Generative AI in Writing Evaluation: Where We Stand and What Lies Ahead

Takahiro IWANAKA

Graduate School of Intercultural Studies, Graduate School of Yamaguchi Prefectural University

Abstract

Generative artificial intelligence is reshaping educational assessment; however, high-stakes evaluations of student writing remain contentious. This study proposes an LLM-derived similarity metric—cosine similarity between essay-level embedding vectors of student essays and expert model texts (e.g., instructor-written benchmark essays)—as an automated indicator of L2 English writing proficiency. Using a longitudinal design, about 35 Japanese university students will produce argumentative essays at three time points over a 15-week semester. Essays will be scored by trained human raters and analyzed for linguistic features, including lexical diversity, syntactic complexity, and cohesion. The author will examine (a) convergent validity via correlations between the similarity metric and human scores, (b) sensitivity to developmental change using repeated-measures models, and (c) incremental predictive validity through hierarchical regression by adding the similarity metric to models based on surface linguistic features. It is hypothesized that the similarity metric will show strong positive associations with human ratings, detect significant longitudinal gains, and explain unique variance beyond traditional feature-based predictors. If validated, this approach could support scalable diagnostics that complement human judgment and improve the reliability and pedagogical utility of L2 writing assessment.

1. Introduction and Research Background

The rapid emergence of Generative Artificial Intelligence (GAI) is fundamentally reshaping the educational landscape, presenting unprecedented opportunities and significant challenges, particularly within higher education. Advanced AI systems are no longer speculative; they are actively fostering personalized, adaptive, and interactive learning environments that challenge traditional pedagogical models (Întorsureanu et al., 2025; Wu & Zhang, 2025). Among the many domains impacted by this technological shift, the evaluation of student writing stands at a critical inflection point, demanding a reconceptualization of long-standing assessment practices.

The quest to automated writing assessment is not new. The history of Automated Essay Scoring (AES) dates back to 1966 with Ellis Page's pioneering Project Essay Grader (PEG) [1], which used statistical methods to correlate surface-level textual features with scores assigned by human raters. For decades, subsequent AES systems followed this paradigm, evolving with advancements in computational linguistics but remaining rooted in the principle of extracting and weighting predefined linguistic features. However, these systems have been persistently criticized for their inherent limitations. Their inability to capture the nuanced and holistic aspects of proficient writing—such as creativity, contextual appropriateness, and argumentative coherence—has led to enduring concerns within the educational assessment community about their reliability, validity, and fairness (Iwanaka, 2016; Shermis & Burstein, 2003; Shermis & Wilson, 2024).

The recent advent of Large Language Models (LLMs) like ChatGPT represents a paradigm shift in this field. Unlike traditional AES tools that rely on manually engineered features, LLMs leverage deep learning architectures trained on vast and diverse text corpora. This allows them to process language holistically, demonstrating a remarkable sensitivity to context, semantics, and rhetorical structure that qualitatively distinguishes them from their predecessors (Mizumoto & Eguchi, 2023; Ouyang et al., 2022). This technological leap moves beyond mere feature counting to a more integrated, contextual representation of a written text, opening new frontiers for automated assessment. Despite this profound potential, the

1 For the details of PEG, see https://ourl.jp/aIxYP.

application of LLMs to the high-stakes domain of writing assessment, particularly for second language (L2) learners, raises critical questions about reliability and validity that demand urgent, rigorous empirical investigation.

While preliminary studies have explored LLMs' capacity to replicate human holistic scores, a critical oversight remains in the current research agenda. The unique internal mechanics of LLMs—specifically their ability to represent the semantic and structural essence of a text as a high-dimensional vector—offer a novel and powerful avenue for assessment. Therefore, the potential of using LLM-derived similarity metrics (e.g., cosine similarity based on vector embeddings) as a stable, objective, and interpretable index of L2 writing development remains underexplored. Such a metric could move beyond a single, often opaque, score, providing a quantifiable measure of a student's text in relation to an expert model, offering a transformative tool for tracking progress and diagnosing areas for improvement. This proposal outlines a research plan designed to rectify this critical gap. The following literature review establishes the empirical and theoretical foundation for this study, synthesizing decades of research on automated writing evaluation to situate the current inquiry at the forefront of educational technology and assessment innovation.

