Article

How to Use LLMs Ethically in Academic Writing?

Tiantian Yu [1,*]

[1] Beijing Academy of Social Sciences, Beijing, China
[*] Correspondence: Tiantian Yu, Beijing Academy of Social Sciences, Beijing, China

Abstract: This paper presents an experimental study based on selected Large Language Models (LLMs) and Artificial Intelligence Generated Content (AIGC) detection systems, conducted within a mixed-methods research paradigm that combines empirical validation and Qualitative Content Analysis (QCA). The empirical validation process consists of both a condition optimization experiment and the main experiment, while the materials for qualitative content analysis are directly derived from these experimental outputs. In the experiments, six LLMs are evaluated using four different AIGC detectors. Through the analysis of the contents generated by these LLMs, the existing theoretical framework, which is referred to as the authors' checklist, for the application of LLMs in academic writing is revised. The updated framework refines the checklist step for assessing and amending the accuracy of AI-generated content. The updated framework contains five steps, Intellectual Contribution, Accuracy of Conceptions, Accuracy of Demonstrations, Academic Competency, and Transparency, for authors' academic writing with the assistance of LLMs. Additionally, it emphasizes the importance of authors' innovation and proficiency in prompting LLMs when ethically using LLMs in academic writing.

Keywords: large language models; academic writing; control experiment; qualitative content analysis

1. Introduction

With the evolution of LLM (Large Language Model) technologies and AIGC (Artificial Intelligence Generated Content) detection systems, it has become increasingly feasible to use LLMs as writing assistants in academic writing. However, various studies have identified numerous issues associated with LLMs' application, including inaccurate content and information, risks of academic misconduct, and increased administrative costs.

To address these challenges brought by the emerging technologies, scholars have explored potential solutions from multiple perspectives — such as users' acceptance, LLMs prompting, the performance of AIGC detection systems, and institutional management strategies. Among these, the educational application of LLMs is influenced by a range of factors, leading to diverse opinions within the academic community [1]. These views vary greatly and are sometimes even contradictory. At the core of this debate lies a fundamental question: Is it ethical?

Nevertheless, from scientific ethics to human-centered policies, the use of LLMs in academic writing is inherently tied to two essential processes: content generation and content verification [2]. Based on these two essential processes, this study designed experimental validation to examine existing theoretical frameworks. As a mixed-methods research attempt in the field of educational technology, this study aims to contribute to reconciling current controversies surrounding LLMs. Ultimately, this paper seeks to answer one central question: How can we make it ethical?

Published: 18 May 2025

2. Literature Review

2.1. LLMs as writing assistants

LLMs as writing assistants can be unrecognizable compared with human writing. A. Molinari and E. Molinari revealed that, within the specific domains of literature review and academic writing, LLM-generated papers not only received higher scores compared to human-written ones but also proved equally difficult to differentiate from student-authored texts during assessment [3].

In recent studies, the bias and misleading nature of content generated by large language models have drawn significant attention. Alvero and colleagues found that human-authored texts exhibited greater individual variability compared to those produced by AI. Notably, AI-generated writing most closely resembled essays written by male students with higher social privileges. The study also highlighted significant inconsistencies and biases in the evaluation of writing, calling for further research to improve the alignment between human and AI authors [4]. Fang and colleagues found that all tested LLMs exhibited notable disparities in the representation and sentiment of certain social groups within the generated news content. These disparities were reflected in reduced visibility, more frequent negative tone, and limited thematic diversity for specific groups [5]. Sun and colleagues conducted a systematic classification and in-depth analysis of distorted information in AI-generated content. They identified eight primary error categories and thirty-one secondary error types, providing a foundational framework for studying the risks associated with AIGC [6].

In terms of writing performance assessment, LLMs generally demonstrate good internal scoring reliability and high validity [7]. When it comes to improving writing performance, LLMs can serve as a valuable educational tool [8].

Some studies have explored collaborative writing with AI assistance. For instance, Wiboolyasarin et al. underscore the positive impact of a collaborative writing intervention enhanced by AI feedback on specific aspects of writing quality [9].

In the context of English education in the United Arab Emirates, Anna Dillon and colleagues employed an interpretive grounded theory approach to analyze focus group interview data. Their study addressed key questions regarding the application of LLMs in relation to academic integrity and similar issues. The research highlights the need for broader future investigations into how different countries and regions are responding to the complex realities of AI-assisted collaboration [10].

As academic research assistants, LLMs can provide comprehensive support for scientific inquiry, including hypothesis generation, method design, data analysis, and manuscript preparation [11]. LLMs such as ChatGPT have shown significant potential in enhancing the English academic writing skills of non-native-speaking medical students and may contribute to the process of educational assessment, particularly in contexts where English is not the primary language [12].

Moreover, large language models demonstrate the capacity for reflective writing — a feature that, when integrated into academic writing practices, may significantly improve the rigor and precision of scholarly output. Studies indicate that models such as ChatGPT can generate high-quality reflective responses, yielding favorable results in various pharmacy education contexts. Given the increasing prevalence of AIGC-generated texts that are difficult for human evaluators to detect, the development of LLMs specifically tailored for reflective writing offers potential to support the responsible and pedagogically aligned use of AI technologies in educational settings.

Research on users' acceptance of LLMs indicates that positive affective experiences related to AI interaction can partially mitigate concerns regarding its use [13]. Within the educational domain, Q. Ma, P. Crosthwaite, D. Sun, and D. Zou explored the challenges teachers may encounter when integrating ChatGPT into their pedagogical practices [14]. The authors suggest that language educators should actively leverage ChatGPT’s distinctive capabilities to advance language teaching toward a more technologically enriched and innovative direction. For learners, the integration of LLMs into writing tasks necessitates both the motivation and the ability to critically revise AI-generated texts. As such, developing proficiency in effectively refining and adapting AIGC outputs is becoming an essential skill in AI-enhanced learning environments [15].

