# WIP: Impact of generative AI on learning: Foci and parameters for an assessment tool

Masood M Khan
Faculty of Science and Engineering
*Curtin University*
Western Australia
https://orcid.org/0000-0002-2769-2380

Chris Ford
School of Civil and Mechanical Engineering
*Curtin University*
Western Australia

Yu Dong
School of Civil and Mechanical Engineering
*Curtin University*
Western Australia
https://orcid.org/0000-0003-1774-1553

Nasrin Afsarimanesh
School of Civil and Mechanical Engineering
*Curtin University*
Western Australia
nasrin.afsarimanesh@curtin.edu.au

***Abstract*—This work in progress, innovative practice paper proposes a Learning and Teaching Impact Assessment Tool (TALIAT) that would help in assessing the sways of Generative Artificial Intelligence (GAI) on important aspects of Learning and Teaching (L&T). The focus of TALIAT is to identify the most relevant elements of GAI grammar and establish how they would reflect on Learning and Teaching practices. GAI-supported systems are generally perceived beneficial as they ensure resource availability, quick and ongoing feedback, higher level of interactivity, and continued engagement. GAI tools also inculcate the ability to perform independent research through progressive and systematic probing and consequent delivery of information. Nonetheless, issues pertaining to trust, liability, transparency, honesty, fairness in judgement and trustworthy reporting of student success need to be addressed for proper adoption of GAI applications. It is being acknowledged now that various stakeholders like system developers, administrators, policymakers, psychologists, educators, sociologists, and legal experts need to work together and agree upon standards and frameworks for producing professionally, ethically, and legally acceptable GAI systems. For these stakeholders to work together, an easy to follow and adopt assessment tool needs to be designed. This would facilitate establishing methods, and processes for confidently determining impacts of GAI technologies on Learning and Teaching. The current literature has yet to provide such facilitation. Such enablement would: improve teacher's perception of, and users' confidence in, GAI systems as tools like TALIAT would assist in assessing impacts of GAI tools prior to their deployment. The proposed TALIAT could also safeguard quality of information, student competence reporting, academic culture, and social and ethical values.**

**TALIAT projects sixteen elements of grammar on seven important aspects of GAI implications namely, handling biases, prompt tuning and evaluation, output fairness, regulatory compliance, error detection, accountability and explainability. These seven aspects are deemed important for investigating the purpose, impetuses, scope, and L&T impact mitigation strategies of GAI applications. A set of thirty-two questions is included in TALIAT for GAI tools' impact assessment.**

**TALIAT will be useful for the GAI stakeholders to work together as it enables informed decision making about adopting GAI systems and places rail guards for development of trustworthy GAI systems.**

***Keywords—generative artificial intelligence, large language models, impact assessment tool, GAI impact grammar, GAI reflections, Learning and Teaching paradigms***

## I. INTRODUCTION

Methods of objectively evaluating instructional software have come to an age now. Ontological and procedural frameworks of these methods are being extended to assess Generative Artificial Intelligence (GAI) applications and supported software. Objective evaluation of instructional software requires determining if the software objectives are consistent and complementary to the Learning and Teaching (L&T) objectives. During the early days of computer application in L&T, only 5% of the assessed software were found to be exemplary [1]. However, description of the term 'exemplary' wasn't provided in the publication.

Generative Artificial Intelligence (GAI) tools and one of their particular genre, Large Language Models (LLM) such as: BERT [2], ChatGPT [3], Llama [4], Titan [5], LaMDA [6] and Claude [7], are becoming ubiquitous in all spheres of Learning and Teaching (L&T). Their impacts on students, educators, society and the ecology of L&T is still being investigated [8].

A recent study attempted to investigate "how does ChatGPT (a GAI application) answer questions related to science education," "how would educators utilise ChatGPT in L&T" and, "how would an educator reflect on ChatGPT as a L&T tool" [9]. The author found it difficult to reflect on ChatGPT's influence on L&T and student lives [9].

In order to assess the impact of Large Language Models (LLM) on learners and their individual and collective lives, participants in one of the two streams of the 2023 National Technology Leadership Summit identified five themes of Generative Artificial Intelligence (GAI) for academics to investigate and reflect upon. These themes included (i) truth and verisimilitude, (ii) equity and justice, (iii) professional works, mindsets, tasks, and skills, (iv) the intersection of GAI in the context of teacher preparation program administration,

