Decision-Making

Applying the Critical Decision Method to Artificial Intelligence for Aviation Event Analysis

**Giovane de Moraes**

*Presenter – [giovane@ita.br](mailto:giovane@ita.br)*

**Ingrid K. L. Strohm**

*[ingridstrohm@ita.br](mailto:ingridstrohm@ita.br)*

**Prof. Dr. Moacyr M. Cardoso Jr.**

*[moacyr@ita.br](mailto:moacyr@ita.br)*

**Prof.<sup>a</sup> Dra. Emília Villani**

*[evillani@ita.br](mailto:evillani@ita.br)*

**Guilherme V. da Rocha**

*[guilherme.vieira@dcta.br](mailto:guilherme.vieira@dcta.br)*

**Nickolas B. M. Machado**

*[nickolas.machado@dcta.br](mailto:nickolas.machado@dcta.br)*

**Guilherme M. B. Moreira**

*[guilherme.moreira@dcta.br](mailto:guilherme.moreira@dcta.br)*

**Instituto Tecnológico de Aeronáutica  
Brazil**

# Critical Decision Method Drives Aviation Safety Insights

Analyzing complex incidents reveals deep crew cognition but demands intensive manual effort, prompting automation for scalable, consistent investigations.

CDM systematically reconstructs timelines, cues, and decision rationales to deeply analyze crew behavior during rare or complex aviation incidents.

Human reference: 72 participants (36 pilots, 36 novices) responding to a 53-item CDM-mapped questionnaire.

Manual qualitative coding of large datasets is labor-intensive and time-consuming, often requiring expert collaboration over hours or days.

This laborious process limits scalability and consistency in safety investigations.

We evaluate a two-model local LLM pipeline (Phi-3 generator + Zephyr-7B judge) on one anonymized incident.

Automation offers potential to enhance scalability, reduce workload, and improve consistency in applying CDM to aviation safety analysis.



Automate CDM by summarizing text, extracting facts, and generating structured responses



Reduce repetitive human effort to accelerate root-cause analysis



Support consistency across agencies to enhance collaborative decision-making



Risks include hallucinations, omission of nuances, and misleading ethical interpretations



Necessitates careful design and validation for safety-critical aviation applications

# Unlocking Efficiency and Managing Risks with LLMs

Harness Large Language Models to automate CDM tasks while ensuring safety in aviation operations

# Advancing Automation in Aviation Investigations

From Early CDM Automation to LLM Integration Challenges and Solutions



Early CDM automation relied on shallow text matching and manual validation, limiting scalability



LLMs trialed in legal and forensic interviews to support thematic coding



Domain mismatch causes LLM hallucinations and missed contextual signals



Current study integrates LLMs with rigorous aviation frameworks for reliable automation

# Incident Analysis: Design, Participants, & Questionnaire

Benchmarking AI Against Expert and Novice Pilot Assessments

## Incident Details

Single anonymised aviation incident analyzed

Gold standard coding by five expert investigators

## Participant Groups

72 total participants

36 experienced pilots

36 volunteers lacking formal aviation training (novices)

## Questionnaire Structure

53-item questionnaire completed by all participants

Items mapped to CDM themes: context, timeline, cognition, counterfactuals

Enabled benchmarking of LLM pipeline against experts and novices

# Dual LLM Pipeline Drives Reliable Aviation Safety Analysis

Generator model produces structured responses; judge model evaluates answer confidence and completeness to reduce hallucinations. Low temperature (0.1) for consistency; prompts calibrated with early-stopping to reduce drift. No training data from the test incident was used (to avoid leakage).

---

	<i>Phi-3-Mini-Instruct</i>	<i>Zephyr-7B-Beta</i>
<b>Model Size</b>	<b>3B parameters</b>	<b>7B parameters</b>
<b>Batch Size</b>	<b>8</b>	<b>4</b>
<b>Learning Rate</b>	<b><math>2 \times 10^{-5}</math></b>	<b><math>1 \times 10^{-5}</math></b>
<b>Epochs</b>	<b>3</b>	<b>2</b>

---

## Phi-3-Mini-Instruct Generator

Generates structured CDM responses including MCQ, Likert scales, and open-ended answers with domain-specific precision.

## Zephyr-7B-Beta Judge

Evaluates generated answers for confidence, completeness, and groundedness to identify questionable outputs and mitigate hallucinations.

# Optimized Pipeline Workflow & Algorithm Interactions

Five-step process integrating Phi-3 and Zephyr-7B for automated analysis and evaluation. Outputs scored on **Confidence, Completeness, Groundedness (0–1)** to support downstream comparisons with human data.

## 1. Data Ingestion

Import incident narratives and questionnaire responses for analysis.



## 3. Response Evaluation

Apply Zephyr-7B to judge and score generated answers.



## 2. Answer Generation

Use Phi-3 model to generate answers based on ingested data.



