Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 88 (2025) 270–277

Transportation Research Procedia

www.elsevier.com/locate/procedia

European Association for Aviation Psychology Conference EAAP 35

From Turing Test to Chinese Room Argument: How to Apply Artificial Intelligence in Aviation

Declan Saunders[a], Wen-Chin Li[a] and Thomas Wang[b]

[a]*Cranfield University, College Road, Bedfordshire, MK43 0AL, United Kingdom*
[b]*Taiwan Transportation Safety Board, Taiwan*

Abstract

The emergence of artificial intelligence (AI) with advanced large language model (LLM) offers promising approaches for enhancing the capacity of textual analysis. Both the Turing Test and Chinese Room Argument explore AI's understanding of human language, although both methodologies have dissimilar interpretation on AI's 'intelligence'. Current AI systems have demonstrated the capacity for achieving defined test goals for 'intelligence'. The aviation industry is increasingly interested in adopting AI to improve efficiency, safety, and cost efficiency; with the Generative Pre-trained Transformers' (GPT) capability to reduce resource-intensive analytics in accident causation classification. This study investigates the potential and challenges of using AI to analyze human factors involved in aviation accidents based on the Human Factors Analysis and Classification System (HFACS). Six subject-matter experts in aviation human factors and AI domain participated in this research. All participants were familiar with the HFACS framework to analyze aviation accident reports and the output of GPT which were based on the prompt engineering developed by the research team. This research creates a framework to perform its initial generation and training using GE 235 accident investigation report from Taiwan Transportation Safety Board (TTSB). Initial discoveries demonstrated that the AI model could populate the sub-dimensions of Level 1 HFACS framework with moderate accuracy, although there remains a high presence of hallucinations in generated outputs, with a lack of reproducibility in consecutive outputs with consistent input data. There are still different opinions on AI applications in real-world operations with ethics and safety concerns. While there is clear potential for GPT models to supplement accident analysis within the HFACS framework, there is still more work to integrate HFACS framework into GPT modelling for effective generation to accident data.

© 2024 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0)
Peer-review under responsibility of the scientific committee of the European Association for Aviation Psychology Conference EAAP 35

Keywords: Accident Analysis; Artifical Intelligence; Chinese Room Argument; Human Factors Analysis and Classification System; Large Language Model

2352-1465 © 2024 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0)
Peer-review under responsibility of the scientific committee of the European Association for Aviation Psychology Conference EAAP 35
10.1016/j.trpro.2025.05.033

1. Introduction

Alan Turning, mathematician and logician, proposed the idea of an artificial brain; one capable of processing large amounts of data simultaneously, and learn in real-time, to solve the most complex problems beyond human capabilities (Turing., 1936). As part of the cryptologists department in Bletchley Park, Turing and his colleagues designed and built the first Turing Machine, one capable of deciphering the infamous Enigma Code; consisting of 150 quintillion possibilities to be tested within an 18-hour window. However, Turing envisioned the idea of his Turing machine capable of mimicking human-like intelligence and one day becoming omniscient. In 1950, Turing provided a methodology to determine if a machine could exhibit human-like intelligence in its thought process during conversation, known as the Turing Test (Duke et al., 2007) (Mei et al., 2023). The Turing Test required an interrogator to ask questions of a varied nature, and assess the response on the commonality between a human response, to determine if their counterpart was a machine or human. While addressing the possibility of machine 'intelligence', Turing was quoted in the following:

"A computer would deserve to be called intelligent if it could deceive a human into believing that it was human"

