AI-Assisted Debrief: Automated Flight Debriefing Summarization and Competency Assessment

Mathijs Henquet[a] and Thomas Bellucci[b]
NLR - Royal Netherlands Aerospace Centre, Amsterdam, The Netherlands

Keywords: Aviation Training, Speech Recognition, Large Language Models, Performance Indicators, Pilot Competency Assessment, Human-Machine Interaction, Artificial Intelligence in Aviation, Qualitative Evaluation, Intelligent Training Systems.

Abstract: This paper seeks to explore the use of speech recognition and large language models (LLMs) to support the reporting process of flight debriefings in aviation training. We develop a system called AI-Assisted Debrief (AAD), which automatically transcribes and summarizes flight debriefings, thereby improving reporting and, in turn, improving knowledge transfer and pilot competency development. In addition, AAD assesses pilot competencies by identifying associated Performance Indicators (PIs) from the debriefs, yielding an automatic assessment of desired competencies to guide future training. We qualitatively evaluate the performance of our system using a dataset of five representative debrief recordings from flight training sessions, which are analysed by AAD and evaluated by experienced flight instructors. Our evaluation shows AAD to be capable of automatically extracting feedback and crucial information, recognizing desired pilot competencies. Future versions of AAD could enable tracking of competency development over time, offering a new method to guide aviation training. We envision AAD evolving into an interactive system which learns from human oversight to improve its accuracy and effectiveness. Propelling aviation training into the AI era, AAD paves the way for a more accurate, efficient, and comprehensive approach to pilot training, setting a new standard for excellence in the skies.

1 INTRODUCTION

Post-flight debriefing stands as a vital component of aviation training, serving as a conduit for knowledge transfer and skill refinement between flight instructors and trainees, while ensuring safety and operational standards within the aviation domain. During flight debriefing, the flight instructor provides verbal feedback to the trainee, covering various aspects, from take-off and landing procedures to regulatory requirements and the decision-making process employed by the pilot. This feedback in turn provides valuable insights to the pilot to correct performance (Mavin, Kikkawa, & Billett, 2018).

The efficacy of flight debriefing is enhanced by its fluid structure, enabling the instructor to tailor their feedback to the student and act responsively to their needs; however, while the unstructured nature of conventional debriefings has been found to aid the learning process (Roth, 2015), it may hamper systematic documentation and reporting. The resulting lack of documentation limits the ability to track progress over multiple sessions, identify overlooked areas, and reinforce learning outcomes. In practice, session reports typically encompass only the instructor's prior observations during the flight, neglecting the nuanced learning moments that emerge during the debriefing. Without a third-party documenting these sessions, important details may go unrecorded or be missed due to information overload, interruptions, and distractions. This gap highlights the need for a more structured approach to document flight debriefings to capture the full scope of learning moments and discussions.

Recent advancements in Artificial Intelligence (AI), particularly Large Language Models (LLMs) (Josh Achiam, 2023), have demonstrated the potential to revolutionize various areas of society. LLMs have shown near-human proficiency in tasks that require

[a] https://orcid.org/0009-0005-9688-2558
[b] https://orcid.org/0000-0002-9044-585X

Henquet, M. and Bellucci, T.
AI-Assisted Debrief: Automated Flight Debriefing Summarization and Competency Assessment.
DOI: 10.5220/0012970200004562

In *Proceedings of the 2nd International Conference on Cognitive Aircraft Systems (ICCAS 2024)*, pages 72-79
ISBN: 978-989-758-724-5