However, not all research findings fully embrace the use of LLMs in academic writing. Scholars have raised concerns regarding the risks and challenges associated with employing LLMs in the writing of research papers and theses. D. E. O'Leary argues that researchers and students must remain accountable for their work and should not misuse AI technologies to replace their own critical thinking and creativity [16].

The study introduces the Dunning-Kruger effect — a cognitive bias in which individuals with limited knowledge or competence in a domain tend to overestimate their abilities, while those with greater expertise often underestimate their skills. In addition, O'Leary conducted a sentiment analysis using LIWC-22 and found that many comments expressed cautious attitudes toward the involvement of AIGC in doctoral dissertations, with many perceiving the use of AI-generated writing as unethical [16].

In response to the potential risks associated with LLM applications, Van Niekerk, Delport, and Sutherland propose an active learning intervention, propose an active learning intervention aimed at reducing students' inappropriate reliance on AI tools. Through a multi-stage intervention design, the study helps students recognize the limitations of ChatGPT in academic writing, thereby discouraging its misuse and promoting more responsible engagement with AI-assisted writing technologies [17].

In the light of the most recent studies concerning the use of LLMs in academic writing, it is apparent that two critical areas require focused consideration in this study:

First, the ethical framework intended to be developed in this research does not have to be tailored for application within a Chinese language context. Both the creation and validation of this framework can cater generally to all authors speaking different languages.

Second, the management framework proposed by this study must facilitate the enhancement of users' abilities to identify errors in AIGC-generated content and to effectively revise materials produced by large language models. This will ensure that users can critically engage with AI outputs and improve the quality of their work.

2.2. AIGC Detectors

The application of LLMs in academic writing raises fundamental questions regarding the reliability and normativity of AI-generated content. These concerns necessitate the development and evaluation of AIGC detection systems to address critical challenges such as academic integrity, the trustworthiness of information, and the development of practical guidelines for academic writing assessment.

In this context, Ibrahim examined the phenomenon of AI-facilitated plagiarism in English writing and proposed the application of fine-tuned language models like RoBERTa as viable tools for identifying machine-generated content [18]. Furthermore, Y. Li, L. Sha, et al. demonstrated that BERT-based classifiers achieved high accuracy in differentiating between student-written and ChatGPT-generated texts, indicating their effectiveness as AIGC detection solutions [19].

While current research on AIGC detection in educational settings has largely centered on algorithmic techniques involving textual features, corpus construction, and classification strategies, this study aims to move beyond theoretical exploration and focus on the real-world performance of leading AIGC detection platforms. Specifically, we will assess the capabilities of systems such as CNKI AIGC Detector, Mitata, VIP Paper Check System, Turnitin, and Chat Zero, providing empirical insights into their applicability and limitations in academia.

2.3. Frameworks of AI-assisted Academic Writing

In contemporary market-oriented contexts, LLM developers are encouraged to enhance user experience, establish consistent pricing models, and provide a range of algorithmic solutions to better serve both individual and institutional users. Empirical validation helps refine the theoretical framework and contributes to shaping the underlying philosophical orientation guiding the educational application of Large Language Models [20].

In a recent systematic review, M. Khalifa and M. Albadawy applied the PRISMA framework to guide the process of study selection and inclusion [21]. Through this rigorous methodological approach, they proposed a comprehensive theoretical model identifying six domains in which AI can contribute to academic functions and its broader potential to transform academic practices. One of the central insights from their work is the significant role of AI in enhancing academic writing across six distinct dimensions: idea generation, content structuring, literature synthesis, data management, editing, and ethical compliance.

In a recent study, Qureshi et al. introduced a secure and scalable framework for AI-assisted academic writing, built upon Microsoft Azure cloud infrastructure. The research systematically reviewed current AI writing platforms, identifying their core functionalities, advantages, and limitations. The proposed framework was subsequently tested and empirically supported through a series of controlled experiments, demonstrating its potential for real-world implementation [22].

Meanwhile, A. Cheng, A. Calhoun, and G. Reedy explored the ethical integration of generative AI tools into academic research [23]. Their work outlines three defined domains in which such tools can be responsibly utilized. More importantly, the authors propose four guiding principles aimed at ensuring high standards of academic integrity and quality in AI-supported research: Intellectual Contribution, Academic Competency, Accuracy of Content, and Transparency.

Overall, recent theoretical developments in AI-assisted writing provide valuable insights for this study. Notably, M. Khalifa and M. Albadawy identified six key areas in which AI can systematically improve academic writing performance. This theoretical model will inform the development of our management framework, with particular attention to adjusting its components to better align with AIGC detection requirements [21].

Equally significant is the secure and empirically validated framework developed by Qureshi and colleagues, which demonstrates strong applicability and reliability. The secure and empirically validated framework proposed by Qureshi et al. will be referenced to ensure the robustness and practical feasibility of our proposed system [22].

Moreover, the four guiding principles — Intellectual Contribution, Academic Competency, Accuracy of Content, and Transparency — introduced by A. Cheng, A. Calhoun, and G. Reedy, will serve as a normative basis for ensuring the responsible and ethical integration of generative AI into academic writing [23].

3. Methodology

3.1. Empirical Validation

Drawing upon previous research, experimental validation in the field of AIGC detection often involves comparative studies among multiple large language models (LLMs) and qualitative or quantitative content analysis. Ibrahim employed a comparative research design to investigate the effectiveness of two AI-generated text detection platforms — GPT-2 Output Detector and Crossplag AI Content Detector — in identifying machine-generated texts [18].