2. Literature Review

This review critically examines the evolution of automated writing evaluation, from its statistical origins to the sophisticated capabilities of modern Large Language Models (LLMs). By summarizing the performance of established Automated Essay Scoring (AES) systems and detailing the current state of research into LLM-based assessment, this section establishes the empirical and theoretical justification for the proposed study. It highlights a critical gap in the existing literature and argues for a new direction in AI-driven writing assessment that leverages LLMs' unique internal representations.

2.1 From Statistical Models to Natural Language Processing in AES

Research into the efficacy of established AES tools provides a crucial baseline for evaluating modern AI. Studies on the pioneering Project Essay Grader (PEG) found that although it outperformed the predictive accuracy of two human raters, it was less predictive than an aggregate of four human raters, particularly for holistic scores and stylistic traits (Page, Poggio, & Keith, 1997). This early finding underscored a persistent theme: the trade-off between automated efficiency and the nuanced judgment of multiple human experts. Subsequent systems demonstrated incremental improvements. Research on e-rater, for instance, showed that it achieved correlation rates with human raters that were comparable to inter-rater correlations between two humans, providing strong support for its validity in large-scale assessment contexts (Burstein, Tetreault, & Madnani, 2013). Similarly, extensive validation of IntelliMetric™ [2] across numerous studies has consistently reported high agreement rates with human scores. Comparative studies have generally found no statistically significant differences between the mean scores assigned by IntelliMetric™ and those from human faculty raters, suggesting a high degree of alignment in scoring outcomes (Eliot, 2003; Rudner, Garcia, & Welch, 2006). These systems, while effective within their design parameters, primarily rely on quantifiable surface features and have been criticized for their inability to assess deeper qualities such as creativity and originality.

2.2 Current Investigations into LLM-Based Writing Evaluation

The emergence of LLMs has prompted a new wave of research focused on their reliability and the quality of their feedback. These investigations reveal a more complex and nuanced performance profile compared to their predecessors.

2.2.1 Scoring Reliability Analysis

Rigorous studies have begun to examine the consistency of LLM scoring. A recent study employing Generalizability

2 For the details of IntelliMetricTM, see https://ourl.jp/uZyeW.

Theory (G-theory) to analyze LLM performance on the AP Chinese Language and Culture Exam [3] found that generalizability coefficients—a measure of reliability—were consistently higher for trained human raters than for AI raters (Lee et al., 2024). This suggests that human experts currently exhibit greater consistency in their judgments. However, the same study found that hybrid models combining human and AI raters could improve overall scoring reliability, pointing toward a future of collaborative, rather than purely automated, assessment.

Further investigation into the intra-rater reliability of LLMs provides a complementary perspective. A large-scale study using the GPT-3 text-davinci-003 model to score over 12,000 essays from the TOEFL11 corpus [4] reported "moderate" intra-rater reliability (Mizumoto & Eguchi, 2023). The study, which used Quadratic Weighted Kappa to assess consistency, found that the model's scores typically varied by only 1-2 points upon re-scoring. While it achieved a high adjacent-agreement rate of 89.15% with the original benchmark levels, this result indicates stability but also highlights a non-trivial margin of variability that warrants caution in high-stakes applications.

2.2.2 Feedback Quality Analysis

Beyond scoring, the quality of automated feedback is a critical concern for pedagogy. A comprehensive study by Steiss et al. (2024) directly compared feedback from trained human educators with that from ChatGPT across five dimensions: criteria-based, clarity of directions, accuracy, prioritization of essential features, and supportive tone. They clarified that trained human educators provided higher-quality feedback across all dimensions, except "criteria-based," where the AI excelled at explicitly referencing the scoring rubric.

Crucially, the study uncovered instances in which ChatGPT generated factually inaccurate feedback, including confusing historical figures central to the essay's topic. This finding underscores a fundamental limitation: LLMs operate on predictive algorithms to generate plausible text and lack genuine comprehension, making them susceptible to producing authoritative-sounding misinformation that could mislead student writers.

