الجمهورية الجزائرية الديمقراطية الشعبية
جامعة الجيلالي ليابس بسيدي بلعباس
وزارة التعليم العالي والبحث العلمي

كلية الآداب اللغات و الفنون

إذن بطبع مذكرة ماستر

قسم: لغة انجليزية

تخصص Linguistics - لسانيات

أنا الممضي أسفله الأستاذ(ة): ..... دلوسع نبية .....

المشرف على مذكّرة ماستر بعنوان:

..... Algorithms as Language Managers: The Linguistic Penalty and the Rise of Anti-Bot language in Algeria EFL Academic Writing.

للطالب(ة): ..... صنعوني رياض سيدي .....

.....

أوافق على طبع المذكّرة وتقديمها للمناقشة أو التقييم (الايداع) بعد أن استوفت الشّروط الضّروريّة لذلك.

ختم وتوقيع الأستاذ المشرف

.....

Dr. DARGBA Nebia

ختم مكتبة الكلية

(بعد إيداع نسخة من المذكرة بصيغة pdf)

تودع هذه الوثيقة بعد ملئها وختمها على مستوى القسم وهي شرط لإيداع المذكرة أو برمجة المناقشة

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Djillali Liabes University - Sidi Bel Abbes
Faculty of Letters, Languages and Arts
Department of English

Algorithms as Language Managers: The Linguistic Penalty and the Rise of Anti-Bot Language, in Algerian EFL Academic Writing

Dissertation Submitted to the Department of English in Partial Fulfilment of The Requirement for the Degree of Master in Linguistics and Language Dynamics

Written by: Mansouri Rayane Moussa

Supervised by :Dr. Darseba Nebia

Board for examiners

President: Pr. Mostari Hind Amel Djillali Liabes University

Examiner: Dr. Boualem Talha Djillali Liabes University

Supervisor: Dr. Derseba Nebia Djillali Liabes University

Academic Year: 2025- 2026

Dedications

In the name of Allah, the most glorious, the most merciful; all praise and gratitude to Allah SWT for every blessing and mercy; His grace has guided me through this journey and I'm nothing but a grateful servant .

To the one who is lost to my sight, yet forever present in my heart ... To the one who gave of her very soul so that I might live. To those eyes closed by death before they could witness the harvest of what they had sown. To my beloved mother, whose body is now veiled by the desolate earth,praying that my devotion and loyalty reach you.

To the free spirit,I only know the most... To that who was my solace in the loneliness of the road, and a herald of hope in the deepest glooms of despair. To the one who shared my joys and sorrows over the span of eight long years.. To the one who believed in me when the days turned their back on me, and tightened his grip on my hand when my own faltered... to, *The Unshackled*.

To the unwavering pillar of support... To the shoulder burdened by the days so that my own burdens might be lightened. My dear father, who spent the bloom of his youth and borrowed from the light of his own eyes so that I might clearly see my path, standing forever as my shield and fortress.

To the beautiful solace bestowed by fate... To them who inherited our mother's tenderness, being sisters by blood and a mother in their boundless care. To the comfort and hurdles that we shared , my dear sisters, the joy of my heart,... *Asma and Imane*.

To those united with me by the shared pulse ... your kindness knows no distance or boundaries , thank you , your kindness shall always be etched in my heart.

Acknowledgments

Reaching the end of this research would not have been possible without the help and direction of several people .

I owe my deepest and inmost thanks to my dear supervisor, Dr.Dersba Nebia. Thank you for steering me in the right direction, for your constructive criticism, and for encouraging me at every stage of the writing process.

I would also like to extend my appreciation to the Jury members Dr Mostari Amel Hind and Dr. Boualem Talha for their time, their close reading of the text, and their thoughtful assessment of this project.

Lastly, a special word of thanks to the participants of this study. Thank you for your honesty and for trusting me with your responses. Your contribution is the true foundation of this research.

ABSTRACT

Universities across the world are deploying Artificial Intelligence (AI) detection tools to manage academic integrity. In Algeria, where English is taught as a foreign language (EFL), this policy has created an unexamined problem: the same writing features that EFL instruction builds into students, formulaic phrasing, standard transitions, and structured argument, are the features that detection software flags as machine-generated. This study investigates that contradiction, examining what the anti-bot language register looks like in student writing, and how Algerian higher education students and educators respond to detection pressures. The research employed a mixed-methods design, collecting data through a 21-item Likert questionnaire completed by 23 EFL students (Cronbach's $\alpha = .772$), semi-structured interviews with a group of four students and two educators, non-participant observations at three academic sites, and writing sample analysis using aidetector.com. Two groups of student writers were compared: one that produced natural, unmonitored academic writing, and one that wrote knowing their text would be scanned. Results indicate that naturally written EFL essays were classified as AI-generated , despite being fully human-authored. Conversely, texts written under detection pressure, full of deliberate errors and structural disruptions, students reported consciously changing their writing style when scanned, and that writing for a detector feels fundamentally different from writing for a reader. Educators confirmed these patterns from the classroom perspective. Ultimately, this study demonstrates that better EFL academic writing is more likely to be flagged as AI-generated than deliberately degraded writing. This is not a minor technical error. It represents a structural consequence of how detection tools work, disproportionately panelizing the students who have worked hardest to master academic English.

Keywords: AI detection, Academic Writing, Academic Integrity, Higher Education Policies , Language policy , Anti-bot Language.

Table of Contents

CHAPTER ONE: LITERATURE REVIEW

CHAPTER TWO: METHODOLOGY

CHAPTER THREE: DATA ANALYSIS AND INTERPRETATION

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

ANOVA: One-Way Analysis of Variance

EFL: English as a Foreign Language

ESL: English as a Second Language

L2: Second Language

LLM: Large Language Model

M: Mean (Statistical Average)

N: Sample Size / Number of Participants

OE1: Open-Ended Item 1

OE2: Open-Ended Item 2

Q01–Q21: Questionnaire Items 01–21

SLA: Second Language Acquisition

GENERAL INTRODUCTION

Higher education institutions around the world have responded to the rise of generative AI, tools like ChatGPT and Gemini, by deploying AI detection software to scan student submissions. The logic is simple: if students can use machines to write, universities should use machines to detect. What has not been examined carefully enough is whether detection tools actually do what they are supposed to do, and who gets hurt when they do not.

AI detection works by measuring two statistical properties of text. Perplexity measures how predictable the vocabulary is to a language model. Burstiness measures how much variety in sentence complexity exists across a piece of writing. Low scores on both measures are associated with machine output. High scores suggest a human hand. This sounds reasonable in principle. In practice, it targets a specific group of writers with particular force.

EFL students, students whose first language is not English, are trained to write in ways that look statistically similar to AI-generated text. They learn to use standard transitional phrases: "First and foremost," "In conclusion," "On the other hand." They write with consistent grammar and reliable structure. They follow the genre conventions they have been taught. The result is writing that is predictable, which is exactly what detection tools flag.

The student who has worked hardest to master academic English is, under this system, the most likely to be accused of not having written her own work. This is what scholars have begun to call the linguistic penalty. It does not fall because students have done something wrong. It falls because the tool cannot tell the difference between a trained EFL writer and a machine.

The academic conversation about AI detection has developed quickly, but it has left important questions unanswered.

Researchers have tested the accuracy of detection tools and found serious problems. Weber-Wulff et al. (2023) examined fourteen commercially available detection platforms and concluded that none performed reliably enough for use as sole evidence in misconduct decisions. Liang et al. (2023) demonstrated that detection tools produce false-positive rates of approximately 61% for non-native English writing, compared to near-zero rates for native-speaker texts. These are significant findings. They tell us the system is broken.

What they do not tell us is what happens to students and educators when they have to live inside that broken system.

No existing study has characterised anti-bot language as a distinct linguistic register with identifiable markers. Researchers have noted that students feel pressured to modify their writing, but no one has asked what, exactly, those modifications look like at the sentence level, which specific strategies students use, or whether those strategies actually work. That question has a textual answer, and this study provides it.

Educators are also largely absent from this literature. Most studies focus on the technical performance of detection tools or on student survey responses. The perspectives of the teachers who encounter detection results daily, who must decide whether to trust a score or override it, and who watch student writing change year by year, have not been gathered in any systematic way.

Finally, no study has examined the specific situation of EFL students in Algerian higher education. Algeria's academic context has its own shape. Students arrive at university with French and Arabic as their primary academic languages and encounter English instruction that is formal, prescriptive, and heavily convention-based. This produces a particular kind of interlanguage, one that is especially vulnerable to algorithmic misclassification. That specific vulnerability deserves a study of its own.

Alongside the controversy between de jure and de facto policies, has produced significant linguistic consequences. The most notable of these are the emergence of a new writing register called anti-bot language or survival language, and the linguistic penalty experienced by students and educators.

This dissertation addresses all three of these gaps. It investigates anti-bot language at the textual level. It documents student behavioral adaptation through survey and interview data. It brings educators into the conversation. And it does all of this within the specific context of Algerian EFL higher education, Consequently, this study seeks to answer the following exploratory research questions:

RQ1: What are the main linguistic and structural characteristics of the new anti-AI writing style, and do these adaptations align with traditional standards of academic writing quality?

RQ2: How do students and educators perceive the ethical implications of the linguistic penalty in AI detection?

The dissertation is organised around three chapters. Chapter One reviews the literature that shapes this study. It defines the technologies at the center of the argument: generative AI, large language models, and AI detection tools. It examines how detection tools work, what they measure, and why those measures create problems for non-native writers. It explains Selinker's (1972) theory of interlanguage and connects it to the specific statistical features that detection tools flag. It applies Spolsky's (2004) Language Policy and Planning framework, with its three components of language beliefs, language management, and language practices, to the question of how algorithmic surveillance reshapes institutional writing culture. And it closes by identifying the three specific gaps this research addresses.

Chapter Two describes the method. The study uses a sequential mixed-methods design. It began with three non-participant observations at academic events in Oran, Sidi Bel Abbes, and an online session, which provided contextual understanding and shaped the focus of later instruments. The main quantitative instrument was a 21-item Likert questionnaire completed by 23 EFL students, measuring awareness of detection, self-reported behavioral change, and attitudes toward academic voice and identity. The qualitative instrument was a series of semi-structured interviews conducted with a group of four students and individually with two educators. The fourth instrument was a writing sample experiment comparing texts produced under two conditions: naturalistic writing, where students did not know detection would follow, and AI-aware writing, where students knew they would be scanned and could use any resources they chose.

Chapter Three presents and discusses the findings. It moves through the questionnaire section by section, connects the statistical patterns to the theoretical frameworks from Chapter One, and then layers in the interview testimony, the observational record, and the writing sample analysis. The chapter does not separate findings from discussion. Each data source is analyzed and interpreted in context, so that the reader sees not just what the numbers show but what they mean.

CHAPTER ONE: LITERATURE REVIEW

1.1 Introduction

Higher education is undergoing a significant transformation due to the growing integration of artificial intelligence (AI), particularly large language models (LLMs) and generative AI (GenAI). These technologies have altered student workflows, challenged academic writing practices, and forced institutions to reconsider their language-use policies. A new institutional challenge has emerged, one that disrupts traditional approaches to academic integrity and creates new patterns of linguistic behavior among students.

This chapter is organized into eight sections. Section 1.2 defines the key AI technologies involved: GenAI, LLMs, and AI detection tools. Section 1.3 examines academic writing in higher education in relation to AI use. Section 1.4 explores the algorithmic turn and the gap between formal and applied institutional policy. Section 1.5 explains the statistical metrics AI detection relies on, along with their limitations. Section 1.6 examines how second language acquisition research reveals a structural vulnerability for EFL writers. Section 1.7 applies Spolsky's Language Policy and Planning framework to algorithmic surveillance. Section 1.8 identifies the research gap that this study aims to address.

1.2 Artificial Intelligence in Higher Education

Within the current literature, Artificial Intelligence (AI) in higher education is rarely treated as a neutral technological tool. When "essays are graded by AI," for instance, AI functions as an automated manager of student assessment and monitoring (Foltz et al., 2013). Similarly, "machine-to-student" interactions facilitated by commercial AI applications have become increasingly common as a result of the rapid shift toward digital learning, a trend that accelerated notably during the COVID-19 pandemic (Rof et al., 2022). Describing AI only in terms of its technological capabilities, however, ignores its broader social implications. As Bearman et al. (2022,p. 370) argue, "AI is not just a matter for technological innovation but also represents a fundamental change in the relationship between higher education and broader socioeconomic interests." In short, AI has become a powerful institutional force that shapes how universities design their policies, not only a reflection of technological progress. The two most relevant AI technologies in the current academic context are Generative AI (GenAI) and Large Language Models (LLMs).

1.2.1 Defining Generative Artificial Intelligence (GenAI)

According to Shah (2023) and Zhuhadar and Lytras (2023), Generative Artificial Intelligence (GenAI) represents a major development in AI, focused specifically on the creation of "new and original content," including text, images, music, and video. Chan and Hu (2023) further define GenAI as a group of machine learning algorithms designed to generate data samples that mimic existing datasets. These systems draw on machine learning, deep learning, and neural network architectures to identify patterns in large datasets and simulate human creativity. Well-known examples include OpenAI's DALL-E for image generation and ChatGPT for text-based interactions.

1.2.2 Defining Large Language Models (LLMs)

Language is a complex system of human communication governed by grammatical rules. While it has traditionally been considered a uniquely human capability, the field of language modeling has advanced significantly in recent decades, making it possible for machines to produce text that closely resembles human writing. Traditional pre-trained language models (PLMs) demonstrated strong performance in natural language processing tasks, but researchers discovered that scaling model size to tens or hundreds of billions of parameters led to a qualitative leap in capability, giving rise to what are now known as Large Language Models (LLMs).

Yigci et al. (2024,p. 1) define LLMs as "artificial intelligence (AI) platforms capable of analyzing and mimicking natural language processing," while Aydin et al. (2024,p. 1) describe them as "sophisticated algorithms that analyze and generate vast amounts of textual data, mimicking human communication." What distinguishes these models is their emergent capacity for reasoning, planning, and in-context learning, abilities that appear only at sufficient scale and were absent in smaller models. These systems can understand complex linguistic patterns and generate coherent, contextually appropriate responses (Raiaan et al., 2024). In the context of higher education, Harrington et al. (2025) characterize LLMs as "accessible, personalized, innovative, and interactive tools that create revolutionary learning experiences, often leading to enhanced cognitive and skill development."

Despite this potential, LLMs present serious challenges. They are prone to hallucinations, producing responses that sound convincing but are factually incorrect. They also carry the capacity to "inherit and amplify societal biases" from their training data, which is why researchers emphasize the need for "alignment tuning" to keep these systems helpful, honest, and safe (Zhao et al., 2023; Naveed et al., 2023).

1.2.3 Defining AI Detection Tools and Their Integration in Higher Education

As LLMs become more capable, educators and institutions have sought ways to determine whether student submissions were written by a human or generated by AI. This demand has led to the widespread adoption of AI detection tools.

AI detection tools are designed to identify the origin of written text, specifically whether it was produced by a human or an AI system. Unlike conversational chatbots, these tools function as text classifiers: they take written content as input and return a probability score indicating the likelihood that the text is AI-generated (Weber-Wulff et al., 2023). In a systematic study of fourteen commercially available detection tools, Weber-Wulff et al. (2023) found that none of the tested tools performed with consistent accuracy across different text types, concluding that current AI detection technology remains too unreliable for use as the sole basis of academic misconduct decisions.

From a linguistic perspective, AI detection tools look for patterns associated with machine-generated language. On the other hand, human-authored text, with its unique stylistic habits and occasional irregular phrasing, shows subtle but identifiable differences from AI output. Large language models are trained to predict the most probable next word or phrase, which often results in text that is highly consistent and predictable in its word choice, sentence structure, and overall rhythm. AI detection tools are built to identify these statistical regularities, focusing on two key measures: perplexity and burstiness.

The widespread adoption of these tools has significantly changed how academic integrity is managed. While the aim is to protect academic standards, this reliance on technology introduces new complications. A 2025 ABC News article by Bergin reported that a major Australian university had used AI detection technology to accuse approximately 6,000 students of academic misconduct, many of whom had not used AI at all. This real-world case demonstrates that the gap between what these tools claim to detect and what they actually identify in practice can have serious consequences for students.

1.3 Academic Writing in Higher Education in the Age of AI

Academic essay writing demands careful research, systematic reasoning, and clear expression. In higher education (HE), students are expected to develop core writing competencies: constructing coherent arguments, writing structured abstracts, producing literature reviews, and following strict citation conventions. Mastery of specialized vocabulary and

adherence to disciplinary style standards are essential for effective academic communication (Malik et al., 2023).

A central feature of this writing culture is the use of disciplinary conventions and genre expectations. As Hyland (2009, p. 1) observes, academic discourse is "a social process" in which writers must learn "the forms and purposes that structure the disciplines they are seeking to enter." This process is particularly demanding for students writing in a second or foreign language, who must master not only the content of their discipline but also the highly specific linguistic and rhetorical conventions that academic English demands.

The growing use of AI tools by students has created a challenge for these standards. Recent evidence suggests that nearly 47% of students are already using LLMs in their coursework, with 39% using them to answer exam questions and 7% submitting entire AI-written assignments (Beale, 2025). This creates what institutions describe as an academic integrity crisis, prompting universities to regulate AI use more strictly.

A key limitation of the current institutional response is that AI is primarily detectable only through written text. Because academic essays and written reports are the main tangible products of student research, institutional scrutiny is almost entirely focused on writing. As a result, AI detection tools are fundamentally limited to analyzing written content.

The dominant institutional response to this challenge has been techno-solutionism: treating the complex, socially situated act of writing as a technical problem that software can fix. The clearest example of this approach is the heavy reliance on AI and plagiarism detection tools, such as Turnitin, to enforce academic integrity. This creates what can be described as a detect-and-punish framework, in which student writing is treated as a technical output to be validated by a machine rather than as an act of intellectual expression to be assessed by a teacher.

1.4 The Algorithmic Turn, De Jure vs. De Facto Policies, and Their Linguistic Outcomes

Over-reliance on AI detection tools shifts academic governance toward algorithmic surveillance and compliance-based thinking, at the cost of trust, human-centered assessment, and institutional accountability. This shift is described as the "algorithmic turn" in education. As Selwyn (2016, p. 106) argues, "digital data do not offer a neat-technical fix to education dilemmas no matter how compelling the output might be." He further warns that the turn to data-driven systems in education can reinforce existing inequalities rather than address them, a warning that proves especially relevant in the context of AI detection and EFL writers.

This shift creates a gap between de jure and de facto policies in higher education institutions. De jure refers to what is formally written in official policy and law; de facto refers to what actually happens in practice (Metych, 2025). In the context of AI detection, de jure policies promote fairness, accuracy, and academic integrity. De facto enforcement, however, relies on proprietary tools with significant error margins of up to 12% (Paustian and Slinge, 2024).