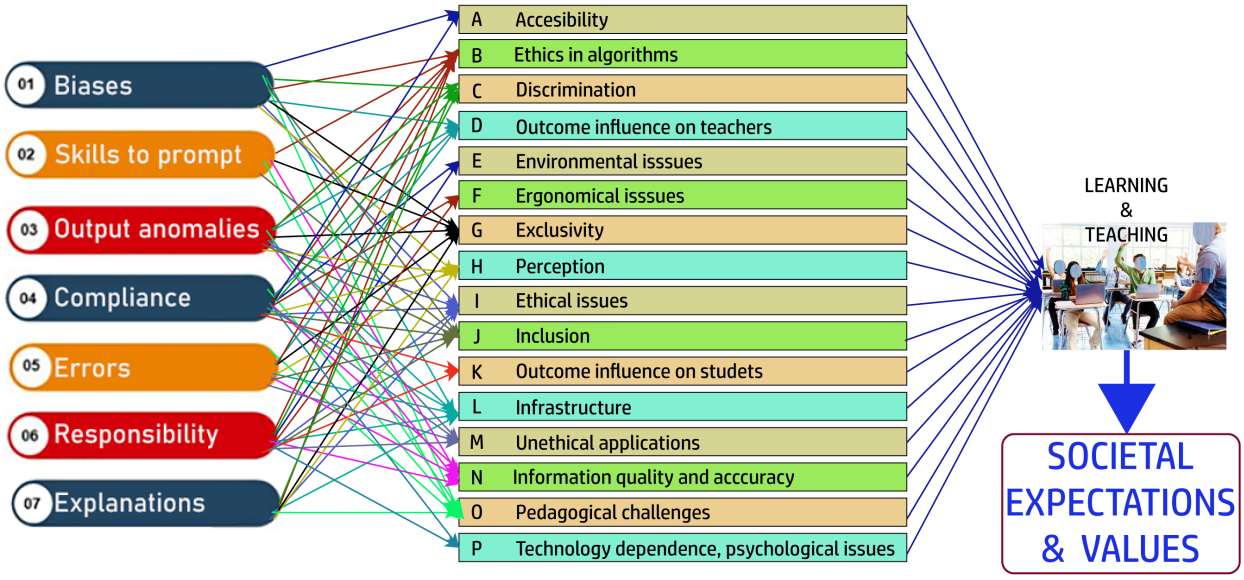


Fig. 1. Major problems built into GAI systems and their perceived effects on L&T and the society.

evaluation, and curriculum, and (v) curriculum for teaching Artificial Intelligence (AI) and society [10]. Another recent work urged educational policymakers, AI system developers, teachers, administrators, and students to probe how AI could be used effectively, ethically, and equitably [11].

Authors in [12] emphasized on teaching ethical use of AI to students and developing contemporary AI literacy. They suggested ensuring that students are aware of GAI's agency, authenticity, and accountability related issues. They suggested that GAI has provided opportunities of reimagining assessment design and pedagogy [12]. A previous work, having realized that GAI tools enable and often support academic plagiarism, proposed extending the scope of authentic assessment[13].

While assessing the impacts and challenges of GAI in medical education, the author appreciated GAI's abilities to generate synthetic data, simulate various conditions in a risk-free manner, and enable swift clinical diagnosis have been noted [14]. However, the author identified ensuring realism in simulations and assessing the accuracy of data synthetic data as major challenges[14].

Another recent study investigated the societal impact of GAI and suggested taking a balanced path while assessing the GAI applications. They recommended avoiding both, the utopian and the dystopian approaches [15]. The paper suggested that predicting the long-term impacts of GAI would be rather unrealistic [15].

A recent study [16] asserts the need for a conceptual understanding of the likely implications of GAI applications in real situations. The authors investigated as to how the adoption of GAI applications would influence the current practices. They identified important challenges and several ethical issues that warrant new development of new skill in users of the technology [16].

The above-cited works highlight the following important issues:

- GAI technologies and applications have the potential to reform and reshape science and engineering education. Their adoption cannot be avoided but their short and long-term influences on academic endeavours and educational activities need to be understood.
- People involved in all spheres of science and technology education need to be aware of the potential impacts of GAI on L&T. Especially, students need to be aware of GAI tools' social and ethical implications and feel their professional responsibilities albeit the emerging landscape of GAI technologies. The student should be able to appreciate their limits and liberties as GAI tools won't be able to impose the needed restrictions on adoption and use of GAI in various academic endeavours and scenarios.
- All stakeholders need to collaborate and work together for developing standards, frameworks and tools for assessing short- and long-term impact of GAI technologies.

This paper contributes to the ongoing efforts of building a generic, practical, easy to use and adopt tool for assessing impacts of GAI applications. The proposed Learning and Teaching Impact Assessment Tool (TALIAT) includes an seven-dimensional framework to let stakeholders assess important aspects of GAI applications. The included aspects are: Handling biases, prompt tuning and evaluation, output fairness, regulatory compliance, reasoning, error detection, accountability and explainability. TALIAT hypotheses that proper assessment of these aspects of a GAI system would ensure reviewing the purpose and impetuses, scope, and L&T impact mitigation strategies of a GAI applications. The overall architecture of TALIAT makes it useful during and after the development of a GAI application.

The following paragraphs present information on the approach we used for developing an impact assessment tool and categorically understanding the role of each of the seven aforementioned aspects associated with GAI tools' application in L&T.

## II. Mapping GAI Problems on Societal Practices

In order to develop an understanding of how the perceived threats and problems of GAI systems affect L&T practices and societal expectations, an extensive literature review was carried out. In particular, the cited literature examined the influence of GAI systems on teachers' perception and practices and students' progress and performance. Based on the cited works, GAI systems' challenges were mapped onto L&T and societal expectations. Figure 1 shows the GAI-posed challenges and their potential influence on L&T and society.

