FT2025
SWEDISH AEROSPACE AT THE CROSSROADS
The 12th Swedish Aerospace Technology Congress
FT2025 in Stockholm
October 14-15, 2025

Cognition and Computation in Decision-Making: Applying the Critical Decision Method to Artificial Intelligence for Aviation Event Analysis

Giovane De Morais[1], Ingrid Kawani Leandro Strohm[2], Moacyr Machado Cardoso Júnior[1], Guilherme Vieira Da Rocha[3], Nickolas Batista Mendonça Machado[3], Guilherme Micheli Bedini Moreira[4], and Emilia Villani[1]

[1]*Manufacturing Competence Center, Instituto Tecnológico de Aeronáutica (ITA), Brazil*
E-mail:, giovane@ita.br, moacyr@ita.br, evillani@ita.br
[2]*Pro-Rectory of Research and Institutional Relations, Instituto Tecnológico de Aeronáutica (ITA), Brazil*
E-mail:, ingridstrohm@ita.br
[3]*Institute for Flight Testing and Research, Departamento de Ciência e Tecnologia Aeroespacial (DCTA), Brazil*
E-mail:, guilherme.vieira@dcta.br, nickolas.machado@dcta.br
[4]*Administrative Division, Departamento de Ciência e Tecnologia Aeroespacial (DCTA), Brazil*
E-mail:, guilherme.moreira@dcta.br

Abstract

This paper examines how local Large Language Models (LLMs) can partially automate the Critical Decision Method (CDM) in aviation safety investigations. The CDM, while widely respected for its ability to elucidate human factors and decision-making processes in rare or complex scenarios, often requires labor-intensive qualitative coding. To address this challenge, we developed a pipeline employing two specialised models: *Phi-3-Mini-Instruct* for generating structured responses and *Zephyr-7B-Beta* as a "judge" to evaluate confidence, completeness, and groundedness. A single anonymised incident served as our pilot case. Seventy-two participants (36 aviation professionals and 36 novices) responded to a 53-item CDM-inspired questionnaire, creating a human reference dataset. The pipeline's performance was benchmarked against both this human data and a classical NLP baseline (TF-IDF + SVM). Results revealed that the LLM matched 78% of majority-human multiple-choice answers and achieved a mean absolute error (MAE) of 0.38 on Likert-scale questions. Its open-ended responses, although moderately accurate, occasionally exhibited factual hallucinations (e.g., referencing non-existent systems) and role misattributions. Further stratification showed that the LLM outperformed novices but did not match pilots' domain expertise, underscoring the importance of operational familiarity for nuanced decision analyses. Despite the single-incident scope limiting statistical generalisation, these findings suggest that LLM-based tools can substantially expedite repetitive data processing and facilitate consistent categorisation tasks that often consume investigators' bandwidth. Future work will expand to multiple incidents, integrate flight data recorder (FDR) and cockpit voice recorder (CVR) information to reduce speculation, and refine both self-evaluation mechanisms and ethical safeguards.

Keywords: Local Large Language Models, Critical Decision Method, Aviation Safety, Ethics and AI, Automated Qualitative Analysis

1 Introduction

The Critical Decision Method (CDM) offers a structured approach for reconstructing the cognitive and contextual factors underlying complex or rare aviation incidents [1, 2]. By eliciting timelines, cues, mental models, and decision rationales, CDM probes deeply into the *why* and *how* of crew behaviour [2]. However, the vast volume of qualitative data in

large-scale safety contexts renders manual coding extremely labour-intensive, often requiring expert analysts to collaborate for hours or even days [1, 3].

Recent advances in Large Language Models (LLMs) open the way for partial automation of this process [4, 5]. LLMs are capable of summarising textual inputs, extracting factual details, and generating structured prompts, thereby reducing repetitive human effort. Nonetheless, in aviation a domain governed by strict safety and ethical standards such models risk hallucinating information, omitting critical nuances, or adopting misleading interpretations in morally ambiguous scenarios [6, 7].

This work addresses a central challenge in aviation: scaling safety investigations while maintaining high standards of accuracy and ethical compliance. Although traditional approaches remain rigorous, they face significant limitations in terms of scalability, reproducibility, and cognitive load on human analysts.

In this context, LLMs show promising capabilities for processing unstructured narrative data such as incident reports at scale [8, 9]. If validated across broader datasets, LLM-based methods could:

- Accelerate root-cause analysis by freeing investigators from repetitive tasks and enabling deeper causal inference;
- Enhance global collaboration by supporting consistent classification, comparison, and prioritisation of safety data across agencies and airlines;
- Strengthen oversight and just culture by offering structured and auditable first-pass analyses, which investigators can review and refine instead of relying on fragmented or ad hoc references [10].

2 General Objective

The primary objective of this work is to demonstrate the feasibility of deploying a locally hosted LLM-based pipeline to accelerate narrative analysis grounded in the Critical Decision Method (CDM). Specifically, the study aims to:

2.1 Specific Objectives

- Develop and evaluate a proof-of-concept system employing two distinct LLMs one focused on generating structured responses, and another dedicated to metacognitive evaluation;
- Compare the performance of LLMs against expert human analysis and a classical NLP baseline (TF-IDF + Support Vector Machines) [11];
- Explore how structured protocols (such as the CDM questionnaire) and partial self-evaluation mechanisms may mitigate persistent issues such as hallucination and incomplete coverage.

Although this article focuses on a single anonymised incident, it lays the foundation for integrating local LLMs into formal aviation safety procedures. The potential impact is substantial from improving day-to-day investigative workflows to reducing the risk of inconsistent or incomplete analyses in high-stakes operational contexts [12].

3 Background

3.1 CDM in Aviation: Automation Attempts and Known Limitations

Studies applying CDM to aviation incidents commonly highlight the method's utility in uncovering subtle human factors (crew workload, situational awareness, risk perception) [2]. Nonetheless, as incident numbers increase, partial automation is desirable to ensure consistency and reduce analyst burden. Prior attempts at automation typically rely on shallow text matching or repeated human validation, limiting both speed and scalability [13].