In another study, Y. Cheng, Y. Fan, X. Li, G. Chen, D. Gašević, and Z. Swiecki, conducted an experiment comparing participants' questioning behaviors when using generative AI versus human mentors as writing supports [8]. The differences were quantified through Principal Component Analysis (PCA) and Epistemic Network Analysis (ENA).

Similarly, Pack et al. evaluated the performance of four prominent LLMs — including Google's PaLM2, Anthropic's Claude 2, and OpenAI's GPT-3.5 and GPT-4 — in automatically scoring English learners' writing, providing empirical insights into the capabilities of these models in educational assessment [7].

In alignment with these methodological approaches, this study will conduct comparative experiments and content analysis on widely used AIGC detection systems currently in practice, including CNKI AIGC Detector, MitataAI, and Turnitin AIGC.

The experimental design consists of two main stages:

Condition optimization experiment: Comparing the outputs of selected LLMs, this part of experiments aims to identify LLMs that generate content with high authenticity in Chinese. Based on the selection, a set of AIGC samples will be created, which present a higher level of detection difficulty.

Main experiment: Using the generated AIGC samples, the detection performance of each AIGC detector will be analyzed and compared. The goal of the main experiment is to identify the most effective AIGC detectors, which will serve as a foundational basis for the subsequent development of the theoretical framework.

3.2. Condition Optimization Experiment

The sample used in this experiment is the argumentative section of my master's thesis. The original Chinese text and its English translation are provided in Appendix A. This document was selected as the research sample because the entire thesis was written in Chinese without any assistance from large language models throughout the writing process.

In the first stage of the experiment, I selected six LLMs that are commonly used for Chinese text generation: DeepSeek, Kimi, Tongyi Qianwen, Xinghuo Spark, ChatGLM, and Baidu Yiyan. We assigned identification codes to texts generated by these models as follows: content generated by DeepSeek was labeled D1, D2, D3, D4, and so on; content from Kimi was labeled K1, K2, K3, K4, etc.; content from Tongyi Qianwen was labeled T1, T2, T3, T4, and so forth; content from Xinghuo Spark was labeled X1, X2, X3, X4, etc.; content from ChatGLM was labeled Q1, Q2, Q3, Q4, and so on; and finally, content from Baidu Yiyan was labeled B1, B2, B3, B4, etc.

In the second stage, I provided each LLM with prompts based on the theme of the argumentative section in my original thesis. These prompts instructed the models to generate texts comparable in length to the original human-written content. Using consistent instructions across all models, I engaged in multiple rounds of revision and refinement with each LLM. Each revised version of the generated texts was saved and labeled according to the coding system established in Stage One. The exact wording of the prompts or instructions used during these interactions is provided in Appendix B.

In the third stage, I selected the texts labeled with "1" — the outputs from the first instruction — and submitted them to three AIGC detection systems for analysis. This process yielded baseline detection scores for each model. Based on these initial results, a second prompt was issued to generate the texts labeled with "2". The two prompts used in this phase were as follows:

1) Please generate a 1200-word article on the topic "Talents in Colleges and Universities" as the argumentative section of a master's thesis. Prior to writing, please review relevant literature and ensure that your content reflects scholarly viewpoints. The language should meet the standards of a master's thesis, and references should be properly cited.
2) Very good. Next, please revise the content you generated by referring to the text provided (see Appendix A), so that it appears more human-like.

In the fourth stage of the experiment, the two generated samples with the lowest AIGC detection scores were selected for further processing. These outputs — from the two LLMs demonstrating the highest ability to evade detection — were combined into a single text. The newly integrated text was then submitted to AIGC detection systems for re-evaluation. Subsequently, based on the detection results, modification instructions were issued to the two selected LLMs, and the prompts were as follows:

3) Very good. Next, please integrate the two similar passages generated under Instruction 2 into one coherent 1200-word article.
4) Very good. Next, please revise the content within the parentheses so that it appears more human-written.

In the final stage of the condition optimization experiment, the manually selected, edited, and proofread the texts labeled "4" or "5" — that is, the outputs generated after multiple rounds of revision by the two selected LLMs. The content produced by these models was then synthesized into a single 1200-word finalized text.

This finalized text, refined through both AI-generated iterations and human post-processing, was used as the main experimental sample in the subsequent stages of the study. It represents a hybrid form of AIGC with high authenticity and reduced detectability, making it suitable for evaluating the performance of current AIGC detectors in academia.

3.3. Main Experiment

3.3.1. Control Experiment I: Performance of Different AIGC Detection Systems

This study adopts the theoretical framework of AI hallucination error types proposed by Y. Sun et al. to assess the accuracy of AIGC detection systems. Based on this classification, I designed a measurement scheme to evaluate how well each detection system identifies and categorizes different types of AI-generated content [6].

A generated text that does not contain any distorted information as defined in the measurement scheme (as Table 1) receives a full score of 10 points. To validate the reliability and consistency of the measurement, all generated texts from the previous stages were evaluated using this scheme. The evaluation covered all model iterations and instruction rounds, identified by their numerical codes.

Table 1. Measurement Scheme for AIGC Detection.

Distortion Information Category		If yes	If no
Overfitting	Illusions of confidence	0	0.5
	Falling into traps	0	0.5
	Flattery	0	0.5
Logic errors	Causal uncorrelation	0	1
	Contradictions	0	1
Reasoning errors	Spatial reasoning errors	0	0.5
	Temporal reasoning errors	0	0.5
	Hypothetical reasoning error	0	0.5
Unfounded fabrication	False proof	0	0.5
	Pseudoscience	0	0.5
	False academic information	0	0.5
Factual errors	Common sense mistakes	0	0.5
	Objective fact errors	0	1
	Authorship errors	0	1
Text output errors	Repetition and redundancy	0	1

A generated text that does not contain any distorted information as defined in the measurement scheme (as Table 1) receives a full score of 10 points. To validate the reliability and consistency of the measurement, all generated texts from the previous experimental stages — across all model iterations and instruction rounds — were evaluated by this scheme according to their numerical identifiers.