These error margins produce false-positive accusations, where students who wrote their own work are incorrectly identified as AI users. This undermines the very integrity that the formal rules are designed to protect. The impact is not evenly distributed. Second-language (L2) writers are disproportionately affected because their natural interlanguage features, including predictable vocabulary and simplified sentence structure, often resemble the statistical patterns of AI-generated text. The consequences extend beyond individual cases, producing a chilling effect across the entire student population: students modify their writing not because they have done anything wrong, but because they fear being wrongly accused.

This dynamic reshapes the university's Language Policy and Planning (LPP) environment. The de facto language manager is no longer the educator assessing communicative competence, but an algorithm that treats statistical unpredictability as the primary marker of human authorship. In response, students develop what this study terms anti-bot language: writing strategies deliberately designed to satisfy the machine's criteria, even at the cost of academic quality.

1.5 Statistical Language Modeling and the Metrics of Humanity

Natural Language Processing (NLP) is the branch of AI concerned with the interaction between computers and human language. It allows machines to understand, analyze, and generate text. As Gustilo et al. (2024) explain, LLMs do not process language as a vehicle for human thought or intention, but as a series of computational and statistical operations. These models are trained on large datasets to identify patterns between words, learning the structure of language through statistical analysis. In practice, LLMs generate text by making a series of predictions about which word, or "token," is most likely to follow a given prompt. Language, for these systems, is probability, not meaning.

AI detection tools rely on two principal metrics to evaluate text: perplexity and burstiness.

1.5.1 Perplexity

Perplexity is a measure of how unexpected or surprising a word is to a particular language model. When a model finds a piece of text unpredictable based on its training data, the perplexity is high; when the text closely matches patterns the model has seen frequently, the perplexity is low, as Jurafsky and Martin (2021) tackled in their research.

A simple example illustrates this. In the sentence "For lunch today, I ate a Japanese food, sushi," the word "sushi" has low perplexity because it is a common, expected completion. In the sentence "For lunch today, I ate an Algerian food, couscous," the word "couscous" has higher perplexity because such a phrase appears less frequently in most language model training data.

AI-generated text tends to have consistently low perplexity, because language models are designed to select the most probable next word at each step. Human writing, by contrast, includes unexpected word choices, shifts in register, and stylistic variations that raise perplexity.

1.5.2 Burstiness

On the other hand, burstiness refers to the variation in perplexity across a document. It measures the rhythm of "surprising" content throughout a text. Human writing typically shows high burstiness: it contains a mix of predictable, conversational sentences and sudden moments of creative, complex, or unusual expression. AI-generated text tends to have low burstiness because language models consistently produce uniformly probable output with little variation in style or rhythm. As Perkins et al. (2023) observe in their "Game of Tones" study, AI-generated text often exhibits "a uniform, unfocused flow" that statistical classifiers recognize as machine-generated, lacking the "personality" found in authentic student writing.

1.5.3 Limitations of Detection Metrics

Despite their usefulness, perplexity and burstiness are not always reliable indicators of human authorship, particularly in high-stakes academic settings. Three key limitations are worth noting.

First, the training data problem: LLMs are optimized to minimize perplexity on the texts they were trained on. Famous human-written documents, such as the Declaration of Independence or widely accessed Wikipedia articles, are sometimes falsely flagged as AI-generated because the model finds them highly predictable due to repeated exposure.

Second, the model-dependency problem: perplexity is not a universal measure. What GPT-4 considers "surprising" may be entirely different from what Claude or Gemini finds

unexpected. As new AI models are released, detection tools based on perplexity metrics can quickly become outdated.

Third, and most significant for this study, is the black box problem. Most commercial AI detection tools do not disclose the specific algorithms, training data, or decision criteria they use. As Weber-Wulff et al. (2023, p. 1) conclude in their systematic evaluation, "the tools are simply not reliable enough for use in high-stakes environments where incorrect results could seriously harm students." Researchers and educators therefore cannot scrutinize, challenge, or fully understand the basis of individual flagging decisions, a situation that places the most vulnerable student populations at particular risk.

1.6 Second Language Acquisition and the Linguistic Penalty

The tendency of AI detection tools to misidentify the writing of second-language (L2) students as AI-generated is not accidental. It is rooted in the cognitive and pedagogical characteristics of second language acquisition. To understand this structural bias, it is necessary to examine Larry Selinker's (1972) theory of Interlanguage (IL).

Selinker proposed that adult language learners do not move directly from their first language (L1) to the target language (TL). Instead, they construct what he called an interlanguage: a distinct, rule-governed linguistic system on a continuum between the two languages. As Selinker (1972, p. 214) described it, interlanguage is "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL norm." This system is shaped by cognitive strategies, most notably simplification, the tendency to avoid complex or ambiguous structures in favor of safer, more predictable forms.

Within academic writing contexts, this tendency toward simplification is heavily reinforced by standard EFL pedagogy. L2 instruction often emphasizes the minimization of creative linguistic risk to maintain a structural form. Consequently, learners are systematically trained to rely on formulaic structures or lexical chunks, such as standardized transitional phrases like "First and foremost," "On the other hand," or "In conclusion." While these formulaic expressions provide a secure and coherent framework for academic arguments, they inherently produce texts characterized by a highly predictable vocabulary and syntactic structures.

This predicament is not limited to the classroom. As Canagarajah (2002, p. 13) argues, non-native scholars and students are already positioned at a structural disadvantage within English-medium academic discourse, facing what he describes as "unequal power relations" that shape the kinds of language considered legitimate in academic settings. When AI detection tools layer algorithmic judgment on top of these existing inequalities, the disadvantage is compounded: EFL writers are penalized not only by native-speaker norms enforced by human assessors but now also by statistical norms enforced by machines.

This predictability is exactly what AI detection tools flag. Large language models generate text by calculating statistical probabilities; their output is inherently uniform, with low perplexity and low burstiness. Because L2 writers naturally produce similar patterns in their interlanguage, favoring conventional phrasing over stylistic variation, their writing can appear statistically indistinguishable from AI-generated output to a computational classifier.

Empirical research confirms this overlap. Liang et al. (2023), in a study examining 91 TOEFL essays written by non-native English speakers, found that AI detection tools produced false-positive rates of approximately 61% for these texts, compared to near-zero rates for essays written by native English speakers. Their study concluded that "GPT detectors consistently misclassify non-native English writing samples as AI-generated," demonstrating that the bias is systematic rather than coincidental (Liang et al., 2023, p. 1).

Similarly, Flowerdew (2001, p. 121) documented that non-native English-speaking scholars are already subject to "an additional and unfair burden" compared to native speakers in academic writing contexts, a burden that now takes a new and more severe form through algorithmic assessment. The student who has worked hardest to learn and apply academic English conventions is, paradoxically, the student most likely to be flagged by a tool that interprets those very conventions as evidence of machine authorship.

The outcome is a profound institutional irony. The specific writing strategies that L2 students learn in order to succeed academically are the same features that cause detection tools to classify their work as AI-generated. The deployment of these tools effectively penalizes non-native English speakers for having mastered the academic conventions they were explicitly taught. This is what the literature has begun to describe as the linguistic penalty (Liang et al., 2023).

1.7 Spolsky's Language Policy and Planning Framework and the Algorithmic Turn.

To understand how AI detection tools change the linguistic environment of higher education, it is useful to examine the institution through the framework proposed by Bernard Spolsky in his 2004 work Language Policy. Spolsky argues that language policy in any community is not simply a set of formal rules. Instead, it consists of three interconnected components: Language Beliefs, Language Management, and Language Practices. As Spolsky (2004, p. 5) explains, "the language policy of a community is to be found first in its language practices," meaning that what actually happens in everyday communication often reveals more about a community's real values than any official document.

This framework was originally designed for human-to-human linguistic environments. However, the integration of algorithmic tools into academic assessment requires a significant revision of Spolsky's model. Proprietary detection software has now become an active participant in the relationship between the institution, the educator, and the student. Furthermore, where language was once produced only by humans, LLMs can now generate it too.

1.7.1 Language Beliefs: The Ideology of Algorithmic Humanity

In Spolsky's model, language beliefs, also described as language ideologies, refer to the values that a community assigns to particular linguistic forms. In a traditional EFL context, the dominant institutional belief emphasizes communicative competence: writing is valued for its clarity, coherence, and ability to convey ideas effectively.

The adoption of AI detection tools, however, introduces a competing and largely hidden ideology. Because these tools treat high perplexity and high burstiness as the markers of genuine human authorship, the institution's de facto language belief shifts. Statistical unpredictability, a feature more typical of native-speaker prose, becomes the new standard for "authentic" writing. The structured, formulaic interlanguage of L2 writers is implicitly categorized as artificial. This reflects what Milroy and Milroy (1985) call Standard Language Ideology: the belief that one particular variety of language, typically associated with educated, native-speaking norms, represents the correct or legitimate form. Applied algorithmically, this ideology equates linguistic safety with inauthenticity and punishes writers for the features they were taught to produce.

1.7.2 Language Management: The Algorithm as De Facto Manager

Language management refers to the deliberate efforts of specific actors, traditionally teachers, curriculum designers, and parents, to shape how learners use language. In the pre-AI era, the classroom teacher served as the primary language manager, offering feedback that was context-sensitive, pedagogically informed, and responsive to individual development. The algorithmic turn disrupts this role. When universities adopt tools like Turnitin, they effectively transfer language management from human educators to non-human, proprietary algorithms. These tools function as context-free managers. Unlike a teacher who understands the specific educational background of an individual student, for example an Algerian EFL student, the algorithmic manager evaluates text purely on statistical variance, without regard for context, development, or intent. This creates a sharp conflict between de jure management, what official policy says about promoting equity and fairness, and de facto management, an algorithm that systematically produces false positives for L2 writers.

1.7.3 Language Practices: The Emergence of Anti-Bot Language

The final component of Spolsky's framework is language practices: the actual linguistic habits and choices that members of a community make in their everyday communication. According to Spolsky (2004), language practices develop in response to both the environment and the management that learners receive.

Because students are highly sensitive to the penalties that the new algorithmic managers can impose, their writing practices undergo a significant and largely negative adaptation. To satisfy the AI detector's preference for statistically unpredictable text, students abandon the academic clarity they were taught to prioritize. Instead, they engage in what this

study terms algorithmic survival: adjusting their writing not to communicate better, but to evade detection. This results in a new pragmatic register: anti-bot language. Students may deliberately inject awkward phrasing, break up sentence structures, or introduce artificial complexity into their writing, specifically to achieve the appearance of human unpredictability. In this situation, the student's language practice is no longer shaped by the goal of communicating effectively with a human reader. It is shaped instead by the need to satisfy the statistical expectations of a machine.

1.8 Identifying the Research Gap

The literature reviewed in this chapter establishes a solid foundation for understanding how AI detection tools operate, why they produce false positives for L2 writers, and how existing frameworks in language policy and second language acquisition can help explain the resulting institutional dynamics. However, a careful reading of this body of scholarship reveals three significant areas that remain largely unexplored, gaps that this study is specifically designed to address.

1.8.1 Anti-Bot Language as an Unexplored Linguistic Register

While the concept of students modifying their writing in response to AI detection has appeared in general discussions of academic integrity (Perkins et al., 2023; Selwyn, 2016), no existing study has systematically investigated anti-bot language as a distinct, emergent pragmatic register with identifiable linguistic markers. Existing scholarship has acknowledged that students feel pressure to change how they write, but it has not yet examined what exactly those changes look like at the level of sentence structure, vocabulary choice, syntactic variation, or stylistic register. The question of how students write differently, and what makes anti-bot writing recognizable as a specific linguistic practice, remains open. This study addresses that gap by collecting and analyzing authentic student writing samples under two contrasting conditions: naturalistic human writing and AI-aware writing produced with full knowledge that detection will follow.

1.8.2 The Linguistic Penalty: Student Behavior and Perception in a Specific Context

The linguistic penalty has been documented at the statistical level by Liang et al. (2023), who demonstrated high rates of false-positive detection for non-native English writers. However, the behavioral dimension of this penalty, specifically how students in a particular institutional context respond to the threat of false detection, has not been studied. It remains unclear how widespread behavioral adaptation is among EFL students, what specific strategies students employ, how aware they are of the mechanisms driving false-positive detection, and how their sense of academic identity and writing confidence is affected. These questions require a mixed-methods approach that combines quantitative survey data with qualitative interview and writing sample analysis. This study provides exactly that, grounding the investigation in the specific context of Algerian EFL higher education, where French-Arabic-English multilingualism and specific pedagogical traditions produce a distinctive interlanguage profile with particular implications for algorithmic misclassification.

1.8.3 The Perspectives of Academic Staff

A further and particularly significant gap concerns the perspectives of academic educators. The existing literature on AI detection in higher education has focused almost entirely on the technical functioning of the tools themselves and, to a lesser degree, on student experiences and perceptions. The professional responses of academic staff, including their awareness of detection tools' limitations for EFL writers, the way they integrate or override detection results in their assessment judgments, and the degree to which they have adjusted their pedagogical practices in response to AI, remain largely unexamined. As Bearman et al. (2022) note, AI represents "a fundamental change in the relationship between higher education and broader socioeconomic interests," yet the views of the educators who must mediate that change in the classroom have not been systematically gathered. This study addresses this gap through semi-structured interviews with both students and educators, making the educator perspective a core analytical strand rather than an afterthought.

Taken together, these three gaps point toward a single overarching research need: an empirically grounded, contextually situated, mixed-methods study that investigates anti-bot language as a linguistic register, maps the behavioral consequences of the linguistic penalty for EFL students, and brings academic staff voices into the analysis. This is the study that the chapters that follow are designed to deliver.

1.9 Conclusion

This chapter has reviewed the key literature relevant to understanding how AI detection tools are reshaping the academic writing environment in higher education. Beginning with definitions of GenAI, LLMs, and detection tools, it traced how institutional responses to AI, particularly the widespread adoption of automated detection, have created a significant tension between formal policy and actual practice.

Through the lenses of Selinker's interlanguage theory and Spolsky's Language Policy and Planning framework, it becomes clear that the current reliance on AI detection tools creates a structural disadvantage for EFL writers, penalizing the very linguistic features they have been taught to produce. It also gives rise to a new form of student writing behavior, anti-bot language, that prioritizes algorithmic evasion over genuine academic expression. The three research gaps identified in Section 1.8 define the specific contribution of this study: an investigation of anti-bot language as a register, of student behavioral responses to the linguistic penalty, and of academic staff perspectives on algorithmic surveillance in the EFL writing classroom. The following chapter outlines the methodological approach taken to address these gaps through empirical research in the Algerian higher education context.

CHAPTER TWO: METHODOLOGY

2.1 Introduction

This chapter describes the research methodology used to investigate how artificial intelligence (AI) and large language models (LLMs) affect academic writing and educational policies in higher education. The study is guided by the tension between institutional de jure and de facto policies. It uses a mixed-methods approach to examine this shift. By combining quantitative and qualitative data, the research investigates how AI detection tools change the standards of acceptable academic English, including the emergence of anti-bot language and its characteristics, as well as the attitudes of educators and students toward the linguistic penalty. The following sections detail participant selection, data collection tools, ethical considerations, and the study limitations and delimitations.

2.2 Research Design

The study uses a sequential explanatory and exploratory mixed-methods design. The goal is to address the research problem without relying on a single methodological approach. The design moves through two main phases.

In the first phase, the study begins with qualitative instruments, specifically non-participant and unstructured observations. This allows the research setting to be understood in its natural form before any data is quantified. This exploratory step recognizes that the context of the phenomenon must first be understood before any attempt at measurement can follow.

In the second phase, quantitative instruments are used to map the patterns of student and educator responses to AI detection tools in higher education. This is then followed by a second qualitative phase involving interviews and writing sample analysis. The research problem involves both measurable and interpretive dimensions. A purely quantitative approach would miss the lived experience of students. A purely qualitative approach would not support claims about broader patterns. The sequential design resolves this tension: quantitative data establishes context, and qualitative data explores it.

Philosophically, this study follows a pragmatist paradigm in the tradition of Creswell and Plano Clark (2018). This tradition prioritizes the research questions over any single methodological approach, and selects the design that best fits the complexity of the problem.

2.3 Quantitative Method

The quantitative part of this study operates within a post-positivist framework. It uses measurement instruments and sampling to pursue two linked objectives. The first is to establish

how common AI detection anxiety and AI tool use are among the target population. The second is to identify significant relationships between key variables, including language proficiency level, institutional policy exposure, and self-reported changes to academic writing behavior.

Data from the Likert-scale questionnaire will be analyzed using descriptive statistics, including measures of central tendency, standard deviation, and frequency distributions. The study will also use inferential procedures, starting with a pilot study and then the full data collection, including one-way analysis of variance (ANOVA), to determine whether observed differences in writing adaptation and policy awareness are statistically significant. Where relationships between continuous variables are expected, Pearson r coefficients will be used to assess the direction and strength of association. Even when statistical results are significant, they remain incomplete without the context that the qualitative phase provides.

2.4 Qualitative Method

The qualitative part of this study is grounded in an interpretivist position. This means that realities are understood as constructed rather than discovered, and that they depend on the perspectives of the people who experience them. The qualitative phase does not aim to test hypotheses or produce generalizations. Instead, it aims to understand how individual students and educators experience and respond to algorithmic governance in academic writing.

The method used is reflexive thematic analysis, following Braun and Clarke (2022). This approach was chosen because of its flexibility. It does not impose a fixed structure on the data, but allows themes to emerge from repeated and careful engagement with participant responses. Semi-structured interviews are the main data source for this phase. Each interview is treated as a space for meaning-making. The researcher own interpretive position is acknowledged and documented throughout the process through reflective memos.

2.5 Mixed Methods Integration

The quantitative and qualitative data in this study are integrated at the level of interpretation, not data collection. Findings from Phase One are used to shape the focus of Phase Two: patterns found in the quantitative data, including unexpected results, directly inform the interview questions and the selection of participants for the writing sample study.

In the final stage, findings from both phases are compared through convergence coding. This process distinguishes where both methods agree and where they produce tension. Points of disagreement are not treated as methodological problems; they are treated as analytically useful.

The overall conclusions are drawn only after this synthesis stage, which preserves the value of each phase while producing the strongest possible combined interpretation.

2.6 Data Collection Tools

Data collection in this study uses four instruments, each chosen for its fit with specific research objectives.

The first instrument is a non-participant observation protocol. It focuses on behavioral data collected in real institutional settings, specifically at AI-focused academic workshops where detection tools are discussed or used. The second instrument is a structured Likert-scale questionnaire, used in the quantitative phase to collect stable and comparable data from a large participant group. The third instrument is a semi-structured interview protocol, used in the qualitative phase to allow participants to speak in depth about their experiences with AI and AI detection in higher education. The fourth instrument is a written text corpus, used to analyze authentic student-produced writing under two contrasting conditions.

These four instruments were selected as part of a deliberate triangulation strategy. No single instrument is treated as sufficient on its own. Each one captures a different layer of the phenomenon: the questionnaire captures attitudes and distributions, the interviews capture personal experience, the observations provide institutional context and research direction, and the writing corpus captures the actual linguistic effects of algorithmic pressure. Together, they produce a more defensible account than any single instrument could provide.