2.3 Synthesis and Justification for the Current Study

The existing literature reveals a clear trajectory. Traditional AES tools, built on the analysis of surface features, have been thoroughly validated but are fundamentally limited in scope. Current research on LLMs in writing assessment has focused mainly on two areas: (1) evaluating their ability to replicate the holistic scores assigned by human raters and (2) analyzing the qualitative characteristics of the formative feedback they generate. While valuable, this focus fails to leverage the core technological advancement that LLMs represent: a shift from surface-feature analysis to deep semantic representation. In contrast to traditional AES pipelines that rely on hand-crafted features and a single predicted score, LLMs yield contextualized, high-dimensional vector embeddings that encode both semantic content and salient structural patterns. This observation reveals a significant and critical gap in the research: an absence of investigation into using these internal representations to generate quantitative, interpretable metrics of writing quality.

A similarity metric, calculated between a student's essay and a "gold standard" model text, represents the logical next step in AES, leveraging the very technology that makes LLMs so powerful. Such a metric could offer a more stable and diagnostic indicator of a text's alignment with proficient writing than a single holistic score, providing a more objective measure of development over time. This proposal is designed to address this critical oversight, moving beyond the question of whether an LLM can score like a human to ask whether its internal representations can provide a novel, valid, and reliable metric for L2 writing assessment. The following sections outline the specific research questions and methodological approach designed to explore this promising new paradigm.

---

3 The AP Chinese Language and Culture Exam is a college-level assessment designed to test high school students' proficiency in Mandarin Chinese and their understanding of Chinese culture.

4 The TOEFL11 corpus is a large-scale dataset specifically designed for Native Language Identification (NLI). Released by ETS in 2013, it contains 12,100 English learner essays from the TOEFL iBT test.

3. Research Aims, Questions, and Hypotheses

This section outlines the central objectives and guiding questions of the proposed research. Flowing directly from the gaps identified in the preceding literature review, this study aims to explore a novel application of LLM technology for L2 writing assessment that moves beyond holistic scoring to introduce a more objective and diagnostic metric of writing proficiency.

3.1 Primary Aim

The primary aim of this research is to investigate the validity and reliability of LLM-derived similarity metrics as objective measures of L2 English writing proficiency and for tracking developmental progress over time.

3.2 Research Questions

1. How strongly do LLM-derived similarity scores—calculated between student essays and expert-level model essays—correlate with holistic and analytic scores assigned by trained human raters?
2. Can longitudinal changes in LLM-derived similarity scores effectively and reliably track the developmental trajectory of L2 writers over an academic semester?
3. How does the predictive validity of an LLM-derived similarity metric compare to, and potentially complement, a model based on traditional linguistic features (e.g., lexical sophistication, syntactic complexity) in explaining the variance in human-assigned scores?

3.3 Hypotheses

1. There will be a strong, statistically significant positive correlation between LLM-derived similarity scores and the holistic proficiency scores assigned by trained human raters.
2. A longitudinal analysis will demonstrate that learners undergoing writing instruction exhibit a statistically significant increase in their essays' similarity scores from the beginning to the end of the term, corresponding to proficiency gains identified by human evaluators.
3. A multiple regression model combining the LLM similarity metric with traditional linguistic features will explain significantly more variance in human scores than a model using only traditional linguistic features.

The methodology detailed in the following section provides a rigorous framework for empirically testing this new paradigm in automated assessment, addressing these questions and testing the corresponding hypotheses.

4. Methodology

This study will employ a longitudinal quantitative design to provide a robust and comprehensive investigation into the proposed research questions. The approach is designed to validate the proposed LLM-derived similarity metric against established human judgments and track its sensitivity to developmental change in L2 writing over time.

4.1 Research Design

The research will utilize a longitudinal design with data collected at three distinct time points: the beginning (Week 1), middle (Week 8), and end (Week 15) of a standard academic semester. This structure will enable robust statistical analysis of developmental change, providing insight into how the similarity metric tracks student progress alongside formal instruction.