## III. Handling Biases

Bias in a GAI system and software is defined as an error or a systematic propagation of error/(s) which would lead to an unfair, imprecise process of inference and eventually a biased system output. As shown in Fig. 1, biases can creep into GAI systems at different stages of the system life cycle. For example, at the data acquisition stage, modelling stage, training stage, and even at the final stage where a user would bring in interpretation biases [17]. As a majority of GAI systems are trained on the datasets available via the World Wide Web, they carry Anglocentric monoculture biases [18]. Generally speaking, the GAI system tend to ignore Asian, Middle Eastern and even East European cultures and practices [19, 20]. Unsupervised learning is also considered a typical source of data bias which can easily bring unfiltered, less reliable and even fictitious information during the model-building stage of a GAI system. Another potential source of bias in the GAI system is human feedback when used for reinforcement learning [21]. Biases associated with model building may come from different sensitivity attributes and weights used by the system developers [22]. Figure 2 illustrates how various biases creep into a GAI system. Some interesting examples of how biases seep in software systems have been cited in [17]. The problem of manipulation and bias in GAI applications have been investigated in many recent studies [23]. In particular, harms and effects of societal biases encoded in GAI systems are evidenced in fear of affecting marginalised communities [24].

A good characterization of biases is given in [17] classifies biases as algorithmic biases, confirmation biases, generative biases, interaction biases, measurement biases, representation biases, and sampling biases.

The biases in a GAI tool crept in during the development and implementation stages must be assessed and their potential effects must be properly determined. All stakeholders of a GAI tool should be well aware of the biases and their mitigation strategies prior to its application in an educational setting.

### *A. How biases affect GAI applications in education?*

Biases in a GAI system can result in significant effects and can further perpetuate through various learning stages. Effects of GAI biases can amplify during the progressive phases L&T. For example, when prompted to comment on the evolution of Denavit-Hartenberg (D-H) parameters, the GAI tools would completely ignore the Russian origin of this analytical method resulting in a learner's inability to fully appreciate evolution of D-H parameters.

Biases in a GAI system could also limit access to learning and comprehension leading to misinformation and knowledge deficit. Different learners using different GAI systems might experience inequalities in multiple levels of comprehension and competence, work readiness and performance.

### *B. Mitigating biases in GAI application*

The AI literature provides an array of techniques to mitigate biases in GAI systems. These techniques include pre-processing of data, model refinement and selection and, post-processing of decisions [17]. Developing moderate size datasets with cultural and geographical orientations is also suggested [18]. human assessment of models and classifiers [25], testing models and classifiers using adversarial and simulated data [25].

## IV. Prompt Tuning and Evaluation

Tweaking and repositioning GAI systems enable a foundational model to carry out new tasks. This capability of a GAI system, is often referred to as "seed text tuning" or "prompt tuning." Prompt tuning can prepare and tweak a model to perform new and (usually) more complex tasks without retaining the underlying model. Hence, prompt training saves system developers from retraining the model while using prompts integrated into the input data. This is achieved by reformulating the downstream tasks such that they appear to be similar to the ones that have been solved earlier [26]. Many highly applicable and useful prompt tuning methods lack explainability as their insertions are usually abstract and the host transformer models demand a high level of self-attention computation [27].

TALIAT posits that prompt tuning should be regarded as an important part of a GAI system used in L&T environments. Having an efficient prompt tuning methodology embedded, a GAI system would readily learn student difficulties and would be able to handle often complex evolving questions. For example, after learning the attributes of an element X, a learner might use a prompt, "X is found in_____" and hope the GAI tool to complete the prompt [26]. This becomes doable because a prompt is considered as a sequence of additional tokens systematically pre-pended to a sequence of tokens [28]. Prompt-supported learning in GAI systems

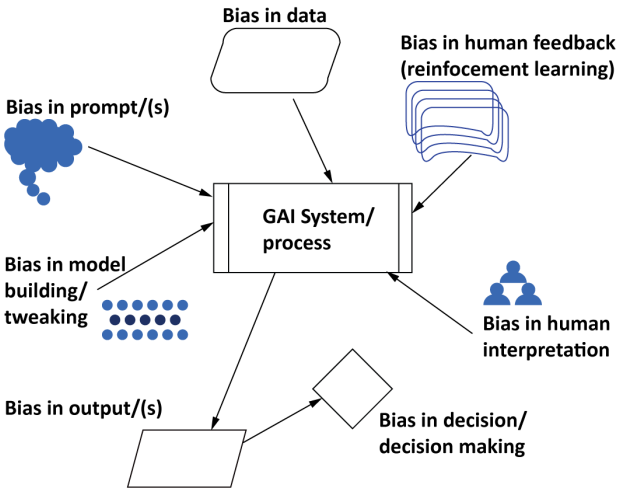


Fig. 2. Sources of biases in a GAI system and the resulting biases

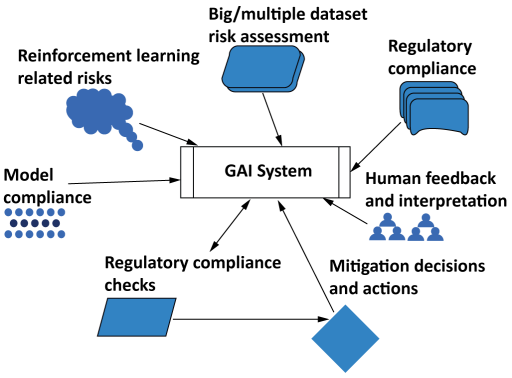


Fig. 3. A generic representation of a GAI system's regulatory compliance related issues.

enables random writing, language translation, coding and conversation [29].