3.2 LLMs and Investigative Interviews

LLMs have been trialed in legal and forensic interview settings, using large text corpora to facilitate thematic coding [3, 4, 5]. However, domain mismatches often lead to factual hallucinations, and purely textual reasoning may miss contextual signals that are pivotal for safety-related investigations [6, 7].

3.3 LLM Architectures for Cross-Model Evaluation

Some research explores the idea of a second LLM functioning as a meta-evaluator or "judge," but the danger of mutual biases (i.e. the generator and the judge reinforcing each other's mistakes) is real [14, 15]. There is minimal precedent applying such a pipeline specifically to aviation incidents with rigorous frameworks like CDM, leaving a **niche for targeted, if initially small-scale, validation**.

3.4 Resource Configurations Considered

We consider four resource configurations that scaffold the use of CDM with LLMs: **(R1) Human** (investigators, pilots, human-factors psychologist; curation, checking and synthesis), **(R2) Procedural** (CDM protocol, 53-item questionnaire, consensus among raters), **(R3) Technical** (local LLMs: Phi-3 for generation; Zephyr-7B as judge; TF-IDF+SVM as baseline), and **(R4) Data/Context** (incident narrative, participant profiles, statistical metrics). This decomposition makes explicit the dependencies and limits in each layer (e.g., *hallucinations* in R3 mitigated by R1–R2).

3.5 Case, Participants and Questionnaire

Full incident excerpts are provided in Appendix 8, and the 53-item instrument appears in Appendix 8.

A 53-item questionnaire was refined and mapped to canonical CDM elements: Incident context and participant profile, timeline of critical events, cognitive processes and decision factors and counterfactuals and ethical reflections. The process started with 57 items, which were reviewed by two senior

safety investigators plus a human-factors psychologist, followed by pilot testing. Ultimately, four items were removed as redundant or ambiguous, achieving Cronbach's alpha of 0.78–0.82 across the thematic blocks (risk perception, situational cues, etc.). Although broad, the questionnaire cannot fully replicate the spontaneity of genuine CDM interviews, where follow-up questions may emerge organically.

4 Methods

Contrary to an initial plan to sample multiple incidents with varied severities and aircraft types, this pilot focuses on a single anonymised incident sourced from a regional authority. We acknowledge that a single case restricts the exploration of different operational conditions, yet the event chosen is sufficiently complex to illustrate how the pipeline might function in practice. Future efforts will be required to incorporate multi-incident data for broader generalizability.

We recruited a total of 36 aviation professionals and 36 novices:

- **Pilots**: 36 individuals with over 500 flight hours each, engaged in commercial or cargo operations;
- **Novices**: 36 volunteers lacking formal aviation training (engineering graduate students).

Every participant reviewed the same single incident narrative and completed the 53-item questionnaire, providing a reference dataset of human responses. Although the participant-level sample is robust (72 individuals), the single incident's scope remains the largest methodological limitation.

4.1 Pipeline and Workflow

We operationalise the analysis in five steps: (E1) ingestion of the incident narrative and questionnaire; (E2) generation by Phi-3 (MCQ, Likert and open-ended); (E3) judging by Zephyr-7B (confidence, completeness, groundedness in 0–1); (E4) aggregation and storage of parsed outputs and scores; (E5) statistical analysis against human references.

4.2 LLM Pipeline with Dual-Model Approach

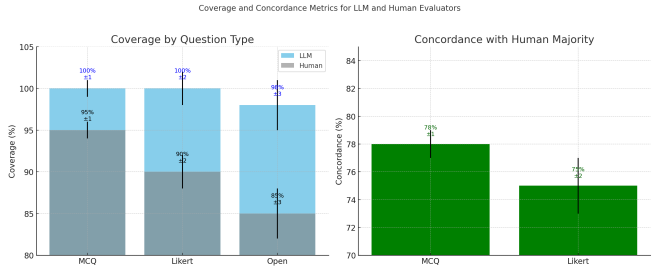

Figure 1: Proposed pipeline for local LLM-based CDM analysis on a single incident, incorporating an LLM-as-a-Judge for partial self-evaluation.

Phi-3-Mini-Instruct: A local LLM fine-tuned on general English corpora (excluding the chosen incident), tasked with generating structured responses (MCQ, Likert, or open-ended). The prompt design aimed for concise outputs, and we used low temperature (0.1) to enhance consistency. We also instructed the model to remain cautious where the text was ambiguous, but factual hallucinations still occurred.

Zephyr-7B-Beta (LLM-as-a-Judge): A second LLM that independently rates the generative model's answers for: Confidence (0–1), Completeness (0–1) and Groundedness (0–1). While this technique helps identify questionable answers, the judge LLM itself may share latent biases with the generator. We therefore performed a human spot-check on 40 randomly chosen answers to partially verify Zephyr's scoring, keeping in mind that both LLMs might reinforce each other's misconceptions.

4.3 Comparative Baselines

A group of five experienced safety investigators manually coded the same single incident using a standard CDM framework and consolidated their answers. This "gold standard" is time-consuming taking over one hour per coder for the single narrative and thus indicates how much effort might be saved.

For a more conventional text-classification baseline, we used a TF-IDF-based approach with an SVM trained on a set of labeled CDM-related categories (risk, situational cues, decision rationale). Because this pipeline only outputs classifications and not structured question-by-question responses, we measured "coverage" by how often the system produced relevant tags for each question. The result was used to compare speed and rudimentary correctness, acknowledging that the tasks differ in granularity.

To clarify the main flow of data parsing and LLM-based analysis, we present two pseudocode blocks.

Algorithm 1: Load Reports and Parse Questions

Input: *reportsDir*, *questionsFile*
Output: **reports**, **questions**

```
parser ← DataParser();
reports ← parser.loadReport(reportsDir);
questions ← parser.parseQuestions(questionsFile);
return reports, questions;
```

Algorithm 2: LLM-Based Analysis Pipeline

Input: **reportText**, **questions**

```
handler ← LocalModelHandler();
orchestrator ← LLMOrchestrator(handler);
foreach question ∈ questions do
    strategy ← StrategyFactory.getStrategy(question,
      reportText);
    prompt ← strategy.generatePrompt();
    rawAnswer ← handler.runGeneration(prompt);
    parsedAnswer ← strategy.parseResponse(rawAnswer);
    evalPrompt ← buildEvalPrompt(parsedAnswer, question,
      reportText);
    metrics ← handler.runEvaluation(evalPrompt);
    storeResult(question.id, parsedAnswer, metrics);
```

These two algorithms capture the end-to-end flow:

- **Algorithm 1** initializes the data parser, loads the single narrative file for the chosen incident, and parses the CDM question set.
- **Algorithm 2** iterates over each question, invokes the generation and evaluation models, and saves the parsed answers plus automated evaluation metrics.