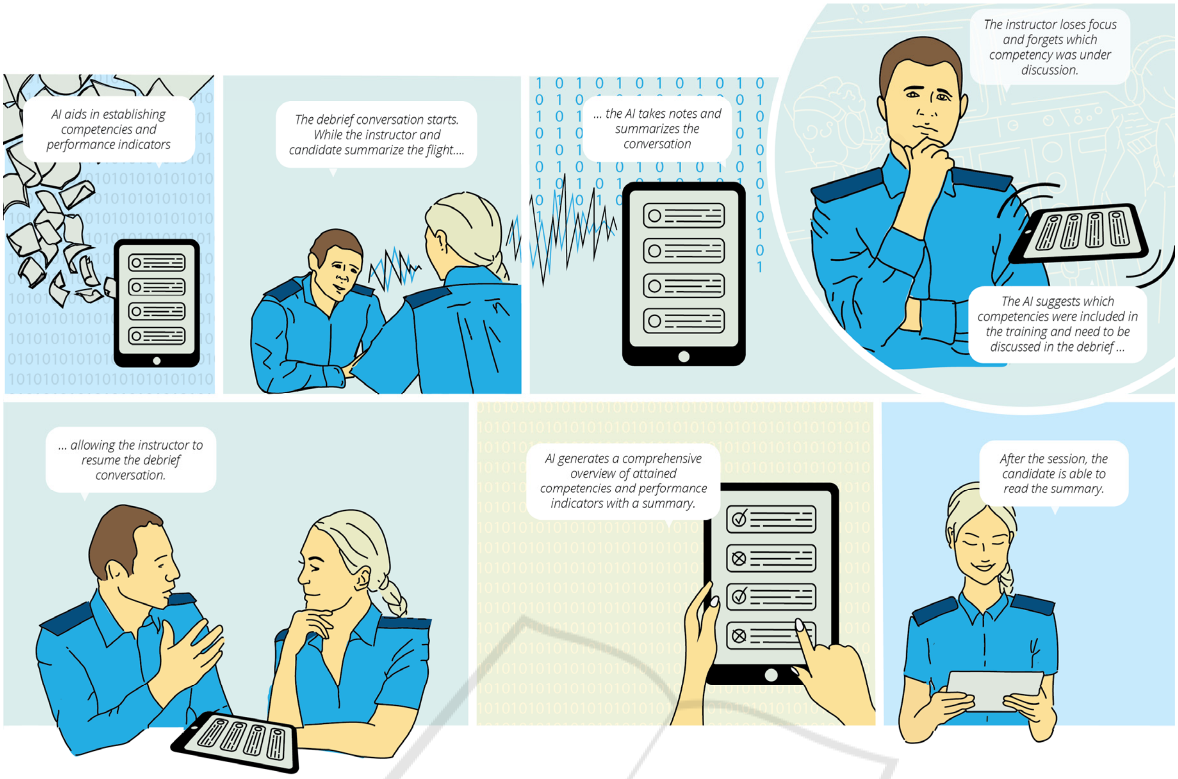

Figure 1: Envisioned future use of AI-assisted debriefing (AAD), enabling automated reporting, instructor support and assessment of pilot competencies during flight debriefing. In this paper we examine the technical feasibility of this scenario.

common sense reasoning and language understanding, making them well-suited for language analysis. Moreover, speech recognition systems, such as OpenAI's Whisper (Radford, et al., 2022), also referred to as *speech-to-text*, have achieved human-level performance in automatic transcription of spoken language, opening up potential applications in aviation training.

This paper seeks to explore the use of speech recognition and large language models to support the reporting process of flight debriefings. We investigate the extent to which debriefs can be automatically transcribed and subsequently analysed by LLM models to distil summaries and extract pertinent information, such as competency assessments and performance indicators. Through this examination, we aim to determine the viability of AI as a tool for supporting the documentation and assessment of flight debriefings.

2 BACKGROUND

2.1 Aviation Debriefing

Over the past decade, the aviation industry has recognized the need for a strategic overhaul of recurrent and type rating training to enhance commercial aviation safety. This shift has led to a gradual adoption of Evidence-Based Training (EBT), which focuses on developing and accessing pilots' competencies, including both technical and non-technical skills, through a framework of behavioural competency descriptions and performance indicators (PIs) (SkyBrary, 2023). The International Civil Aviation Organization (ICAO) supports EBT, emphasizing core competencies such as procedure application, communication, and leadership.

Debriefing sessions, a staple in military and civil aviation training, play a crucial role in this competency-based approach. These sessions, which can occur immediately after a flight or be scheduled later, cover flight performance, decision-making processes, leadership, teamwork, and regulatory requirements. They aim not just to highlight successes but also to identify areas for improvement, fostering an environment of constructive feedback. Effective debriefing involves active self-learning, a clear developmental intent, reflection on specific events, and input from multiple information sources (Tannenbaum & Cerasoli, 2013).