2.6.1 Participant Selection

Participant recruitment follows a purposive sampling logic in the qualitative phase and a stratified convenience sampling logic in the quantitative phase. These two approaches reflect the different goals of each phase.

For the questionnaire phase, participants are undergraduate and postgraduate students enrolled in English-medium academic courses at the target institution.

For the interview phase, participants are selected from both students and educators. Student participants are drawn from questionnaire respondents who reported significant changes to their writing behavior in response to AI detection tool use. Educator participants include lecturers, writing instructors, and academic integrity officers. This targeted selection, rather than random sampling, ensures that all participants have direct and relevant experience with the study research questions.

A sample of eight to twelve interviewees across both groups is expected to be sufficient to reach thematic saturation, the point at which additional interviews no longer produce new conceptual insights. Writing sample participants form a separate sub-sample, recruited specifically for the textual analysis phase. Details of their selection and group assignment are provided in Section 2.7.6.

2.6.2 Questionnaire

The questionnaire is the primary data collection tool for the quantitative phase. It also provides supplementary qualitative data through two open-ended items. The questionnaire is titled "Algorithmic Surveillance, Academic Writing, and the De Facto Language Manager." It was designed to address the study main research questions across three areas: awareness of and attitudes toward AI detection (Section A), self-reported behavioral changes related to algorithmic surveillance and anti-bot language (Section B), and attitudes toward academic voice, linguistic identity, and institutional fairness (Section C).

The questionnaire contains twenty-three closed Likert-scale items and two open-ended prompts. Completion takes between ten and fourteen minutes. All twenty-three closed items use a five-point scale from "Strongly Disagree" (1) to "Strongly Agree" (5). Two items, Q08 and Q21, are reverse-scored to reduce acquiescence bias and check response consistency. Before analysis, these items are recoded using the formula (6 minus raw score) to ensure that all items point in the same direction when added into subscale scores. This recoding will be carried out during data cleaning in SPSS and reported clearly in the results chapter. Internal consistency for each of the three subscales will be assessed using Cronbach alpha, with a threshold of a >= .70 considered acceptable.

2.6.3 Semi-Structured Interviews

Semi-structured interviews were conducted with both student and educator participants. This format was chosen because it balances structure and openness. It allows consistent comparison across participants while also giving participants space to share experiences and ideas that were not anticipated in the interview guide.

The main purpose of the interviews was to explore participants views on AI in higher education, including its benefits, risks, and ethical dimensions, as well as their specific experiences with AI detection tools.

For students, interview questions focused on three areas: their direct experience using, avoiding, or adapting to AI tools in academic writing; their awareness of and emotional responses to AI detection, including any instances of being flagged or deliberately changing their writing; and their views on institutional fairness, academic integrity, and the legitimacy of algorithmic assessment.

For educators, questions addressed the rationale behind their institution adoption of

AI detection tools, their assessment of those tools accuracy and fairness implications, and their observations of changes in student writing behavior related to detection pressure.

Student interviews were conducted individually in most cases, with the option of a paired format available for students who preferred it. Educator interviews were conducted individually to encourage candid responses about institutional matters and professional judgment.

2.6.4 Non-Participant Observations

Non-participant observation was the first data collection tool used in this study. It was added to the design because questionnaires and interviews alone cannot fully capture the real-world context of the phenomenon. The purpose of the observational phase was exploratory: to understand how AI and academic integrity are experienced and discussed in real educational settings, and to sharpen the theoretical focus of the later stages of the study. Watching how institutional actors discuss and respond to AI in academic writing provided a layer of contextual understanding that informed the analysis of interview and survey data.

Three observations were conducted across different settings. The first was at an AI-focused academic workshop in Sidi Bel Abbes, Algeria. The second was at a similar event in Oran. The third was a virtual observation conducted through an online platform hosting a professional academic discussion session. The choice to observe three different sites, two in person in Algeria and one online, was deliberate. It was intended to capture the phenomenon across a range of institutional and social settings, rather than drawing conclusions from a single location.

In all three sessions, the researcher observed without intervening, asking questions, or directing the discussion. Data collection consisted mainly of detailed field notes taken during and immediately after each session. These notes focused on how participants talked about AI tools and detection practices in academic writing. After each session, the researcher also wrote brief reflective memos to record initial analytical impressions while they were still clear. This practice helped preserve the connection between observation and subsequent analysis.

2.6.5 Writing Sample Collection and Experimental Groups

In addition to the questionnaire, interviews, and observations, this study collects authentic student writing samples as a fourth data source. This part of the study goes beyond self-reported perceptions to examine the actual linguistic features of student writing under different conditions. Specifically, it investigates whether naturally written EFL texts are falsely flagged as AI-generated, and whether anti-bot language emerges as a real writing behavior when students know their work will be scanned. Two groups of participants wrote under contrasting conditions.

2.6.5.1 Group One: Naturalistic Writing Condition

Students in the first group were asked to write a complete academic text, a short argumentative or expository essay on a topic related to their field of study or any topic they considered academic. They were not allowed to use AI writing tools such as ChatGPT or Gemini. However, they could use the internet for research and could use AI-assisted grammar correction tools for minor edits. Critically, these students were not told that their texts would later be analyzed using AI detection tools. This was a deliberate methodological decision. Telling them in advance would have changed how they wrote, producing exactly the kind of adaptive writing behavior that the second group is designed to capture. It would also have altered the natural writing baseline that this group is meant to provide.

The purpose of this group is to test whether authentic human academic writing by L2 English students is susceptible to false-positive identification as AI-generated content. This group directly tests the study core theoretical claim about the linguistic penalty: that EFL student writing, because of its predictable vocabulary and formulaic structure, resembles the statistical patterns of machine-generated text closely enough to trigger algorithmic flags.

2.6.5.2 Group Two: AI-Aware Writing Condition

Students in the second group were asked to write an academic text using any resources they chose, including AI writing tools, LLMs, paraphrasing tools, or any combination. Unlike the first group, these students were explicitly told before writing that their texts would be examined using an AI detection tool. This transparency was a deliberate methodological choice. Knowing that detection would follow creates the exact kind of pressure this group is designed to study. It encourages students to engage in strategic writing modifications, such as introducing unpredictable syntax, irregular vocabulary, and artificial complexity, which together make up the anti-bot language register that this dissertation theorizes.

The purpose of this group is twofold. First, it documents whether anti-bot language actually emerges as a behavioral response when students know they are being detected. Second, it identifies the specific linguistic markers and stylistic patterns of this register, including syntactic disruptions, unusual word choices, and register inconsistencies, in order to produce an early empirical profile of anti-bot language as a sociolinguistic phenomenon. Together, the two

groups form a comparative dataset that addresses both the diagnostic question, that is whether human texts are falsely flagged, and the behavioral question, that is how students adapt their writing to avoid detection.

2.7 Analytical Framework

The data collected in this study comes from four different instruments: a questionnaire, interview transcripts, observation field notes, and student writing samples. Analyzing this range of data requires more than one analytical approach. The overall framework is built around three connected pillars: sociolinguistic theory, critical discourse analysis, and computational text analysis. These three pillars are not used separately; they are used together to produce a coherent account of the linguistic phenomenon under examination.

2.7.1 Questionnaire Data: Descriptive and Inferential Statistical Analysis

Questionnaire data are analyzed using IBM SPSS. Descriptive statistics, including frequencies, means, standard deviations, and percentage distributions, are computed for each item and subscale. This provides an overview of attitudes and behavioral patterns across the full sample. Inferential procedures, including independent samples t-tests, one-way ANOVA, and Pearson correlation analysis, are used to assess whether observed differences and associations between key variables are statistically significant. Cronbach alpha is computed for each subscale to verify internal consistency. This statistical approach is appropriate because the questionnaire data are numerical, and answering the research questions directed at this phase requires population-level claims that only statistical inference can produce.

2.7.2 Interview and Observational Data: Reflexive Thematic Analysis

Interview transcripts and observation field notes are analyzed using reflexive thematic analysis following Braun and Clarke (2022). This approach was chosen over alternatives such as grounded theory or framework analysis because of its flexibility. It does not impose a fixed coding structure on the data; instead, it allows themes to develop from repeated and careful reading of participant responses. Analysis moves through six steps: familiarization through repeated reading, generation of initial codes, construction of candidate themes, review and refinement of themes against the full dataset, final definition and naming of themes, and production of the written analysis. Throughout this process, Spolsky (2004) Language Policy and Planning framework, specifically the three components of Language Beliefs, Language Management, and Language Practices, serves as a guiding conceptual lens rather than a rigid

structure. This framework is particularly appropriate here because it was already adapted in the Literature Review to account for the role of algorithms in institutional language governance.

2.7.3 Writing Sample Data: Computational and Critical Discourse Analysis

The analysis of student writing samples requires a three-level framework that operates at the computational, linguistic, and discursive levels at the same time.

At the computational level, all texts from both groups are submitted to a publicly available and free AI detection platform, such as GPTZero, to generate AI probability scores and classification outcomes. These scores are treated as data points to be analyzed, not as definitive judgments. The main focus is on how reliably and fairly these tools classify authentic human writing. Detection scores for Group One are examined to identify false-positive rates. For Group Two, detection outcomes are examined alongside the linguistic features of the texts to determine whether adaptive writing strategies succeed in evading detection and at what linguistic cost.

At the linguistic level, all texts are analyzed using quantitative measures of lexical diversity, including the Type-Token Ratio and the Measure of Textual Lexical Diversity, syntactic complexity measures such as mean clause length and subordination index, and cohesive device frequency. These measures give concrete form to the theoretical concepts of perplexity and burstiness, allowing the study to test empirically whether Group One texts show the low-perplexity, low-burstiness patterns associated with both AI-generated output and L2 interlanguage writing. Analysis tools include the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) and AntConc.

At the discursive level, Critical Discourse Analysis (CDA) following Fairclough (1992) is applied to selected texts from Group Two, specifically those showing the strongest signs of anti-bot adaptation. This involves close reading for syntactic disruption, lexical incongruence, register inconsistency, and structural fragmentation. These features are understood as discursive responses to the institutional power of the AI detector as a de facto language manager. CDA connects individual sentence choices to the broader institutional policies and sociolinguistic inequalities that shape the study theoretical argument.

This three-level framework is appropriate because anti-bot language is computationally triggered, linguistically produced, and discursively meaningful. Analyzing it at only one level would produce an incomplete account.

2.8 Analytical Procedures

This section describes the specific steps used to process, code, and interpret the data from each instrument.

2.8.1 Phase One: Quantitative Analysis of Questionnaire Data

After data collection is complete, all questionnaire responses are exported from the survey platform into SPSS. In the first step, incomplete responses are identified and managed according to a set threshold. Responses with more than 20% of items missing are excluded; item-level missing data below this threshold are replaced using mean substitution. In the second step, internal consistency is assessed for each subscale using Cronbach alpha. Any subscale falling below .70 is reviewed for revision before analysis continues. In the third step, descriptive statistics are computed for every item and subscale, and frequency distributions are produced for key variables. Effect sizes (Cohen d) are reported alongside significance values to show the practical scale of observed differences. In the fourth step, bivariate correlations are computed among key continuous variables using Pearson r, and the results are examined for theoretically meaningful patterns. All findings are tabulated and interpreted in relation to the study theoretical framework.

2.8.2 Phase Two: Qualitative Analysis of Interviews and Field Notes

Immediately after each interview, audio recordings are transcribed word for word. Paralinguistic features such as extended pauses, emphasis, and laughter are noted where relevant. Each transcript is read in full at least twice before coding begins. Initial notes are written to record first impressions and emerging questions. In the first step, an inductive codebook is developed from the data. Meaningful units, including phrases, sentences, or longer passages, are assigned descriptive codes that capture their content without imposing early interpretation. In the second step, codes are grouped into preliminary candidate themes, which are then reviewed for internal coherence. In the third step, candidate themes are tested against the full dataset to confirm they are well-supported. Themes that are too broad, too narrow, or insufficiently evidenced are refined, merged, or divided as needed. In the fourth step, each final theme is defined and named precisely, and representative quotations are selected to support the written analysis. Observation field notes are analyzed through a similar but shorter process: key moments of institutional talk about AI, and patterns of compliance or resistance, are identified, coded, and integrated into the thematic account produced from the interviews.

2.8.3 Phase Three: Analysis of Writing Samples

Writing samples from both groups are immediately anonymized after collection. Each text is assigned a unique alphanumeric code indicating its group (G1 for naturalistic writing, G2 for AI-aware writing) and participant number, with no identifying information attached.

In the first step, all texts are submitted to the selected AI detection platform and detection scores are recorded in a structured data spreadsheet. False-positive rates for Group One are

calculated by recording the number of human-written texts classified as AI-generated, expressed as a percentage of all Group One texts. For Group Two, detection outcomes are recorded alongside participants self-reports of AI use, to allow comparison between texts that evaded detection and those that did not. In the second step, all texts are processed through corpus analysis software to compute lexical diversity scores, syntactic complexity measures, and frequency profiles of formulaic language features. Scores for Group One and Group Two are compared. In the third step, texts from Group Two that show the strongest signs of anti-bot adaptation are selected for detailed CDA. Each selected text is examined for specific discursive strategies: intentional syntactic fragmentation, register inconsistency, lexical incongruence, and structural disruption. These strategies are mapped onto the theoretical construct of anti-bot language from the Literature Review, producing a descriptive profile of the register main features.

2.8.4 Phase Four: Cross-Strand Integration and Meta-Inference

In the final analytical phase, findings from all three parts of the study, the questionnaire data, the interview and observation data, and the writing sample analysis, are brought together for comparison. Where statistical patterns from the questionnaire align with thematic findings from interviews, and where both are supported by the writing sample analysis, the overall conclusions are strongest. Where findings from different phases diverge or contradict each other, these differences are examined rather than ignored. They often point to important nuances and help produce more precise and honest conclusions. The overarching findings produced through this integration process represent the study's main contribution to the empirical literature on AI detection, language policy, and second-language writing in higher education.

2.9 Ethics of the Research

This study follows the ethical guidelines of the British Educational Research Association (BERA, 2018). Before any data collection began, the study design, instruments, and recruitment procedures were reviewed and approved through the host university institutional ethics process. This ensured independent ethical review before contact with any participants.

Informed consent is treated as an ongoing commitment throughout the study. All participants receive a plain-language information sheet before taking part. This document clearly explains the purpose of the study, what participation involves, how long it will take, how data will be stored and anonymized, and the participant unconditional right to withdraw at any point without giving a reason or facing any consequence. Written consent is collected before interviews and other data collection sessions.

One specific ethical consideration concerns Group One of the writing sample study. These participants were not told at the time of writing that their texts would later be analyzed

using AI detection tools. This limited non-disclosure was methodologically necessary to preserve the natural quality of their writing. However, full disclosure was provided immediately after data collection through a structured debriefing session. At that point, participants were reminded that they could withdraw their text from the study without any consequence. This deferred consent procedure is consistent with established ethical practice in naturalistic social science research and was reviewed and approved as part of the institutional ethics application.

All participant data is anonymized throughout the study. Names, institutional affiliations, geographic details, and any other identifying information are replaced with alphanumeric codes as early as possible in the process. Writing samples are stripped of identifying metadata before analysis. All data is stored securely and will be retained for one to three months, as stated to participants, before secure deletion. Because the subject matter of this study involves AI detection accusations and academic misconduct concerns, special care is taken to ensure that no individual can be identified from the way findings are presented.

2.10 Limitations

Every methodological design has constraints that limit how widely its findings can be applied. This study is no exception.

The most significant limitation is the specific institutional context. The study is conducted within Algerian higher education institutions, and the findings cannot be generalized to higher education globally without replication across different national, linguistic, and institutional settings. The Algerian EFL context is shaped by French-Arabic-English multilingualism, a distinct post-colonial academic history, and institutional structures that differ from Anglophone systems. These specific conditions are acknowledged as both a limitation on generalizability and a source of unique analytical value.

A second limitation concerns the self-report nature of much of the data. Questionnaire responses and interview testimony depend on participants' memories, their willingness to be honest, and the limits of accurate recall. Participants who have experienced false-positive detection accusations may over-report anxiety or behavioral adaptation. Those who have not may underestimate how much algorithmic surveillance has affected them. Triangulation with observation data and writing samples partially compensates for this, but cannot fully eliminate these reporting gaps.

A third limitation concerns the writing sample collection. The non-disclosure used with Group One, while methodologically necessary and ethically managed through debriefing, means some participants may have felt unsettled when they learned of the AI detection analysis during debriefing. No participant reported distress, and all retained the right to withdraw. However, this element of the design represents a departure from fully transparent consent at the moment of

completing the writing task. Finally, AI detection technology changes rapidly. The tools used in this study reflect detection capabilities at the time of data collection. Findings concerning accuracy, bias patterns, and linguistic effects may need to be revisited as newer detection systems are released.

2.11 Conclusion

In summary, this study methodology is designed to provide a thorough and honest account of the anti-bot language register and the attitudes of students and educators toward AI use and related activities. While the limitations above are acknowledged, the study aims to maintain high ethical standards and produce findings that are reliable and useful for the field.

CHAPTER THREE:

DATA ANALYSIS AND

INTERPRETATION

3.1 **Introduction**:

This chapter presents and interprets the data generated through four of the study's instruments: Observation,the student questionnaire , the semi-structured interviews and the group study. The analysis moves through the questionnaire in order, section by section, from the pre-questionnaire awareness items through Section A (awareness of AI detection), Section B (writing behavior under surveillance), and Section C (academic voice and identity), before addressing the open-ended responses. It then turns to the interview data, reading the student group discussion and the individual educator interviews in the context of what the questionnaire established.Then the observation and lastly the most vital experiment which is the writing sample analysis. The chapter does not describe the data and stop there. It insists on discussing why each pattern exists, what institutional conditions produce it, and what it means for the students, educators.

3.2 Questionnaire Analysis

Before analysis begins, a brief word on the sample. Twenty-three students completed the questionnaire in full. All identified as EFL writers studying within Algerian higher education institutions. This is not a coincidental demographic detail; it is theoretically important, because the study's central argument concerns the unequal impact of algorithmic surveillance on exactly this kind of writer. The Cronbach's alpha for the full 21-item Likert instrument, calculated after reverse-coding Q21, was a = .772. This figure confirms that the questionnaire functions as a coherent, internally consistent measure.the items are measuring the same underlying set of attitudes and experiences from different angles.

Throughout the analysis, the symbol M refers to the Mean, which is the average score of all 23 participants on a given questionnaire item. The questionnaire uses a five-point scale where 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree. The midpoint of this scale is 3.0, which represents a neutral position. A mean above 3.0 means the group tended toward agreement on that item; a mean below 3.0 means they tended toward disagreement. For example, M = 3.87 means the average response was close to Agree, while M = 2.43 means it was closer to Disagree. These mean scores are reported alongside percentage figures throughout the chapter so that every statistic can be read in plain terms without requiring statistical background.