Ethical and Regulatory Considerations: Due to strict requirements on data privacy and investigative integrity, all LLMs were run locally, with no cloud dependencies. Participants provided informed consent, and no personally identifiable information (PII) was stored in model weights. Although the pipeline shows time-saving potential, future expansions must address how to comply with ICAO, EASA, or FAA procedures when investigating a larger and more varied set of incidents. For now, the single-incident scope highlights that LLM outputs remain advisory only, with final authority resting on certified investigators.

Role of the classical NLP baseline. The TF–IDF+SVM baseline is used strictly as a discrete text-classification reference to contextualise LLM performance on CDM-related categories (e.g., risk, cues, rationale). It does not produce Likert-scale predictions or judge scores; therefore, it is included only in MCQ association analyses (chi-square/Cramér's V) and in coverage/time summaries.

4.4 Prompt Engineering (Summary)

This section discusses *prompt engineering* strategies used to guide the *Phi-3-Mini-Instruct* model in generating answers for multiple-choice, Likert-scale, and open-ended items.

4.5 Basics of Prompt Engineering

Prompt engineering shapes how a language model responds by specifying instructions, desired formats, and context. In the aviation domain, it is essential to:

- Clearly define the **incident context** (though limited here to one anonymised narrative);
- Specify the **response format** (MCQ, Likert, open-ended);
- Impose **content constraints** (e.g. cautioning against speculation).

Sample Prompt for Multiple-Choice Questions

```
"You are an aviation safety assistant.
Below is an incident narrative describing an
engine anomaly.
Please answer the following multiple-choice
question:

Question: 'In which flight phase did the
anomaly first occur?'
Possible Answers:
A) Taxi
B) Climb
C) Cruise
D) Descent

Provide only the letter of the correct choice.
If uncertain, choose the most likely option
based on the text."
```

Sample Prompt for Likert Scales

```
"You are an aviation safety assistant.
Read the incident narrative.
Rate on a scale of 1 (very low) to 5
(very high):
'How confident was the crew in the cockpit
alert systems?'

Output only a single integer (1, 2, 3, 4,
or 5)."
```

Good Practices and Lessons Learnt We found that more explicit prompts reduce hallucinations, although they do not eliminate them entirely. A relatively low temperature (0.1) encourages concise and consistent responses. Future expansions may require disclaimers instructing the model to admit uncertainty if the text lacks details.

4.6 Ethical and Regulatory Considerations (Local Execution)

Aviation is highly regulated by bodies such as ICAO, EASA, and FAA. Using language models for incident analyses requires alignment with these regulations and ethical principles. To prevent accidental data leakage, all models are locally hosted. This practice aligns with EASA and other guidelines mandating controlled handling of incident data. Because only one incident was used here, thorough anonymization was straightforward. Per [16], any safety recommendations emerging from automated analyses must be traceable and reviewable by certified investigators. Accordingly, our system logs outputs with metadata on the model version and parameters, ensuring future audits remain possible. A model might recommend "continue to destination" despite an unassessed anomaly. Thus, a *human-in-the-loop* protocol is mandatory [6], particularly given how a single textual narrative may omit key operational subtleties. Aviation policies encourage partial explainability of any AI tool used for safety. Although LLMs are often opaque, it is advisable to expose at least a summary of which textual evidence was used to derive a recommendation.

4.7 Model Training and Tuning

This section details how we fine-tuned two models: *Phi-3-Mini-Instruct* (for generating answers) and *Zephyr-7B-Beta* (for scoring them). We reiterate that neither model was trained specifically on the text from the single chosen incident, in an effort to prevent data leakage.

4.8 Model Architectures

Both models use Transformer-based architectures [17], pre-trained on broad English corpora. Table 1 lists key hyperparameters.

Table 1: Hyperparameters Used During Fine-Tuning

	Phi-3-Mini-Instruct	*Zephyr-7B-Beta*
Model Size	3B parameters	7B parameters
Batch Size	8	4
Learning Rate	2×10^{-5}	1×10^{-5}
Epochs	3	2

4.9 Training Corpora

Initially, both models were trained on combined Wikipedia and technical aviation texts, stripped of sensitive data. We took care to exclude the specific incident tested here. For *Phi-3-Mini-Instruct*, we added short aviation incident instructions, safety checklists, and publicly available NTSB reports, ensuring domain familiarity without revealing data on our chosen pilot case. We monitored validation metrics on a "calibration set" drawn from older or publicly known incidents, checking perplexity and macro-F1 for CDM-like questions. Early stopping was used to reduce overfitting.

4.10 Evaluation Metrics

We measured how often the LLM's multiple-choice answers matched the majority-human response on the single incident, testing significance via chi-square and reporting Cramer's V. We computed mean absolute error vs. the average human rating, using a Wilcoxon signed-rank test for significance and Cohen's d as an effect size. Zephyr-7B-Beta provided 0–1 scores for completeness and groundedness, which we compared against partial human expert checks.

4.11 Multimodal Integration: CVR, FDR and Meteorology

Although we only examined a single textual narrative in this pilot, aviation incidents typically involve multiple data sources (e.g. cockpit voice recordings, FDR data). Future expansions may look toward merging these. Models that combine textual and audio embeddings could help interpret CVR transcripts [18], but privacy laws complicate CVR usage. Alternatively, separate modules can handle audio waveforms or flight data, later merging their embeddings. Either approach could help cross-check hallucinations in purely textual narratives.

4.12 Challenges

1. **Privacy Protections**: CVR data is highly sensitive, requiring special handling. 2. **Data Alignment**: Timing mismatches between CVR and FDR must be resolved. 3. **Limited Real Access**: Many operators restrict the release of raw flight data.