## V. OUTPUT FAIRNESS ASSESSMENT

Please note that we have used the term "output fairness" rather than fairness. This was necessary to allow assessing the fairness of a GAI system based on output alone but not on the basis of a system's outcome. A general understanding of the connection between bias and fairness would lead to the assumption that reducing biases in a software would automatically increase fairness of a system. However, this assumption would result in oversimplification of the issue of fairness in the context of a GAI system. Fairness in a GAI system is a complex issue. Biases cannot be equated against fairness in a GAI system such that all biases are cancelled by fairness [17, 30]. Five major types of fairness have been identified in the literature [17]. They are described below:

- Causal fairness – The system can ensure obstructing all forms of historical biases.
- Counterfactual biases – the system can ensure that the same decision is made in various situations, and conditions.
- Group fairness – the system can ensure demographic parity and would provide equal opportunity to various groups and individuals. It would effectively avoid disparate judgements.
- Individual fairness – the system can ensure individuals of the same attributes are treated alike. For example, the GAI system's output would not vary in terms of the nature and frequency of output production.
- Procedural fairness – the system would ensure transparency and consistency of the decisions and the decision-making process.

Traditionally, testing and evaluation of AI systems is based on systematic assessments against discriminatory instances. Practitioners use adversarial data generated under certain assumptions for testing and evaluation. Weaknesses of this approach have been highlighted in [31]. Particularly, deviation from the real data distribution was considered a serious drawback. A previous study [31] envisaged a novel approach for generating rather natural individual discriminatory instances using a Generative Adversarial Network (GAN). Authors imitated the decision boundary of

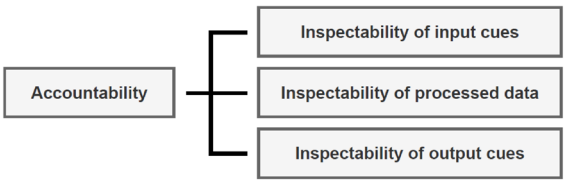


Fig. 4. Components of accountability [44, 45]

the target model in the semantic latent space of GAN and achieved better fairness assessment performance [31].

Our assessment tool, as will be shown later, benefits from both the traditional and GAN assisted approaches for assessing the output fairness of a GAI system.

## VI. REGULATORY COMPLIANCE

Ethical issues in development and use of GAI systems related to specific domains are less understood and haven't been fully examined yet [32]. Any algorithm-based system or software is expected to follow its internal regulatory mechanism vis-a-vis remain compliant to external regulations under all conditions [33]. The frequent use of human oracle-assisted reinforcement learning, training on multiple massively large datasets and a lack of global and local mechanisms to ensure regulatory compliance of GAI systems have been recognized in the recent literature [32, 34]. Figure 3 shows a generic model that includes major issues and risks associated with a typical GAI system. However, the regulatory compliance and legal limits of L&T focused tools differ from finance, marketing and law related GAI tools in terms of their implications and impacts. While finance and business related GAI tools are often required to provide comprehensive documentation and detailed description of the tool [35], such requirements couldn't be typically imposed on L&T systems. With the emergence of European Union's AI act and similar initiatives in other parts of the world, this situation might change soon [36, 37]. In the domain of L&T, system developers and users need to ensure accuracy, currency, relevance and permanence of the information at both, input and output sides of the GAI tool. Furthermore, ethical compliance (included in regulatory compliance box of Figure 2) has lateral and longitudinal issues related to all L&T spheres. Hence, the L&T stakeholders need to be aware of the health, governance, scope and purpose of the fundamental sources and the underly knowledge reinforcement mechanisms of a GAI tool.

An important question regarding compliance assurance in a GAI system stems from the trustworthiness of information and the authenticity of data available to the GAI tool. This issue would often require establishing the quality and methodological framework of the sources used during the training and modelling stages [38, 39].

Another important aspect of compliance with regulatory issues is embedded in differences between the local and global compliance requirements and frameworks [40]. Thus, a practical yet universally applicable compliance framework needs to established. TALIAT attempts to address the aforementioned problems and presents a doable set of checks for the GAI system.