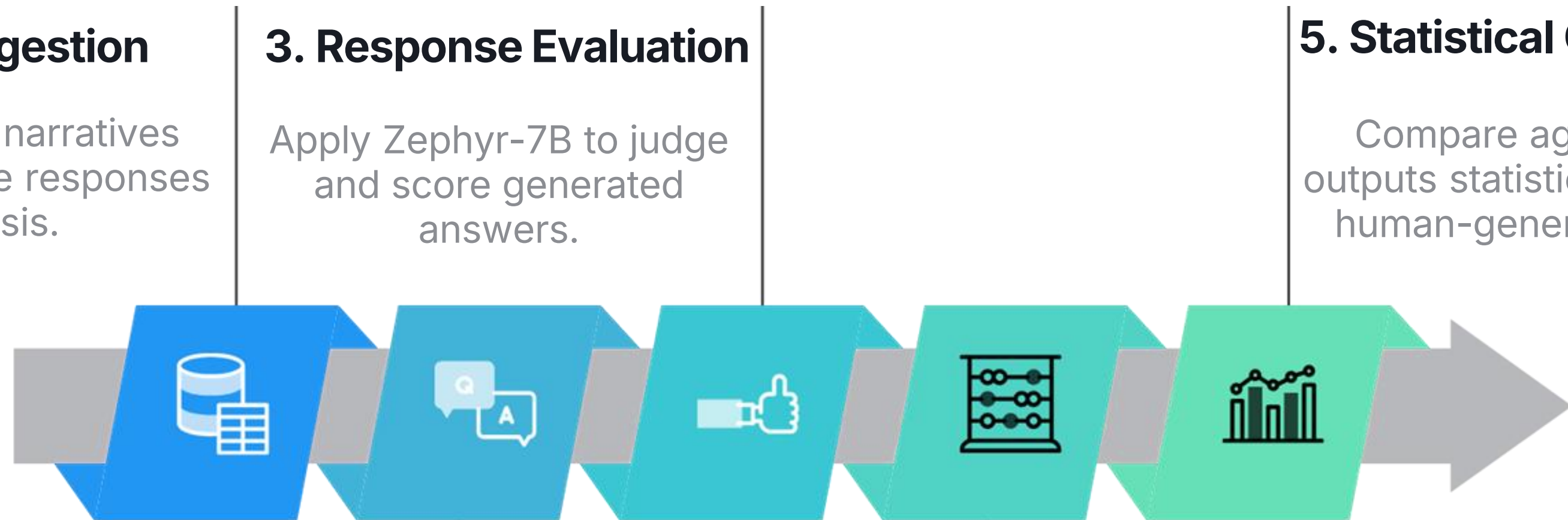
## 4. Output Aggregation

Collect and consolidate model outputs for comprehensive review.



## 5. Statistical Comparison






Compare aggregated outputs statistically against human-generated data.





# Prompt Engineering & Ethical Safeguards

Designing precise LLM prompts and embedding strong privacy and compliance measures.

-  Tailored prompt instructions ensure concise, cautious outputs across MCQ, Likert, and open-ended formats
-  Low temperature setting (0.1) improves output consistency and reliability
-  Local model execution protects sensitive data by preventing external exposure
-  Data anonymization and human-in-the-loop review enhance ethical oversight
-  Compliance with aviation regulatory standards maintains investigative integrity. We log model versions, parameters, and prompts for full audit traceability. Recommendations remain advisory and require certified investigator approval before dissemination.

# Evaluating LLM vs Human Judge Scores

Zephyr-7B-Beta excels in completeness but trails in groundedness; 78% judge alignment verified

Zephyr-7B-Beta rated higher in completeness than human judges, demonstrating thorough response generation.