However, John Searle proposed the counter-view of machine intelligence with a simple, yet still engaging, thought experiment in 1980, known as the Chinese Room Argument. Searle believed the Turing Test had fallibility in the machines 'understanding', stating that consciousness on semantics is associated with true intelligence, where AI is primarily focused on syntax. The argument was such: There are two rooms, one contains a fluent Chinese speaking individual and the second contains a non-Chinese speaking individual. The non-Chinese speaking individual is provided an English translation guide on Chinese symbology. Both rooms receive the same Chinese text as input and both produce the same output text in Chinese, with the human evaluator unable to identify the non-Chinese individual. The system has successfully passed the Turing Test, however there is no semantic awareness of the text produced in the second room, rather the syntax of word association to symbology (Sepetyi., 2017). This is the consideration of modern AI, where tokens of words are given weightings, and logarithmic probability determines the next syntax in word sequencing, although no true semantic has been identified and processed by the AI. This argument proposes that a machine/system capable of passing the Turning Test is not necessarily an indication of consciousness or intelligence, but rather a complex interaction between a human and machine programmed with specialized instructions/prompts (Hamid., 2023).

This counter-view on machine intelligence and machine learning was widely controversial within the industry of computer science and AI science, as it failed to foresee the ability of non-native language speakers to learn and grow accustomed to foreign language, which would appear originally as 'meaningless sounds'. If considering Searle's argument as true, the idea of learning would be impossible, as conscious awareness of syntax and semantics is unknown to the individual (system) initially. Despite this, the Searle argument has fueled debate on how we define machine intelligence and how we demonstrate this capacity, as AI becomes more capable of replicating human interaction; holding a steady conversation and providing suggestions to questions asked (Gozzano., 1995).

In the modern day, an AI chatbot named Eugene Goostman was successfully able to pass the Turing Test in 2014, convincing the judges that it was a 13-year-old boy; demonstrating AI's potential for achieving human-like conversations (Mei et al., 2023). This has fueled interest within the aviation domain to its future applications, as AI has the capacity of objective trend-analysis in large-data streams in real-time, opening the opportunity for integration into resource-intensive analysis applications within the aerospace field (Hejji et al., 2021). Despite this, AI integration concepts into aviation practises is limited to the security risks of malicious intentions by external operators, as well as the limitation of 'creative thinking' required to adapt to real-world threats, to the same standards as human operators (Gao et al., 2024).

Considering the potential limitations and risks associated to high level integration of AI within the aviation industry (flight deck), this report considers the use of AI within the field of causal accident investigation and data analysis. This is attributed to the properties of AI functions like Generative Pre-Trained Transformers (GPT) and Large Language Model (LLM) to assimilate and disseminate between human language and computing language (Chen et al., 2023). The goal of this integration is to reduce the human resources required to conduct complex

analysis into human factor causal events in accident investigation, summarized in Human Factors Analysis and Classification System (HFACS), and to promote the unison of AI into aviation safety management for a proactive approach to trend analysis, commonly considered a limitation to the human capacity in data analytics (Ziakkas & Pechlivanis., 2023).

This study prioritized the use of human evaluation methods to train a GPT-4 AI model for the explicit purpose of HFACS analysis, using TransAsia Airways Flight 235 (GE 235) as a case study report. This accident report was selected based on its human factor contributions towards the failure model for the accident sequence, as described by Endsley (1995) in the Endsley Model of Situational Awareness, and supported by the accident investigators involved. Training was centred around fine-tuning practices, over iterative conversations between GPT and the authors, and tailored Prompt Engineering practises that focused the chain-of-thought on the process of HFACS within the scope of the multi-page reports. These would be reviewed by the investigation teams associated to the accident report to summarise the characteristics of each output into four distinct reflection criteria: Accuracy, Comprehensiveness, Satisfaction and Hallucinations, weighted between 0-100%.

2. Literature Review

2.1 Large Language Model

Large Language Models (LLM's) are generative mathematical models of the statistical distribution of tokens in the corpus of human language, incorporating: words, parts of words or individual characters and punctuation. The term 'generative' is applied to LLM's as we are able to ask questions and sample from it, although a specific format is required to sample from this dataset. Additionally, this generation is not based on any pre-existing knowledge or awareness of factual information, rather is the statistical significance of certain words or phrases, from the vast public corpus of English language, to the sequence of the questions asked. For example, when we ask an LLM "*Who was the first person to walk on the Moon?",* we know the answer is Neil Armstrong. However, the LLM cannot disseminate based on this information, instead follows a chain-of-though more inclined to "*Given the statistical distribution of words in English language, what words are most likely to follow the sequence "Who was the first person to walk on the Moon?"* (Shanahan., 2022).