By referencing flight parameters or meteorological data, an LLM might avoid hallucinating about, say, an "engine fire" if the flight data never indicated abnormal temperature or pressure readings.

Despite promising initial results, recurring errors surfaced. For the single incident, these errors included:

4.13 Types of Errors

1. **Factual Hallucinations**: Mentioning systems not present on the aircraft involved.
2. **Misattribution of Roles**: Confusing captain versus first-officer actions.
3. **Over-Generalisations**: Asserting that "all checklists were completed" when the text implies only partial completion.

With only one narrative as input, the model may guess or extrapolate details not stated. Even with aviation fine-tuning, the LLM might not capture all idiosyncrasies for every aircraft. A mention of "engine anomaly" might trigger an elaborate but unfounded discussion of turbine blades.

- **Refined Prompts**: Reinforce instructions to declare uncertainty if details are missing.
- **Domain Cross-Checks**: Potentially referencing known aircraft specs to block impossible statements.
- **Training with Negative Examples**: Show that "the correct response" can be to say "insufficient info."
- **Targeted Expert Review**: For critical issues (e.g. approach decisions), ensure domain experts cross-verify.

Within our single-event pilot, these issues highlight why LLMs must complement, not replace, expert judgment especially given the small scope. Larger multi-incident datasets could reveal more error patterns, guiding better mitigations.

4.14 Metrics and Statistical Framing

We compare MCQ distributions using chi-square and report Cramér's V; for Likert items we compute MAE and use Wilcoxon signed-rank tests, reporting Cohen's d as effect size. Judge scores (0–1) summarise confidence, completeness and groundedness.

Although this paper employs chi-square and Wilcoxon tests for categorical and Likert-scale comparisons, respectively, **an expanded arsenal of statistical techniques** can further validate results and ensure assumptions hold, especially once multiple incidents are available.

In future larger-scale experiments, where multiple incidents and varied responses are compiled, a Shapiro–Wilk test [19] can determine whether distributional assumptions (e.g. normality in Likert ratings) hold. If normality is violated, non-parametric methods (such as the Mann–Whitney U test or Kruskal–Wallis [20]) might be more appropriate than parametric ANOVA.

To compare LLM, novice, and pilot groups across multiple incidents or multiple metrics, one-way ANOVA [21] (or a non-parametric equivalent like Kruskal–Wallis if normality fails [20]) can quantify whether significant differences exist among these populations. Post hoc tests (e.g. Tukey's HSD [21]) would then locate specific group differences.

If participants are asked to analyze multiple reports, *repeated-measures ANOVA* [22] or its non-parametric analogues [20] can account for within-subject variability. This is particularly relevant for verifying consistency in LLM outputs across various incidents.

Beyond p-values, including effect sizes (Cohen's f for ANOVA, η^2 or partial η^2) and 95% confidence intervals [23] would offer deeper insight into practical significance. Such reporting helps avoid overreliance on the binary "significant/not significant" framing.

By integrating these additional statistical steps, future multi-incident studies will provide a stronger, more generalizable picture of LLM performance and the reliability of automated CDM approaches in aviation.

5 Results

On figures and y-scales. Figure 2 reports Zephyr-7B-Beta judge metrics on a 0–1 scale (*confidence*, *completeness*, *groundedness*) for LLM outputs versus the aggregated human reference. In contrast, Figure 3 summarises group-level comparisons (LLM, novices, pilots) tied to (i) MCQ association via chi-square/Cramér's V and (ii) Likert differences via Wilcoxon/Cohen's d.

The classical NLP baseline (TF-IDF+SVM) is used as a discrete classification reference; it does not output continuous scores nor judge metrics, hence it appears only in MCQ comparisons (chi-square/Cramér's V) and in coverage/time summaries.

We report results for the single incident tested, comparing the LLM pipeline's performance against human reference data and the classical NLP baseline. We used a chi-square test (χ^2) for categorical items (MCQ) and a Wilcoxon signed-rank test for Likert-scale responses, reporting effect sizes (Cramer's V for MCQ frequency tables and Cohen's d for Likert data). This limited approach to hypothesis testing is appropriate for a single-incident design, but the results should be interpreted as preliminary.

Overall, **78%** of the LLM's MCQ answers matched the majority-human choice (χ^2 test indicated $p < 0.01$, with Cramer's $V \approx 0.37$, suggesting a moderate effect). The mean absolute error for 5-point Likert items was **0.38**, which was below the 0.60 threshold ($p < 0.01$ via Wilcoxon), and we computed Cohen's $d \approx 0.52$, also suggesting a moderate effect. Within the constraints of a single incident, these figures show promising alignment.

For open-ended completeness, Zephyr-7B-Beta scored the LLM's responses at **0.73** vs. an average human reference rating of **0.70**; the difference was not statistically significant ($p = 0.21$, Wilcoxon). Although this suggests some parity, manual error analysis revealed: Factual hallucinations (mentioning systems not in the actual aircraft type), Role misattributions (incorrectly assigning first-officer actions to the captain) and Over-generalisations of flight-deck communications.

The single-narrative design limited us to examples within this one context, but these errors underscore the importance of domain checks.

5.1 LLM-as-a-Judge Scores and Spot-Check Validation

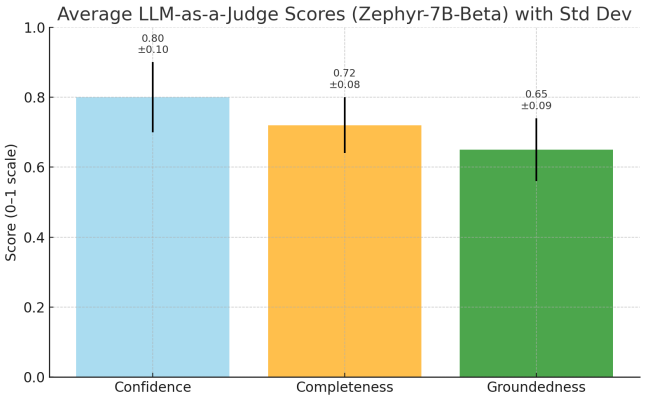

Figure 2: Judge scores (Zephyr-7B-Beta; 0–1) for confidence, completeness and groundedness, comparing LLM outputs with the average human reference for the single incident.