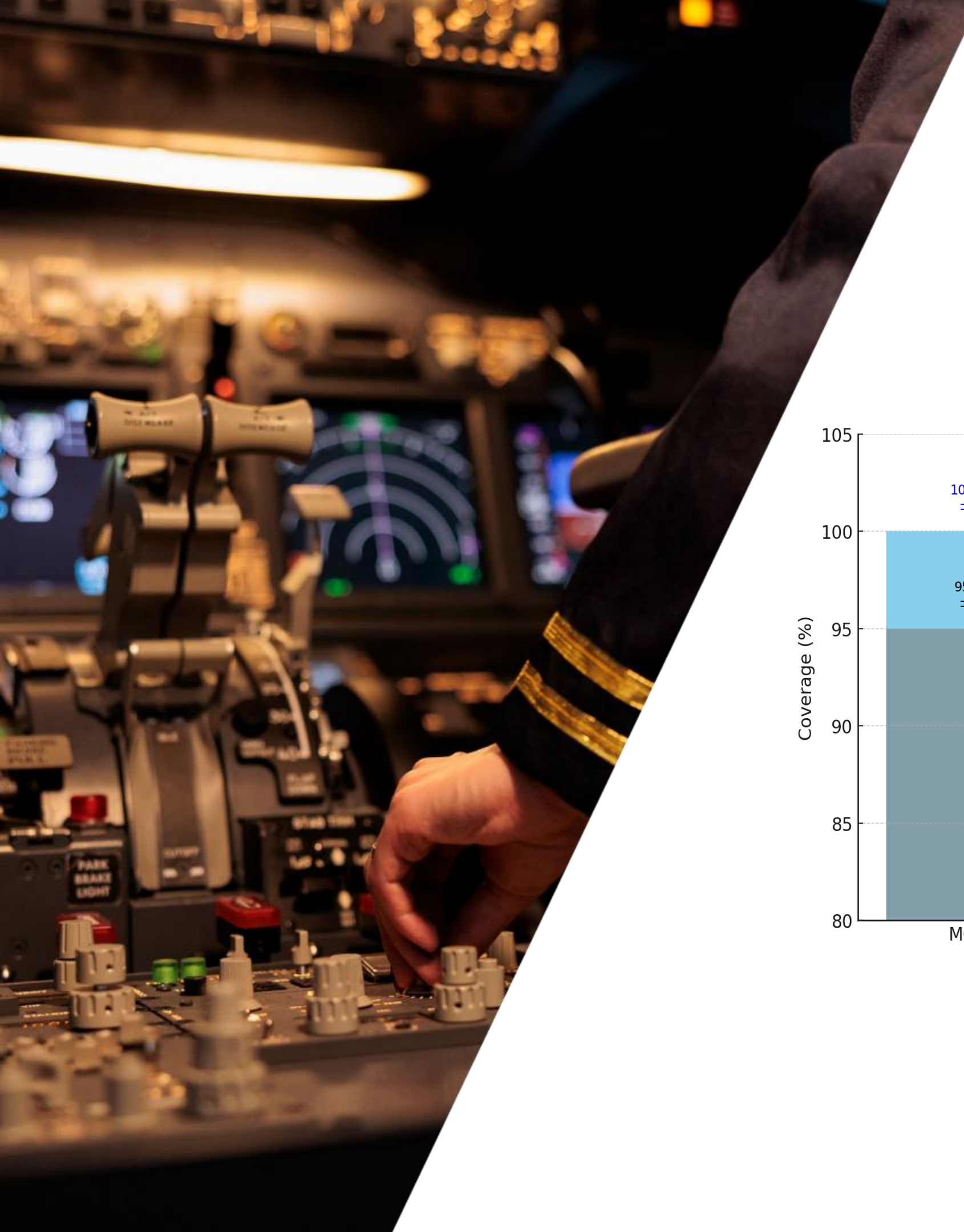
Groundedness scores were lower for the LLM compared to humans, indicating room for factual accuracy improvements.

78% of spot-checks aligned with judge scores, confirming consistent evaluation reliability.