The resulting score distributions are presented in Figure 1. The line chart demonstrates a logically coherent trend: as LLMs receive refined and effective instructions, they can generate increasingly accurate and contextually appropriate content.

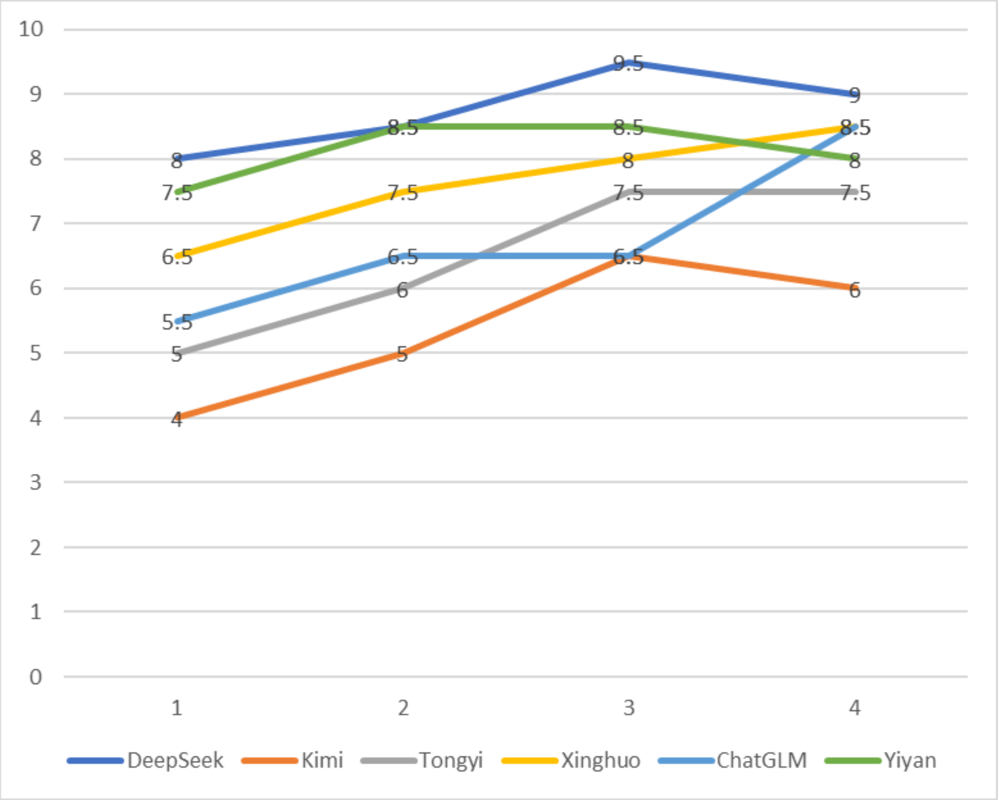

Figure 1. Score Trends of LLM Outputs under Progressive Instruction Optimization.

3.3.2. Control Experiment II: Performance Across Different Language Contexts

Having examined the detection metrics of LLM-generated content and the capabilities of various AIGC detectors, a follow-up question arises: Will the detection results be affected when the text is translated into another language? As a critical component of training data and model architecture, language may act as a barrier or a modulating factor in the performance of AIGC detectors. Therefore, it is plausible that AIGC detection indicators vary depending on the language used.

To investigate this potential confounding variable, we conducted a controlled cross-lingual experiment. Experimental texts from the Condition Optimization Experiment were translated into nine target languages using the Youdao Translation System. Youdao was selected for this study due to its strong performance in handling Chinese input — being a China-developed translation platform — and its demonstrated accuracy in translating academic-style content. The nine target languages were English, French, German, Portuguese, Spanish, Russian, Arabic, Korean, and Japanese.

Each translated version of the original text was then submitted to three widely used AIGC detection platforms: CNKI AIGC Detector, Chat Zero, and Turnitin, to evaluate whether detection scores vary significantly across multilingual contexts.

This control experiment aims to assess the language sensitivity of current AIGC detection tools and explore whether multilingual transformations can affect the reliability of AIGC identification — particularly in scenarios involving cross-cultural or international academic communication.

3.4. Qualitative Content Analysis

To rigorously examine the performance of AIGC detectors and AI-based revisions, this section presents a detailed analysis of 26 experimental samples generated and revised by LLMs. This analysis involves three main stages: coding, categorization, and theoretical integration.

For coding the texts, manual annotation is performed to identify similarities and differences in the texts generated and revised by different LLMs. The primary focus is on pinpointing key distinctions between the outputs of various models and how these differences manifest in terms of content quality, coherence, and adherence to academic standards.

For categorization, I utilize the four core principles of academic integrity proposed by Cheng, A., Calhoun, A., & Reedy, G. as the theoretical framework for classifying the coded data. These principles include Intellectual Contribution, Academic Competency, Accuracy of Content, and Transparency. Each sample is evaluated against these criteria to determine its alignment with ethical standards of academic writing [23].

In theoretical integration, I compare the current theoretical assumptions with the actual classification framework derived from the empirical data. Based on the empirical findings, an Ethical Framework is developed specifically for contemporary AIGC detection technologies. This framework provides a structured approach to ensure that AI-generated content aligns with academic integrity standards, while also identifying areas where current detection systems require improvement.