## VII. ERROR DETECTION

Evaluation of trustworthiness is critical for confidently using a GAI tool [41]. Errors in GAI systems are interpreted

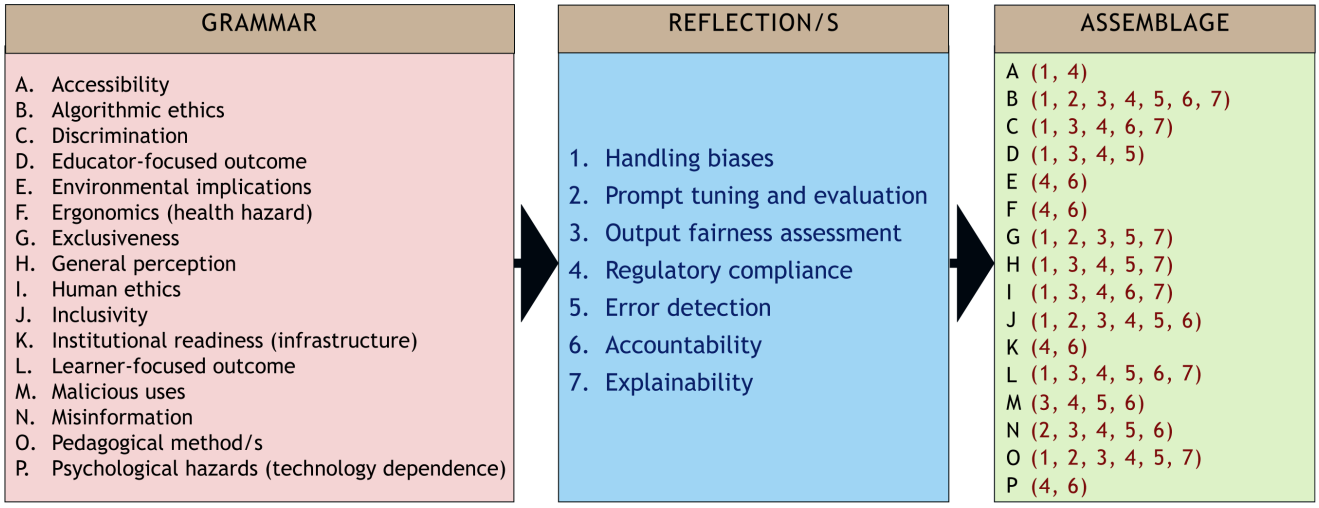


Fig. 5. Development of the impact assessment framework.

in many ways, so they are referred to as using various terms like hallucination, misinformation and lack of reliability [42]. Being a generative system, a GAI tool would not be able realize the authenticity or accuracy of the facts, information and inferences. This is a complex and multifaceted issue that needs further investigation.

In a human-to-human conversation, a person would easily highlight a lack of reliability in a statement by saying "*I am not very sure*" or "*I am fifty percent confident*." This type of uncertainty statement were easily incorporated in expert systems via an explicit mention and expression of confidence level [43, 44]. In order to address the issue of reasoning and errors in AI systems, a recent article suggested, '*The key is to replace "reasoning by association" with "causal reasoning" —the ability to infer causes from observed phenomena'* [23].

A widely used approach used to estimate errors in GAI systems rely on GAI systems' built-in ability to garner data and information from multiple sources which, the system and system developers and users find trustable. In the recent literature, several deviations from this approach have been reported. For example, a two-step framework was proposed in [41] which first requires a LLM to provide a justification of its answer and then aggregates the justification for confidence estimation.

In TALIAT, a novel idea is being proposed to estimate the trustworthiness of answers and the probability of having erroneous answers.

## VIII. ACCOUNTABILITY

Accountability in a GAI system reflects on the ownership of outcomes, the responsibility for results and the obligatory response of various stakeholders. The idea behind establishing responsibilities and obligations is not new to software systems but its implications are wider in GAI systems [45]. As shown in Figure 4, authors in [45] broken down accountability of an AI system into three measurable components viz., inspectability of input cues, quality of data being processed and the output information. In [46] examples of incorporating these three measurable components was demonstrated. Authors in [47] identified and prioritized enablers for GAI system to be used in medical education. Their investigations revealed that credibility was the first and accountability was the second most important enabler of GAI systems. Being an important enabler, establishing accountability of the actions, outcomes and answers of a GAI system must be easy and consistent across various situations [35, 48].

TALIAT proposes an easy to assess and universal approach of assessing GAI systems' accountability.

## IX. EXPLAINABILITY

Almost all genres of GAI systems need to have explainability embedded in them. Knowing their internal decision-making and inference mechanism may help users in deciding about their applicability. This is often referred to as 'lack of transparency' which connotates missing components of explainability from the GAI systems impede understanding of their reasoning and inferences [23, 49]. Factors such as the mere size of data used in training them, their almost unlimited scalability and their ability to regenerate new information make it very difficult to express GAI systems' explainability. More details have been reported in [45] that highlight reasons and remedies for bridging the gap between practitioners' requirements and system developers' understanding of explainability in AI systems. The literature asserts that such gaps would cause a lack of acceptance and less utility of AI systems. Other factors impeding adoption of LLM models have been reported in [50]. Overall, the literature asserts that some degree of explainability should be embedded in GAI systems to take full advantage of these systems and truly reflect on their limitations in L&T arena. Interestingly, authors in [51] provide a grouping of methods under two categories: fine-tuning based method and prompt-based method. used for incorporating explainability.

Building upon these previous studies, TALIAT provides a novel method of determining common impacts of a lack of explainability in GAI systems on their L&T application.