Reflection is a critical component of debriefing, with models like Mavin's reflective debriefing model

(Mavin T. J., 2016) guiding pilots to self-assess their performance. The European Union Aviation Safety Agency (EASA) and other experts provide guidelines for facilitating debriefing sessions, emphasizing the importance of crew participation, avoiding instructor-centred sessions, and ensuring all critical topics are covered (EASA, 2023) (McDonnell, Jobe, & Dismukes, 1997).

2.2 Developments in AI

Automatic distillation of summaries and recognition of competencies and associated performance indicators from debrief recordings is a challenging task; however, recent advances in AI technology might be profitably combined to tackle this problem.

Large Language Models (LLMs). LLMs have made significant contributions to natural language processing (NLP) in recent years (Floridi, 2020). LLMs are trained on large corpora of text data, allowing them to generate human-like text, answer questions, and complete other language-related tasks with high accuracy (Kasneci, 2023). Recent developments include ChatGPT, an LLM trained on a web-scale dataset, which has demonstrated state-of-the-art performance on a wide range of natural-language tasks, including summarization, question answering, essay writing, and computer programming (Team, 2020). By leveraging additional fine-tuning on human feedback, LLMs can learn to follow human instructions, making them promising tools for problems that require language analysis and generation (e.g. summarization).

Speech-to-text. With recent advancements in NLP and machine learning, the field of speech processing has witnessed significant progress, which has resulted in greatly enhanced accuracy and efficiency of speech recognition systems. Automated transcription can streamline the process of documenting and analysing instructor-trainee communication, which is crucial for training and safety reviews. Whisper (Radford et al. 2022), developed by OpenAI, represents a leap forward in speech recognition technology. This cutting-edge model is proficient in deciphering various accents, dialects, and coping with ambient noise and variation in recording devices. Furthermore, Whisper's robust multilingual capabilities (Radford et al. 2022) make it an ideal candidate for the global aviation industry, where pilots often communicate in a technical language mixed with English terms. This system can be used to transcribe the spoken debriefing discussion into written text. This ensures an accurate and detailed textual record of the conversation. The efficiency and accuracy of these tools allow instructors to focus on the discussion without the distraction of manual note-taking, or it can supplement and complete the possibly terse notes taken by an instructor. Transcripts created by speech-to-text tools provide accessible and shareable records, enabling pilots to revisit the feedback at their own pace and reinforcing the learning process.

2.3 Text-to-Text with LLMs

LLMs are trained to predict which word is likely to follow after a given sequence of preceding words in a text, known as its *context*; this predictive ability can then be harnessed to generate text by repeatedly sampling the most likely following word, one at a time. This task is known as *autoregressive language modelling,* or *completion*.

To make an LLM perform a task such as summarization, a technique known as *prompting* is used. Here, a user provides a textual prompt – a written instruction or question – to guide the model's generation process. A user might, for example, prompt the model with `"Summary of take-off procedures:"`; enticing the model to complete the prompt with a summary of take-off procedures.

Simple prompting can sometimes lead to counter intuitive results. For example, the most likely completion to a question could be just another question (completing a list of questions). Therefore, models are commonly finetuned for *instruction following*. By this process the model is tuned to always complete a question with a relevant and helpful answer. With an instruction following model the above prompt for example can be replaced with the instruction `"Give me a summary of the take-off procedures"`.

Despite the power of instruction following, it is important to realize that LLMs are fundamentally word-by-word completers of text. Moreover, the time spent 'thinking' about individual words is the same, so it cannot spend a lot of time on difficult words.

To use LLMs effectively, one has to be aware of a few important pitfalls, specific techniques, and selection criteria (Deng, 2022).

Hallucination. When the context expects an answer, fact or definite connection, a large language model generally prefers to generate something which looks correct over breaking off and admitting that it does not know. This is natural behaviour from the point of view of text prediction while it is not how humans behave. It is therefore important to always check the answers to LLMs and not ask it suggestive questions.