It is commonplace to anthropomorphise LLM's, given the resemblance of human characteristics when interacting with them. However, human beings have evolved to co-exist over millions of years, with culture evolving to facilitate this co-existence over thousands of years, which allows for a degree of mutual understanding between humans. To apply the same reflection onto the interactions between humans and AI systems in a serious mistake that can inherently lead to the spreading of non-factual information (Shanahan, 2022) (Jayasurya et al., 2023).

2.2 Generative Pre-trained Transformers

Generative Pre-trained Transformer (GPT) are a framework of neural networks that utilize Natural Language Processing (NLP) to perform tasks within a dataset, primarily those associated with natural language task, such as conversation generation and data analysis (Zhou & Jiang., 2024). The configuration of GPT models decouples the need for LLM's to be associated with them, albeit the majority of GPT models consist of multi-modal LLM to improve their capability for text and image processing. The generation of human-like responses are based on the logarithmic probability of the next sequence of tokens to follow the input dataset, re-configurable based on user-defined selection of hyperparameters. Dependent on the pre-training of the model prior to output generation, the model is capable of complex analysis and scrutiny of a user-defined dataset, although the consecutive generation in output can vary significantly (Aithal & Aithal., 2023). In recent years, GPT-3.5 and subsequent models has set new standards for the linguistic capabilities of AI systems, pushing the boundaries of what is possible with deep learning (Bender et al., 2021). This gives organizations and industries the ability to integrate GPT models into human-centered tasks, such as Q&A chatbots, text summarization and content generation (Kumar et al., 2023a).

2.3 Human Factors Analysis and Classification System

The Human Factors Analysis and Classification System (HFACS) is the most widely utilised human factors-coding framework used for accident/incident analysis. HFACS was developed from James Reason's '*Swiss Cheese Model'* in 1997, which superseded the '*Generic Error Modelling System'* (GEMS) in 1990 (Duggan & Harris., 2001). Developed as an expansion to the systematic analysis of causal chain of events within accident sequences, HFACS gives scope of the organisation level influences to lower-level operators that impact the decisions and actions in an unpredictable and uncharacteristic way to the individual (Shappell & Wiegmann., 2000). The multi-layered structure of HFACS can introduce complexities that demand additional time for accident/incident investigation, as this is dependent on the relationships between two or more level's. Attributing these relationships and organising the influences effecting Aeronautical Decision-Making (ADM) inherently is time-consuming and requires large corpus of data availability (Hsieh et al., 2018).

2.4 Prompt Engineering techniques

GPT models require specific and informative instructions to perform the desired analysis requirements of the end user. The method of these instructions is heuristic of human reading strategies; how we derive information from a dataset (Chen et al., 2023). When given basic instruction to user inputs, the model faces a plethora of options and makes the results broad. When developing these instructions to GPT, the most effective approach is converting these expectations into prompts; goal-orientated instructions to elicit specific outputs from the LLM (Rodriguez et al., 2023). Prompt Engineering involves the training of a GPT model to use existing data to enhance its categorization of text-based information to suit the requirements of the end user (Rodriguez et al., 2023). In the scope of this study, prompt engineering techniques focused on improving the relationship of GPT to human behaviour and actions, derived from the HFACS framework and the definition of 'human error' from the U.S Department of Defence (DOD), 2022. Chain-of-thought (CoT) prompting is relatively new to the field of AI development for LLM's, improving the accuracy for logical reasoning tasks. Analysis is conducted on intermediate reasoning steps which guide the models response, facilitated by simple prompts such as: 'Lets think step by step', or through manual demonstrations on the reasoning. This method benefits from a clear structured approach for the model to conduct reasoning tasks, promoting less ambiguity behind the models conclusions to the end user (Chen et al., 2023).