## X. IMPACT ASSESSMENT FRAMEWORK

In order to develop an impact assessment too, a large group of elements of grammar (antecedents) was selected from the GAI literature. The selected grammar was examined to see if the terms in the set of antecedents reflect on major aspects of GAI related issues. As shown in Figure 5, seven aspects of GAI's were carefully examined to verify if the antecedents would fall under at least one of the aspects. Once the antecedents and their reflections were fully established, an

assemblage process was carried out. The idea was to see which of the selected antecedents would reflect on one or more aspects of the GAI systems. Figure 5 depicts the framework development process. An exhaustive and iterative assessment of the process and the resulting frame work helped including appropriate questions in TALIAT, our assessment tool. In order to simplify the tool, redundancies were removed and a simple yet comprehensive tool was developed.

## XI. The Proposed Tool

After the removal of redundancies and overlaps, a set of thirty-two questions was incorporated in the proposed assessment tool. All questions are appended below.

1. Were the potential users or groups of users correctly identified?
2. Was it easy to identify and understand problems the GAI tool was supposed to solve?
3. Is the GAI tool compatible with common online language translators?
4. Is the user-interface compliant with CSS and HTML standards[52]? For example page and text scaling features are ensured.
5. Do the user interface and various input/output backgrounds and fonts have excellent colour contrast and complement to support vision impaired demographics[53]?
6. Is text to speech and speech to text conversion possible?
7. Would the system use require any special device, instrument, hardware or software?
8. Were the word association tests conducted? This is to ensure that gender, race, age, geography and skin colour related biases are removed from the underlying algorithms.
9. Were the models trained on multiple datasets which were inclusive and exhaustive? This is to ensure that a GAI model isn't heavily tilted toward a particular set of publishers or resources.
10. Were various classes of data included in the training datasets? For example, attention was paid to the nature of data used in developing the model. For example, personal, public, proprietary, non-commercial, commercial sources of data were examined.
11. Were the possible data connectivity tests conducted?
12. Were connections with third-party models examined and reported?
13. Were the training protocols sensitive to the possibility of manifesting name, origin, and social status related biases?
14. Was a hierarchical responsibility structure ensured to identify, select and include various sources of information?
15. Did the system outputs consistently exhibit the desired level of bias-free behaviour?
16. Was the relevance and accuracy of outputs extensively tested against the input prompts?
17. Were any data/ information filters incorporated to reject noisy, irrelevant and unreliable data during the reinforcement learning process?
18. Were the relevant legislations identified? Were complicities ensured?
19. Were the relevant regulatory bodies identified? Were their requirements understood and followed?
20. Would the outputs properly cite sources of data and information? Would it be possible to move between the cited sources for verification and validation of outputs?
21. Would it be possible to report and/or cite any evaluations of the sources, data and information in the output?
22. Is the tool capable of distinguishing between real and fake data/information? Would it be able to warn users about any potential harms of using 'bad' or 'mix of good and bad' data?
23. Was it possible to identify compliable ethical frameworks and guidelines during the data acquisition and model development stages?
24. Were legacy issues investigated? Were the negative effects of the system use-history, previous prompts, and the legacy-caused restrictions reported?
25. Was the privacy of prompts, data, information and outputs ensured?
26. Were the data and information accuracies assessed in terms of stability, repeatability and scale?
27. Is the system capable of assessing prompts and suggesting any possible improvements to the prompts?
28. Was it possible to use some elements of fairness for assessing the system outputs?
29. Does the system compromise on accuracy for being fair? For example, does it accept and mix 'bad' data with 'good' data for being fair.
30. Was any self-correction mechanism incorporated in the system? This feature should be available at both the input and output ends.
31. Were the ownerships of data and information assessed before acquiring them? Were there any ownership related restrictions on data acquisition and model training?
32. Were the legacy related system access issues examined?

### *A. Using the tool*

Each element of GAI reflection is assessed on a scale of 0-5 scale. Such a scale-based assessment would identify strong and weak elements of reflection in a GAI tool. For example, if each question pertaining to accuracy is marked as 5 but questions pertaining to graphics and text are marked as 2 and 3, the GAI tool would have strong accuracy but weak accessibility. Hence, an ideal and perfect GAI tool would be

able to score the maximum (160) points and a complete failure type of tool would score minimum (0) points.

NOTE: We are in the process of assessing three LLM's viz., BERT, ChatGPT and Claude using the TALIAT. The results are anticipated to be published in a following paper.

## XII. Conclusion

Based on an extensive review of the recent literature, we have developed a simple GAI impact assessment tool. Our proposed tool projects sixteen antecedents of the GAI grammar on seven widely cited and globally accepted aspects of GAI viz., handling biases, prompt tuning and evaluation, output fairness, regulatory compliance, error detection, accountability and explainability. The seven implications of adopting GAI tools in the L&T arena were examined through thirty-two relevant questions. Answers and associated weights to these questions will help in assessing impacts of a GAI system. The proposed impact assessment schema is simple and practical. Depending on the level of objectivity in an assessment, each one of the thirty-two questions will be evaluated on a scale of 0 to 5. This scale-based evaluation of features would enable distinguishing between weak and strong aspects of the GAI system under investigation. Hence, TALIAT and would let system developers, software assessors and users focus on remedial actions and remove the undesired features from the GAI tool under question.