4.2 Participants

Participants will be recruited from university-level EAP (English for Academic Purposes) courses. The target sample will consist of approximately 35 L2 learners of English whose L1 is Japanese. Recruitment will be conducted through in-class announcements; students will receive a stipend for their participation.

4.3 Materials and Corpus Development

4.3.1 Writing Prompts

A set of standardized argumentative essay prompts will be developed. These prompts will be structured similarly to those used in high-stakes assessments like the TOEFL and IELTS, requiring students to formulate a clear position on a given topic and support it with logical reasoning and evidence (Refer to Appendix A for sample prompts).

4.3.2 Model Texts

For each prompt, a "gold standard" model essay will be professionally written. These texts will exemplify high-proficiency writing (equivalent to IELTS Band 9), demonstrating superior task fulfillment, coherence, lexical resource, and grammatical range and accuracy. These model texts will serve as the benchmark against which student essays are compared.

4.3.3 Evaluation Rubric

A comprehensive analytic scoring rubric will be adapted from established frameworks used in recent writing assessment research (e.g., Mizumoto & Eguchi, 2023; Steiss et al., 2024). The rubric will guide human raters in assessing four core criteria: Lexical Resource, Coherence and Cohesion, Grammatical Range and Accuracy, and Depth of Argument (Refer to Appendix C for a sample rubric).

4.4 Procedure

The study will proceed through four distinct stages.

4.4.1 Data Collection

Students will compose essays in response to the standardized prompts at the three specified time points during the semester. All writing sessions will be conducted in a controlled environment to ensure consistency in writing conditions and timing.

4.4.2 Human Rating

All collected essays will be anonymized and independently scored by two expertly trained human raters using the analytic rubric. A subset of essays will be scored by both raters to calculate inter-rater reliability, thereby ensuring high scoring consistency. Disagreements will be resolved through discussion to arrive at a consensus score.

4.4.3 LLM-Based Similarity Analysis

A pre-selected state-of-the-art LLM (e.g., a GPT-family model) will be used to generate a high-dimensional vector embedding for each student essay and its corresponding model text. A vector embedding represents a text as a point in a high-dimensional space, so that texts with more similar meanings are positioned closer together. In this study, embeddings serve as numerical representations that capture semantic content and salient structural characteristics of the texts.

The primary metric, Cosine Similarity, will be calculated between the vector for each student's essay and the vector for the corresponding model text. This calculation measures the angle between the two fingerprints; a smaller angle (a score closer to 1) indicates greater semantic alignment between the texts. This yields a score between 0 and 1 that represents their proximity to the expert model.

4.4.4 Linguistic Feature Analysis

As a point of comparison, all essays will be processed using established computational linguistics tools to extract a suite of traditional linguistic features. These features will include measures of lexical diversity, syntactic complexity, and cohesion, reflecting the metrics used in prior AES research (Mizumoto & Eguchi, 2023).

4.5 Data Analysis Plan

The data collected will be analyzed using a sequence of statistical procedures, each mapped to a specific research question:

To address RQ1/H1, Pearson correlation coefficients will be calculated to determine the strength and direction of the linear relationship between the human-assigned scores (both holistic and analytic sub-scores) and the LLM-derived cosine similarity scores.

To address RQ2/H2, a repeated-measures ANOVA or a linear mixed-effects model will be used to test for statistically significant differences in mean cosine similarity scores across the three time points (beginning, middle, and end). This will determine if the metric can effectively track developmental change.

To address RQ3/H3, a hierarchical multiple regression analysis will be conducted. In the first step (Model 1), the traditional linguistic features will be entered as predictors of human-assigned scores. In the second step (Model 2), the cosine similarity score will be added to the model. The analysis will focus on the change in R-squared (ΔR^2) to determine whether the LLM similarity metric explains a significant amount of additional variance in human scores beyond that captured by traditional features. This comprehensive methodological approach will allow for a thorough evaluation of the proposed metric, setting the stage for discussing its potential contributions to the field.