3. Methodology

3.1 Research Framework

This study explored the capacity of GPT-4 to conduct a comprehensive and accurate HFACS analysis, through prompt engineering practices and human-evaluation. Based on domain-expert feedback, the prompt would be revised, and new iterations would be generated with careful manipulation of GPT hyperparameters. Temperature was manipulated throughout the trials; user-defined fine tuning to balance generation of deterministic and creative outputs, ranging from 0 – 1 respectively. Top_P refers to nucleus sampling on possibilities of words to be considered in output sequencing. These trials used a low Top_P to restrict word sequencing to the highest probability of words, to restrict hallucinations/non-factual statements. Presence penalty encourages the model to embrace diversity in word sequencing, by reducing the logits score of words based on their presence in the output. Its presence in the text tells the model to look for other words, instead of penalizing based on its frequency in the

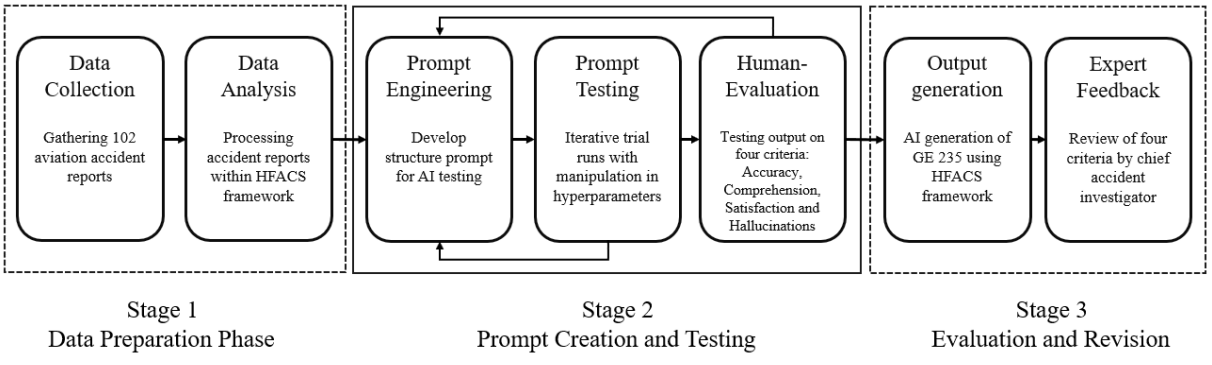


Fig 1: Research design for preparation and review of GPT conversation output

text. Frequency penalty aims to reduce repetition by reducing the logits score for the frequency of the word in generated text. A higher penalty score will reduce the logit score more significantly and decrease the likelihood of repetition in the following word sequencing (Sayak., 2018).

The scope of the research was to investigate the interaction, modelling, conversation outputs and future training potential for a GPT-4 AI-HFACS model (Chen et al., 2023). The aim was to reduce the generation of 'hallucinations' in the output while improving the overall accuracy, by constraining the analysis on the content provided by the operator. The use of HFACS framework dictates a strict systematic approach to achieve categorization and classification on the causal chain of events, which influenced the selection of hyperparameters for these trials. The process involved can be seen below in Fig 1, separated into three separate but interconnected stages.

Currently, the pre-trained process remains largely supervised, meaning the human operator must continuously update the dataset and indicate the area of interest for the model to generate an HFACS framework output (Ray., 2023) (Zhou & Jiang., 2024). However, for GPT to be considered as an effective analysis tool for investigation purposes, the pre-training process would have to be modelled as unsupervised, allowing GPT to 'learn' from previous training and coagulate information from historical data to aid in prediction of future or real-time events, based on trend analysis and probability of occurrence (Kumar et al., 2023b).