Chain of Thought. It is much better to ask a large language model LLM to first give a step-by-step reasoning and then a definitive answer. This allows the model to first synthesize useful information from the context which it can then use to answer the question. Giving a definitive answer first would force it to commit to a potentially wrong answer which it then tries to rationalize. After all there are not a lot of text documents where the author second guesses themselves.

Synthesising over Analysing. When asking the system analysing questions, especially leading ones, it is prone to hallucinating connections where there are none. Synthesizing tasks, which are more open ended in nature, are much more stable.

Language Proficiency. A crucial component of the proposed tool is its ability to comprehend Dutch dialogue as spoken in the aviation domain. Thus it was important to select a language model that can accurately interpret our domain language and instances of code switching, where English terminology is used seamlessly within otherwise fully-Dutch phrases.

Context Window Size. Typical flight debriefs involve discussions between one or more trainees and an instructor lasting upwards of 20 minutes; to ease summarization, it is best if the totality of the conversation can fit within the model's context window. The context window represents the maximum number of tokens (i.e. word fragments) the LLM is capable of processing. An inadequate context window may result in a loss of crucial information for summarization and limit the tool's ability to provide a comprehensive and coherent summary of the conversation as a whole.

3 METHOD

An AI-Assisted Debrief (AAD) tool has been developed, employing an speech-to-text system and LLM to summarize flight debriefs into concise textual summaries and identify the presence of pilot competencies and PIs, as used in EASA's EBT. A high-level diagram of this system is illustrated in figure 2. First, an audio recording of the debrief is transcribed by a speech-to-text system, resulting in a plain-text transcript of the debrief conversation. As segments of speech from the debrief can originate from either the instructor or trainee, we employ a speaker identification, or diarization, algorithm to identify the source of each utterance. We then prefix each utterance in the transcript with a speaker marker, such as "Instructor:" or "Trainee:", enabling the LLM to consider the speaker in its subsequent processing. Then the- resulting speaker-annotated transcripts are summarized by an LLM to obtain a succinct summary of the debrief. Through careful prompting of the LLM, the system can identify main points of feedback from the instructor and list key take-aways from the debrief.

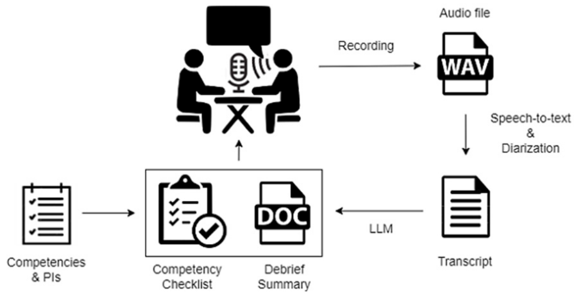

Figure 2: High level diagram of software architecture.

Moreover, as trainee's competencies are identifiable by a set of measurable PIs as used in EASA's EBT, we additionally prompt the LLM to assess the presence of a list of predefined competencies by relating their associated performance indicators to the debrief transcript.

In order to support the flight instructor in an effective manner, it is desirable to obtain concise summaries of the debrief (e.g. in the form of 5-10 bullet points) which encapsulate the primary points of feedback and main take-aways from the debrief. In light of the criteria of Section 3, two multilingual language models were examined in our experiments:

- *Llama-2 70B* (Touvron, 2023), a state-of-the-art LLM developed by Meta AI
- *Yi-34*, developed by 01-AI.

In light of privacy and security concerns, our experiments were limited to open-source language models only, hosted on local machines.

Preliminary assessment of Llama-2 and Yi-34B showed these models to be proficient in understanding Dutch texts and respond well to instructions. The Yi-34 model boasts a large context window upwards of 200k words, allowing the model to summarise and analyse a debrief transcript in one sweep, eliminating the need to analyse a transcript in sections. Moreover, their extensive training on a diverse range of text domains and languages, including English and Dutch, make Yi-34 and Llama-2 well-suited for processing aviation-related terminology, allowing them to understand English jargon while capturing contextual nuances specific to flight debriefings in Dutch.