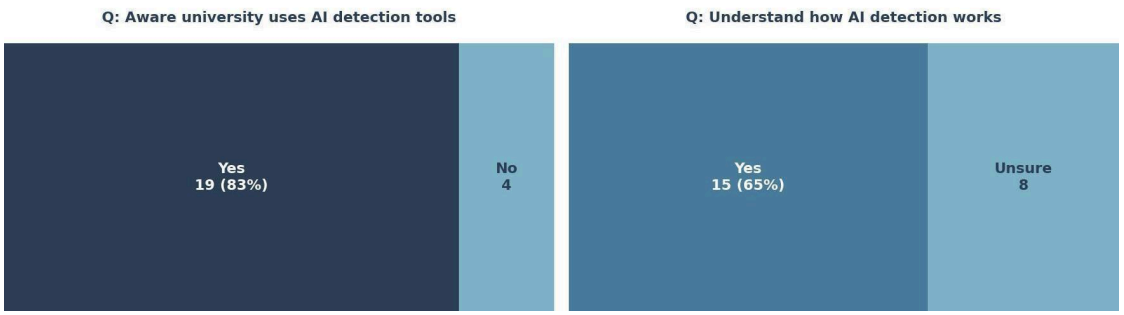
Before reaching the Likert-scale questions, participants were asked to confirm two things. The first was that they understood the difference between AI detection tools and AI writing tools. This distinction matters enormously for the study: a student who confuses Turnitin

or any detection tool with ChatGPT would not be giving meaningful responses about how detection affects their writing. All twenty-three participants confirmed this understanding. None chose to withdraw. This unanimous confirmation means the behavioral and ideological findings that follow cannot be attributed to basic conceptual confusion about what a detection tool is.

The second item asked whether participants were aware that their university uses AI detection tools. Nineteen out of twenty-three said yes (82.6%), and four said no (17.4%). The third item asked whether they understood how these tools work. Fifteen out of twenty-three said yes (65.2%), and eight were unsure (34.8%). The gap between those who know surveillance exists (82.6%) and those who understand how it works (65.2%) is worth pausing on. Students are being judged by a system that a meaningful portion of them cannot fully explain. This is not simply a gap in technical knowledge. It is the starting condition for a particular kind of institutional anxiety: you know you are being evaluated, but you do not fully know the basis of that evaluation.

Figure 1

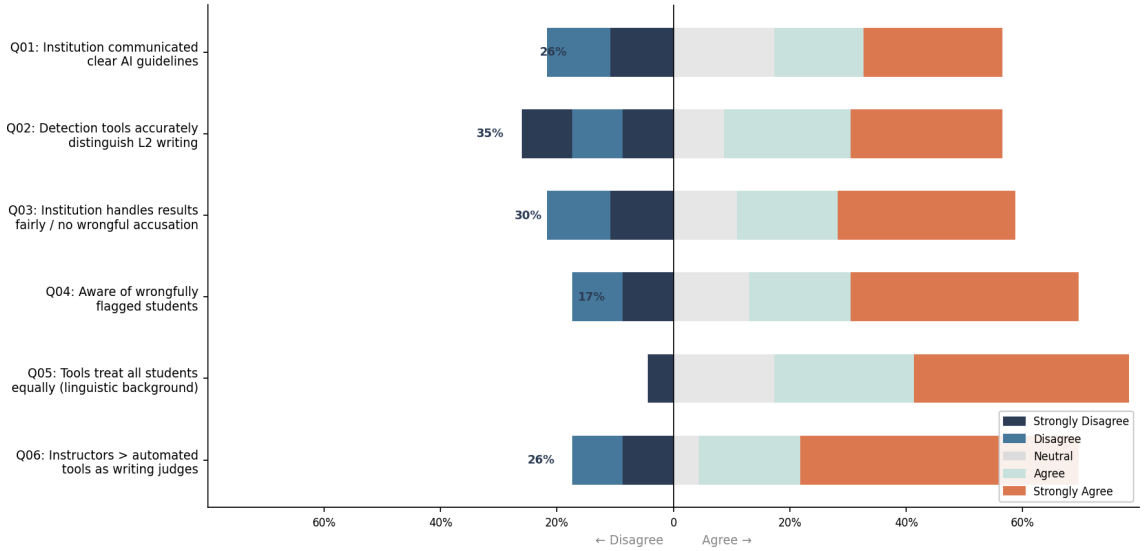
Pre-Questionnaire Awareness Items (N=23). Dark blue =Yes / Aware Light blue = No / Unsure.



3.2.1 Section A: Awareness of AI Detection Tools (Q01 to Q06)

Section A was built around six items that map how clearly students understand their institution's AI policies, how much they trust the institution to handle detection fairly, and whether they think the tools treat everyone equally. In Spolsky's (2004) terms, this section addresses language management: the explicit, institutional mechanisms through which language use is regulated. What the data show, taken together, is a group of students who are partially informed, partially trusting, and clearly aware that something in the system is not working as it should.

Figure 2:*Section A diverging response distribution. Bars extending left = disagreement; bars extending right = agreement. Percentages shown where total exceeds 15%.*



Q01: Institution communicated clear AI guidelines (M = 3.17)

This item asked whether participants felt their institution had communicated clear and official guidelines about the use of AI in academic writing. The mean of 3.17 sits just above the neutral midpoint of 3.0. In practical terms, this means the average response was slightly above neutral, with 39.1% agreeing, 34.8% neutral, and 26.1% disagreeing. This is not the profile of an institution that has communicated clearly and consistently. If the guidelines had been clear to everyone, the mean would sit closer to 4.0 or above. Instead, the spread of responses tells us that the same policy, if it exists, is reaching different students in different ways, or in some cases not reaching them at all. That inconsistency matters because a student who does not know the rules cannot follow them, and a student who is penalized for breaking rules they were never given has been treated unfairly.

Q02: Detection tools accurately distinguish L2 writing from AI (M = 3.00)

Item Q02 asked whether participants believed AI detection tools can accurately tell apart AI-generated text from the writing of a non-native English speaker. The mean of exactly 3.00 is the closest this questionnaire comes to a perfectly split result: 47.8% agreed, 17.4% were neutral, and 34.8% disagreed. This distribution reflects genuine uncertainty in the student population, not indifference. Crucially, the 34.8% who actively doubt accuracy are not wrong to do so. The empirical research on this question, including the experimental findings of Liang et al. (2023), confirms that AI detectors produce significantly higher false-positive rates for ESL writers than for native English speakers. The students in this study who doubt the tool's fairness are, in other words, accurately reading a real problem. The students who believe in the tool's accuracy may be

unaware of that research. This unevenness in what students know is itself a form of inequity within the population.

Q03: Institution handles detection results fairly (M = 3.22)

Q03 measured institutional trust: whether students believed their institution would handle detection results fairly and protect them from wrongful accusation. The mean of 3.22 and 47.8% agreement may appear encouraging at first, but the 30.4% who actively disagreed is a significant number. Nearly one in three students does not trust the institution they attend to protect them if a detection tool makes a mistake. In an educational context where the consequences of a wrongful accusation include formal misconduct charges, academic penalties, and reputational damage, a trust deficit of this scale is not a minor satisfaction issue. It is a governance problem. And it is one that no software upgrade will fix, because it is rooted in a structural question: who is accountable when the machine is wrong?

Q04: Aware of wrongfully flagged students (M = 3.61)

This item asked whether participants were aware of cases, real or reported, in which students had been wrongly flagged as AI users despite having written their own work. The mean of 3.61 and 56.5% agreement make this the first clearly above-midpoint item in Section A. What this tells us is that false-positive incidents have become shared knowledge among students. More than half of this sample know that the system has made mistakes. They do not need to experience a false accusation personally to be affected by knowing that it happens. A student who sits down to write an essay while knowing that peers have been falsely accused will approach that essay differently, will worry more, and will be more likely to modify their writing to protect themselves from a system they know can be wrong. The behavioral effects of this knowledge are exactly what Section B will document.

Q05: Tools treat all students equally regardless of language (M = 3.65)

Q05 produced a mean of 3.65 and 60.9% agreement. At first, this appears to suggest that students broadly believe the tools are equitable. But the distribution tells a more careful story. Of the 60.9% who agreed, only 13% chose strongly agree, while 47.8% selected agree. This is a soft majority, not a confident one. And the 34.8% at neutral or below represent students who already suspect structural disadvantage. When this item is read together with Q02, where 34.8% doubted the tool's accuracy for L2 writers, a picture forms of a population divided between those who trust the system and those who have genuine reason not to. That divide, as will be seen, maps closely onto who reports behavioral adaptation in Section B.

Q06: Instructors, not tools, are the right judges of student writing (M = 3.61)

The final item of Section A asked whether human instructors, rather than automated tools, are the most appropriate judges of whether student writing is genuinely the student's own. With a mean of
3.61 and 65.2% agreement, this records the strongest endorsement in the section.

Two thirds of students prefer to be assessed by a person who knows their academic context, their writing history, and their intellectual development. This is a reasonable preference. A teacher who has read a student's work across an entire semester brings knowledge to a judgment that an algorithm simply does not have. The fact that this preference exists so strongly in the data, alongside the institutional reality that algorithmic scores are being used as primary assessment evidence, is not a trivial gap. It is a sign that the institution is providing something its students did not ask for, in place of something they clearly value.

Table 1:*Section A descriptive statistics. M = mean on a scale of 1 to 5 where 3.0 is neutral.*

Item	What it measures	M	Agree % (4+5	Neutral %	Disagree % (1+2)
Q01	Clear institutional AI policy communicated	3.17	39.1%	34.8%	26.1%
Q02	Detection accurate for L2 writer	3.00	47.8%	17.4%	34.8%
Q03	Institution handles results fairly	3.22	47.8%	21.7%	30.4%
Q04	Aware of wrongfully flagged students	3.61	56.5%	26.1%	17.4%
Q05	Tools treat all students equally	3.65	60.9%	34.8%	4.3%
Q06	Instructors better judges than to	3.61	65.2%	8.7%	26.1%

3.2.2 Reading Section A as a whole

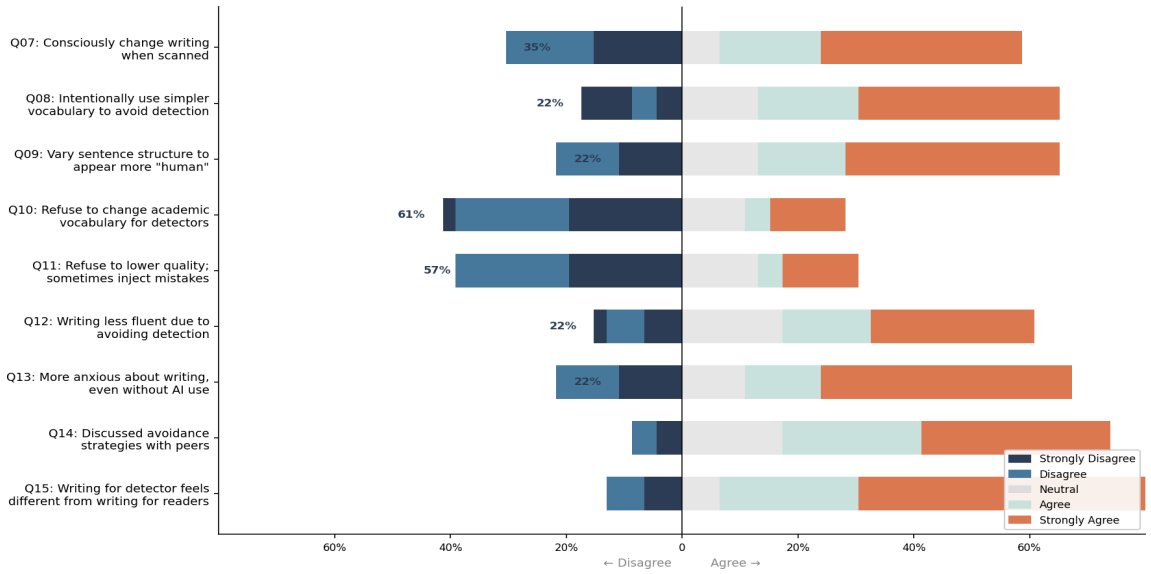
Two patterns run through Section A consistently. The first is a cluster of qualified institutional doubt: students are uncertain about policy communication (Q01: M=3.17), split on detection accuracy (Q02: M=3.00), partially distrustful of institutional fairness (Q03: M=3.22), well-informed about the possibility of false positives (Q04: M=3.61), and ambivalent about equal treatment (Q05: M=3.65). Not a single item in this section reaches a mean of 4.0. Not one. This is not a student body that is broadly satisfied with the institutional governance of AI detection. The second pattern is the preference gap captured in Q06 (M=3.61, 65.2% agree): students want human judgment; they are getting algorithmic judgment. This gap between what students want from assessment and what they currently receive represents the practical face of the de jure versus de facto divide that this study has been building toward since Chapter Two. Formally, institutions speak about integrity and fairness. In practice, they have handed the first

judgment of student writing authenticity to a machine that a significant proportion of their own students neither fully understand nor fully trust.

3.2.3 Section B: Writing Behavior Under Algorithmic Surveillance (Q07 to Q15)

If Section A established the conditions that make behavioral adaptation predictable, Section B documents the adaptation itself. The nine items of this section measure whether and how students change their writing when they know a detection tool will evaluate it. The section subscale alpha of
.755 confirms strong internal consistency, meaning these nine items are reliably measuring the same underlying behavioral reality. The overall Section B mean of 3.275 sits above the neutral midpoint, which tells us that the dominant tendency in this population is toward adaptation rather than resistance. But the range of individual means, from 2.43 to 3.87, shows that not all adaptations are equally common.

Figure 3: *Section B diverging response distribution for Q07 to Q15.*



Q07: Consciously change writing when scanned (M = 3.30)

This item functioned as the opening question for the behavioral section. It asked simply: when students know their writing will be scanned, do they consciously write differently? The mean of 3.30 and 52.2% agreement tells us that more than half of this sample do exactly that. This finding, even before the more specific items, carries considerable weight. Academic writing education assumes a degree of behavioral consistency: a student learns to write a certain way and carries those habits from one context to the next. What Q07 documents is a split in that consistency. A majority of students in this study have developed two different writing approaches: one for writing that will reach a human reader, and one for writing that will first pass through a machine. The existence of that split is the foundation on which everything else in Section B rests.

Q08: Use simpler vocabulary to avoid detection (M = 3.35)

With a mean of 3.35 and 52.2% agreement, more than half of respondents report intentionally using simpler vocabulary than they would otherwise use, specifically to avoid being flagged. Consider what this means in practice. Students have spent years building academic vocabulary through reading, coursework, writing practice, and instructor feedback. They now have access to more precise, more sophisticated, more disciplinary language than when they began university. And they are choosing not to use it in assessed writing, not because they have forgotten it, not because it is wrong, but because a detection tool might interpret sophisticated academic vocabulary as a sign that the writing came from a machine. The academic vocabulary that took years to acquire becomes a liability in the very context where it should be an asset.

Q09: Vary sentence structure to appear more human to the detector (M = 3.52)

Q09 produced the highest mean of the adaptation cluster at 3.52, with 52.2% agreement. This item asked whether students deliberately mix short and long sentences specifically to make their writing appear more human to a detection tool. In academic writing, variation in sentence structure is a natural feature of good prose. Writers alternate between complex, elaborated sentences and shorter, direct ones because their ideas demand it, not because they have calculated what a statistical classifier expects. What Q09 reveals is that this natural feature has become a deliberate, self-conscious strategy for a majority of students. They are engineering variation not because it serves their argument, but because they have learned, through peer knowledge or trial and error, that machines tend to produce uniformly structured text, and that varying structure signals humanness. Language has stopped being a tool for thinking and has become a set of parameters to manage.

Q10: Refuse to change academic vocabulary for the detector (M = 2.43)

Q10 introduced a crucial reversal. It asked whether students prioritize their traditional academic discourse markers and refuse to change their vocabulary to satisfy AI detectors, even when their writing gets flagged. The mean of 2.43 and the distribution of 17.4% agreement against 60.9% disagreement make this the lowest-scoring item in the entire questionnaire. The

clear majority reject this resistance stance. They do not hold their academic register firm in the face of algorithmic pressure. This is not a failure of courage or a sign of poor values. It is a rational response to an institutional environment where the cost of being flagged is concrete and serious, while the cost of writing more poorly is silent and unmeasured. Students are doing the calculation that anyone in their position would do. The result is that the academic vocabulary they were taught is being abandoned not because of any pedagogical decision, but because of a detection system the institution itself deployed.

Q11: Inject deliberate mistakes to lower detection score (M = 2.52)

Q11 measured the most extreme behavioral consequence in the section: whether students deliberately introduce grammatical irregularities or informal phrasing into their academic writing to lower their AI detection score. With a mean of 2.52 and 17.4% agreement against 56.5% disagreement, this is a minority behavior. But a minority of 17.4% in a group of twenty-three students means roughly four people. Four students in this sample have, at some point, deliberately made their academic writing worse in order to satisfy a machine. In any educational context, that figure should prompt serious reflection. Writing education exists to improve the quality and precision of student expression. A student who introduces intentional errors into assessed work has not failed that education; the institution has failed them by creating conditions in which writing badly is the rational thing to do.

Q12: Writing has become less fluent due to detection avoidance (M = 3.26)

Q12 shifted from reporting a strategy to reporting its felt consequence. It asked whether participants feel their writing has become less fluent or less natural as a result of trying to avoid AI detection. With a mean of 3.26 and 43.5% agreement, nearly half of respondents report experiencing a real decline in the quality of their own writing, and they attribute that decline directly to the pressure of algorithmic surveillance. This item is particularly valuable for the study because it connects the behavioral findings of the previous items to a subjective, meta-linguistic judgment. Students are not only changing their writing; they are aware that those changes are making it worse. They feel the unnaturalness of writing that has been engineered for a machine.

Q13: More anxious about writing even without using AI (M = 3.65)

Q13 measured surveillance anxiety among students who have not used AI assistance. The question was direct: has the existence of AI detection tools made participants more anxious about writing assignments, even when they have done nothing wrong? The mean of 3.65 and 56.5% agreement produce one of the clearest findings in the entire study. More than half of these students are anxious about writing because of a detection system, despite having no reason to fear it. This is not irrational behavior. It is a precise response to a situation that Section A already documented: students know the system makes mistakes (Q04: 56.5%), they do not fully trust the

institution to handle those mistakes fairly (Q03: 30.4% distrust), and they are aware that people like them have been wrongly accused. The anxiety is the entirely predictable outcome of accurate information about a fallible system. A policy that generates this level of anxiety among students who did nothing wrong has not made writing assessment more reliable; it has made the educational environment less safe.

Q14: Discussed avoidance strategies with other students (M = 3.57)

Item Q14 asked whether participants had discussed strategies for avoiding AI detection flags with other students, whether in group chats, study groups, or informally. The mean of 3.57 and 56.5% agreement confirm that anti-bot adaptation has moved beyond individual coping behavior and become a shared, socially circulating practice. When more than half of students have talked with each other about how to evade a detection tool, that tool is no longer governing their writing at the individual level only; it is reshaping the norms of an entire student community. Students compare notes, share what works, and develop a collective practical knowledge about how to write for a machine. In Spolsky's (2004) framework, this is language practice being reshaped at the community level by a new and powerful institutional actor.