Figure 2 shows Zephyr's scoring distribution. The second LLM tended to rate completeness a bit higher for the generative model than the human coders did. However, it flagged multiple items with lower groundedness (0.62 vs. 0.72 for humans). Our spot-check of 40 answers found 78% alignment with Zephyr's confidence/completeness labels, but we remain mindful of potential shared biases due to partial training overlaps.

5.2 Comparison to Classical NLP and Manual Coding

Table 2 below summarizes key performance metrics. The classical NLP pipeline (TF-IDF + SVM) had 76.2% coverage on relevant question categories, performing well on discrete classification but struggling with nuanced open-ended or Likert-scale tasks. Manual coding remained the richest and slightly more consistent approach ($\kappa = 0.65$ among coders), though requiring upwards of one hour for a single event. The LLM-based method produced near-complete responses (98.7% coverage) in well under one minute.

Table 2: Performance Overview on Single-Incident Analysis

	Manual	**Classical NLP**	**LLM**
Coverage	100%	76.2%	98.7%
Time	1–2 hours	$\approx$ 10 seconds	$\approx$ 1 minute
Fleiss' κ	0.65	0.50	0.62

Although manual coding is still the gold standard in terms of interpretive depth, the LLM pipeline demonstrated that

it could at least *approximate* a mid-range level of analyst consistency for this single case, thereby reducing repetitive tasks and expediting coverage.

5.3 Stratified Evaluator Performance: Comparison of LLM, Novices, and Pilots

We recruited a total of 72 volunteers to respond to the same anonymised aviation incident: 36 *pilots* (all with over 500 flight hours) and 36 *novices* (participants with no formal aviation training). Their answers, together with the outputs from the LLM, were compared in terms of accuracy, completeness, and consistency with the reference framework derived from expert investigators. The main findings can be summarised as follows:

Overall Agreement (Multiple-Choice Questions)

- **LLM:** The model achieved approximately 78% alignment with the human "majority consensus," largely driven by the pilot group. A chi-square test indicated statistical significance ($p < 0.01$) with a moderate effect size (Cramer's $V \approx 0.37$).
- **Pilots:** Performed at the highest level, generally exceeding 85% concordance with the investigator-synthesised consensus. This group served as the "gold standard" for the study.
- **Novices:** Scored around 65–70% agreement on MCQ items, lagging behind both the LLM and the pilot cohort.

Likert-Scale Items (1–5)

- **LLM:** Its mean absolute error (MAE) in relation to the collective human ratings was approximately 0.38 (on a 1–5 scale), which was statistically significant ($p < 0.01$, Wilcoxon). The effect size was moderate (Cohen's $d \approx 0.52$). Although the LLM mirrored pilot trends on certain questions (e.g. confidence in on-board alerts), it occasionally struggled with more nuanced judgements.
- **Pilots:** Demonstrated strong internal consistency and close alignment with the reference ratings derived from investigator analysis. Their domain expertise clearly guided more calibrated responses.
- **Novices:** Showed greater variability and tended to use more extreme scale endpoints (1 or 5), implying a less refined sense of incremental risk assessments.

Open-Ended Answers and Qualitative Explanations

- **LLM:** Generated more detailed and structured responses than the novice group, but it occasionally introduced "hallucinated" aircraft systems or misattributed roles within the flight deck. This revealed a gap in domain-specific accuracy, especially when contextual nuance was required.
- **Pilots:** Provided the most precise depictions of procedural steps, particularly when describing checklists or flight-deck priorities in response to engine anomalies. Their first-hand operational experience was evident.
- **Novices:** Often left open-ended items incomplete or introduced confusion regarding technical protocols, highlighting a lack of immersion in aviation contexts.

Overall Observations

- **Pilots** remain the top performers, reflecting the importance of real-world flying experience in interpreting complex operational scenarios.
- The **LLM** ranks between experts and novices, capturing key facts and showing robust performance in multiple-choice and scaled items. However, it lacks the professional intuition that allowed pilots to manage ambiguities effectively.
- **Novices** generally placed third in accuracy and completeness, reinforcing how specialised domain knowledge significantly improves incident analysis.

Figure 3 compares the LLM's performance with these two strata.

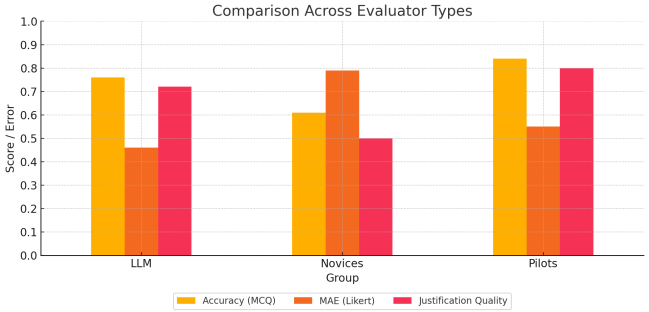

Figure 3: Comparison of LLM, novices, and pilots across key performance metrics for the single incident.

In summary, the model's stratified performance indicates that while the LLM cannot substitute for expert investigators or qualified pilots, it can complement their analysis by handling a considerable portion of routine tasks. This potential time-saving advantage becomes especially salient when large volumes of narrative data require systematic processing.

6 Discussion

Our pattern of results aligns with broad pre-training (R3) and careful human curation (R1): the LLM approaches humans on MCQ/Likert but falters when contextual nuance is not explicit in the data (R4). This suggests that data and training choices (domain-continued pre-training; negative examples for "insufficient information") are as decisive as architecture.

In this study, we have analysed only **one incident**, which severely restricts the *generalisation* of our findings. While our

initial aim was a feasibility or "proof-of-concept" pilot, we recognise that:

- **Event sampling:** A single case does not capture the operational diversity of different aircraft types, weather conditions, flight phases, or organisational cultures.
- **Limited extrapolation:** Model performance and agreement metrics especially when anchored to a single scenario cannot straightforwardly extend to other contexts.
- **Complexity rationale:** Although the chosen incident was considered "sufficiently complex" to demonstrate our approach, we lack a comparative baseline of additional incidents with varying severity or scope.