4. Results and Discussion

4.1. LLMs and AIGC Detector

The results of the first AIGC detection on the experimental samples are referred to as Result I. The samples for the second round of detection were revised by AI and subsequently proofread manually, with their detection results recorded as Result II. The third round used samples in which the AIGC-detected sections are modified manually, and the corresponding results are labeled as Result III. As shown in Table 2, the three results show a decreasing trend, with Result III being closer to the standards of academic writing compared to Result II and Result I. This indicates that both AI and human revisions reduce detection rates, but human editing is more effective in reducing the AIGC detection rate. Notably, the detection results from Mitata deviate significantly from those of the other systems, suggesting that there are substantial technological barriers in update efficiency among current AIGC detectors, while LLMs continue to evolve in text generation capabilities.

Table 2. Detection Results Across Three Rounds of Revision.

	Result I	**Result II**	**Result III**
CNKI AIGC Detector	96.49%	84.66%	19.32%
Mitata	3%	3%	1%
VIP Paper Check System	77.63%	65.38%	14.95%
Turnitin	89.6%	78.74%	23.4%

The results of Control Experiment II are shown in Table 3. Turnitin demonstrates sensitivity to English and Western languages, but is relatively less responsive to texts written in Asian languages. The CNKI detection system performs well in identifying direct translations from English and is primarily optimized for language patterns common in Chinese academic writing. Chat Zero performs consistently across multilingual contexts, though its results show a certain deviation from the highest level of AIGC detection accuracy.

Table 3. Detection Results of Translated Samples Across Different Languages.

	CNKI	Chat Zero	Turnitin
English	96.49%	41.03%	99.07%
French	-	39.72%	88.2%
German	-	37.95%	91.32%
Portuguese	-	40.18%	86.89%
Spanish	-	38.59%	87.14%
Russian	-	40.22%	70.34%
Arabian	-	37.60%	65.03%
Korean	-	42.78%	76%
Japanese	-	42.15%	79.74%

4.2. Coding and Categories

This qualitative content analysis adopts the four key principles of academic integrity proposed by A. Cheng, A. Calhoun, and G. Reedy — Intellectual Contribution, Academic Competency, Accuracy of Content, and Transparency — as its theoretical framework. Table 4 lists the codes generated during the coding phase, along with their corresponding explanations [23].

Table 4. Codes and Explanations from QCA.

Code	Explanation
a1: Non-academic explanation of core concepts	The LLM first explains the core concept, with data sourced from general web searches rather than academic databases.
a2: Presentation of related concepts	After explaining the core concept, the LLM integrates related concepts into a single paragraph to enrich the content.
a3: Substitution without logic	The LLM connects and combines related content superficially, lacking internal logic and rigor.
a4: Explanation and expansion of other relevant terms	The LLM generates text related to the core concept but lacks academic coherence and logical structure.
a5: Generalized or vague examples	The LLM cites broad or ambiguous examples in the text without elaborating on specific real-world cases.
a6: Formulaic responses in humanities	The LLM produces clearly templated structures within the overarching framework.
a7: Unsubstantiated viewpoints and conclusions	The LLM generates stereotypical conclusions without supporting evidence or data.
a8: Stereotyping	The LLM draws stereotypical conclusions about individuals related to the core concept.
a9: Oversimplified causal reasoning	The LLM proposes solutions or strategies based on overly simplified understandings of complex situations.
a10: Template-like sentence patterns	The LLM uses formulaic expressions in introductions and conclusions, showing signs of templated writing.
b1: Matching literature	The LLM identifies relevant literature based on the core concept and summarizes its content.
b2: In-depth dissection of themes	The LLM deeply analyzes the instructed theme, locates corresponding materials, and integrates them effectively.
b3: Matching associated themes	The LLM accurately matches content related to the main topic.

b4: Confusion of thematic concepts	The LLM introduces concepts that do not align with the instructed theme.
b5: Misuse of academic language	The LLM generates content that appears academic but is irrelevant or meaningless in context.
c1: Connecting themes through case studies	The LLM uses relevant cases to explain the intersection of two complex topics.
c2: Transformation of syntax and sentence structure	The LLM restructures grammar and sentence patterns based on the original material to generate new content.
c3: Repurposing ideas without attribution	Without citing sources, the LLM presents other scholars' views and conclusions to explain relationships between complex topics.
d1: Colloquialization of academic language	The LLM mimics human conversational style to evade AIGC detection systems.
d2: Phrase deconstruction and reconstruction	The LLM deconstructs and rephrases original phrases to avoid detection.
d3: Overcomplicating simple expressions	The LLM imitates explanatory academic language by overfitting on simple statements.
d4: Adding connectives for appearance	The LLM inserts linking words between two cases or topics to simulate coherence, though the internal logic remains weak.
d5: Superficial editing	The LLM does not verify the accuracy of cases or arguments, only modifying sentence-level structures.
d6: Localized revisions	The LLM edits content without considering the overall structure, altering key local information and disrupting the flow of the entire passage.
d7: Imitation of writing styles	The LLM uses rhetorical devices to mimic the linguistic style of academic writers.

According to the codes and explanations in Table 4, I categorize the 25 codes within the predefined theoretical framework. The specific analytical process is shown in Table 5, where the 25 codes are grouped into six categories, each linked to one of the four principles of academic integrity.

Table 5. Theoretical Integration and Categorization.