This criterion plays into the strengths of GPT and LLM's when utilising the weighted probability of words to generate textual outputs. Developing a GPT model capable of unsupervised learning can be deployed into a safety management role and can provide recommendations or warnings based on historical data and the likelihood of occurrence (Ziakkas & Pechlivanis., 2023).

3.2 Designing the Prompt

The prompt was developed using the Chain-of-Thought (CoT) methodology, a highly structured framework for logical reasoning tasks. Consisting of a step-by-step framework, CoT prompting gives the model a 'thought' process to achieve a targeted goal, typical achieved through logical/successive steps pre-defined in the prompt, of increasing levels of complexity. In this study, we prompted the GPT into definitions of sub-dimension terminology within HFACS, before prompting the model to categorize and classify these meanings within the framework. This allowed for multiple actions to be completed within a single conversation, reducing the human interaction involved to conduct the analysis. Validation would still rely on the domain-experts evaluation of output conversations. CoT prompting was selected to constrain the degrees of freedom for the model, primarily its creativity in conversation generation (Chen et al., 2023).

The HFACS model would be applied using this prompting methodology to GE 235 case study, after reviewing over 100 accident cases as seen in Fig 1. Selection of this accident was due to its contributions to all level classifications in the HFACS framework. Using texts from Wiegmann and Shappell, 2012, a HFACS framework for Level 1 classification was developed, to correctly identify and provoke the GPT model in search criteria for more accurate classification reporting.

The rationale of CoT prompting; while restricting creativity and freedom, allows the model to coagulate information not explicitly in the prompt, providing greater application across multiple accident reports in future trials. This methodology also incorporated the use of information generation, giving the model the ability to formulate its 'thoughts' before providing answers (Chen et al., 2023).

4. Results and Discussion

Prior to testing, the prompt was configured with the specific hyperparameters as recommended by previous studies, for complex and factual analysis on a dataset (Hejji et al., 2021) (Aithal & Aithal., 2023) (Gupta et al., 2024). The hyperparameters would restrict the creativity and information sourcing of the model, without over-constraining the model. This would cause the model to revert to the next suitable hyperparameter setting where it could generate an output. The initial hyperparameters were selected as follows: (1) Temperature 0.01, (2) Top_P 0.2, (3) Presence Penalty 0.8, (4) Frequency Penalty 0.2, (5) Max_Token 4096, and (6) N as 1 (Agarwal et al., 2024).

The HFACS framework was applied to the GE 235 accident report, focused on the Level 1 classifications, using the configurations mentioned earlier. Temperature was manipulated within three dimensions, reflecting more conformity to the prompt instructions (0.01), a more neutral approach (0.05) or a more creative approach (0.1).

Trial one (Temperature 0.01):

- **Decision Error**: Accuracy 10%, Comprehensiveness 10%, Satisfcation 10%, and Hallucinations 30%.
- **Skill-Based Error**: Accuracy 70%, Comprehensiveness 60%, Satisfaction 60%, and Hallucinations 20%.
- **Perceptual Error**: Accuracy 25%, Comprehensiveness 30%, Satisfcation 20%, Hallucinations 30%.
- **Violation**: Accuracy 30%, Comprehensiveness 50%, Satisfcation 30%, and Hallucinations 30%.

Trial two (Temperature 0.05):

- **Decision Error**: Accuracy 20%, Comprehensiveness 20%, Satisfcation 20%, and Hallucinations 30%.
- **Skill-Based Error**: Accuracy 60%, Comprehensiveness 60%, Satisfaction 60%, and Hallucinations 30%.
- **Perceptual Error**: Accuracy 30%, Comprehensiveness 30%, Satisfcation 20%, Hallucinations 30%.
- **Violation**: Accuracy 30%, Comprehensiveness 50%, Satisfcation 30%, and Hallucinations 30%.