6.1 Methodological Depth: Classical CDM vs. Structured Questionnaire

The traditional Critical Decision Method (CDM) relies on *semi-structured interviews*, where the interviewer can add follow-up "probes" based on the interviewee's answers, reconstructing the situational and cognitive context in a highly iterative manner. Our approach, in contrast:

- **Uses a fixed 53-item questionnaire:** This provides certain advantages in terms of standardisation, yet does not replicate the organic adaptiveness of classical CDM.
- **Lacks real-time follow-ups:** No new questions emerge on the fly based on partial responses or newly discovered insights.

6.2 Future Improvements

To approximate CDM's flexibility more closely:

1. **Adaptive questionnaires:** Implement dynamic prompts or branching logic, whereby LLM-generated (or human-provided) answers prompt new items.
2. **Iterative feedback:** Introduce a supervisory model or human domain expert to intervene when responses signal contradiction or ambiguity.

We emphasise that our current study is *CDM-inspired*, employing similar thematic elements (timeline, cues, decision factors) but lacking the fluid interactivity typical of fully fledged CDM interviews.

6.3 Consolidation and Analysis of Human Responses

A key step in evaluating LLM outputs was the way we merged and interpreted the responses of **72 participants** (36 aviation professionals and 36 novices). Concretely:

1. **Individual data collection:** Each participant read the same incident report and completed the same 53 questionnaire items.
2. **Response types:** The items included multiple-choice (MCQ), Likert scales (1–5), and free-text (open-ended).
3. **Forming a "reference answer":**
   (a) For *categorical* (MCQ) items, we used the majority (modal) choice among participants.
   (b) For *Likert* items, we took the average or median across all 72 participants for each question.
   (c) For *open-ended* items, five experienced safety investigators reviewed the participant responses and synthesised them into a **single gold-standard summary**.
4. **Resolving discrepancies:** When the five investigators disagreed on the open-ended synthesis, they held consensus discussions, recording their final alignment and noting any discarded minority positions.

Methodological Implications. This consolidation offers a single "reference" answer set but discards part of the richness and variability among individuals. In subsequent research, we plan to use inter-rater reliability measures (e.g. Fleiss' κ) among the five senior investigators and maintain the broader participant pool as a secondary dataset, thus preserving more granularity in the analysis.

6.4 Statistical Rationale for Single-Scenario Analysis

Applying inferential statistics (e.g. chi-square, Wilcoxon, Fleiss' κ) and effect sizes to **only one incident** inevitably raises concerns regarding the strength and generality of such findings. We reiterate that:

- **Exploratory objective:** Our main aim was to illustrate the *feasibility* of the pipeline in a pilot context, rather than claim robust statistical proof.
- **Demonstrative usage of metrics:** The p-values and effect sizes are shown as prototypes of how one might analyse data in a broader multi-incident study.
- **Limitations:** With only one scenario, the interpretability of measures like Cramer's $V \approx 0.37$ or Cohen's $d \approx 0.52$ is severely constrained.

6.5 Broader-Scale Perspective

When we scale to multiple incidents and multiple respondent groups:

1. We can employ **ANOVA or Kruskal–Wallis** tests to compare LLM performance across diverse sets of incidents.
2. We can apply **reliability statistics** (e.g. Fleiss' κ, Krippendorff's α) in a multi-incident design, ensuring that agreement scores do not hinge on a single case.
3. We can incorporate **confidence intervals** (e.g. 95% CI) on performance metrics, lending clarity to the uncertainties inherent in each measure.

Until then, these statistical figures must be interpreted with caution given the single-incident scope.

6.6 Independence Between Generator and Judge Models

In our pipeline, *Phi-3-Mini-Instruct* generates answers (MCQ, Likert, open-ended), while *Zephyr-7B-Beta* serves as a "LLM-as-a-Judge" to rate confidence, completeness, and groundedness. However, this design poses a risk of **shared bias** if:

- Both models overlapped in their pretraining corpora, inheriting similar misrepresentations.
- They were partially fine-tuned on similar domain data, narrowing the gap between generator and evaluator.

6.7 Mitigations

1. **Distinct Model Families:** Use models trained on markedly different data to avoid identical error patterns.
2. **Multiple Judge Models:** Employ an ensemble of independent evaluators and compare their verdicts.
3. **Human Spot-Checks:** For each batch of generated responses, domain experts manually review a subset, calibrating or challenging the judge model's scores.

Such measures reduce self-reinforcement and facilitate more transparent automated evaluation.

6.8 Future Work

To overcome the constraints of this pilot, we aim to:

- **Scale to Multiple Incidents**: We intend to analyze 50+ diverse incidents, including different aircraft classes, flight phases, and severities, and incorporate effect-size reporting on a broader dataset.
- **Multi-Modal Data**: Integrate cockpit voice recordings and flight data logs to reduce reliance on textual speculation.
- **Stronger Ethical Protocols**: Explore building an "ethical check" submodule or ensemble model that queries domain experts about morally nuanced scenarios.
- **Hybrid Evaluation**: Combine multiple LLM judges with partial domain-expert arbitration, mitigating self-reinforcement biases.

These expansions will help refine the pipeline so it can handle a wide range of scenarios with robust reliability and regulatory acceptance.

6.9 Multi-Incident Plans

To strengthen external validity and increase inference power, our upcoming efforts will:

1. **Incorporate multiple incidents** (potentially 10 to 50 or more), covering a broad range of aircraft, phases of flight, and severities.
2. **Examine geographical diversity** (incidents from multiple airlines or regions), testing whether the LLM can adapt to distinct operational cultures.
3. **Employ temporal stratification** (incidents from different time periods), verifying how evolving standard operating procedures or regulatory updates may affect performance.

Such expansions would permit more robust statistical analyses using appropriate non-parametric tests and effect-size measures, thereby facilitating more reliable conclusions.