Q15: Writing for detector feels different from writing for a reader (M = 3.87)

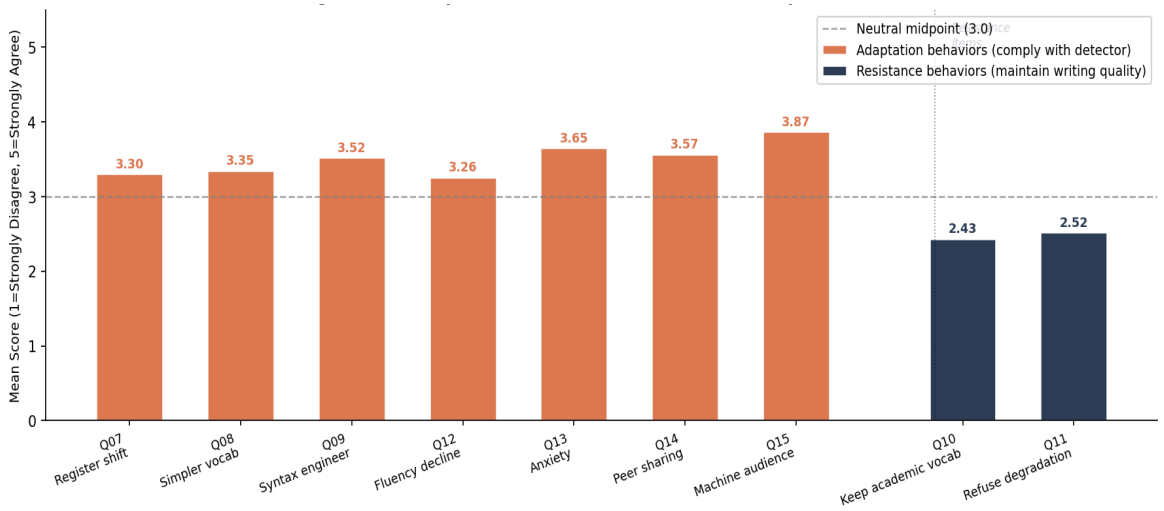
The final item in Section B produced the highest Section B mean of 3.87 and 73.9% agreement, the second highest agreement rate in the whole questionnaire. Q15 asked whether writing for a detector's approval feels fundamentally different from writing to communicate ideas to a human reader. Nearly three quarters of students answered yes. This is the item that most precisely names what this dissertation has been building toward. Academic writing is a communicative act addressed to a human reader. It is built on shared conventions, genre expectations, and the assumption of reciprocal understanding between writer and reader. What Q15 tells us is that this communicative foundation has been disrupted for the majority of students in this study. They are not writing for a reader and also worrying about a machine on the side. They are writing for two different audiences with two incompatible sets of standards, and they feel that tension in the act of writing itself.

Table 2: *Section B descriptive statistics. M = mean on a scale of 1 to 5 where 3.0 is neutral.*

Item	What it measures	M	Agree %	Neutral %	Disagree %
Q07	Conscious writing change und detection	3.30	52.2%	13.0%	34.8%
Q08	Intentional vocabulary simplification	3.35	52.2%	26.1%	21.7%
Q09	Sentence structure variation fo detection	3.52	52.2%	26.1%	21.7%

Q10	Refuse to change academic vocabulary	2.43	17.4%	21.7%	60.9%
Q11	Deliberate error injection	2.52	17.4%	26.1%	56.5%
Q12	Felt fluency decline from avoidance	3.26	43.5%	34.8%	21.7%
Q13	Anxiety without AI use	3.65	56.5%	21.7%	21.7%
Q14	Peer sharing of avoidance strategies	3.57	56.5%	34.8%	8.7%
Q15	Detector audience vs. human audience	3.87	73.9%	13.0%	13.0%

Figure 4: *Adaptation behaviors (orange) vs. resistance behaviors (navy) compared by mean score.*



3.2.4 The pattern in Section B

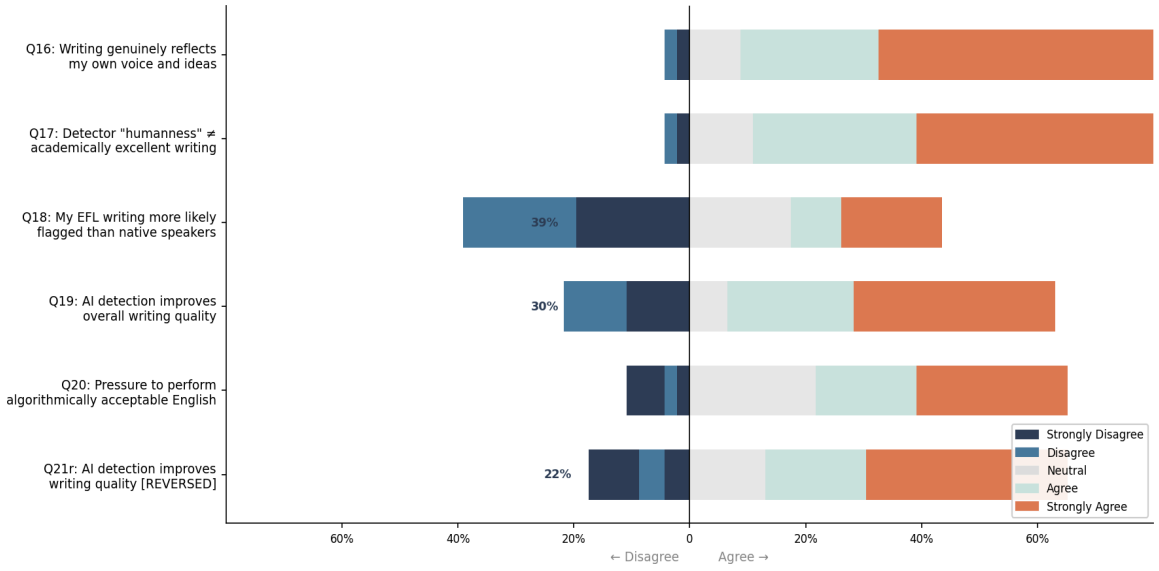
Figure 4 makes the pattern visible in a single image. The adaptation behaviors, Q07 through Q09 and Q12 through Q15, all sit above the neutral midpoint, ranging from M=3.26 to M=3.87. The resistance behaviors, Q10 and Q11, sit clearly below it at M=2.43 and M=2.52.

This is not a random distribution. It reflects a coherent institutional logic: when the cost of being flagged is concrete and serious, students adapt; they do not resist. The Cronbach alpha for Section B (.755) confirms that these items are measuring a single, real behavioral phenomenon from nine different angles. That phenomenon is adaptation to algorithmic surveillance as the primary shaping force of assessed writing.

3.2.5 Section C: Academic Voice and Linguistic Authenticity (Q16 to Q21)

Section C moves from behavior to belief. Where Section B asked what students do when facing detection, Section C asks what they think and feel about who they are as writers within a surveilled environment. This corresponds to Spolsky's (2004) deepest layer of language policy: ideology, the beliefs and values through which a community decides what counts as legitimate, valued, and authentic language use. The section's lower internal consistency (alpha = .223) reflects the fact that language ideology is multi-dimensional by nature. These six items are not measuring one thing from different angles; they are each touching a different face of the same underlying set of questions about identity, fairness, and educational value.

Figure 5: *Section C diverging response distribution for Q16 to Q21 (Q21 reverse-coded before analysis).*



Q16: My writing genuinely reflects my own voice and style (M = 4.04)

Q16 produced the highest mean in the entire questionnaire: 4.04, with 78.3% agreement and only 4.3% disagreement. Nearly four in five students feel that their academic writing genuinely reflects their own voice, ideas, and style. At first sight, this seems to contradict the behavioral findings of Section B. If students are simplifying their vocabulary, engineering their sentence structure, feeling less fluent, and writing for machines rather than readers, how can they simultaneously feel that their writing reflects their own voice? The answer lies in context. Q16 is asking about the student's sense of writing identity in general, not specifically about assessed, surveilled writing. A student who has a strong sense of their own academic voice in tutorials, drafts, and informal writing, but who suppresses that voice in high-stakes assessed work to avoid algorithmic penalty, can honestly answer yes to Q16 while also endorsing the behavioral adaptations of Section B. Read alongside Q15 (M=3.87, 73.9% agree that the detector audience feels different from the human audience), Q16 does not produce a contradiction. It produces something more troubling: students know who they are as writers, and they are being asked by the institution to perform something different.

Q17: Detector approval is not the same as academic excellence (M = 3.87)

Q17 asked whether writing that satisfies a detector's criteria for humanness is not necessarily the same as writing that is academically excellent. The mean of 3.87 and 73.9% agreement show that nearly three quarters of students clearly reject the idea that passing a machine is equivalent to writing well. They understand, as an intellectual conviction, that the two standards are separate and that the detector's standard is the lesser one. This finding would be reassuring if it led to different behavior. But it does not. Students who disagree with the detector's authority (Q17: 73.9%) are still simplifying their vocabulary (Q08: 52.2%), engineering their syntax (Q09: 52.2%), and directing their writing at a machine rather than a reader (Q15: 73.9%). They are complying with a standard they know to be educationally inferior because the institutional consequences of non-compliance are real, while the educational cost of compliance is silent. This gap between conviction and practice is one of the most important patterns the data produce.

Q18: As an EFL writer, my style is more likely to be flagged than a native speaker's (M = 2.96)

Q18 asked EFL writers directly whether they feel their natural writing style is more likely to be flagged as AI-generated compared to a native English speaker's writing. The mean of 2.96 falls just below neutral, with 26.1% agreeing, 34.8% neutral, and 39.1% disagreeing. This is the most evenly distributed item in Section C. The majority who disagree or are neutral might suggest equity concerns are not widely felt. But neutral on an equity question like this can mean uncertainty about the comparison rather than genuine belief in equal treatment. The 26.1% who do agree are particularly significant: these are students who, despite social pressure to believe the system is fair, are naming the structural disadvantage they perceive. Their perception matches

what the SLA research literature confirms experimentally, specifically that EFL writing characteristics, including formulaic structures and lower lexical variation, produce higher false-positive rates in detection tools. They are right to feel what they feel.

Q19: AI detection improves overall writing quality (M = 3.30)

Q19 asked whether AI detection tools improve the overall quality of student writing. The mean of
3.30 and 56.5% agreement make this the most divided ideological item in the section. Some students can see, or are willing to accept, a potential benefit: detection as a deterrent against AI-assisted cheating might, in theory, encourage more genuine engagement with writing. Others, drawing on their direct experience reported in Section B, actively reject this proposition. This ideological split reflects the fact that the student population in this study is not uniformly hostile to all institutional use of AI detection. What they are hostile to, as the behavioral data confirm, is the specific form that use has taken: an unsupported, unexplained, and poorly communicated detection system that treats their writing as suspect by default.

Q20: Pressure to perform algorithmically acceptable English rather than develop my own voice (M = 3.30)

Q20 asked whether participants feel pressure to perform a kind of acceptable academic English that satisfies institutional software, rather than to develop their own authentic scholarly voice. The mean of 3.30 and 43.5% agreement, with an unusually large neutral cluster of 43.5%, suggests a complex response. The large neutral group may reflect students who feel this pressure without having the conceptual vocabulary to name it in the terms the item uses. But the 43.5% who agree are making a precise and significant claim. There is a meaningful difference between performing acceptable English for a machine and developing your own scholarly voice. Performance satisfies a metric; development builds a mind. The students who endorse this item are reporting not merely a writing inconvenience, but an educational loss.

Q21 reversed: AI detection improves writing quality, reverse-coded (M = 3.35)

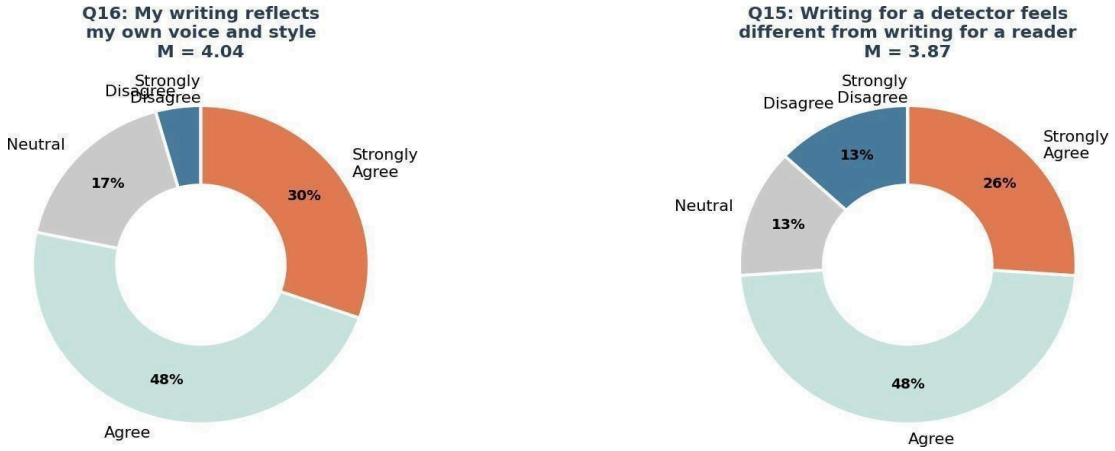
Q21, reverse-coded before analysis, functioned as an internal consistency check on the attitude measured in Q19. The reverse-coded mean of 3.35 sits close to Q19's mean of 3.30. The near-identical results across both items, presented in different positions in the questionnaire, confirm that the instrument captured stable attitudinal orientations rather than momentary or

arbitrary responses. Students answered both formulations of the same underlying question with consistency, and this consistency contributes to the overall instrument reliability of a = .772.

Table 3: *Section C descriptive statistics. M = mean on a scale of 1 to 5 where 3.0 is neutral. Q21 was reverse-coded (formula: 6 minus original score) before analysis.*

Item	What it measures	M	Agree %	Neutral %	Disagree %
Q16	Writing reflects own voice and style	4.04	78.3%	17.4%	4.3%
Q17	Detector approval is not academic quality	3.87	73.9%	21.7%	4.3%
Q18	EFL writing more likely flagg than native	2.96	26.1%	34.8%	39.1%
Q19	Detection improves writing quality	3.30	56.5%	13.0%	30.4%
Q20	Pressure to perform algorithm English	3.30	43.5%	43.5%	13.0%
Q21r	Detection improves quality (reversed)	3.35	52.2%	26.1%	21.7%

Figure 6: *The voice paradox.*



Note.Q16 (left) shows students feel strongly that writing reflects their own voice (M=4.04, 78.3% agree). Q15 (right) shows they also feel writing for a detector is fundamentally different from writing for a human reader (M=3.87, 73.9% agree). Both can be true because they describe different writing contexts.

3.2.6 Reading Section C as a whole

The dominant tension in Section C is between two things that coexist in the same students. On one side: a strong sense of voice ownership (Q16: 78.3%) and a clear rejection of the detector's authority over writing quality (Q17: 73.9%). On the other: pressure to perform algorithmically acceptable English (Q20: 43.5%) and real uncertainty about whether EFL writers bear a disproportionate burden (Q18: 26.1% agree, 34.8% neutral). These are not contradictory positions. They are the positions of a student community that has not given up on its own sense of what good writing is, but that is being pushed, daily and institutionally, toward a different and lesser standard. The longer that push continues without being named, examined, or challenged by the institution itself, the more likely the ideological tension is to resolve in the direction of compliance.

3.2.7 Full Questionnaire Overview: Means, Distribution, and Reliability

Figure 7: *Mean scores across all 21 Likert items. The dashed line marks the neutral midpoint of 3.0. Items in navy = Section A, blue = Section B, orange = Section C.*

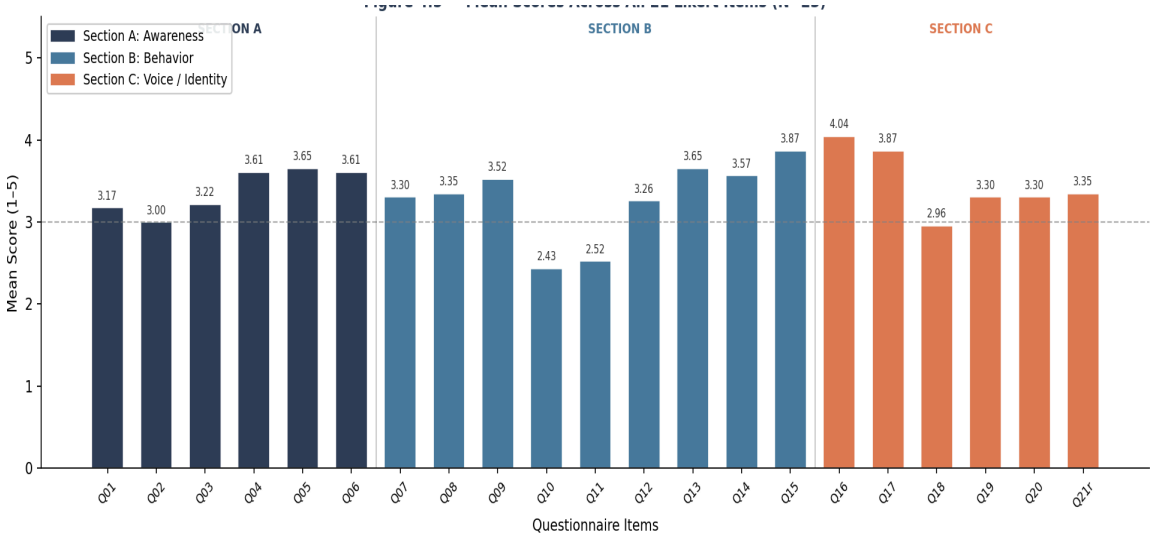


Figure 8: *Left panel shows the three section subscale means compared. Right panel shows the full distribution of all responses across all 21 items and 23 participants.*

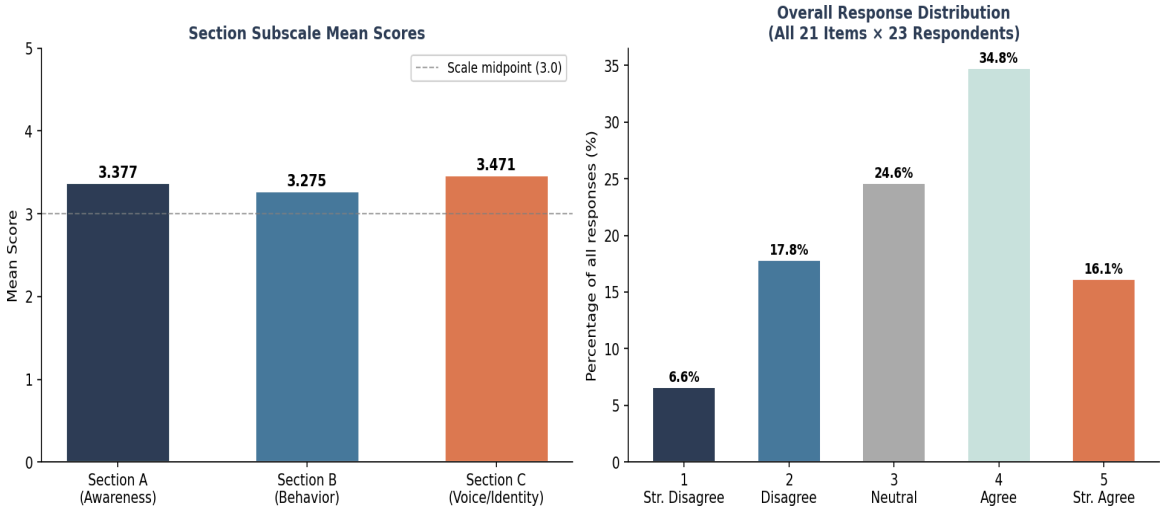


Figure 9: *Response heatmap. Each row is one questionnaire item; each column is one participant. Red indicates disagreement; blue indicates agreement. The white horizontal lines separate sections. Visual patterns of consistently warm or cool rows correspond to items with clear majority positions.*