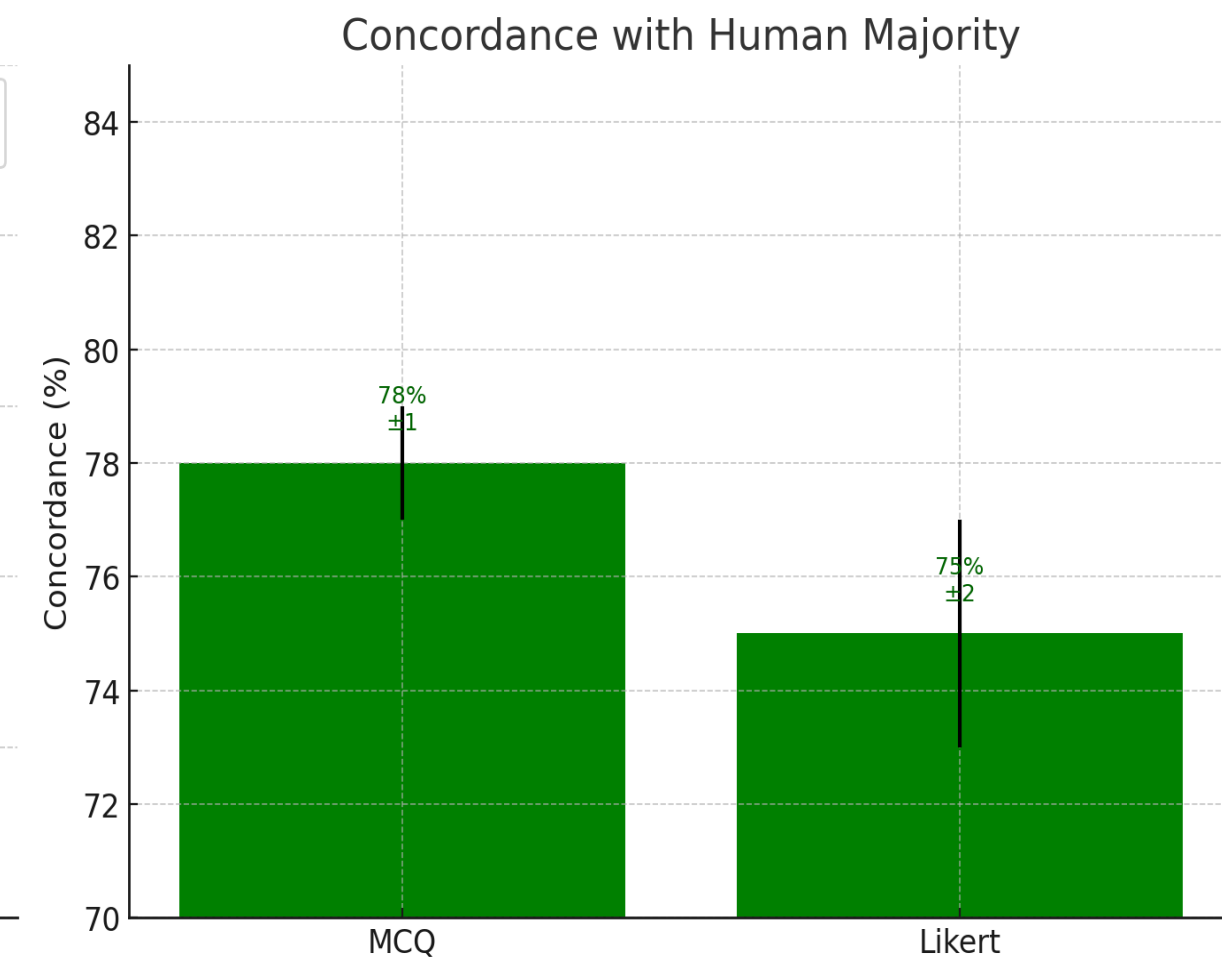
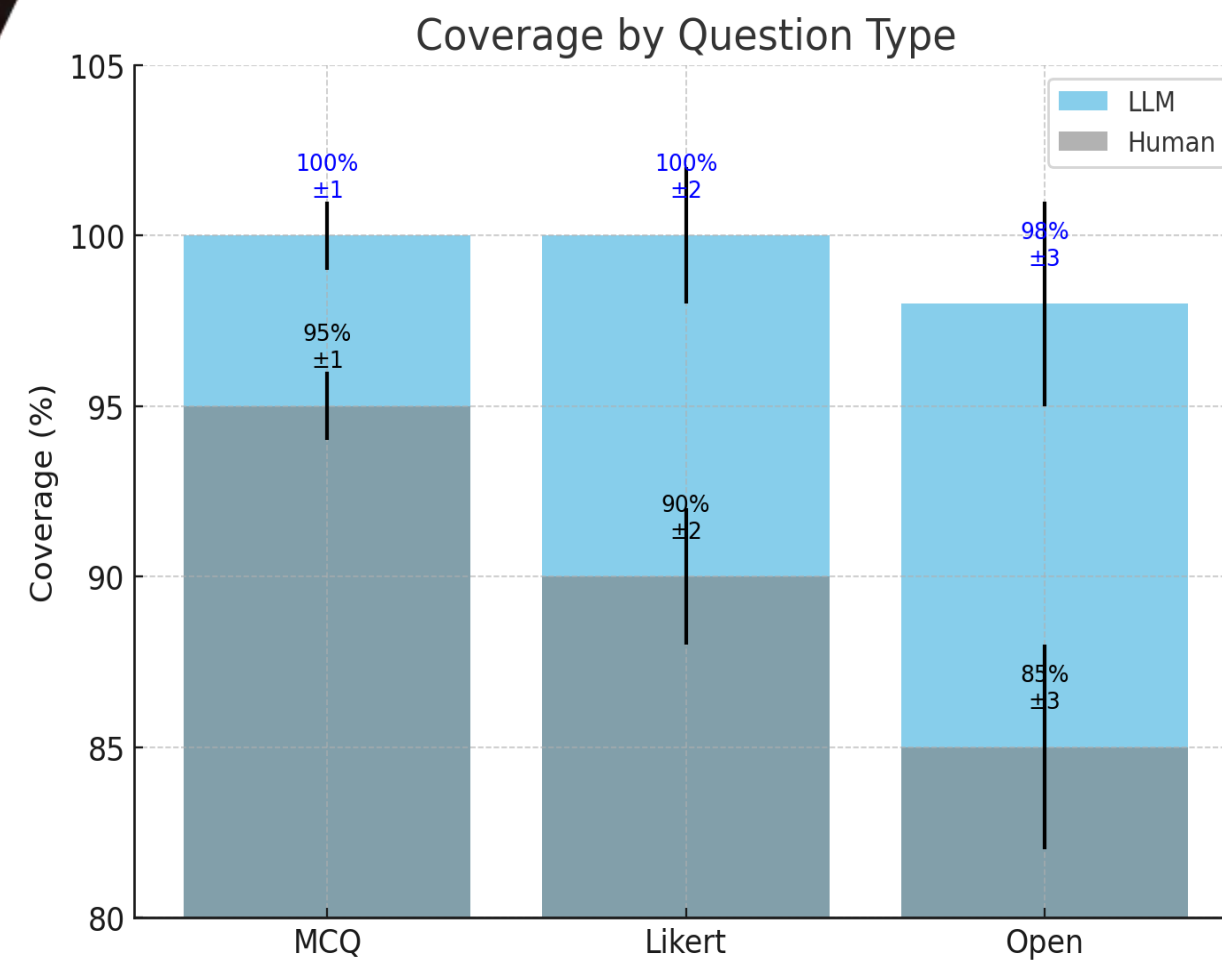
Classical NLP baseline (TF-IDF+SVM) offered discrete classification but lacked nuanced scoring detail.