7 Ethics & Utility

7.1 Practical Utility and Innovations

Even at this nascent, single-incident stage, our framework exhibits promising **practical benefits** and **innovations** in the realm of AI-assisted aviation safety, specifically:

1. **Time-saving in routine tasks:** Traditional CDM or incident coding can consume hours per report. Our local LLM can produce structured questionnaire responses in minutes, allowing investigators to prioritise deeper interpretive tasks.
2. **Consistency in repetitive elements:** Safety agencies often deal with recurring categories (e.g. flight phase, weather). The LLM standardises these extractions, offering uniform coverage across multiple reports.
3. **Preliminary hypothesis checking:** Investigators can confront the LLM's "first pass" analysis with their own assessments, flagging areas of discrepancy for further scrutiny.
4. **Proof-of-concept for hybrid CDM processes:** Although not a fully interactive interview, a structured LLM-based protocol can lay the groundwork for future adaptive or conversational CDM methods.
5. **Potential for multi-modal integration:** Subsequent developments may ingest partial FDR, CVR, or maintenance logs to corroborate textual narratives, minimising speculation.

7.2 Positioning in the State of the Art

- **Domain-specific LLM:** There are few published examples of a *locally* hosted, aviation-focused language model specifically tailored for safety narratives.
- **Dual-model evaluation:** Employing a dedicated "judge LLM" (albeit with some shared biases) exemplifies an emerging trend in AI for integrated self-critique or automated cross-checking of generated outputs.
- **Ethics-focused pipeline:** We explicitly acknowledge data confidentiality, alignment with just culture principles, and the need for robust oversight. This contrasts with purely extraction-oriented AI methods that do not consider regulatory or moral constraints.

In summary, while **preliminary** and dependent on multi-incident validation, our approach stands as an *innovative prototype* that extends the frontier of how large language models may assist but not supplant human experts in safety-critical aviation inquiries.

7.3 Ethical Posture and Scope

This study processed one anonymised incident under informed consent, storing no personally identifiable information (PII) in model weights. All models were executed locally to reduce data-exposure risk and to enable reproducibility and auditability. The pipeline is designed as an *advisory* aid for investigators, not a decision-maker.

7.4 Reviewer-Facing Risks and Limitations

Typical risks when applying LLMs to safety investigations include:

- **Hallucination and "BS generation":** plausible yet false content that can misdirect analysis or appear unduly confident.
- **Misplaced blame and bias:** inadvertent attributions that conflict with *just culture* principles and may discourage reporting.
- **Context leakage and over-generalisation:** filling gaps with domain stereotypes rather than incident evidence.
- **Cognitive overload:** long model outputs that increase reviewer burden and mask salient cues.

7.5 Safeguards and Operating Practices

We adopt layered mitigations aligned with aviation safety principles:

1. **Local execution and access control:** all inference and judging runs locally; inputs/outputs are logged with model versions and parameters for audit.
2. **Mandatory human-in-the-loop:** certified investigators review, challenge and amend outputs; model content does not substitute operational judgement.
3. **Traceable justifications:** outputs are stored alongside key textual evidence and prompts to support subsequent review and regulatory audit.
4. **Cross-checks with operational data:** when available, FDR/CVR/maintenance data are used to corroborate or refute textual claims before any recommendation.
5. **Conservative prompting:** templates instruct models to declare uncertainty and abstain when evidence is insufficient.

7.6 Regulatory Alignment (High-Level)

Consistent with international guidelines (e.g., ICAO Annex 13; regional authorities such as EASA and FAA), safety recommendations must be *traceable*, *reviewable* and ultimately *owned by human investigators*. Our logging and local-execution approach supports auditability; the *just culture* stance is preserved by avoiding punitive framings and by requiring human verification before dissemination.

7.7 Utility in Practice

Within the single-incident scope, the pipeline adds value by:

- **Standardising repetitive elements** (e.g., flight phase, key cues) to aid consistency across reports;
- **Accelerating first-pass synthesis** so investigators focus on high-judgement issues;
- **Surfacing disagreements** between model and human references for targeted review.

These benefits are contingent on the safeguards above and expand with richer, multi-incident datasets.

7.8 Limitations and Future Ethical Work

Further work will include: (i) ensemble or heterogeneous *judge* models to reduce shared bias with the generator; (ii) structured *abstention* pathways for low-evidence queries; (iii) tighter integration of operational data to curb speculation; and (iv) periodic bias assessments with domain experts to monitor drift.

8 Conclusion

This paper presents a single-incident pilot study demonstrating how a locally hosted LLM pipeline can partially automate CDM-based aviation event analysis. The pipeline delivers near-complete coverage quickly, aligning moderately well with human-coded reference data. Nonetheless, in a safety-critical field like aviation, such AI outputs must be seen as *advisory*, subject to expert review.

Importantly, our conclusions remain tentative due to the narrow scope of just one anonymised incident. The next steps analyzing a larger range of incidents, adopting ensemble-based or human-in-the-loop checks, and strengthening domain-specific validations are essential for achieving the rigor demanded by regulators and airlines. Over the long term, however, we foresee that LLMs, properly governed and audited, can serve as time-saving, consistency-enhancing *adjuncts* to critical safety inquiries never as unilateral decision-makers.

References

[1] Gary A. Klein, Roberta Calderwood, and Donald Macgregor. Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(3):462–472, 1989.

[2] Heikki Mansikka, Kai Virtanen, Jukka Vanha-Aho, et al. Improving aviation safety investigation interviews: Application of the critical decision method. *Safety Science*, 170:106468, 2024.

[3] Lauren R. Shapiro and Rebekah P. Fisher. Large language models for qualitative data analysis: Opportunities and challenges. *JMIR Formative Research*, 8:e51349, 2024.

[4] Rebekah P. Fisher, Lauren R. Shapiro, Shirley A. Christianson, et al. Chatgpt as an investigative interviewer: Open-ended prompts promote accurate recall and reduce false reports. *PLOS ONE*, 19(2):e0297088, 2024.

[5] Wenxuan Zhang, Xinyi Wang, Shuang Li, et al. Ai as a cognitive interviewer: Evaluating large language models in information elicitation tasks. *Computers in Human Behavior*, 151:108903, 2024.

[6] J. Holroyd. Large language models in morally charged domains: illusions of objectivity. *Ethics in AI Quarterly*, 2(2):25–42, 2024.