In this study, we opted to have AAD generate outputs in English, while processing the Dutch transcripts. This decision was based on preliminary tests that demonstrated improved grammaticality and factuality with English. The difference in performance between languages is a well-known phenomenon in the field which occurs due to factors such as the availability of training data.

```
### Input:
[TRANSCRIPT]

### Instruction:
Make a summary of the flight de-
briefing below. Focus on learning
points for the candidate.

### Output:
The candidate had a challenging
flight debriefing, where they
identified several areas for im-
provement in their flying and
procedures execution. The in-
structor provided constructive
feedback and suggested an addi-
tional session to help the can-
didate improve their skills. Here
are some key learning points from
the debriefing: […]
```

```
### Input:
[TRANSCRIPT]

### Instruction:
Is anything said in the flight debrief dialogue above related to
the performance indicator (PI) "[PERFROMANCE INDICATOR]".

Start your response with "Good." if the PI is mentioned and the
pilot did well;

Start with "Bad." if it was mentioned but the pilot did not do
well;

Start with "Unknown" if nothing is stated related to the PI at
all. Always explain your reasoning

### Output:
Good. The performance indicator (PI) "Demonstrates practical and
applicable knowledge of limitations and systems and their inter-
action" is mentioned in the flight debrief dialogue, and the
candidate demonstrated a good understanding of it. […]
```

Figure 3: Left the prompt used for summarization, and right the prompt used for performance indicator extraction, with the LLM completion in bold.

3.1 Evaluation Dataset

To develop and assess our system, a dataset of audio recordings was created under the supervision of two experienced instructor pilots with an average 30 years of commercial flight experience and 12 years of flight instruction experience. The pilots were tasked with re-enacting several representative scenarios of flight debriefings in which one acted as the instructor and the other as a pilot-in-training, alternating their roles between sessions. In total, five debriefings with varying scenarios were created.

Audio was recorded in a closed room using a Zoom H2n recorder – a stereo audio recorder. The instructor and trainee were positioned along the left-to-right stereo axis of the recorder, respectively, to enable identification of the speaker.

To obtain ground-truth transcriptions for evaluation, the audio recordings were first transcribed using Whisper and manually corrected. For speaker identification, we employed a simple stereo heuristic that identified speakers based on the dominant audio channel. The resulting transcripts were then lightly post-processed, assigning each utterance the corresponding speaker role, *'Instructor:'* or *'Candidate:'*, and merging adjacent sentences by a single speaker into longer paragraphs.

The LLM was subsequently used to do various tasks such as summarization and competency identification using prompts like in Figure 3. To make the LLM perform these tasks we used zero-shot prompting (Radford, et al., 2019) as described in section 2.3, where the model is directly asked a question about the transcript.

For recognition of individual performance indicators, the LLM was instructed to look at the performance indicator and rate the candidate on it. This prompt (see Figure 3) was executed for each of the 8-10 performance indicators associated to the 9 competencies as defined by the *Evidence Based Training Pilot Competencies* competency framework (EASA, 2023).

We will quantitively evaluate the individual steps of our approach on this dataset. For transcription we will determine the word error rate (WER) and the diarization error rate (DER) which is simply the rate of incorrectly transcribed words or misattributed speakers. The large language model outputs will be evaluated on the corrected transcripts so that it's performance can be judged in isolation. The LLM outputs will be judged qualitatively by expert evaluation.

3.2 Interactive Debrief Tool

To streamline the instructor's interaction with AAD, the speech-to-text system and language models were integrated into a single debrief application. Given an audio file of the debrief dialogue, the tool first generates the transcriptions after which a summary of the debrief is generated using a language model of choice