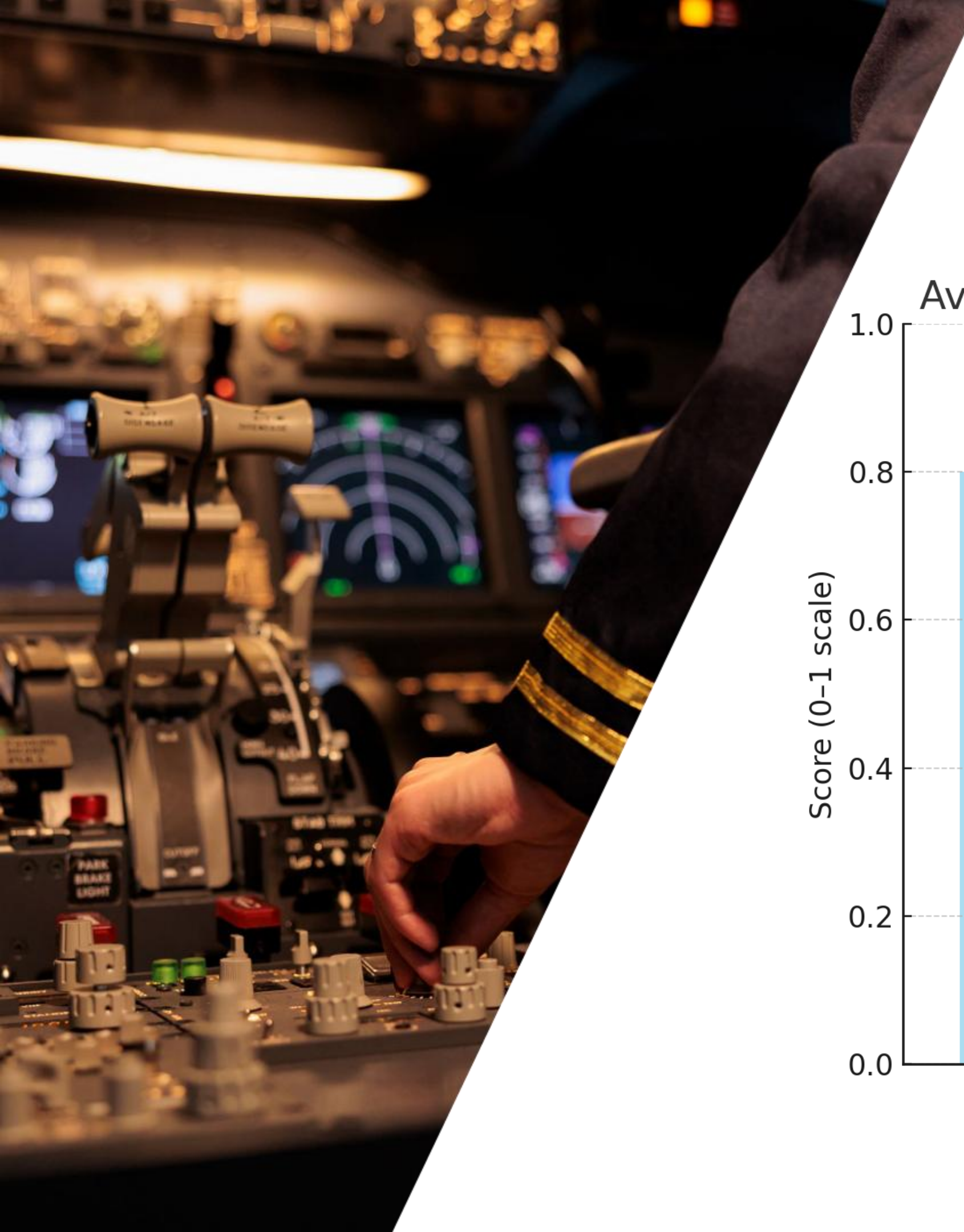
LLM pipeline provided near-complete, rapid responses, showing promising consistency with human references.