Trial three (Temperature 0.1):

- **Decision Error**: Accuracy 10%, Comprehensiveness 20%, Satisfcation 10%, and Hallucinations 40%.
- **Skill-Based Error**: Accuracy 30%, Comprehensiveness 50%, Satisfaction 30%, and Hallucinations 50%.
- **Perceptual Error**: Accuracy 10%, Comprehensiveness 30%, Satisfcation 10%, Hallucinations 50%.
- **Violation**: Accuracy 30%, Comprehensiveness 50%, Satisfcation 30%, and Hallucinations 40%.

Reflecting on the weightings provided during expert feedback, there is potential application of an AI-HFACS GPT model for low-level accident investigation. The weighting provide support to the future development of AI-HFACS modelling aimed at high-level integration to the investigation process. This would involve the coagulation of safety data, historical accident data, and well refined prompting to enable less restrictive information generation from the GPT model, increasing the application of the model to numerous safety and accident investigations.

The model could provide legible content that was moderately associated with the accident report, as supported from comments and feedback by the chief investigator for GE 235 investigation. Despite this, there was a general lack of cohesion to categorization of findings and the Level 1 sub-dimensions. This was made in general reflections during expert feedback session, where the classification of decision error and skill-based error had some overlap. Much of the analysis done by GPT remains largely on the surface of the investigation process, meaning there is a lack of context or evidence to support the conclusions made. In some instances, GPT would report of findings and evidence that was not considered for the final report, either because there was a lack of solidarity behind the conclusions drawn from them or because there was contradictory opinions on what the evidence was indicating. The tendency for GPT to report such findings perplexed many domain experts while reviewing the output conversations, specifically to AI's 'understanding' and 'thought-process' to reach such conclusions. This influenced the relatively low accuracy weighting for decision error across all three trials, while skill-based error received a more generous weighting. Reflecting on the definition and differences between decision error and skill-based error will prioritise the separation in semantic for the AI. Refining the prompt to reflect these semantics should enable the AI to correctly identify the 'actions' for each instance more accurately, but will require further testing and evaluation to determine its effectiveness at reducing these hallucinations.

Further constraints of the GPT model were related to the referencing accuracy to the accident data, especially when considering some of the output data reported by GPT when using the HFACS tool. Notably, this was related to the exact citation reference to the accident report: page, paragraph, heading, and section numbering. Rationale for this inconsistency could be related to the formatting and style of the report when published, as each accident investigation agency has there own reporting format. This influences a limitation within the GPT model for image analysis, which is not supported in this trial. This provides a potential explanation to the lack of consistency and accuracy in the referencing throughout the report. Moving forward, the model of GPT used for future trials should focus both on the textual data analysis and image analysis, with further trials aimed at identifying the reduction in hallucinations generated in report referencing.

Finally, it was noted that the variation of temperature across the three trials had significant variation to the weightings of sub-dimensions in the Level 1 HFACS framework. The variation of temperature demonstrated that incorporating more creativity into the model would increase the hallucinations generated, while minimal effect would be had on the other three criteria. After reviewing the results, the optimum temperature configuration is seen

to favour the 0.05 selection, being a more neutral approach between conformity and creativity on the initial prompt. This can be explained in part by the nature of the analysis conducted, with strict retention to the initial prompt dependent on the quality, with a more open approach having less restriction to report primarily on the report data provided. This creates ambiguity in the weightings, as the data is not incorrect but also not elicit in the report, leading to lower weightings than originally expected. This effect to the significance of analysis has yet to be classified, but it is assumed that the implementation of fine-tuning will remove these discrepancies and consolidate the output conversation accordingly.

5. Conclusion

In conclusion, this study has demonstrated the capacity and capability of a developed prompting strategy to extract key details from a dataset, to provide a legible, comprehensive, and moderately accurate HFACS assessment to an accident investigation, using the report data. Although, there is evidence of significant development and training required to the GPT model to reduce hallucinations in output conversation and improve the retention of factual data generation in these outputs, based on the reported data. Future research will continue to explore new prompting techniques and conducting iterative trailing to refine the optimum hyperparameter setting as well as improving the quality of information provided in the initial prompt. The goal is to develop a refined and sophisticated GPT model for specifically applying HFACS framework to aviation accident reports, with scope to expand to safety data and safety trend analysis and monitoring in future iterations.