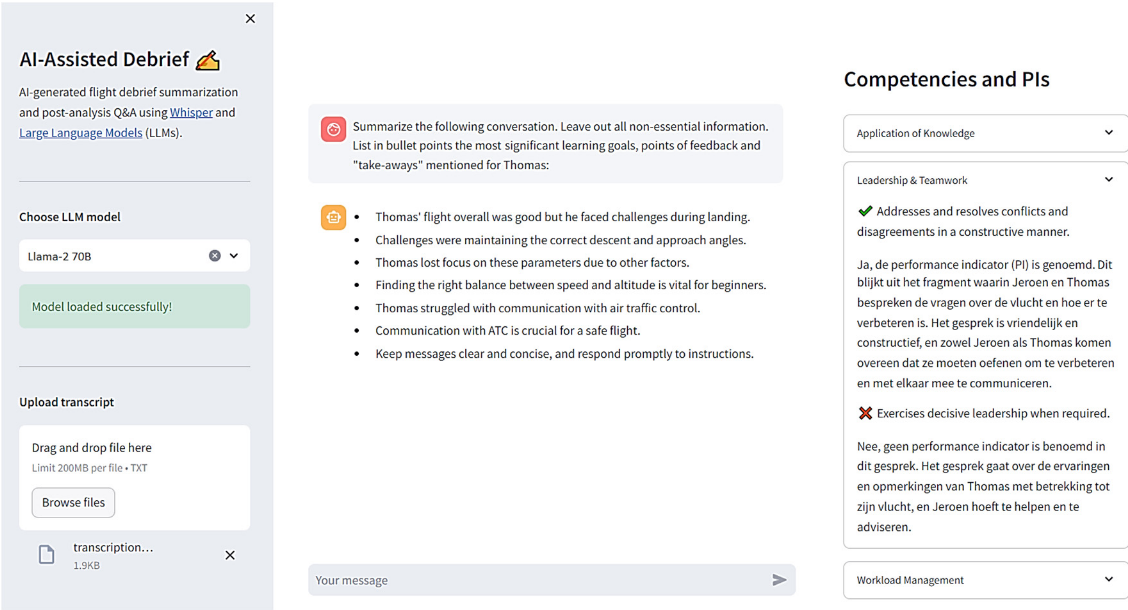

Figure 4: The debrief tool showing a summary of a post-flight debrief in an interactive chat panel (centre column) and a list of competencies with associated performance indicators (PIs) as identified by the Llama-70B LLM (right).

(centre). The resulting summary is then visualised inside an interactive chat panel, allowing the instructor/trainee to further inquire information about the debriefing if desired.

Moreover, to identify the competencies expected of the trainee, the interface enumerates the competencies listed in the *Evidence Based Training Pilot Competencies* framework along with their associated PIs (right); it then verifies, using the LLM, whether the behaviour of the candidate exhibited signs of the desired competencies by evaluating each PI belonging to a competency of interest against the debrief dialogue, as shown in Figure 4.

4 RESULTS

Our results are summarized in Table 1. Displayed is various information about the recorded scenarios with various quantitative and qualitative evaluations of the applied techniques.

Transcriptions by Whisper yielded an average Word Error Rate (WER) of around 2-3%. This is better than the reported WER on Dutch by OpenAI, which is 5% for Whisper v3 and 8% for Whisper v2. As such, quantitative error rates indicate accurate and reliable dialogue transcription; nonetheless, errors were observed which would have likely been detrimental to downstream performance, while others do not seem to modify the semantics of the text. Our relatively simple speaker identification technique, based on the dominant stereo channel, performed reasonably with error rate (DER) of around 5%. This is likely unacceptable for the downstream LLM tasks as important semantic information is lost.

The LLM was evaluated on the corrected transcripts where it performed best on the summarization task. The system can almost always pick out the general main points in the ground truth, only sometimes getting details of individual points wrong. For the competency task the model was able to identify the major competency categories but was prone to making up details or mixing up student and instructor. Finally, for the instructor evaluation, the model seemed unable to criticise the instructor in the last scenario when they failed to show a lack of attention to the workload of the trainee.

5 DISCUSSION

5.1 Evaluation

Overall, Whisper was found robust in transcribing domain-specific jargon and code-mixed phrases involving Dutch with English words, although sometimes this bilingualism also caused it to miss the mark. For example, Whisper displayed a preference in transcribing ‘deicing’ to Dutch as ‘deijsen’ and similarly ‘flightpath’ to ‘flypad’ which are hybrid Dutch-English composite words with similar semantics, see figure 5.