Predefined Theory	Categories	Codes
Intellectual Contribution	Receiving the Instructional Theme	a1; b1
	Shallow Concept Expansion	a2;a4;a5; b2;b3; c1
Accuracy of Content	Superficial Content	a7;b5; d5; a8
	Misleading Information	a3; b4;c3; d3;d4;d6
Academic Competency	Formulaic Writing	a6;a9;a10
Transparency	Imitative Human Editing	c2; d2;d7;d1

4.3. Ethical Framework for Using LLMs

After conducting a reverse analysis of the categorization process presented in Table 5, I find that Accuracy of Content can be further operationalized into multiple strategies. These strategies can be integrated and adapted by scholars when revising or translating AI-generated content. Particularly, they highlight practical approaches for improving the factual reliability and coherence of AI-assisted outputs.

Furthermore, Academic Competency encompasses not only a scholar's ability to demonstrate academic innovation, but also their capacity to prompt LLMs. Moreover, it includes the ability to manage and critically evaluate AI-generated content — a skill that is becoming increasingly essential in the age of AI-assisted writing.

By integrating the findings from the empirical validation of the study, I construct an Ethical Framework. This framework is visually represented in Figure 2: Ethical Checklist for the Use of LLMs in Academic Writing, which outlines key ethical considerations and best practices for AI assistance in scholarly work.

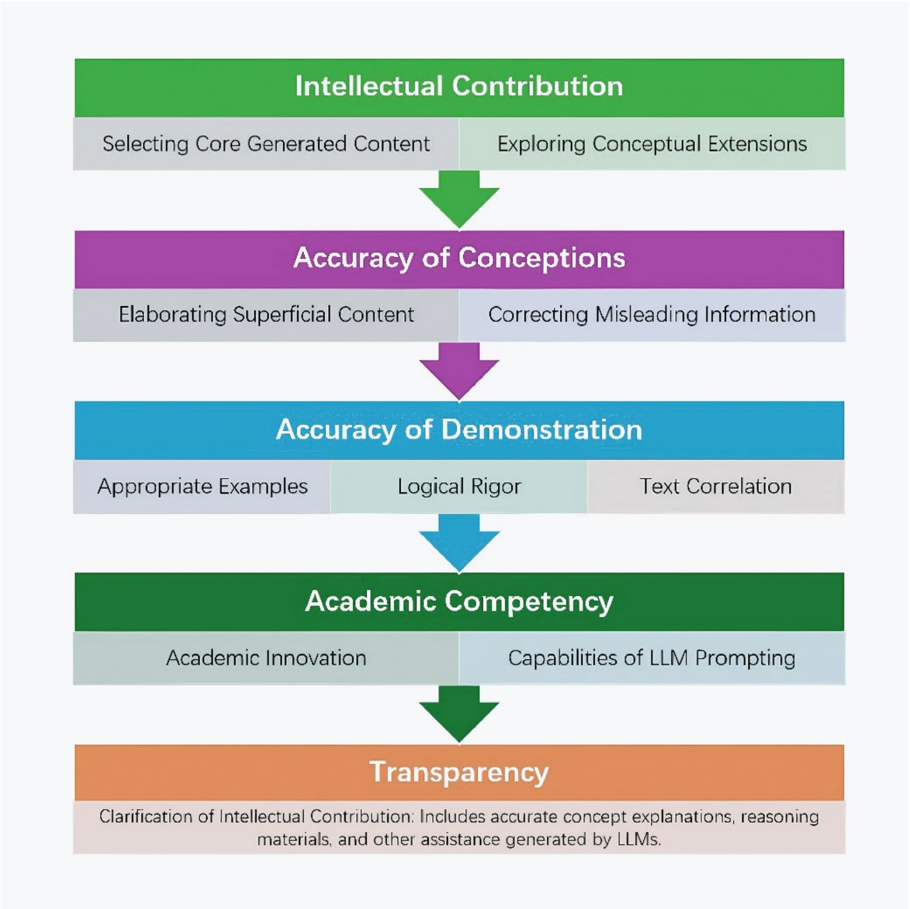
Intellectual Contribution
Selecting Core Generated Content
Exploring Conceptual Extensions
Accuracy of Conceptions
Elaborating Superficial Content
Correcting Misleading Information
Accuracy of Demonstration
Appropriate Examples
Logical Rigor
Text Correlation
Academic Competency
Academic Innovation
Capabilities of LLM Prompting
Transparency
Clarification of Intellectual Contribution: Includes accurate concept explanations, reasoning materials, and other assistance generated by LLMs.

Figure 2. Ethical Checklist for the Use of LLMs in Academic Writing.

As A. Cheng, A. Calhoun, and G. Reedy suggested, the four key principles of academic integrity include Intellectual Contribution, Academic Competency, Accuracy of Content, and Transparency. Based on this, I have updated the author's checklist for LLMs in academic writing by proposing five main steps, each with detailed guidelines [23].

The first step of the ethical framework is to select the core concept(s) explanation(s) that the author would like to focus on, while allowing the LLM to generate related topics and concepts.

The second step of the procedure is to revise the conceptions in the previously AI-generated content, including elaborating on superficial content and correcting misleading information.

The third step of the checklist is to revise the AI-generated demonstrative content, which can be guided by three main criteria:

1) Are the examples appropriate? If yes, the author is advised to annotate the intellectual contribution. If not, the author must delete or replace the AI-generated examples.
2) Is the logic of the demonstration rigorous and consistent? If yes, the author is advised to annotate the intellectual contribution. If not, the author must correct the logical errors.

3) Are the texts and paragraphs logically connected? If yes, the author is advised to annotate the intellectual contribution. If not, the author must rethink and clarify the internal logic and coherence within the content.

The fourth step of the checklist is to remind authors to focus on academic innovation and to develop the ability to craft effective prompts for LLMs.

The last step is to clarify the intellectual contributions of LLMs, including accurately generated content, correct explanations and demonstrations, revised logical structure, and linguistic translations, among others.