To the best of our understanding and knowledge, TALIAT is the first simple and comprehensive tool that would facilitate quick and holistic assessment of a GAI tool. However, use of TALIAT in its current form is not immune to assessors' understanding and biases. Nonetheless, TALIAT provides a candid assessment regardless of the domain of a GAI system's application. We intend to automate the assessment process in a future work. Our possible automation schema would be based of extraction of data on users' feedback and TLIAT-initiated experimental data gathered through online and real time sessions of using a GAI application. We anticipate that TALIAT's grammar and reflective connotations will help all stakeholders concerned about GAI's influence on L&T activities.

## References


[1] S. Bayram and A. P. Nous, "Evolution of Educational Software Evaluation: Instructional Software Assessment," *Turkish Online Journal of Educational Technology-TOJET,* vol. 3, no. 2, pp. 21-27, 2004.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[3] J. Kocoń *et al.*, "ChatGPT: Jack of all trades, master of none," *Information Fusion,* vol. 99, p. 101861, 2023.

[4] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971,* 2023.

[5] A. Zeng *et al.*, "Glm-130b: An open bilingual pre-trained model," *arXiv preprint arXiv:2210.02414,* 2022.

[6] J. Holmes *et al.*, "Evaluating large language models on a highly-specialized topic, radiation oncology physics," *Frontiers in Oncology,* vol. 13, 2023.

[7] S. Wu *et al.*, "A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology," *arXiv preprint arXiv:2308.04709,* 2023.

[8] D. Beerbaum, "Generative Artificial Intelligence (GAI) Software–Assessment on Biased Behavior," *Available at SSRN 4386395,* 2023.

[9] G. Cooper, "Examining science education in ChatGPT: An exploratory study of generative artificial intelligence," *Journal of Science Education and Technology,* vol. 32, no. 3, pp. 444-452, 2023.

[10] C. Mouza, "Editorial: A report on the 2023 National Technology Leadership Summit. Contemporary Issues in Technology and Teacher Education," *Contemporary Issues in Technology and Teacher Education Journal,* vol. 23, no. 4, Editorial, 2023.

[11] M. Searson, E. Langran, and J. Trumble, "Exploring New Horizons: Generative Artificial Intelligence and Teacher Education," 2024.

[12] J. Mao, B. Chen, and J. C. Liu, "Generative Artificial Intelligence in Education and Its Implications for Assessment," *TechTrends,* vol. 68, no. 1, pp. 58-66, 2024.

[13] M. M. Khan, Y. Dong, and N. A. Manesh, "Authentic assessment design for meeting the challenges of Generative Artificial Intelligence," in *2023 IEEE Frontiers in Education Conference (FIE)*, 2023: IEEE, pp. 1-8.

[14] J. P. Singh, "The Impacts and Challenges of Generative Artificial Intelligence in Medical Education, Clinical Diagnostics, Administrative Efficiency, and Data Generation," *International Journal of Applied Health Care Analytics,* vol. 8, no. 5, pp. 37-46, 2023.

[15] R. Sabherwal and V. Grover, "The Societal Impacts of Generative Artificial Intelligence: A Balanced Perspective," *Journal of the Association for Information Systems,* vol. 25, no. 1, pp. 13-22, 2024.

[16] R. Gupta, K. Nair, M. Mishra, B. Ibrahim, and S. Bhardwaj, "Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda," *International Journal of Information Management Data Insights,* vol. 4, no. 1, p. 100232, 2024.

[17] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci,* vol. 6, no. 1, p. 3, 2023.

[18] A. Singh, "Diverse Yet Biased: Towards Mitigating Biases in Generative AI (Student Abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 21, pp. 23653-23654.

[19] D. Gal, "Perspectives and approaches in AI ethics: East Asia," *Oxford Handbook of Ethics of Artificial Intelligence, Oxford University Press, Forthcoming,* 2019.

[20] N. Berberich, T. Nishida, and S. Suzuki, "Harmonizing artificial intelligence for social good," *Philosophy & Technology,* vol. 33, pp. 613-638, 2020.

[21] X. Zheng, J. Li, M. Lu, and F.-Y. Wang, "New Paradigm for Economic and Financial Research With Generative AI: Impact and Perspective," *IEEE Transactions on Computational Social Systems,* 2024.

[22] C. T. Teo, M. Abdollahzadeh, and N.-M. Cheung, "FairTL: a transfer learning approach for bias mitigation in deep generative models," *IEEE Journal of Selected Topics in Signal Processing,* 2024.

[23] J. M. Bishop, "Artificial intelligence is stupid and causal reasoning will not fix it," *Frontiers in Psychology,* vol. 11, p. 513474, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7874145/pdf/fpsyg-11-513474.pdf.

[24] A. Chan, "GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry," *AI and Ethics,* vol. 3, no. 1, pp. 53-64, 2023.