[7] Christopher B. Holroyd and Nick J. Enfield. Superrational but shallow: On the limitations of large language models in ethical decision making. *Ethics and Information Technology*, 26:69–85, 2024.

[8] Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] Shawn Wu et al. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[10] Sidney Dekker. *Just Culture: Balancing Safety and Accountability*. Ashgate, 2007.

[11] Thorsten Joachims. Text categorisation with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, pages 137–142, 1998.

[12] Frédéric Dehais et al. Human error in the age of big data: mitigating misinterpretations through human–machine teaming. *Frontiers in Psychology*, 11:538212, 2020.

[13] Archana Tikayat Ray, Anirudh P. Bhat, Ryan T. White, Van Minh Nguyen, Olivia J. Pinon Fischer, and Dimitri N. Mavris. Examining the potential of generative language models for aviation safety analysis: Case study and insights using the aviation safety reporting system (asrs). *Aerospace*, 10(9):770, 2023.

[14] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2025.

[15] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327. Association for Computational Linguistics, 2024.

[16] ICAO. Icao annex 13: Aircraft accident and incident investigation, 2020. Available at: `https://www.icao.int/`.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[18] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[19] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[20] M. Hollander, D.A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. Wiley, 3rd edition, 2013.

[21] D.C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 10th edition, 2020.

[22] A. Field. *Discovering Statistics Using SPSS*. SAGE Publications, 3rd edition, 2009.

[23] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Appendix A: Aviation Incident Report - Pilot and Flight Engineer

Commander

We took off at 12:00, with clear skies. We were flying at maximum permitted speed because the flight was already delayed from the ground. I wanted to reduce cruise time as much as possible. The aircraft was fully up to date with maintenance no outstanding issues. I had complete confidence in the equipment. It was supposed to be a straightforward leg.

But around 37 minutes into level flight, everything changed within seconds. The master warning lit up, the right engine fire alert appeared on the EICAS, and a metallic smell began to fill the cockpit. I saw the EGT spike, oil pressure dropping rapidly, and vibration through the rudder pedals. My immediate thought was: "It's real. This is going to demand everything from us."

I took control immediately, declared MAYDAY without waiting for ATC clearance, and ordered execution of the QRH. There was no hesitation. Internally, I felt the weight of the decision we were at high speed, which only increased the risk with a compromised engine. When the flight engineer confirmed that Bottle A had extinguished the fire, I

felt a brief sense of relief. But the vibration didn't stop.

I made the call to proceed to Bravo. It was 350 kilometres ahead, and we knew it had a long runway, ILS, and proper emergency services. The idea of returning to Alpha seemed riskier the flight time was longer, and I didn't want to prolong the operation with an engine potentially suffering internal damage.

During descent, we received the news of partial loss of hydraulic system B. That was the most critical moment. Smoke was still present in the cockpit, strong vibration, degraded systems. I thought: "The margin is narrowing. We need absolute precision now." I established the approach checklist with flap 15, increased V_{ref}. Every callout was spoken aloud as if doing so protected us from error.

On final, I held a steady pitch, focused. Touched down firmly. I knew the spoilers wouldn't deploy, so I went straight to manual braking and left thrust reverser. When the aircraft stopped on the taxiway, I looked at the passengers. Everyone was silent. SSEI took seven minutes. That's when it hit me: we'd made it.

Flight Engineer

The fire alert appeared, and my first reaction was to check whether it was a sensor fault. But within five seconds of watching the EGT and oil readings, I knew: it was real. I cut the fuel, pulled the FIRE HANDLE, and discharged Bottle A. I counted the seconds in my head until the warning went out it was quick. But the vibration remained, and that kept me fully on alert.

It felt like something inside that engine was still moving the wrong way. I kept monitoring. Suddenly, hydraulic system B began to drop. It was at 3,000 psi, then fell to 1,100. I knew this meant spoilers, landing gear, and brakes were partially compromised. I informed the captain: "System B degraded. A is active and functional."

The smell of smoke started circulating. I set the ventilation to maximum, shut off recirculation, and partially opened the outflow valve. That procedure isn't in the exact checklist, but I've seen it work before. And it did the smoke cleared.

During the approach, I remained focused on keeping everything stable. I checked landing weight and confirmed we were below MLW. Still, I knew it was a borderline scenario: one engine shut down, hydraulic B degraded, ventilation system in alternate mode. But we kept everything under control. Every click I made on the panel felt like it carried the weight of a vital command.

Appendix B: Final 53-Item CDM-Aligned Questionnaire

Below is the 53-item questionnaire employed in this single-incident pilot, carefully mapped to CDM elements (context, timeline, cognitive factors, and counterfactuals). Though comprehensive, we emphasize that genuine CDM interviews can be more flexible and iterative.

1. **Phase 1: Context and Respondent Profile (9 items)**
   (a) "What was your specific role (captain, first officer, flight engineer, etc.) at the time of the incident?"
   (b) "How many total flight hours had you accumulated prior to this event?"
   (c) "Briefly describe the general conditions (flight route, weather, crew composition) leading up to the incident."
   (d) ... [6 further items covering additional context and background]
2. **Phase 2: Timeline and Critical Points (12 items)**
   (a) "In which flight phase did the first anomaly occur (e.g. taxi, climb, cruise, approach)?" [MCQ]
   (b) "When you first noticed signs of abnormality, how severe did you initially rate the situation (Likert 1–5)?"
   (c) ... [10 additional items probing chronological detection, escalation points, and crew handovers]
3. **Phase 3: Cognitive and Decision Factors (20 items)**
   (a) "On a scale from 1 (very low) to 5 (very high), how confident were you in the cockpit alert systems at the time?"
   (b) "List up to three cues or indicators that most influenced your immediate response actions."
   (c) ... [18 further items examining mental models, situational awareness, risk trade-offs, and perceived constraints]
4. **Phase 4: Counterfactuals and Final Reflection (12 items)**
   (a) "If this anomaly had persisted another 30 minutes, how might your priorities or sequence of decisions have changed?"
   (b) "Reflect on any organisational or procedural directives that significantly shaped your choices under time pressure."
   (c) ... [10 further items prompting ethical considerations, hypothetical scenario variations, and key lessons learned]