Further scope will see the research conducting fine-tuning techniques to improve the quality of conversation between user and GPT, to improve the weighted ratings and reduce the ‘hallucinations’ reported by domain experts. It is expected that GPT modelling will become an integral part of industry safety management systems, capable of semi-supervised operation and reporting, benefiting proactive approach to safety assessment and mitigation. While this will require extensive research and successive trailing on GPT models to conduct deep learning and improve the volume of neural network connections prior to analysis, it is envisaged that this will revolutionise the world of safety management and mitigation.

6. References

Agarwal, A., Mittal, K., Doyle, A., Sridhar, P., Wan, Z., Doughty, J. A., Savelka, J., & Sakr, M. (2024). Understanding the Role of Temperature in Diverse Question Generation by GPT-4. SIGCSE 2024 - Proceedings of the 55th ACM Technical Symposium on Computer Science Education, 2. https://doi.org/10.1145/3626253.3635608

Aithal, S., & Aithal, P. S. (2023). Effects of AI-Based ChatGPT on Higher Education Libraries. *International Journal of Management, Technology, and Social Sciences*. https://doi.org/10.47992/ijmts.2581.6012.0272

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. Computation and Language, 4. https://doi.org/10.48550/arXiv.2310.14735 Focus to learn more Submission history

Duggan, S., & Harris, D. (2001). Modelling Naturalistic Decision Making Using an Artificial Neural Network: Pilots’ Responses to a Disruptive Passenger Incident. Human Factors and Aerospace Safety, 1(2), 145–165. https://pure.coventry.ac.uk/ws/portalfiles/portal/19054439/HFAS_Vol_1_2_Duggan_and_Harris_145_166_1_.pdf

Duke, E. L., Vanderpool, C. C., & Duke, W. C. (2007). A turing test for UA operations: Defining operational requirements. Collection of Technical Papers - U.S. Air Force T and E Days, 2007(Test and Evaluation. https://doi.org/10.2514/6.2007-1622

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 32–64. https://doi.org/10.1518/001872095779049543

Gao, S., Lu, Z., Luan, H., Yin, M., & Wang, L. (2024). AI Pilot in the Cockpit: An Investigation of Public Acceptance. International Journal of Human-Computer Interaction. https://doi.org/10.1080/10447318.2024.2301856

Gozzano, S. (1995). Consciousness and Understanding in the Chinese Room. *Informatica*, *19*, 231–235. https://www.academia.edu/273466/The_Chinese_Room_Argument_Consciousness_and_Understanding

Gupta, A., Chadha, A., & Tewari, V. (2024). A Natural Language Processing Model on BERT and YAKE Technique for Keyword Extraction on Sustainability Reports. IEEE Access, 12. https://doi.org/10.1109/ACCESS.2024.3352742

Hamid, O. H. (2023). ChatGPT and the Chinese Room Argument: An Eloquent AI Conversationalist Lacking True Understanding and Consciousness. 2023 9th International Conference on Information Technology Trends, ITT 2023. https://doi.org/10.1109/ITT59889.2023.10184233

Hejji, D., Talib, M. A., Nassif, A. B., Nasir, Q., & Bouridane, A. (2021). AI-based Models for Resource Allocation and Resource Demand Forecasting Systems in Aviation: A Survey and Analytical Study. Proceedings of the 2021 IEEE International Conference on Internet of Things and Intelligence Systems, IoTaIS 2021. https://doi.org/10.1109/IoTaIS53735.2021.9628555