[25] D. McDuff, S. Ma, Y. Song, and A. Kapoor, "Characterizing bias in classifiers using generative models," *Advances in neural information processing systems,* vol. 32, 2019.

[26] L. Yan, C. Han, Z. Xu, D. Liu, and Q. Wang, "Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning," in *Proceedings of international joint conference on artificial intelligence (IJCAI)*, 2023, pp. 1622-1630.

[27] A. Bozkurt and R. C. Sharma, "Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world," *Asian Journal of Distance Education,* vol. 18, no. 2, pp. i-vii, 2023.

[28] K. Sohn *et al.*, "Visual prompt tuning for generative transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19840-19851.

[29] D. Park, G.-t. An, C. Kamyod, and C. G. Kim, "A Study on Performance Improvement of Prompt Engineering for

Generative AI with a Large Language Model," *Journal of Web Engineering,* vol. 22, no. 8, pp. 1187-1206, 2023.
[30] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *2018 IEEE international conference on big data (big data)*, 2018: IEEE, pp. 570-575.
[31] Y. Xiao, A. Liu, T. Li, and X. Liu, "Latent imitator: Generating natural individual discriminatory instances for black-box fairness testing," in *Proceedings of the 32nd ACM SIGSOFT international symposium on software testing and analysis*, 2023, pp. 829-841.
[32] D. Oniani *et al.*, "Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare," *NPJ Digital Medicine,* vol. 6, no. 1, p. 225, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10693640/pdf/41746_2023_Article_965.pdf.
[33] S. Ingolfo, A. Siena, J. Mylopoulos, A. Susi, and A. Perini, "Arguing regulatory compliance of software requirements," *Data & Knowledge Engineering,* vol. 87, pp. 279-296, 2013.
[34] R. Al-Shabandar, G. Lightbody, F. Browne, J. Liu, H. Wang, and H. Zheng, "The application of artificial intelligence in financial compliance management," in *Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing*, 2019, pp. 1-6.
[35] J. Schneider, R. Abraham, and C. Meske, "Governance of Generative Artificial Intelligence for Companies," *arXiv preprint arXiv:2403.08802,* 2024.
[36] J. Faivre, "The AI Act: Towards Global Effects?," *Available at SSRN 4514993,* 2023.
[37] J. Laux, S. Wachter, and B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk," *Regulation & Governance,* vol. 18, no. 1, pp. 3-32, 2024.
[38] A. Smolansky, A. Cram, C. Raduescu, S. Zeivots, E. Huber, and R. F. Kizilcec, "Educator and student perspectives on the impact of generative AI on assessments in higher education," in *Proceedings of the tenth ACM conference on Learning@ Scale*, 2023, pp. 378-382.
[39] S.-C. Kong and Y. Yang, "A Human-Centred Learning and Teaching Framework Using Generative Artificial Intelligence for Self-Regulated Learning Development through Domain Knowledge Learning in K–12 Settings," *IEEE Transactions on Learning Technologies,* 2024.
[40] S. Wang, "Research on Generative Artificial Intelligence Management," *International Journal of Computer Science and Information Technology,* vol. 2, no. 2, pp. 350-356, 2024.
[41] M. Li, W. Wang, F. Feng, F. Zhu, Q. Wang, and T.-S. Chua, "Think twice before assure: Confidence estimation for large language models through reflection on multiple answers," *arXiv preprint arXiv:2403.09972,* 2024.
[42] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," in *2023 IEEE Global Engineering Education Conference (EDUCON)*, 2023: IEEE, pp. 1-9.
[43] J. Jiang, G. Klein, and R. Vedder, "Persuasive expert systems: the influence of confidence and discrepancy," *Computers in Human Behavior,* vol. 16, no. 2, pp. 99-109, 2000.
[44] K. Mehmood, "Diagnostic expert system for troubleshooting hydraulic systems," *International journal of computer application technology,* vol. 8, no. 1-2, pp. 116-120, 1995.
[45] M. M. Khan and J. Vice, "Toward accountable and explainable artificial intelligence part one: theory and examples," *IEEE Access,* vol. 10, pp. 99686-99701, 2022.
[46] M. Khan and J. Vice, "Toward Accountable and Explainable Artificial Intelligence Part Two: The Framework Implementation," *Authorea Preprints,* 2023.
[47] M. Quttainah, V. Mishra, S. Madakam, Y. Lurie, and S. Mark, "Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study," *JMIR AI,* vol. 3, no. 1, p. e51834, 2024.
[48] K. Wach *et al.*, "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT," *Entrepreneurial Business and Economics Review,* vol. 11, no. 2, pp. 7-30, 2023.
[49] J. Sun *et al.*, "Investigating explainability of generative AI for code through scenario-based design," in *27th International Conference on Intelligent User Interfaces*, 2022, pp. 212-228.
[50] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys,* vol. 55, no. 5, pp. 1-31, 2022.
[51] H. Zhao *et al.*, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology,* vol. 15, no. 2, pp. 1-38, 2024.
[52] !!! INVALID CITATION !!! [52, 53].
[53] !!! INVALID CITATION !!! [53].