The entire checklist consists of five main steps: Intellectual Contribution, Accuracy of Conceptions, Accuracy of Demonstrations, Academic Competency, and Transparency. Compared with the theoretical framework proposed by A. Cheng, A. Calhoun, and G. Reedy, the theoretical framework of this study presents a different perspective regarding the order of Accuracy of Content and Academic Competency, as well as a more specific interpretation of Accuracy of Content in the context of LLMs in academic writing. The checklist emphasizes the innovative and decisive role of authors when using LLMs as academic writing assistants, whereas the original theoretical framework focuses more on avoiding misconduct and delinquency in AI-assisted writing [23].

5. Conclusion and Implication

This study, following a creatively mixed-methods paradigm, finds that different LLMs show varying performance, and their generated content would become increasingly standardized in response to manually specific instructions. Furthermore, there are noticeable discrepancies among AIGC detection systems, which requires algorithms that evolve alongside LLMs. In addition, the capacity of AI revision is currently limited to imitating basic human language patterns and has not yet achieved proficiency in learning critical thinking as humans.

However, even the most advanced LLM cannot fully replicate human scholars' writing or replace human contributions to academia. Since human scholars are the users of AI, and they bear the responsibility to use LLMs in an appropriate, systematic, and ethical manner. Within a defined ethical framework, the use of LLMs by scholars is commendable, as our valid inputs contribute to the expansion of LLM training data and promote technological advancement.

I encourage scholars to enhance their own innovative abilities, continuously reflect on current research paradigms and language models, rather than being constrained by rigid standards. While machines follow strict criteria, human beings should not be evaluated in the same way.

The present study has several limitations that warrant improvement. The experimental results were not subjected to reliability or validity testing, and therefore may involve a relatively high degree of subjective judgment. A more complex and comprehensive experimental design is needed in future research.

Moreover, the multilingual generated texts in this study should be proofread by professional translators before being resubmitted to AIGC detection systems, for further assess their accuracy across language contexts.

The number of AIGC detection tools used in this study was limited; therefore, future research should incorporate additional detection systems to conduct more thorough cross-linguistic evaluations.

Since the QCA methodology employed in this study was guided by clear theoretical assumptions, future studies should expand the set of prompts given to LLMs, increase the volume of textual data, and aim to develop a theoretical framework for the application of LLMs that is not bound by existing theoretical presets.

Appendix A. Research Sample (Talents in Colleges and Universities).

Academic Entrepreneurs are innovative individuals whose cultivation is a critical topic in the realm of higher education research in China. The developmental strategy of Chinese higher education system necessitates the construction of an innovative education model. To address the cultivation of innovative talents, an innovative educational paradigm is required — a system based on proactive practice, incorporating socially-oriented problem-solving content, enhancing student learning systems, and maintaining a scientifically balanced curriculum structure. Within this framework for cultivating innovative talents at higher education institutions, AEs translate market and societal issues into research projects from a practical standpoint, presenting these projects to students with foundational principles. This approach enhances the practical relevance of academic instruction. The educational attributes of AEs contribute significantly to the cultivation of innovative talents in higher education institutions and offer fresh perspectives for enhancing pedagogy. Establishing specialized zones for innovative talent development, centered around key universities, is a significant initiative in building institutional interest in China.

In the context of Chinese "Double First-Class" universities, an innovative talent training model features a redefined training objective and a results-oriented evaluation system. The goals and systems for innovative talent training align with the growth pathways of AEs. Consequently, Realization of Scientific Research identified in the first path of AE training, can serve as a practical evaluation metric, replacing some title-based and idealized indicators. When the evaluation system acknowledges that true honors arise from contributions to practical and creative advancements in society, academics are less likely to pursue research and teaching driven solely by prestige. The Educational Attributes emphasized in the third path serve as a talent cultivation orientation, fostering an innovation system for students, encouraging them to transcend quantitative measures such as grades and credits and venture into the realms of innovation, practice, and exploration.

Higher education should transcend being merely rivers and planets, instead becoming a vast ocean and an expansive cosmic galaxy, embodying noble qualities and well-rounded personalities. It aids students in transitioning from a constrained test-taking existence to an expansive limitless life experience. University talents are a vital component of Chinese national talent pool. The quality of higher education in China is largely influenced by governmental directives, which can introduce inherent constraints, potentially impeding talent development. To establish extinctive educational institutions worldwide, we must overcome test-taking limitations, and the introduction of market mechanisms that can unleash the full potential of university development. The adoption of market-oriented mechanisms in universities responds to global trends, with higher education policies reflecting both policy-driven and neoliberal orientations. As hubs of talent training and scientific innovation, First-Class universities exert a growing influence on Chinese talent ecosystem, due to the symbiotic relationship between science, technology, innovation, and talent.

Appendix B. Prompt Templates

1) Please generate a 1200-word passage on the topic "The Application of Academic Entrepreneurship in Talent Development at Higher Education Institutions", serving as the argument section of a master's thesis. Before generating the content, please review relevant literature and ensure that the writing conforms to the academic standards of a master's thesis. Include proper citations where appropriate.
2) Very good. Next, please revise the content you generated based on the provided text, making it sound more like human-written prose.
3) Very good. Next, please integrate the following two similar passages into a single 1200-word coherent article, combining the content from Step 2. Ensure logical flow and structural integrity.

4) Very good. Next, please integrate the following two similar passages into a single 1200-word coherent article, combining the content from Step 3. Maintain clear logic and smooth transitions.
5) Very good. Next, please revise the content within the parentheses to make it sound more like human-written prose — specifically targeting the sections flagged as AIGC-generated by detection tools.