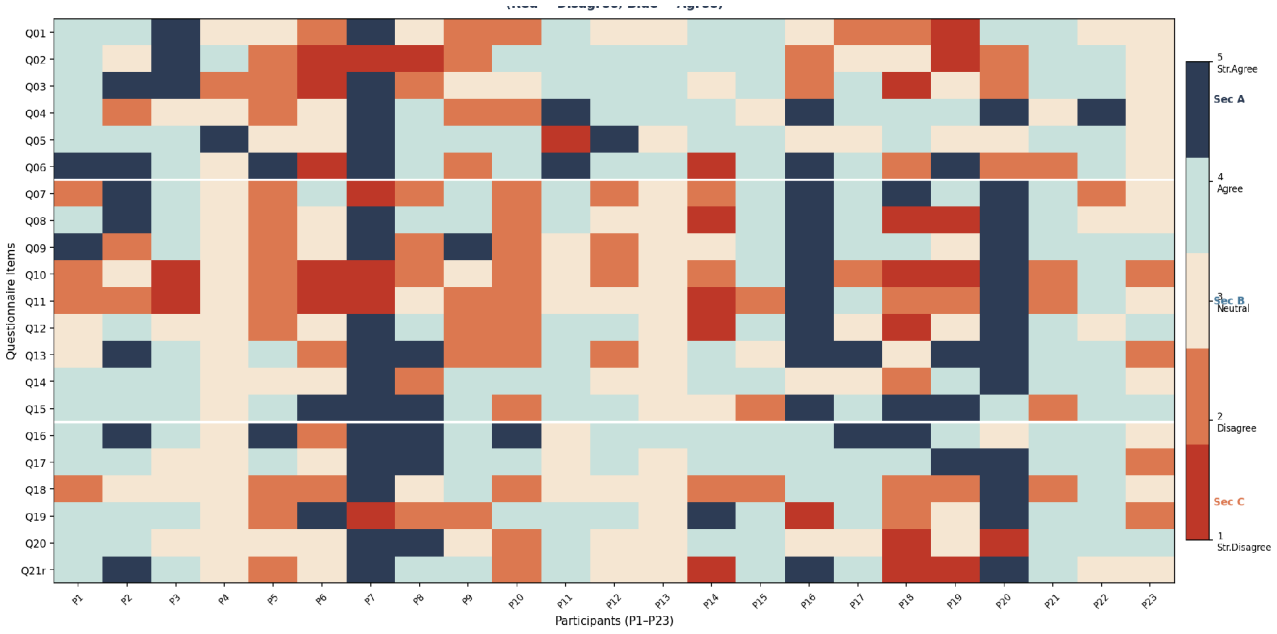


Figure 7 shows the full item-level mean profile and makes the polar structure of the questionnaire visible at a glance. Q10 (M=2.43) and Q11 (M=2.52), the two resistance items,form a clear trough at the left end of Section B. Q16 (M=4.04) and Q15 (M=3.87) form the two highest peaks across the full instrument. The pattern reads clearly: students score lowest on the behaviors that would require them to push back against algorithmic pressure, and highest on the items that affirm their own sense of writing identity and describe the communicative displacement they feel. Figure 8, the response heatmap, adds another layer to this reading by showing individual variation. The warm cluster visible around Q10 and Q11 confirms that disagreement with resistance behaviors was spread across most participants, not driven by a small subgroup. The cool cluster around Q16 confirms that voice ownership was broadly endorsed, not the position of a vocal minority.

The three section subscale means (Section A: M=3.377, Section B: M=3.275, Section C: M=3.471) are all above the neutral midpoint of 3.0. The overall response distribution in Figure 4.6 shows that agree (4) was the most common single response category across all items combined. The Cronbach alpha for the full instrument (a=.772) is acceptable and confirms the questionnaire's overall coherence. The lower subscale alphas for Section A (.549) and Section C (.223) are expected, given that awareness and ideology constructs are by nature multi-dimensional rather than measuring a single unified attitude.

3.2.8 Open-Ended Questionnaire Responses

Two open-ended items followed the Likert sections. The first asked students to reflect on what good academic writing means to them now, and whether that definition has changed since they started university. The second asked them to describe a specific moment when they changed how they wrote because they knew a detection tool would scan their work. Four responses to the first item and two to the second were excluded from analysis because the answers did not engage adequately with the question, leaving nineteen responses for OE1 and twenty-one for OE2. What the remaining responses show, read carefully and in full, is a student body that is not passive in the face of algorithmic surveillance. They are making choices, forming opinions, and in several cases arriving at conclusions about detection tools that match the academic literature without ever having read it.

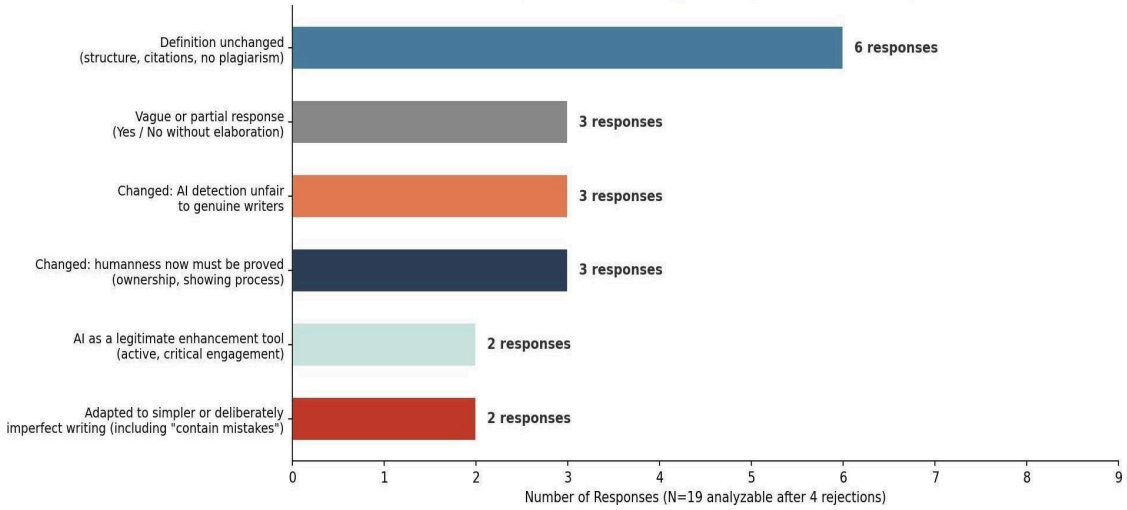
A/ OE1: What Does Good Academic Writing Mean to You Now?

Six broad patterns run through the nineteen analyzable responses to this question. The largest single group, six responses, maintained an unchanged definition: writing that is clear, structured, properly cited, and free from plagiarism. These students have held their ground. For them, the arrival of AI detection has not altered what they believe good writing to be. Response 5 puts this flatly:

"Good academic writing to me is still just what it has always been: clear structure, proper citations, no plagiarism. The AI stuff has not really changed anything for me. Writing is writing." (Respondent 5)

This group is not analytically uninteresting. The fact that some students have not shifted their definition, while the behavioral data of Section B shows the majority adapting their writing practice, suggests that ideological conviction and behavioral compliance do not always travel together. A student can believe writing is writing while still shortening their vocabulary when detection looms.

Figure 10: *OE1 Response Themes. Six students maintained an unchanged definition of good writing; others shifted toward humanness, fairness concerns, or simplified standards.*



Three responses described a definition that has evolved specifically around the concept of provable humanness. Respondent 2 makes this explicit and with some precision:

"Good academic writing used to mean formal tone, correct grammar, structured paragraphs, clear arguments, and proper referencing. But today it is not just about how well you write, but how clearly your thinking belongs to you." (Respondent 2)

This is a significant ideological shift. The criteria for good writing have expanded beyond the communicative and the formal to include a new requirement: demonstrating ownership of thought. Writing must now prove its humanity, not just demonstrate its quality. That additional burden falls unevenly across a student population, and it falls hardest on those whose natural style is most likely to be mistaken for a machine.

Three responses described the situation as one of institutional unfairness. Respondent 8 combines personal frustration with a broader observation:

"At the beginning we sought to write in a very sophisticated way, but now such writing is prone for falsely accusations which is irritating at best. I could say that AI is ruining it for genuinely talented and linguistically smart people." (Respondent 8)

Respondent 17 arrives at an accurate description of the false-positive mechanism without using any of the academic vocabulary for it

"The use of AI detection can be unfair to those who have not used AI, simply because the AI assumes that a text with perfect orthography was generated by AI, when that is not the case at all." (Respondent 17)

This student has independently identified the core problem that Liang et al. (2023) documented experimentally: that linguistic precision and formal correctness, rather than evidence of AI use, are what trigger high detection scores. They have done this from lived experience, not from reading the research.

Two responses described AI as a legitimate writing tool worth engaging with actively. Respondent 9 is the most analytically rich response in the entire open-ended dataset:

"I heavily rely on AI for writing and search, and with repetitive engagement with AI style, I developed a better sounding sense of wording and a better scientific way of writing even when I am not using AI. It depends on the person using AI. Some people literally let AI give them a fish while others ask the AI to teach them how to fish." (Respondent 9)

The fishing metaphor,captures an important distinction that the institutional debate around AI detection has largely failed to make. Passive reliance on AI output and active engagement with AI as a learning tool are not the same behavior, and a detection tool cannot distinguish between them. The student who used AI conversation to develop their understanding of academic style, and who now writes more precisely as a result, is not protected from flagging because they did not copy a sentence.

The most alarming response in the dataset is Respondent 19, whose definition of good academic writing now consists of two words:

"Simple, contain mistakes." (Respondent 19)

This is anti-bot logic fully internalized as a writing value. A student who defines good writing as simple and containing deliberate errors has not just adopted an avoidance strategy; they have replaced the educational goal of writing with its institutional opposite. The fact that this response appears in the same dataset as Respondent 2's thoughtful account of evolving authorship standards shows the range of ideological response to the same institutional pressure. Some students are developing more sophisticated conceptions of writing; others have reduced the goal of academic writing to the mechanics of evading a machine.

Respondent 11 offers what is, in literary terms, the most eloquent account of human writing in the dataset, and does so without using any AI detection terminology:

"When a person is inspired to write something, they will go through periods between one writing session and another, and emotions differ from time to time. Sometimes a sentence comes out clear and understandable, and sometimes it

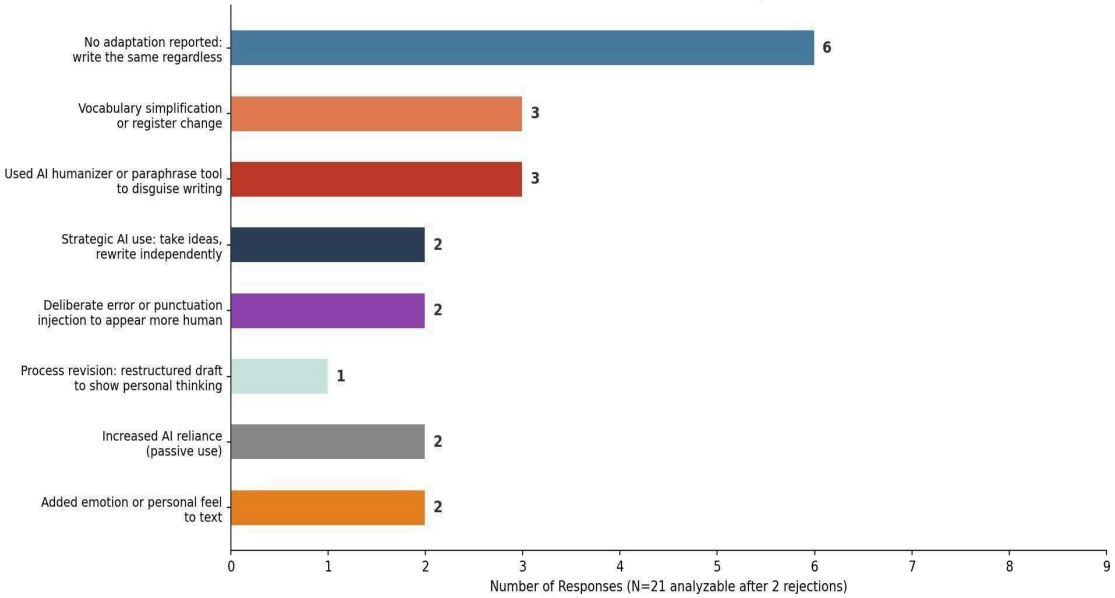
comes out more difficult and complex because the idea itself was hard to reach in the first place. And in between, the pen does not always hit the mark, which is why there may be some gaps." (Respondent 11)

This response describes, in plain and experiential language, the feature that computational linguistics calls burstiness: the natural variation in syntactic complexity and clarity that characterizes human writing across time. The student did not use the word burstiness. But they described it exactly, and their description locates that variation in the emotional and cognitive reality of writing, not in statistical metrics. It is precisely this kind of temporal, emotional, cognitively uneven prose that detection tools claim to identify as human, and yet the three Group 1 writing samples below will show those same tools failing to recognize it.

B/ OE2: A Specific Moment of Adaptation

The second open-ended item was designed to force concrete testimony. By asking for a specific moment rather than a general attitude, it required students to describe actual decisions they had made about actual words, sentences, and paragraphs. The twenty-one analyzable responses divide into five broad behavioral categories.

Figure 11: *OE2 Strategy Types. Six students reported no adaptation. The remaining fifteen describe a range of strategies from vocabulary simplification to deliberate error injection.*



Six students reported no adaptation whatsoever. Several of these responses carry a tone of confident resistance. Respondent 14 states: 'I have not changed my writing, as I trust my own voice is human enough.' Respondent 13: 'I have never changed my writing.' These non-adapters

are not naive; some of the same respondents scored above the midpoint on anxiety items in Section
B. Their refusal to adapt is a choice, not an absence of awareness.

Three students described vocabulary simplification or register change. Respondent 19's account is the most specific and the most instructive:

"I felt unjusticed and I was obliged to change transitional words as well as using simplified words instead of specialized terms despite it being related to my work, in order to not get accused of plagiarism." (Respondent 19)

Three things in this single sentence deserve attention. The student uses the word 'unjusticed,' which is not a standard English form, and that non-standard form communicates something no standard word quite captures: the felt experience of a justice violation. Second, they describe being 'obliged' to change their writing, not 'choosing' to. The adaptation was not voluntary; it was experienced as a form of institutional coercion. Third, they explicitly note that the specialized terms they replaced were appropriate to the subject matter. They were not removing bad language; they were removing correct language because it was too precise to look human.

Respondent 7's account confirms this same dynamic from a different angle. A supervisor, not even a detection tool, told this student that their writing looked like AI:

"I once changed my writing style because my supervisor said it was AI because I used some fancy words such as furthermore, to add to, and so on." (Respondent 7)

This is the formulaic language abandonment documented in Q12 (M=3.26), given a face, a specific relationship, and a specific institutional trigger. The suppression of standard academic transitional phrases is being enforced not only by algorithms but by educators who have internalized the algorithm's suspicion. When a supervisor tells a student that 'furthermore' looks like AI, the institution has, in effect, sent a single message: write less formally, write less carefully, write less like the academic writer we taught you to be.

Three students described using AI humanizer tools after writing. Respondent 3 is direct:

"When I first heard that our work would be scanned, I instantly had the idea of using a text humanizer, which will humanize the AI-written text and make it human-written text. Also, my friend advised me to use a certain website to paraphrase AI-written text to make it fully human writing." (Respondent 3)

Respondent 9 describes a version of the same behavior as a conditional strategy: 'if the AI detection thing is very rigid and mandatory,' they ask AI to make their text sound human. Respondent 15 wrote their response in Arabic, but its meaning is clear: they used an app to make their writing appear as human writing. All three confirm Educator 2's observation from the

interview about students using AI tools to disguise their own human work. The detection apparatus intended to reduce AI use has, in practice, generated a category of AI use that did not previously exist.

Two students described deliberate error or structural injection. These are the responses that connect most directly to the writing sample analysis that follows. Respondent 17:

"When I have to turn in a written assignment, I run it through an online AI detector and sometimes I deliberately leave in orthography mistakes or poorly worded sentences to avoid being wrongly accused." (Respondent 17)

Respondent 20:

"I changed the placement of punctuation marks or removed them altogether. I mixed human-written sentences with AI ones." (Respondent 20)

Respondent 17 describes a systematic pre-submission workflow: write, run through detector, deliberately worsen, resubmit. This is anti-bot practice as a structured personal process. Respondent 20's punctuation disruption is a precise match for one of the anti-bot markers the detector identifies in Group 2's writing samples below. These responses are not just qualitative data. They are the verbal description of the textual behavior that the writing sample analysis will document.

Table 4: *Summary of OE1 and OE2 Response Themes with Theoretical and Cross-Instrument Connections*

Theme	N Respon ses	Key Quotes or Examples	Theoretical Connection
OE1: Unchanged definition	6	Response 5: "writing is writing"	Ideological resistance (Spolsky: belief stability)
OE1: Humanness mu be proved	3	Response 2: thinking must "clearly belong to you"	Spolsky: language belief shift
OE1: Detection is unfair	3	Response 17: "perfect orthography = AI"	False-positive awareness (Liang al., 2023)
OE1: AI as legitimate tool	2	Response 9: fish/fishing metaphor	Active vs. passive AI engageme
OE1: Simplified/error definition	2	Response 19: "Simple, contain mistake	Anti-bot ideology fully internalized
OE2: No adaptation	6	Response 14: "I trust my own voice"	Resistance; non-compliance

OE2: Vocabulary simplification	3	Response 19: removed specialized ter	Q08 (M=3.35) corroborated

OE2: Used humanize tools	3	Response 3: instant use of humanizer	Educator 2 observation confirm
OE2: Deliberate erro injection	2	Response 17: leaves in orthography mistakes	Q13 (M=2.52) corroborated; lin to Group 2
OE2: Punctuation disruption	1	Response 20: changed punctuation placement	Anti-bot marker found in Text 5

3.3 Semi-Structured Interview Analysis

The semi-structured interviews generated two distinct data streams: a group discussion with four student participants, referred to throughout as A1, A2, A3, and A4, and individual interviews with two educator participants, referred to as Educator 1 and Educator 2. The interview questions were pre-planned but not rigidly administered; the conversations followed their own logic while staying anchored in the study's core themes. What follows is not a summary of the transcripts but a reading of them in light of what the questionnaire data has already established.