5. Expected Outcomes and Significance

This study is poised to generate critical insights into the next generation of automated writing assessment. By moving beyond traditional scoring paradigms, this research is expected to validate a novel, theoretically grounded metric for evaluating L2 writing. This section articulates the anticipated findings and argues for their significant contribution to second language acquisition research, language assessment, and pedagogical practice.

5.1 Expected Outcomes

Based on the proposed methodology, the author anticipates three key outcomes. First, the LLM-derived similarity metric is expected to be validated as a reliable proxy for expert human judgment, demonstrated by a significant positive correlation with human-assigned scores. Second, the metric should prove sensitive enough to track longitudinal development in L2 writing proficiency, showing a statistically significant increase in scores over an academic semester. Finally, the study aims to demonstrate that this metric captures unique dimensions of writing quality beyond traditional, surface-level linguistic features, accounting for additional variance in human scores within a regression model.

5.2 Scholarly and Practical Significance

The successful validation of this metric would carry significant implications for multiple stakeholders in the field of language education.

For SLA research, this study will introduce a novel, scalable, and objective methodology for quantifying writing development. The ability to generate a reliable, continuous metric of text quality automatically will enable more extensive and fine-grained longitudinal research into the trajectory of L2 writing skills, which the time and cost of expert human rating have historically constrained.

For language assessment, the similarity metric could serve as a foundational component of new diagnostic and formative assessment tools, rather than providing a single, summative score; such a tool could offer learners and teachers an

intuitive, interpretable index of their writing's proximity to a target model. This shifts the focus from a "right/wrong" paradigm to a more developmental one, highlighting a clear path toward improvement.

For pedagogical practice, the technology explored in this proposal could support a powerful hybrid assessment model. By using AI to provide rapid, data-driven insights on drafts, educators can be freed from more mechanical aspects of evaluation to focus their time and expertise on higher-order feedback, personalized instruction, and fostering critical thinking. This aligns with a constructive approach to learning where technology serves to augment the indispensable role of the teacher, fostering active engagement and more profound understanding (Cooperstein & Kocevar-Weidinger, 2004).

Ultimately, this research aims to provide an empirically validated tool to build more effective, efficient, and pedagogically sound approaches to writing instruction and assessment in the age of AI.

6. Discussion: The Evolving Role of Generative AI in Education

The research proposed herein is not merely a technical validation of a new metric; it is situated within a broader, transformative shift in the role of technology in education. Synthesizing insights from the rapidly growing body of literature, it is clear that integrating generative AI requires reconceptualizing pedagogical roles, learning processes, and ethical frameworks. This study aims to make a critical contribution to ensuring that this integration is both practical and responsible. The most productive future for educational AI lies not in replacing human educators, but in augmenting their capabilities. Evidence points to a hybrid, collaborative model in which AI and humans work in tandem, leveraging their respective strengths (Lee et al., 2024).

In this model, AI can excel at time-consuming but productive tasks such as initial scoring or feature analysis, freeing the human educator to transition from a "knowledge transmitter" to a "learning designer" (Yamauchi, 2025). This partnership enhances teacher productivity and allows them to concentrate on the uniquely human aspects of mentorship, providing nuanced feedback on argumentation and creativity, and tailoring instruction to individual student needs. From the student's perspective, GAI offers powerful new pathways for engagement. Empirical research demonstrates that both the quality of interaction with GAI and the quality of its output can positively influence learning motivation and self-efficacy, which are key mediators for improved learning outcomes (Bai & Wang, 2025).

However, the immense potential of GAI is accompanied by a significant and well-documented risk of producing inaccurate or biased information with a high degree of linguistic fluency (Steiss et al., 2024). This underscores the absolute necessity of human oversight and the cultivation of critical digital literacy. The pedagogical response to the technology's limitations is to design tasks that encourage knowledge construction—where students actively critique, question, and build upon AI-generated content—rather than simple regurgitation (Cooperstein & Kocevar-Weidinger, 2004). This approach leverages GAI to promote learning autonomy and a constructive orientation to knowledge itself.