I: Maar ben je tevreden met hoe het **gegaan** ~~gaas~~ was?

C: Nou, **mwah** ~~moi~~. Nee, voor mijn gevoel had het wel wat strakker gekund.

I: En waar dan? Denk je dat **je steken** ~~het teken~~ hebt laten vallen?

C: Ja, dat vind ik een beetje lastig. Het is meer het **overall** ~~overal~~ gevoel.

I: Ja, oké. Misschien even concreet dan. Toen jullie bij de baan stonden.

C: Na het **de-icen** ~~deijsen~~ bedoel je?

I: Ja. Dus na het **de-icen** ~~deijsen~~ zijn jullie naar de baan gereden.

C: Ja.

I: Wat hebben jullie toen allemaal gedaan? Vanaf het **de-icen** ~~deijsen~~ naar de baan

Figure 5: Debrief fragment between Instructor and Candidate, as transcribed by Whisper large-v2 with manual corrections, in bold are the corrections to the strikethrough errors.

In this study, we investigated the application of current-generation open large language models for summarization and performance indicator detection tasks. Our findings indicate that while not perfect these open models are already quite good. It is not at the level to be trusted blindly, but its output can serve very well as a first draft to be checked and supplemented by the end user.

Automatic recognition of competencies through their Performance Indicators (PIs) can be achieved using LLMs, but we observed several challenges in accurately identifying competencies from debrief statements. These challenges include:

Overgeneralization: Models may determine that the trainee meets or fails to meet competency requirements based on vague evidence. For example, a debrief mentioning difficulties during a task might be interpreted as a failure to verify task completion to the expected outcome, even if the evidence is not direct.

Context Misinterpretation: Despite instructions to evaluate performance based on specific criteria (e.g., during the flight), models might consider competencies in the broader context of the debrief conversation. For instance, active listening and understanding demonstrated in a post-flight debrief might be incorrectly attributed to in-flight performance.

Hallucination: Models may generate false positives by attributing competencies that were not demonstrated. For example, claiming resilience in handling unexpected events during a landing when no such events were mentioned.

Lack of Explanation: Models might not adhere to instructions to provide reasoning for their assessment of a competency's presence or absence. This results in evaluations that lack justification, making it difficult to understand the model's decision-making process.

5.2 Future Work

Recent work has shown that fine-tuning Whisper on air traffic control data can improve its performance on that domain (van Dorn, 2023). This is an easy way to improve domain specific performance but requires training data, on the order of hours, e.g. van Dorn fin-tuned on 10 hours of ATC data. Relatedly, Whisper supports a textual context which is can also be filled with domain relevant terms so that it is nudged towards correctly transcribing these, e.g. by putting 787 in the context the transcription of 787 becomes more likely than the incorrect 78 / 8. The diarization, or speaker identification, can likely be improved by moving to a speaker timbre fingerprinting model like (Bredin, 2023). This should also be useful in many-user debriefs as are typical in the industry.

Table 1: Information about the recorded scenarios with results by expert evaluators. WER is the word error rate, DER is the diarization (speak identification) error rate. LLM analyses were rated correct (✓ green), correct with incorrect details (– yellow), incorrect (× red).

		Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Short description		Candidate struggles with procedures	Candidate is too hasty	The candidate has difficulty with flying.	Candidate is passive, co-pilot forced to intervene	Candidate not well prepared, knowledge lacking
Length (mm:ss)		13:26	13:12	18:05	12:10	12:10
Transcription (Whisper)	WER	2.3%	2.8%	2.5%	0.55%	3.8%
	DER	-	-	-	5.5%	5.6%
Analysis LLM Yi-34B	Summarization	✓ – – ✓ ✓	✓ ✓ – ✓ ✓ ✓	– ✓ – ✓ ✓	✓ ✓ – ✓ –	✓ – ✓ ✓ ✓
	Competencies	✓ – – – –	✓ ✓ – ✓ × –	✓ × – – –	✓ ✓ ✓ – –	✓ ✓ ✓ × –
	Instructor Eval	✓	–	✓	–	×

Analysis performance of the LLM can be enhanced by employing larger models or integrating more task-specific training data. Closed models, like those developed by OpenAI and Anthropic, expected to be larger and equipped with superior training data, may outperform in these tasks. The domain of large language models, encompassing both open and closed models, is evolving at an unprecedented pace, with significant yearly improvements. Future enhancements to our system could be achieved by adopting newer more advanced models.