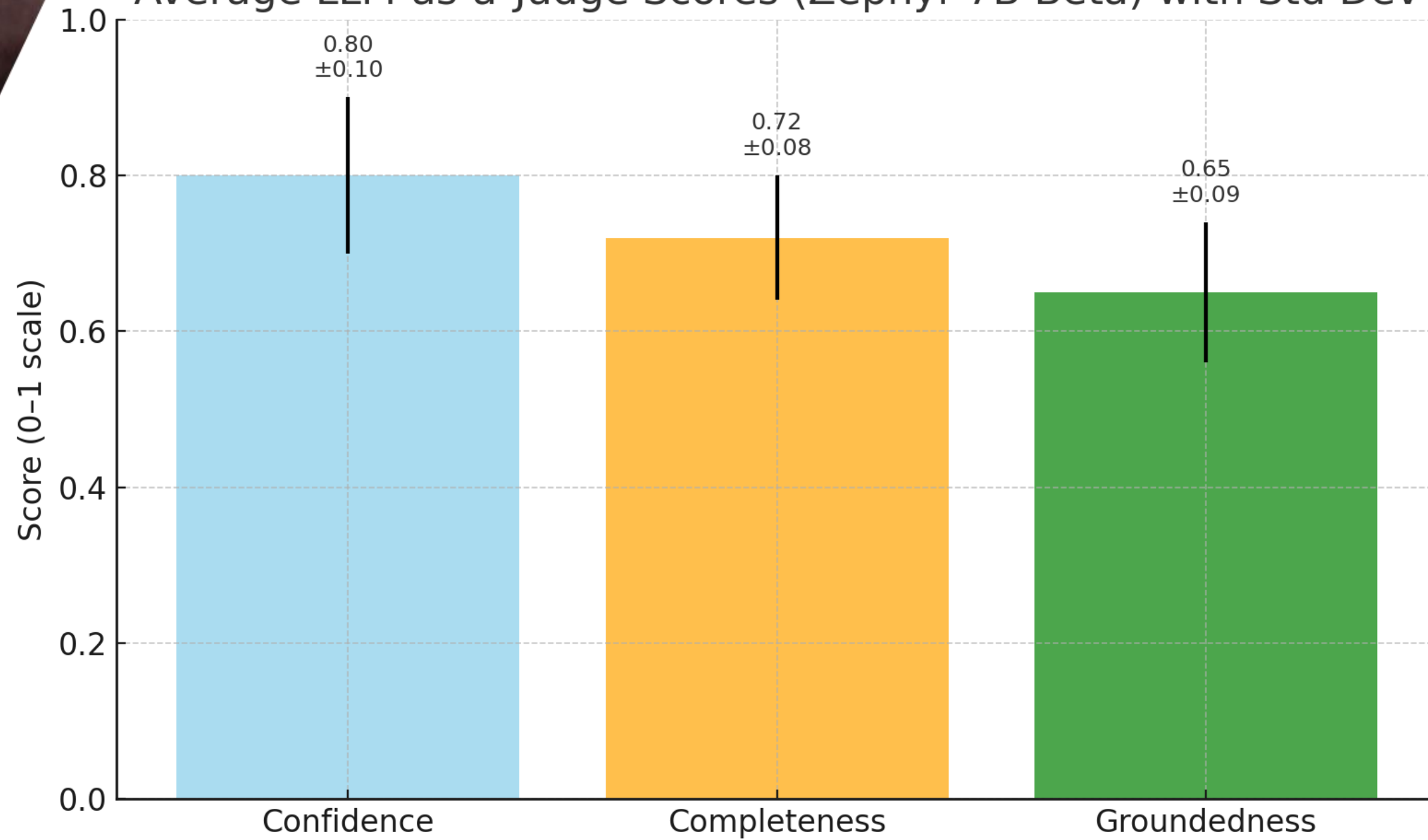


### Coverage and Concordance Metrics for LLM and Human Evaluators

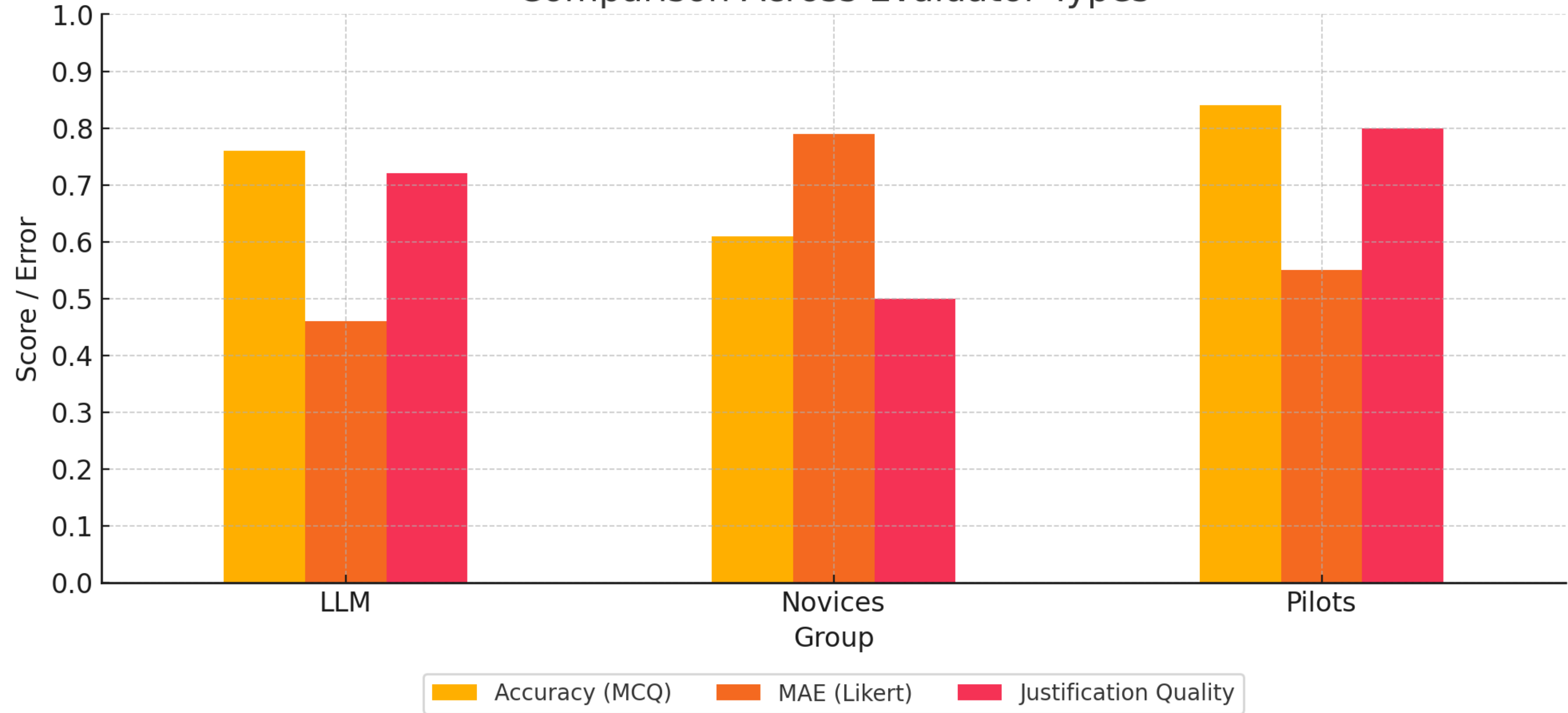




Average LLM-as-a-Judge Scores (Zephyr-7B-Beta) with Std Dev



# Comparison Across Evaluator Types



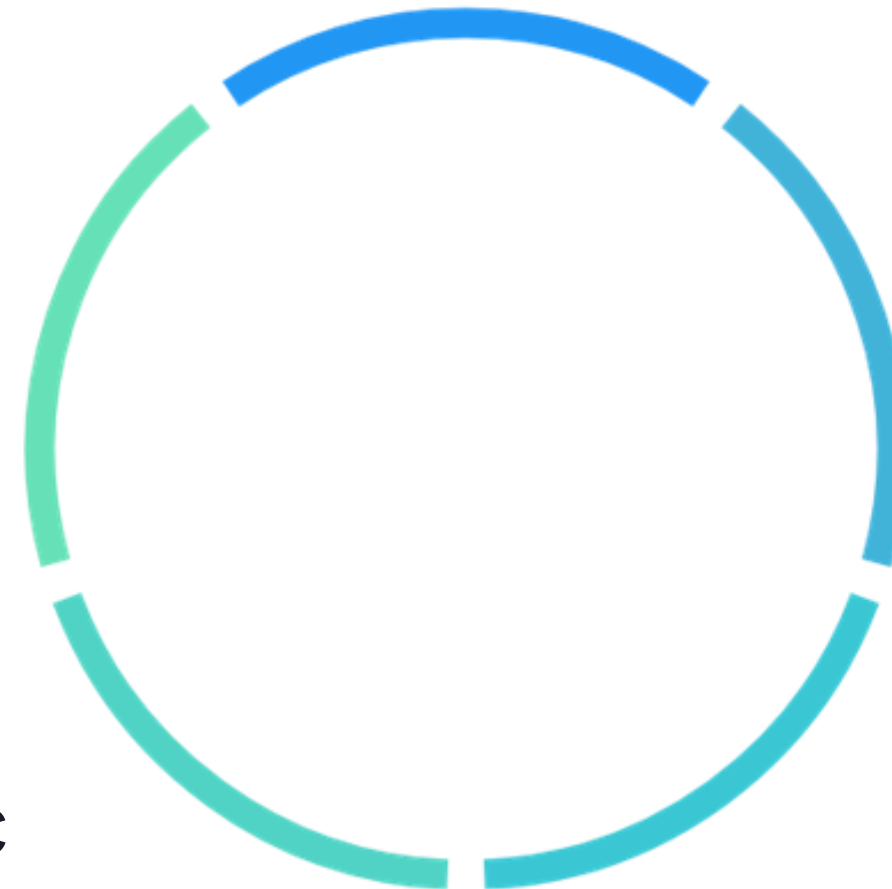
# Evaluating LLM Strengths and Constraints

Balancing performance on structured tasks with domain-specific challenges and expert integration

Excels at structured MCQs and Likert scale items with reliable accuracy

Supports but does not replace expert judgment, especially given limited single-incident scope

Data quality, domain-specific training, and prompt design critically influence performance



Struggles to interpret contextual nuance and maintain domain-specific accuracy

Occasional hallucinations and role misattributions affect output reliability

# Enhancing CDM with Adaptive & Iterative Feedback

Implementing dynamic questionnaires and supervisory intervention to increase automation flexibility

Advance pipeline adaptivity to enhance data quality and decision accuracy

Incorporate dynamic questionnaires featuring branching logic for adaptive data collection

Integrate real-time intervention to approximate classical CDM flexibility within an automated pipeline

Enable supervisory feedback from human or AI agents to resolve ambiguous responses

# Scaling Up LLM Evaluation with Multi-Incident Analysis

Expanding to 10–50+ incidents enables rigorous statistical validation across diverse aircraft and regions

Analyze 10–50+ diverse incidents spanning multiple aircraft types and geographic regions