This research proposal is driven by a vision of generative AI as a supportive force in education. By developing and validating objective, interpretable, and reliable AI-based metrics, the research community can provide educators and institutions with the tools needed to harness this technology effectively. The goal is to contribute to a future where AI does not supplant the core mission of education—the cultivation of critical, creative, and independent minds—but instead serves as a powerful new instrument to support and advance that mission.

References

Bai, Y., & Wang, S. (2025). Impact of generative AI interaction and output quality on university students' learning outcomes: A technology-mediated and motivation-driven approach. *Scientific Reports*, *15* (1), 1–15.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 55–67). Routledge.

Cooperstein, S. L., & Kocevar-Weidinger, E. (2004). Beyond active learning: A constructivist approach to learning. *Reference*

Services Review, *32* (2), 141–148.

Eliot, S. (2003). IntelliMetric: From Mary Had a Little Lamb to the GMAT. In M.D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Lawrence Erlbaum Associates.

Întorsureanu, I., Oprea, S.-V., Bâra, A., & Vespan, D. (2025). Generative AI in education: Perspectives through an academic lens. *Electronics*, *14* (5), 1053. DOI: https://doi.org/10.3990/ electronics14051053.

Iwanaka, T. (2016). Rewriting based on feedback: Automated-computer-based feedback on forms and taylor-made feedback on content. *CASELE Journal*, *41*, 61-70.

Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2* (2), 100050. DOI: 10.31219/osf.io/2uahv.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 27730–27744. DOI: 10.48550/arXiv.2203.02155.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). *Computer analysis of student essays: Finding trait differences in the student profile*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, USA.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric™ scoring system. *Journal of Technology, Learning, and Assessment*, *4* , 1–18.

Shermis, M. D. & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates.

Shermis, M. D. & Wilson, J. (2024). *The Routledge international handbook of automated essay evaluation*. Routledge.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. *Learning and Instruction*, *91*, 101894.

Wu, D., & Zhang, J. (2025). Generative artificial intelligence in secondary education: Applications and effects on students' innovation skills and digital literacy. *PLoS One*, *20* (5), e0323349. DOI: https://doi.org/10.1371/journal.pone.0323349.

Yamauchi, Y. (2025). Seisei AI no onsei taiwa kinō o katsuyō shita eigo supīkingu gakushū no genjō to kadai [Current status and challenges of English-speaking learning using the voice dialogue function of generative AI.] Nihon Kyōiku Kōgakkai Shūki Zenkoku Taikai Happyō Ronbunshū [Proceedings of the JSET Autumn National Conference], 269–270.

Appendices

Appendix A: Sample Writing Prompts

Prompt 1: Privacy vs. Public Security

Topic: In the age of big data and advanced surveillance, should personal privacy be considered an absolute right, or is it a necessity to sacrifice it for national security and public safety?

Directions:

Context: Introduce the current technological landscape, mentioning tools such as facial recognition, data mining, or global surveillance systems.

Thesis Statement: Clearly state your position on whether the protection of individual privacy should outweigh the needs of state security or vice versa.

Argumentation: Support your stance with at least three distinct points, considering ethical implications, potential for government overreach, and the effectiveness of surveillance in preventing crime.

Counter-argument and Rebuttal: Acknowledge a significant opposing viewpoint (e.g., "Those who favor security argue that...") and provide a logical rebuttal to strengthen your own case.

Structure: Ensure a formal academic structure with an introduction, well-developed body paragraphs, and a conclusion that synthesizes your main points.

Prompt 2: AI Integration in Higher Education

Topic: As Generative AI (such as ChatGPT) becomes increasingly capable of performing academic tasks, should universities strictly prohibit its use to protect academic integrity, or should they fully integrate it into the curriculum as an essential tool for the future?

Directions:

Context: Discuss the rapid rise of Generative AI and its impact on traditional methods of assessment in higher education.

Thesis Statement: Present a clear argument for either the restriction or the integration of AI tools within the university setting.

Critical Analysis: Analyze how your proposed approach affects the development of critical thinking skills, creativity, and the preparation of students for the professional workforce.

Ethical Considerations: Address issues such as the "digital divide" (inequality in access to AI) and the definition of original authorship in the 21st century.