Just as with Whisper, an improvement strategy for the LLM involves fine-tuning on domain-specific training data, such as transcribed conversations with high quality summaries or competency assessments. Likely a few hours of high quality data, such as those generated for this paper, would already yield positive results. This approach can further refine the capabilities of an already trained large language model.

Our proposed future iteration of the AI-Assisted Debrief should incorporate user intervention at every stage of the process, enabling correcting of the system's intermediate outputs. For instance, the system could automatically flag potentially misinterpreted words or incorrectly identified speakers, allowing users to manually rectify these errors. Similarly, users should have the ability to adjust summaries and PI identifications as needed.

These corrections made by users would not only improve the immediate output but also contribute valuable data for the fine-tuning of AAD. This creates a dynamic system that progressively improves its performance and accuracy in executing its designated tasks. Through this iterative learning process, AAD would evolve into an increasingly reliable tool.

Given the inherent limitations of current-generation LLMs, particularly their tendency to hallucinate, we posit that the most effective application of these technologies lies in such a human-in-the-loop framework. This approach synergistically combines the unique strengths of both LLMs and human expertise. Human experts possess an unparalleled capacity for critical thinking and the nuanced evaluation of complex scenarios, which LLMs currently cannot match. Conversely, LLMs excel in rapidly processing and analysing vast quantities of data, a task that is time-consuming and labour-intensive for humans.

ACKNOWLEDGEMENTS

We would like to thank Anneke Nabben for pitching this project and facilitating it throughout; Simone Caso for his help researching flight debriefings; Jeroen van Rooij and Thomas Janssen for creating the evaluation dataset; Asa Marjew for the illustrations; Astrid de Blecourt for her leadership; Jelke van der Pal for his helpful guidance and reviews; finally Marjolein Lambregts and Jenny Eaglestone and for their help.

REFERENCES

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *Proc. INTERSPEECH 2023.*

Deng, J. a. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, 81-83.

EASA. (2023, 12 18). *EASA*. Retrieved from EASA EUROPA: https://www.easa.europa.eu/community /topics/post-flight-debrief

Floridi, L. &. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 681-694.

Josh Achiam, e. a. (2023). *GPT-4 Technical Report.* OpenAI.

Kasneci, S. K. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*.

Mavin, T. J. (2016). Models for and practice of continuous professional development for airline pilots: What we can learn from one regional airline. *Supporting learning across working life: Models, processes and practices*, 169-188.

Mavin, T. J., Kikkawa, Y., & Billett, S. (2018). Key contributing factors to learning through debriefings: commercial aviation pilots' perspectives. *International Journal of Training Research*, 122-144.

McDonnell, L. K., Jobe, K. K., & Dismukes, R. K. (1997). Facilitating LOS debriefings: A training manual.

OpenAI. (n.d.). *Whisper Github Repository.* Retrieved from https://github.com/openai/whisper

Radford, A., J.W., K., Xu, T., Brockman, G., C., M., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision.* OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Roth, W. (2015). Cultural Practices and Cognition in Debriefing: The Case of Aviation. *Sage Journals*, 263–278.

SkyBrary. (2023, December 19). *SkyBrary*. Retrieved from https://skybrary.aero/articles/evidence-based-training-ebt

Tannenbaum, S., & Cerasoli, C. (2013). Do Team and Individual Debriefs Enhance Performance? A Meta-Analysis. *Human Factors and Ergonomics Society*.

Touvron, H. e. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* Meta.

van Dorn, J. L. (2023). *Applying Large-Scale Weakly Supervised Automatic Speech Recognition to Air Traffic Control.* Retrieved from http://resolver.tudelft.nl/uuid:8aa780bf-47b6-4f81-b112-29e23bc06a7d