Hsieh, M. chih, Wang, E. M. yang, Lee, W. chiang, Li, L. wen, Hsieh, C. yi, Tsai, W., Wang, C. ping, Huang, J. li, & Liu, T. chu. (2018). Application of HFACS, fuzzy TOPSIS, and AHP for identifying important human error factors in emergency departments in Taiwan. International Journal of Industrial Ergonomics, 67. https://doi.org/10.1016/j.ergon.2018.05.004

Jayasurya, P., Ganta, S. R., Damerla, U., Reddy, M. H., & Mohan, G. K. (2023). Tokenization in GPT Models: Overcoming Challenges for Non-English Languages. *International Journal of Advance Research in Science and Engineering*, *12*(4), 24–30. https://www.researchgate.net/publication/369830735_Tokenization_in_GPT_Models_Overcoming_Challenges_for_Non-English_Languages

Kumar, S., Simran, & Singh, N. T. (2023a). Enhancing Contextual Understanding in GPT through Multimodal Pre-training. 2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management, IC-RVITM 2023. https://doi.org/10.1109/IC-RVITM60032.2023.10435352

Kumar, T., Kait, R., Ankita, & Rani, S. (2023b). Possibilities and Pitfalls of Generative Pre-Trained Transformers in Healthcare. Proceedings - 2023 International Conference on Advanced Computing and Communication Technologies, ICACCTech 2023. https://doi.org/10.1109/ICACCTech61146.2023.00016

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2023). A Turing Test: Are Ai Chatbots Behaviorally Similar to Humans? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4637354

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rodriguez, A. D., Dearstyne, K. R., & Cleland-Huang, J. (2023). Prompts Matter: Insights and Strategies for Prompt Engineering in Automated Software Traceability. Proceedings - 31st IEEE International Requirements Engineering Conference Workshops, REW 2023. https://doi.org/10.1109/REW57809.2023.00087

Sayak, P. (2018). *Hyperparameter Optimization in Machine Learning Models*. Datacamp. https://www.datacamp.com/tutorial/parameter-optimization-machine-learning-models

Sepetyi, D. (2017). Can the Chinese Robot Think? . *International Society of Philosophy and Cosmology* , *19*, 143–153. https://www.researchgate.net/publication/327594092_Can_the_Chinese_Robot_Think

Shanahan, M. (2022). Talking About Large Language Models. Computation and Language, 5. https://doi.org/10.48550/arXiv.2212.03551

Shappell, S. A., & Wiegmann, D. A. (2000). The Human Factors Analysis and Classification System-HFACS. https://www.researchgate.net/publication/247897525_The_Human_Factors_Analysis_and_Classification_System-HFACS

Turing, A. (1936). On computable numbers. *Proceedings of the London Mathematical Society*, *42*. https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf

U.S DOD. (2022). Human Factors Analysis and Classification System U.S. Department of Defense Human Factors Analysis and Classification System (DoD HFACS) version 8.0.

Wiegmann, D. A., & Shappell, S. A. (2012). A human error approach to aviation accident analysis: The human factors analysis and classification system. In A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System. https://doi.org/10.4324/9781315263878

Zhou, Y., & Jiang, R. (2024). Advancing Explainable AI Toward Human-Like Intelligence: Forging the Path to Artificial Brain. Artificial Intelligence , 1. https://doi.org/10.48550/arXiv.2402.06673

Ziakkas, D., & Pechlivanis, K. (2023). Artificial intelligence applications in aviation accident classification: A preliminary exploratory study. Decision Analytics Journal, 9, 100358. https://doi.org/10.1016/j.dajour.2023.100358



https://dspace.org/

From Turing Test to Chinese Room Argument: how to apply artificial intelligence in aviation

Saunders, Declan; Li, Wen-Chin; Wang, Thomas

2025
Elsevier
https://doi.org/10.1016/j.trpro.2025.05.033
Saunders D, Li W-C, Wang T. (2025) From Turing Test to Chinese Room Argument: how to apply artificial intelligence in aviation. Transportation Research Procedia, Volume 88, 2025, pp. 270-277

Downloaded from a DSpace repository. https://dspace.org/