Refutation: Identify the primary concerns of the opposing side (e.g., the fear of cognitive decline or the risk of plagiarism) and explain why your approach is more sustainable in the long term.

Prompt 3: State Intervention for Climate Change

Topic: Environmental experts argue that individual lifestyle changes are insufficient to combat climate change. To what extent should governments impose strict regulations on individual consumption—such as meat consumption, air travel, and energy use—to achieve net-zero goals?

Directions:

Context: Briefly outline the urgency of the climate crisis and the debate between individual liberty and collective environmental responsibility.

Thesis Statement: Define the degree to which you believe government intervention in personal lifestyle choices is justifiable or necessary.

Multidimensional Analysis: Examine the topic from at least three perspectives: economic impact, social justice (how regulations affect different socio-economic classes), and the limits of democratic freedom.

Evidence: Use logical reasoning and hypothetical or real-world examples to demonstrate the potential consequences of such regulations.

Counter-argument: Address the "infringement on personal freedom" argument. How should a society balance the survival of the planet with the rights of the individual?

Appendix B: Assessment Criteria for B2 Learners

Lexical Resource: Use of academic vocabulary and avoidance of repetitive phrasing.

Cohesion and Coherence: Effective use of transitions (e.g., Furthermore, Notwithstanding, In light of this) to connect complex ideas.

Grammatical Range: Correct use of complex structures such as conditionals, passive voice, and relative clauses.

Depth of Argument: The ability to sustain a logical argument over 800 words without relying on fluff or circular reasoning.

Appendix C: Argumentative Essay Assessment Rubric

Criteria	Level 4 Exceptional	Level 3 Proficient	Level 2 Developing	Level 1 Limited
Lexical Resource	Uses a wide range of academic and topic-specific vocabulary naturally and accurately. Demonstrates strong control of collocations and register. Errors are rare and do not impede meaning.	Uses a sufficient range of vocabulary to allow flexibility and precision. Shows good awareness of style and collocation, though some minor inaccuracies in word choice or word formation may occur.	Vocabulary is adequate for the topic but may be repetitive or overly general. Occasional errors in word choice occur, but the meaning is clear. Limited use of high-level academic terms.	Vocabulary is basic or repetitive. Frequent errors in word choice, spelling, and word formation make it difficult for the reader to follow the argument.
Cohesion and Coherence	Information and ideas are logically organized and flow smoothly. Uses a wide variety of cohesive devices (transitions, pronouns, substitution) effectively. Paragraphing is logical and sophisticated.	Information is organized logically, with a clear overall progression. Uses a range of cohesive devices appropriately, though there may be some over-use or under-use in specific sections.	Ideas are generally organized, and the essay follows a standard structure. Some cohesive devices are used correctly, but transitions between paragraphs may feel mechanical or abrupt.	Organization is evident but lacks logical flow. Cohesive devices are faulty, inadequate, or repetitive. Paragraphing may be confusing or absent.
Grammatical Range and Accuracy	Uses a wide range of complex structures (e.g., conditionals, passive voice, nominalization) with high accuracy. Punctuation is used effectively to enhance the argument.	Uses a mix of simple and complex sentence forms. Shows good control of grammar; while some errors may occur in complex structures, they rarely cause misunderstanding.	Produces a range of sentence structures, but complex sentences are often less accurate than simple ones. Grammar errors are present but do not significantly obscure the message.	Uses only a limited range of structures. Frequent grammatical errors in basic tense, agreement, or sentence construction significantly distract the reader.
Depth of Argument	Presents a persuasive, nuanced thesis. Arguments are supported by sophisticated evidence and deep critical analysis. The counterargument and rebuttal are handled with high logical precision.	Presents a clear and consistent position throughout the essay. Main ideas are extended and supported with relevant examples. The counterargument is clearly identified and logically addressed.	The position is clear, though some supporting points may lack depth or be slightly repetitive. The counterargument is included but may be addressed predictably or superficially.	The position is unclear or inconsistent. Arguments are under-developed, purely descriptive, or lack supporting evidence. Fails to address a counterargument effectively.