3.3.1 Student Group Discussion

The four student participants arrived at consistent positions through clearly different experiential routes, and their convergence is itself analytically significant. These are not students who share a single perspective because they have been taught one; they are students who have, through their own separate writing experiences, arrived at the same conclusions about what algorithmic surveillance does to the act of writing.

A/ The anxiety of innocence

A2's opening observation in the discussion is worth quoting directly:

"I do not even use AI to write my assignments, but I still get nervous. Like, what if the detector says I used AI when I did not?" (A2)

This single sentence gives the chilling effect of Q13 (M=3.65, 56.5% agreement) a face and a voice. The anxiety A2 describes is not guilt-based; it is structural. A student who has done nothing wrong and is still anxious about being accused is not being irrational. They are reading their situation

accurately: the system they are subject to makes mistakes, and those mistakes, as Q04 confirms, fall disproportionately on writers who produce the kind of text that detection algorithms associate with machines.

B/ The behavioral specifics

A3 and A4 articulate the adaptation behaviors of Section B with striking directness. A3 describes it this way:

"If I write a sentence that sounds really good, I start wondering if it looks too much like AI. Then I end up changing it." (A3)

This is the burstiness engineering behavior of Q09, described from the inside, with its affective cost intact. A4 goes a step further:

"If I use a fancy word or my grammar sounds too perfect, I change it because I do not want the detector to think it is AI." (A4)

This is vocabulary simplification (Q08) and felt fluency decline (Q12) in a single observation. Both students are aware that they are degrading their own work. A3 names it directly: it feels like we are spending more time trying to avoid being flagged than actually learning how to write better. This is an accurate description of a learning environment in which the primary feedback mechanism has shifted from pedagogical development to algorithmic compliance.

C/ The process-product argument

One of the most analytically important themes in the student discussion is an argument that A1, A2, and A4 develop independently but in parallel: detection tools see only the final text and have no access to the research, thinking, drafting, and intellectual work that produced it. A1 puts it this way:

"The final essay is only the end result. It does not show all the reading, note-taking, and planning that happened before." (A1)

A2 adds: A detector only looks at the final text. It cannot see all the time we spent researching and putting our ideas together. What these students are articulating, without using the academic vocabulary for it, is a fundamental epistemological limitation of text-based detection: a text is not equivalent to the process that produced it, and a tool that can evaluate only the text is an inadequate judge of the authorship question it claims to answer. The students have arrived at this conclusion without being taught it. They know it because they have lived it.

"Sometimes it feels like everyone is being treated as suspicious." (A4)

This observation from A4 names something important about the relationship between the institution and its students under the detect-and-punish framework. Educational institutions are built, at their best, on a presumption of good faith: the teacher assumes the student is trying to learn; the student assumes the teacher is trying to help. A system that treats all submitted writing as potentially fraudulent until proven otherwise has inverted that presumption. It has made suspicion the default. And as Q05 and Q13 confirm, students feel that default clearly and personally.

3.3.2 Educator Interviews

The two educator interviews produced data that complements and extends the student findings by providing the institutional view from the assessment side. Both educators had developed critical positions about the tools their institutions ask them to use, and both independently identified the same problems and proposed the same kinds of solutions.

A/ Educator 1: investigating instead of teaching

Educator 1's opening reflection describes what the algorithmic turn has done to the educator's relationship with student writing:

"Whenever I read a paper, part of me is wondering where the writing came from. Did the student write it? Did AI help? It can be frustrating because sometimes it feels like I am investigating instead of teaching." (Educator 1)

Before AI detection became common, reading student writing was primarily a communicative act: the educator encountered a student's thinking and responded to it. The detection context has inserted a prior question, a question of authenticity and origin, that must be answered before the communicative encounter can even begin. The educator is no longer reading primarily for what the student thinks; they are reading first for whether the student thought. This restructuring of the reading relationship is not neutral. It changes what the educator attends to, how they grade, and how they relate to their students.

Educator 1 then describes a direct experience of a false-positive incident, which is worth examining in detail. A student submitted a paper that the educator, from a full year of classroom observation, recognized as unmistakably that student's own work. The ideas, the examples, the way she expressed herself all matched what I had seen from her all year. But the detector gave it a very high AI score. This single case confirms the theoretical claim of this dissertation more concisely than any statistical analysis can. The tool was wrong. The educator's judgment was right. The institutional danger is not that educators like Educator 1 will be deceived by the tool; it is that not every educator has the pedagogical history with a student that Educator 1 had, the confidence to override a high score with their own judgment, or the institutional support to do so without consequence.

B/ Educator 2: the humanizer paradox

Educator 2 introduces a dimension that neither the questionnaire nor the student discussion fully captures: the emergence of AI humanization tools as a second-order consequence of detection pressure.

"Students are using things like humanizers and paraphrasing tools after they finish writing. Some of them are not even trying to cheat. They wrote the work themselves, but they are worried the detector might flag them by mistake, so they use those tools just to feel safer." (Educator 2)

The irony here is almost too precise to have been designed. The detection apparatus, intended to reduce AI use in student writing, has generated a new category of AI use among students who were not using AI in the first place. Students are using AI to make their human writing look less like AI, so that a machine designed to flag AI will not flag them. Educator 2 identifies this clearly: the detectors were supposed to reduce AI use, but in some cases they are actually encouraging students to use even more software. This is not a fringe behavior; it is a predictable institutional outcome of a surveillance system that creates incentives misaligned with its stated goals.

Educator 2 also speaks directly to the EFL dimension. Some students feel their writing gets flagged because it is more structured or follows patterns they learned in language classes. This observation, made from direct classroom experience, gives practical, observable grounding to the theoretical claim about interlanguage and false-positive rates that the literature review established from experimental research. The convergence between the theoretical claim, the experimental literature, the student questionnaire responses (Q02, Q18), and the educator's observational testimony is not coincidental. They are all describing the same structural problem from different vantage points.

Educator 2's final diagnosis is worth quoting because it matches the quantitative finding of Q15 so precisely:

"Instead of asking: Is my argument clear? They are asking: Will this get flagged? That is not really helping them develop as writers." (Educator 2)

The shift from asking whether an argument is clear to asking whether text will be flagged is the pragmatic rupture that Q15 (M=3.87, 73.9% agreement) measured at the population level. Educator 2 names it from the classroom. Both educators propose the same response to these problems: move institutional focus from detecting the final text to evaluating the writing process, through drafts, outlines, revision histories, and oral components. This proposed shift is not a technical adjustment to how detection is administered. It is a reconceptualization of what academic writing assessment is for, grounded in the recognition that a text and the process that

produced it are not the same thing, and that any system which can only evaluate the former is asking the wrong question.

Table 5: *Cross-instrument convergence. Each row maps a theme across all three data sources.*

Theme	Questionnaire	Student Interview	Educator Interview
Anxiety without use	Q13: M=3.65, 56.5	A2: still nervous without using AI	Implicit in observed student behavior
Vocabulary simplification	Q08: M=3.35, 52.2	A4: changes fancy words	Educator 1: papers sound more basic
Less fluent writing	Q12: M=3.26, 43.5	A3: more time avoiding flags than learning	Educator 1: students holding themselves back
Peer sharing of strategies	Q14: M=3.57, 56.5	Group discussion context	Not directly raised
Machine vs. human audience	Q15: M=3.87, 73.9	Described throughout discussion	Educator 2: wrong question being asked
Process not visible to detector	Not directly measured	A1, A2, A4: detector misses research process	Both educators: need process evidence
False positives happen	Q04: M=3.61, 56.5	A3: just a computer making a guess	Educator 1: direct false-positive experience
EFL writers more vulnerable	Q18: 26.1% agree	Not raised directly	Educator 2: structured patterns get flagged
Humanizer tool use	Not directly measured	Not raised	Educator 2: full description of the paradox

n

3.4 Integrated Discussion

The questionnaire and interview data, produced through different methods and from different positions, arrive at the same picture by different routes. That convergence is not accidental; it is the result of asking different kinds of questions about the same underlying reality. The picture has five features that, taken together, constitute the central empirical argument of this chapter.

First, algorithmic surveillance has penetrated student consciousness deeply enough to produce behavioral consequences across the majority of the sample. The 82.6% who confirmed awareness in the pre-questionnaire, the 52.2% who report consciously changing how they write (Q07), and the 56.5% who experience anxiety even without AI use (Q13) are not describing an edge case. They are describing a normal feature of assessed academic writing in this institutional context.

Second, the behavioral adaptations are coherent, rational, and costly. Students simplify vocabulary, engineer syntax, share avoidance strategies socially, and feel their writing declining in quality. They do these things not randomly but in direct response to the institutional signals they receive: be flagged and face serious consequences, or adapt and pay the silent cost of writing more poorly. Both educators confirm this pattern from the assessment side. The adaptation is real, it is observable in submitted work, and it is educationally damaging.

Third, the writing-only scope of detection creates a structural injustice that no one in this study's observational or interview settings appears to have formally raised. Students use AI to help with presentations, laboratory reports, research planning, and oral examination preparation, none of which is subject to algorithmic surveillance. The penalty for AI-related conduct falls only on students who produce their AI-assisted work in textual form. This is not a principled policy position; it is an artifact of what detection technology currently can and cannot do. Students understand this, as A1, A2, and A4's process-product observations confirm. The institution, as far as this study's data can tell, does not appear to have examined it.

Fourth, both educators independently confirm that the remedy lies in process-based assessment rather than product-based detection. This is a clear and actionable institutional conclusion. Drafts, outlines, revision histories, oral discussions, and in-class writing exercises all provide evidence about authorship that a detection score on a final text cannot provide. The fact that both educators reached this conclusion without prompting, and that it aligns with what students in the group discussion implied they would find fair, suggests a genuine institutional consensus waiting to be acted on.

Fifth, and most enduringly significant, Section C documents the beginning of an ideological shift that has not yet fully resolved. Students still believe in their own voice (Q16: 78.3%) and still reject the detector's authority over academic quality (Q17: 73.9%). But 43.5% already feel pressure to perform algorithmically acceptable English instead of developing their own scholarly voice (Q20), and 73.9% experience writing under detection as a fundamentally different act from writing for a human reader (Q15). These are not compatible long-term positions. If the institutional conditions that produce this tension are not addressed, the resolution is predictable: compliance wins, not because students chose it, but because they were given no institutional support for choosing anything else.

3.5 Observational Field Notes: Three Sites, One Pattern

Non-participant observation was the first data collection tool used in this study, and it served an exploratory rather than confirmatory purpose. The three sessions, one held in Sidi Bel Abbes, one in Oran, and one conducted virtually, were observed before the questionnaire and interviews were administered. The field notes taken at these sessions did not produce quantifiable data. What they produced was a structural understanding of how AI and academic integrity are being talked about in real Algerian academic settings, an understanding that shaped the direction and focus of everything that followed.

Across all three sites, the researcher attended as a non-participant, took detailed field notes during and immediately after each session, and wrote reflective memos to capture analytical impressions while they were still fresh. No questions were asked of participants; no discussion was initiated. The role was strictly observational.

A/ The Oran Workshop

The first observation took place at an academic workshop focused on AI in education. The session brought together educators from the local higher education community. The central organizing question of the workshop, as it emerged across the session, was how institutions should respond to AI use in student assessments. Discussion moved repeatedly toward detection tools: which tools to use, how to configure them, how high a detection percentage should trigger a formal misconduct review. These are legitimate institutional questions. But what the field notes record is an almost total absence of a different set of questions: who bears the cost of false positives, how EFL writers are specifically affected, whether the current policy is producing the writing culture the institution intends.

One moment in the session was particularly instructive. A participant asked how the institution should respond when a detection score was high but the educator personally believed the student had written the work themselves. The answer offered was that the detection score should be treated as the primary evidence, and that if an educator disagreed with it, they would need to produce supporting documentation for their judgment. This is the algorithmic turn described in the literature review made institutional policy: the machine's verdict is the default, and the human educator's judgment requires justification to override it.

B/ The Sidi Bel Abbes Workshop

The second observation followed a structurally similar pattern. The practical focus of the Sidi Bel Abbes session was on the administration of detection results: how to document flagged submissions, how to communicate results to students, what the formal misconduct process looks

like. These are procedural concerns, and they are real. What was absent, again, was any discussion of the accuracy limitations of the tools being administered, or of the populations within the student body for whom those limitations are most consequential.

A brief exchange between two participants caught the researcher's attention in the field notes. One educator mentioned that some students seemed to be producing writing that was deliberately informal or error-ridden, and wondered whether this was a sign of declining standards or something else. The question was not followed up. The session moved on. From the perspective of this study, that unanswered question is the most analytically significant moment of the Sidi Bel abbes observation: an educator noticed the behavioral consequence of detection pressure and then, in the absence of a conceptual framework for interpreting it, let it pass. The anti-bot register was visible to a practitioner. It was not recognized.

C/ The Virtual Session

The third observation was conducted online and drew participants from a wider range of institutional contexts than the two in-person sites. This diversity produced the most varied discussion of the three sessions. Several participants raised concerns about the pedagogical cost of detection-focused assessment, and one voice argued explicitly that institutions should be investing in AI literacy education rather than AI detection infrastructure. This position received a mixed response. Some participants agreed; others pushed back with practical concerns about academic integrity and the difficulty of monitoring AI use without detection tools.

The contrast between the virtual session and the two in-person Algerian workshops is itself a finding. In the two physical Algerian settings, the regulatory framing of AI, treating it as a threat to be managed, was nearly unanimous. In the virtual session, alternative framings circulated alongside it. This does not necessarily mean that Algerian higher education institutions are more regulatory than others; the sample is too small to support that conclusion. But it does suggest that the context within which a practitioner encounters the question of AI in education shapes how they think about it, and that the Algerian institutional settings observed in this study are, at present, primarily regulatory in their orientation.

3.5.1 The Observation That Connects All Three Sites

One structural pattern ran through all three observational sessions without exception, and it is the most important finding from the observational phase. In every session, at every site, artificial intelligence was discussed as a writing problem. The examples used were written essays, the tools discussed were text detection tools, the outcomes discussed were written assignment grades. Oral presentations, laboratory reports, research design, seminar discussion, problem-solving tasks: none of these appeared in any session as sites of concern about AI use.

This is not because AI use in those domains does not exist. Students use AI to prepare presentations, to structure arguments before seminars, to generate ideas for problem sets, to plan the structure of a research project. None of that use is visible to a detection tool, and so none of it generated institutional concern in any of the sessions observed. The entire discourse about AI integrity in Algerian higher education, as represented in these three settings, had collapsed into a discourse about text, because text is the only thing that can currently be detected.

The consequence of this collapse is a structural inequity that no participant in any session raised directly. A student who uses AI to prepare a presentation that receives high marks faces no detection penalty. A student who writes their own essay, but whose natural EFL style produces low perplexity and high formulaic density, faces a detection penalty they did not earn. The penalty is not calibrated to the behavior the institution claims to be governing. It is calibrated to what the available technology can measure. And what it can measure is text. So text-writers pay, and no one else does.

3.6 Writing Sample Analysis

The fourth and final data strand of this study consists of six authentic student-produced academic texts, three from each experimental group, analyzed using aidetector.com. This section explains the detection tool, presents the findings for each group in turn, and then reads the two groups together to produce the study's most direct empirical argument about the linguistic penalty and the anti-bot register.

3.6.1 The Detection Tool: aidetector.com

The AI detection tool used in this study is aidetector.com, a free, publicly accessible platform. The decision to use a free tool was not incidental; it reflects the real resource conditions of this research project, which had no access to institutional subscriptions to paid detection software. This decision is also, in its own way, a methodological argument: if free tools are what students encounter when they run their own writing through a detector before submitting it, then studying free tools is studying the actual conditions of the phenomenon under investigation.

Key technical claims of aidetector.com, as stated on the platform: (1) Accuracy rate of over 90%, using advanced algorithms to analyze text patterns and structures. (2) Detects content generated by ChatGPT, GPT-3, GPT-4, Bard, Claude, and other major models. (3) Breaks text into individual sentences, evaluating each one for AI-generated characteristics to provide a detailed, sentence-level analysis. (4) Widely used by educators and institutions to ensure academic integrity. (5) Uses sophisticated techniques to minimize false negatives while remaining effective against common evasion tactics. *(aidetector.com FAQs/About)*

The most analytically valuable feature of aidetector.com for this study is not the overall detection percentage but the sentence-level breakdown. The tool does not just say whether a text is AI-generated; it explains which sentences it found suspicious and why, in plain descriptive language. This sentence-level reasoning is what allows this analysis to move beyond a single score and into the specific linguistic features that trigger or suppress detection. For a study whose central argument is about which linguistic features get punished and which get rewarded, this feature is indispensable.

3.6.2 Group One: Human Writing That Looks Like AI

Three students were asked to write an academic text without using AI tools. They could use the internet for research. They could use grammar checkers for minor corrections. They were not told their texts would later be analyzed using AI detection. All three texts were submitted to aidetector.com. All three were flagged at or above 73%.