References

1. S. Guizani, T. Mazhar, T. Shahzad, W. Ahmad, A. Bibi, and H. Hamam, “A systematic literature review to implement large language model in higher education: issues and solutions,” *Discov. Educ.*, vol. 4, no. 1, pp. 1–25, 2025, doi: 10.1007/s44217-025-00424-7.
2. S. Porsdam Mann, A. A. Vazirani, M. Aboy, B. D. Earp, T. Minssen, I. G. Cohen, and J. Savulescu, “Guidelines for ethical use and acknowledgement of large language models in academic writing,” *Nat. Mach. Intell.*, pp. 1–3, 2024, doi: 10.1038/s42256-024-00922-7.
3. A. Molinari and E. Molinari, “The added value of academic writing instruction in the age of large language models: A critical analysis,” *IADIS Int. J. WWW/Internet*, vol. 22, no. 1, 2024, doi: 10.33965/ijwi_2024220104.
4. A. J. Alvero, J. Lee, A. Regla-Vargas, R. F. Kizilcec, T. Joachims, and A. L. Antonio, “Large language models, social demography, and hegemony: Comparing authorship in human and synthetic text,” *J. Big Data*, vol. 11, no. 1, p. 138, 2024, doi: 10.1186/s40537-024-00986-7.
5. X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, and X. Zhao, "Bias of AI-generated content: an examination of news produced by large language models," *Sci. Rep.*, vol. 14, no. 1, p. 5224, 2024, doi: 10.1038/s41598-024-55686-2.
6. Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, pp. 1–14, 2024, doi: 10.1057/s41599-024-03811-x.
7. A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100234, 2024, doi: 10.1016/j.caeai.2024.100234.
8. Y. Cheng, Y. Fan, X. Li, G. Chen, D. Gašević, and Z. Swiecki, "Asking generative artificial intelligence the right questions improves writing performance," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100374, 2025, doi: 10.1016/j.caeai.2025.100374.
9. W. Wiboolyasarin, K. Wiboolyasarin, K. Suwanwihok, N. Jinowat, and R. Muenjanchoey, "Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100228, 2024, doi: 10.1016/j.caeai.2024.100228.
10. A. Dillon, G. Chell, N. Al Ameri, N. Alsayed, Y. Salem, M. Turner, and K. Gallagher, "The use of large language model tools such as ChatGPT in academic writing in English medium education postgraduate programs: A grounded theory approach," *J. Educ. Online*, vol. 21, no. 2, p. n2, 2024, doi: 10.9743/jeo.2024.21.2.5
11. J. J. Yang and S. H. Hwang, "Transforming hematological research documentation with large language models: An approach to scientific writing and data analysis," *Blood Res.*, vol. 60, no. 1, pp. 1–11, 2025, doi: 10.1007/s44313-025-00062-w.
12. J. Li, H. Zong, E. Wu, R. Wu, Z. Peng, J. Zhao, and B. Shen, "Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students — the educational application of ChatGPT," *BMC Med. Educ.*, vol. 24, no. 1, p. 736, 2024, doi: 10.1186/s12909-024-05738-y.
13. Y. Cui, "What influences college students using AI for academic writing? — A quantitative analysis based on HISAM and TRI theory," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100391, 2025, doi: 10.1016/j.caeai.2025.100391.
14. Q. Ma, P. Crosthwaite, D. Sun, and D. Zou, "Exploring ChatGPT literacy in language education: A global perspective and comprehensive approach," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100278, 2024, doi: 10.1016/j.caeai.2024.100278.
15. A. Radtke and N. Rummel, "Generative AI in academic writing: Does information on authorship impact learners’ revision behavior?," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100350, 2025, doi: 10.1016/j.caeai.2024.100350.
16. D. E. O'Leary, "Using large language models to write theses and dissertations," *Intell. Syst. Account. Finance Manag.*, vol. 30, no. 4, pp. 228–234, 2023, doi: 10.1002/isaf.1547.
17. J. van Niekerk, P. M. Delport, and I. Sutherland, "Addressing the use of generative AI in academic writing," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100342, 2025, doi: 10.1016/j.caeai.2024.100342.
18. K. Ibrahim, "Using AI-based detectors to control AI-assisted plagiarism in ESL writing: ‘The Terminator Versus the Machines’," *Lang. Test. Asia*, vol. 13, no. 1, p. 46, 2023, doi: 10.1186/s40468-023-00260-2.
19. Y. Li et al., "Can large language models write reflectively," *Comput. Educ. Artif. Intell.*, vol. 4, p. 100140, 2023, doi: 10.1016/j.caeai.2023.100140.
20. T. Yu, "Language models as dissertation assistants: Academic misconduct or efficiency upgrades?," *Int. J. New Dev. Educ.*, vol. 6, no. 11, 2024, doi: 10.25236/IJNDE.2024.061116.

21. M. Khalifa and M. Albadawy, "Using artificial intelligence in academic writing and research: An essential productivity tool," *Comput. Methods Programs Biomed. Update*, p. 100145, 2024, doi: 10.1016/j.cmpbup.2024.100145.
22. U. M. Qureshi, E. K. M. Tong, Z. V. Chen, C. Deng, and G. Nandakumar, "A robust and secure framework for an AI-assisted academic writing platform," in *Int. Symp. Emerg. Technol. Educ.*, Singapore: Springer Nature Singapore, Nov. 2024, pp. 235-249. ISBN: 9789819644063.
23. A. Cheng, A. Calhoun, and G. Reedy, "Artificial intelligence-assisted academic writing: recommendations for ethical use," *Adv. Simul.*, vol. 10, no. 1, p. 22, 2025, doi: 10.1186/s41077-025-00350-6.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of SOAP and/or the editor(s). SOAP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.