Assess LLM domain adaptation robustness across temporal and geographic variations

Calculate confidence intervals to quantify uncertainty in performance and domain adaptation



Apply reliability statistics to measure consistent LLM performance over time

Use ANOVA to compare LLM effectiveness across different incident categories and domains

# Mitigating Bias in Generator & Judge Models

Layered strategies reduce mutual bias from shared training data, improving evaluation reliability

Shared training data risks mutual bias, causing error reinforcement between models

Combined mitigations enhance transparency and reliability of model evaluations



Use distinct model families for generators and judges to diversify perspectives

Integrate human spot-checks to calibrate and verify automated evaluations

Employ multiple judge models to cross-validate and challenge automated assessments

# Maximizing Efficiency with Ethical Integrity

Automating workflows while ensuring compliance and accountability



Achieves significant time savings by automating repetitive tasks



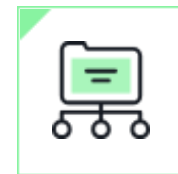
Standardizes data extraction to improve consistency and reliability



Enables investigators to prioritize complex, high-judgement issues



Incorporates human-in-the-loop protocols to ensure oversight and validation



Provides traceable justifications and cross-checks with operational data



Ensures full compliance with aviation safety principles and standards



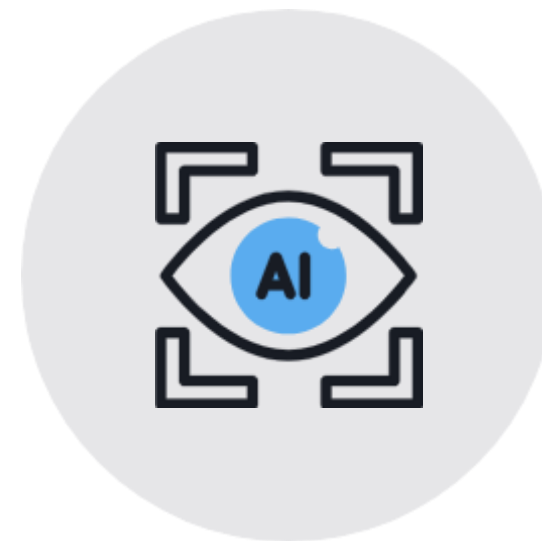
Executes locally to protect sensitive data and maintain privacy

# Mitigating Ethical Risks in Advisory AI Systems

Addressing hallucination, bias, and data privacy to ensure responsible AI use

**Hallucination:** Risk of generating inaccurate information; mitigated by refining abstention pathways

Ongoing improvements focus on bias assessments and integrating operational data for ethical compliance



**Misplaced Blame:** Avoid assigning responsibility to AI by clarifying system's advisory role

**Cognitive Overload:** Design system as an advisory aid, not a decision-maker, to support user clarity

**Context Leakage:** Protect user privacy by processing anonymized data locally with informed consent

# Ensuring Regulatory Compliance and Auditability

Aligning with ICAO Annex 13, EASA, and FAA through traceability, data protection, and human oversight



*Standard*



*Policies*



*Requirement*

- ▲ Meets ICAO Annex 13, EASA, and FAA requirements with traceable, reviewable safety recommendations
- ▲ Safety recommendations are owned and verified by human investigators to ensure accountability
- ▲ Pipeline logs all model versions and parameters for full audit traceability
- ▲ Executes analysis locally to safeguard sensitive data and maintain confidentiality
- ▲ Preserves just culture principles by requiring human confirmation prior to recommendation dissemination

# Conclusions: Advancing AI-Driven Aviation Incident Analysis

Pilot validates local LLM pipelines with expert-aligned outputs and outlines scalable, ethical future enhancements



Demonstrated local LLM pipelines can partially automate CDM-based incident analysis with rapid, near-complete expert-aligned coverage



Emphasized AI outputs as advisory tools requiring certified investigator oversight to ensure reliability



Plan to scale analysis across multiple incidents and integrate diverse multi-modal data sources



Commitment to enhancing ethical safeguards and refining evaluation methods to meet stringent regulatory standards



Overall, this work lays the foundation for robust, scalable AI-assisted aviation safety investigations



# Thank You / Obrigado

**Giovane de Moraes**

*Presenter – [giovane@ita.br](mailto:giovane@ita.br)*

**Ingrid K. L. Strohm**

*[ingridstrohm@ita.br](mailto:ingridstrohm@ita.br)*

**Prof. Dr. Moacyr M. Cardoso Jr.**

*[moacyr@ita.br](mailto:moacyr@ita.br)*

**Prof.<sup>a</sup> Dra. Emília Villani**

*[evillani@ita.br](mailto:evillani@ita.br)*

**Guilherme V. da Rocha**

*[guilherme.vieira@dcta.br](mailto:guilherme.vieira@dcta.br)*

**Nickolas B. M. Machado**

*[nickolas.machado@dcta.br](mailto:nickolas.machado@dcta.br)*

**Guilherme M. B. Moreira**

*[guilherme.moreira@dcta.br](mailto:guilherme.moreira@dcta.br)*

**Instituto Tecnológico de Aeronáutica  
Brazil**