Figure 12: *AI Detection Scores for All Six Writing Samples. Group 1 texts (navy) all exceed 70%. Group 2 texts (orange) all fall below 20%. The dotted red line marks 75%, a common institutional concern threshold.*

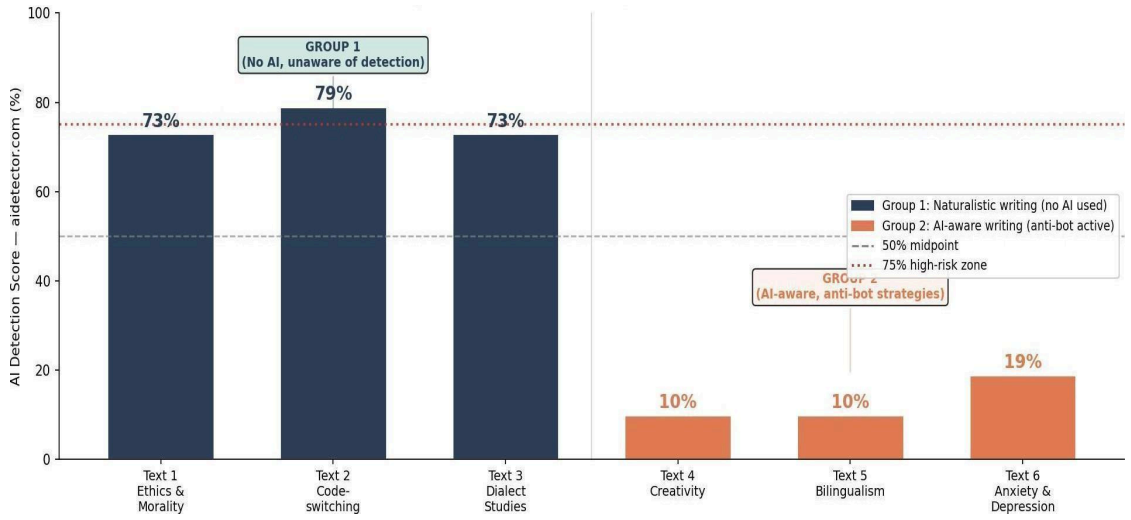
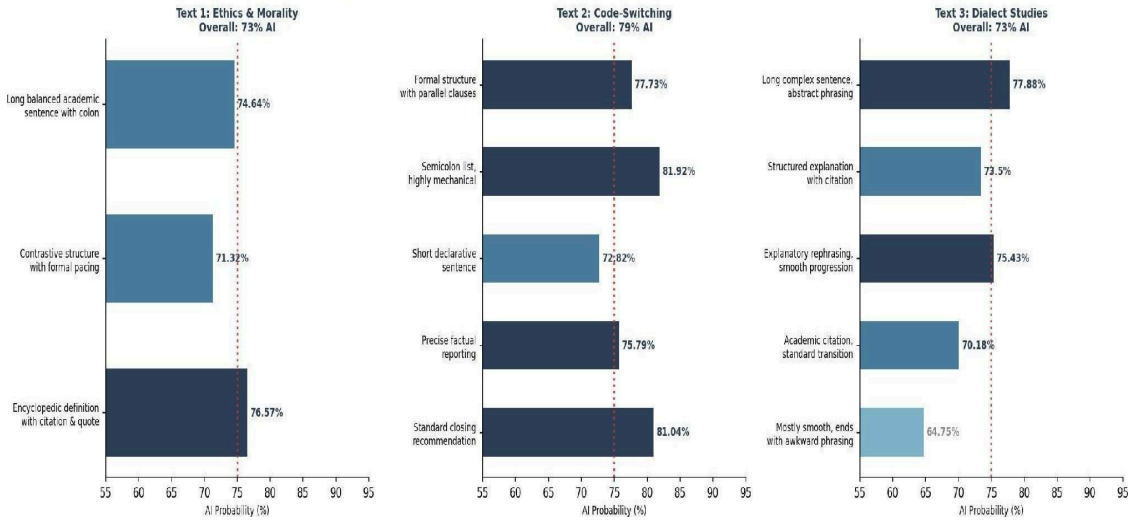


Figure 13: *Sentence-Level Detection Scores for Group 1. Every feature flagged as AI-like corresponds to standard characteristics of EFL academic writing instruction.*



A/ Text 1: Ethics and Morality (73% AI)

Text 1 is a short academic paragraph on the distinction between ethics and morality, written with clear definitional structure, a formal contrastive frame, and a direct citation with quotation. It reads as follows (first sentence): 'Ethics is often used in connection with the activities of organisations and with professional codes of conduct: for instance, medical and business ethics that often set a number of rules and guidelines stating how employees are expected to behave in their workplaces.'

The detector flagged this sentence at 74.64% AI probability, describing it as a 'long balanced academic sentence with colon and example.' The second sentence, a contrastive structure distinguishing morality from ethics, was flagged at 71.32% as a 'contrastive structure with even formal pacing.' The third sentence, which includes a direct quotation from Jackson (1984), was flagged at 76.57% as an 'encyclopedic definition with citation and quote.'

Every single feature that triggered a flag is a feature that academic writing instruction deliberately builds in students. The structured example after a colon. The formal contrast between two concepts. The integration of a cited definition with embedded quotation. These are not AI behaviors; they are the behaviors of a well-trained academic writer. The detection tool cannot tell the difference, and neither can the institutional policy that treats a high detection score as presumptive evidence of misconduct.

B/ Text 2: Code-Switching (79% AI)

Text 2 is a research summary on code-switching patterns, the highest-scoring text in the study at 79% AI. Its structure follows the standard research summary format taught in academic writing courses: methodological description, results, enumerated patterns, and a closing recommendation for further research. The detector's sentence-level findings are particularly detailed. A parallel clause structure was flagged at 77.73%. A semicolon-separated list of code-switching patterns, described by the tool as 'highly mechanical structure,' was flagged at 81.92%. A short declarative sentence in the middle of the results section was flagged at 72.82%. The precise factual reporting style was flagged at 75.79%. The standard closing recommendation sentence was flagged at 81.04%.

The 81.92% score for the semicolon list is worth focusing on. Semicolons used to separate items in a list are a punctuation feature that academic writing courses explicitly teach as a mark of formal academic register. The student who wrote this text was not imitating a machine; they were following a convention. The convention itself is what the machine flags. The 81.04% score for the closing recommendation is similarly revealing. Research summaries routinely close with a call for further research. That formulaic closure is required by the genre. The genre is what the machine punishes.

3/ Text 3: Dialect Studies (73% AI)

Text 3 is an analytical paragraph on the sociolinguistic definition of dialect, drawing on Labov (1966) and Dendane (1994). The paragraph moves carefully through a conceptual argument, using academic transitions and citation integration throughout. The detection scores range from 64.75% to 77.88%. Notably, the final sentence of the text, which the detector describes as 'mostly smooth but ends with an awkward phrasing,' scores the lowest at 64.75%. The slight loss of syntactic smoothness at the end of the paragraph, an imperfection in an otherwise well-constructed piece of academic writing, is what brings the score down. The rest of the paragraph, where the writing is most controlled and coherent, scores in the mid to high 70s.

This is the central paradox of Group 1 in its most visible form. The better the writing, the higher the detection score. The one moment of slight awkwardness, where the writer's phrasing becomes less polished, is the moment the machine grants even partial credit for humanity.

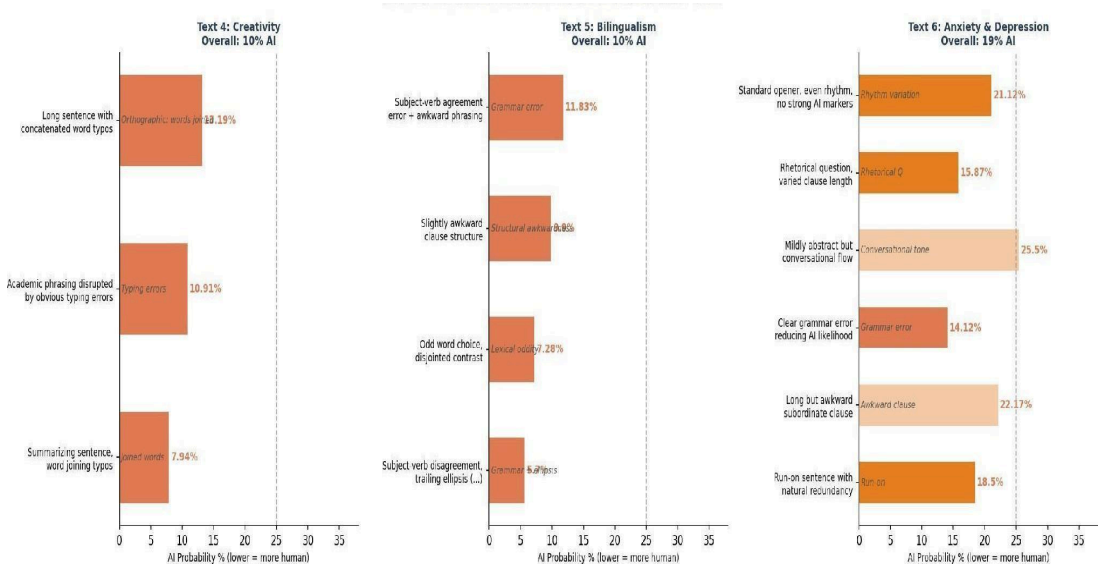
Table 6: *Group 1 Detection Results Summary. All three texts were written by students without AI assistance. All three exceeded 70% AI probability.*

Text	Topic	Overall Detection Score	Highest Sentence Score	Feature That Triggered Highest Flag
Text 1	Ethics & Morality	73%	76.57%	Encyclopedic definition with citation ar quote
Text 2	Code-Switching Research	79%	81.92%	Semicolon list, described as highly mechanical structure
Text 3	Dialect Studies	73%	77.88%	Long complex academic sentence with abstract phrasing

3.6.3 Group Two: Strategically Imperfect Writing That Looks Human

Three students were asked to write an academic text using any resources they chose, including AI tools. They were explicitly told before writing that their texts would be scanned by an AI detection tool. All three texts were submitted to aidetector.com. All three scored below 20%. The contrast with Group 1, which sits at a distance of fifty to sixty percentage points, is not gradual. It is a sharp structural break.

Figure 14: *Sentence-Level Detection Scores for Group 2.*



Text 4: Creativity (10% AI)

Text 4 is an academic paragraph on the definition of creativity, opening with a survey of historical perspectives and incorporating citations from Guilford (1950) and Rogers (1954). The overall score is 10%. The detector's sentence-level findings identify the cause with precision: 'long sentence with multiple concatenated word typos' (13.19%), 'academic phrasing disrupted by obvious typing errors' (10.91%), and 'summarizing sentence marred by multiple word joining typos' (7.94%).

The pattern is clear. The academic content is in place: the theoretical framing, the citation integration, the argumentative summary. But running through the text are typographic irregularities, words joined without spaces or broken in non-standard ways, that the detector reads as human imperfection. The content argues like an academic; the surface of the text looks like a human who typed carelessly. The detector trusts the surface. It does not read for meaning.

Now consider what happens when those typos are corrected. Based on the detector's own sentence-level explanations, correcting the orthographic errors would remove the only features identified as reducing AI probability. A version of Text 4 with perfect spelling and spacing would, by the detector's own logic, score significantly higher. Better writing means more suspicion. Worse writing means less. This is the anti-bot mechanism in its most transparent form.

Text 5: Bilingualism and Code-Switching (10% AI)

Text 5 opens with a long sentence defining bilingualism and plurilingualism, then moves into a discussion of code-switching, borrowing, and the differences between them. The overall score is 10%. The detector identifies four sources of reduced AI probability: a subject-verb agreement error combined with awkward phrasing (11.83%), a slightly awkward clause structure (9.9%), an odd word choice combined with a disjointed contrast (7.28%), and subject-verb disagreement paired with a trailing ellipsis (5.7%).

The subject-verb agreement error deserves specific attention. The relevant sentence reads: 'it must be noted that borrowing and codeswitching is not limited to standard varieties...' The subject 'borrowing and codeswitching' is plural in meaning but takes the singular verb 'is.' This is a grammar error: 'is' should be 'are.' A student writing without awareness of detection might catch and correct this error. A student writing with full awareness that errors lower their detection score has a rational reason to leave it in. The trailing ellipsis at the end of the text, which leaves the sentence unfinished, performs a similar function: it signals human incompleteness in a way that no AI system, trained to produce grammatically closed output, would produce.

If this text were corrected, the subject-verb error fixed, the trailing ellipsis removed, and the awkward clause restructured, the detector's rationale for scoring it at 10% would disappear. The corrected version would read as precise, formally constructed, and well-organized academic prose. It would, in other words, read like Group 1's texts. It would likely score in the same range.

Text 6: Anxiety and Depression in the Modern Age (19% AI)

Text 6 is the highest-scoring Group 2 text at 19%, and it is structurally the most interesting. It opens with a standard academic framing sentence, moves through a rhetorical question, then develops an argument across a series of long, cumulative sentences. The final sentence of the text is a single continuous construction that runs for two full lines, accumulating subordinate clauses in a way that reads as genuinely difficult to parse. The detector identifies this as a 'run-on explanatory sentence with natural redundancy' and scores it at 18.5%.

The grammar error in the text reads: 'the language through which psychological suffering are understood.' The subject 'suffering' is singular; the verb 'are' is plural. The detector identifies this as a 'clear grammar error reducing AI likelihood' and scores the sentence at 14.12%. The error is the evidence of humanity. Without it, the sentence would be grammatically correct, formally composed, and abstract in phrasing: exactly the kind of sentence that Text 1 and Text 3 of Group 1 produced at 73% to 77% AI probability.

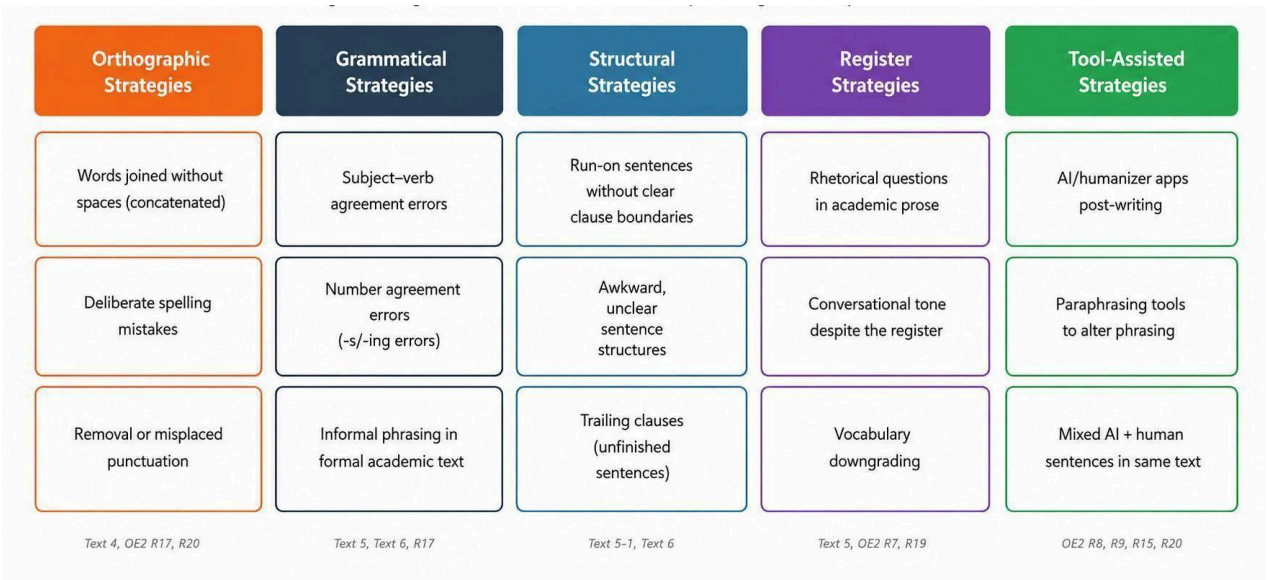
The rhetorical question in the second sentence, 'Does the modern age cultivate greater psychological vulnerability, or has the expansion of media and mental health awareness simply brought to light conditions that have always existed?' scores at 15.87%. Rhetorical questions are not a feature that language models typically produce, because they interrupt the forward propositional logic of AI text. A human writer uses a rhetorical question to engage the reader and create a dramatic pause in the argument. The detector recognizes this as human. But the question could just as easily have been written by a student who learned, through peer knowledge or trial and error, that rhetorical questions signal humanness to detection tools. The tool cannot tell the difference between a genuine rhetorical choice and a strategic one.

Table 7: *Group 2 Detection Results Summary.*

Text	Topic	Overall Score	Key Anti-Bot Markers	Effect If Corrected
Text 4	Creativity	10%	Concatenated word typos, joine words	Score would rise; errors are th only human signals
Text 5	Bilingualism	10%	"is" for "are" (S-V error), traili ellipsis (...)	Score would rise; grammar ellipsis are key human signals

Text 6	Anxiety & Depressior	19%	"are" for "is" (S-V error), run-c final sentence, rhetorical questi	Score would rise if errors corrected and structure tighten

Figure 15: *Taxonomy of Anti-Bot Language Markers. Five categories of deliberate linguistic strategies.*



3.6.4 The Comparative Finding: What the Two Groups Together Argue

Read side by side, the two groups produce a finding that the Likert data and interviews could predict but not demonstrate. Group 1 texts are better academic writing: better cited, better structured, more formally paced, more precisely argued. They score between 73% and 79% on the detection tool. Group 2 texts are worse academic writing: they contain grammar errors, typographic irregularities, run-on sentences, and structural awkwardness. They score between 10% and 19%.

The detection tool, in its current form, rewards worse writing. It penalizes better writing. A student who follows academic writing conventions precisely is more likely to be flagged as a machine than a student who introduces errors, avoids formulaic transitions, disrupts sentence structures, and uses punctuation inconsistently. This is not a marginal finding. It is a direct empirical refutation of the premise on which the detect-and-punish framework rests, the premise that detection tools can function as reliable judges of academic authenticity.

The comparison also reveals why the behaviors documented in Section B are rational rather than naive. Students who simplify their vocabulary (Q08: 52.2%), engineer syntactic variation (Q09: 52.2%), feel less fluent as a result (Q12: 43.5%), and share avoidance strategies

with peers (Q14: 56.5%) are not behaving irrationally. They are behaving exactly as a Group 2 writer would: introducing the features of imperfect human writing into text that would otherwise be flagged for being too good. The questionnaire documented the behavior at the attitudinal level. The writing samples document its linguistic reality at the textual level. Together, they confirm the anti-bot register as both a felt experience and a measurable set of linguistic choices.

One final question arises from this comparison, and it is worth stating plainly: what does it mean for academic writing education that the institutional tool for protecting academic integrity penalizes the linguistic features that academic writing instruction spends years building in students? The answer, supported by both data strands, is that the writing curriculum and the detection apparatus are pulling in opposite directions. Students who want to write well, and who have learned to write well, are more vulnerable to algorithmic misclassification than students who have learned to write strategically for a machine. Until institutions recognize this structural conflict and address it directly, the educational cost of AI detection will continue to fall on the students who can least afford to pay it.

3.7 Discussion of Main Findings

This section brings the findings of the chapter into direct dialogue with the two research questions that guided the study.Each research question is addressed in turn.

Research Question 01: What are the main linguistic and structural characteristics of the new anti-AI writing style, and do these adaptations align with traditional standards of academic writing quality?

The data collected in this study provide a clear answer to this question, it identifies this emerging register through intentional orthographic disruptions, grammatical irregularities, and structural fragmentations, alongside inconsistent register shifts and the use of external modification tools. These adaptations actively contradict traditional standards of academic quality, as the current detection environment rewards linguistic imprecision while penalising the precision and coherence typically expected in scholarly prose. These findings extend the work of Perkins et al. (2023); whereas they identified uniformity as a marker of machine-generated text, this study demonstrates that students are now adopting calculated irregularity as a counter-strategy. Furthermore, the results provide empirical weight to the conclusions of Weber-Wulff et al. (2023) regarding the unreliability of automated detection tools in high-stakes academic settings. This creates an institutional paradox where the disciplinary socialization and genre mastery described by Hyland (2009) have transitioned from academic assets into significant liabilities for the writer.

Research Question 02: How do students and educators perceive the ethical implications of the linguistic penalty in AI detection?

The second research question concerns perception and ethics rather than textual features. The data show that both students and educators perceive the linguistic penalty as fundamentally inequitable and educationally regressive, particularly for those writing in a second language. The ethical implications are manifested in heightened student anxiety and a professional shift for educators from pedagogical support to investigative suspicion. These perceptions corroborate the experimental findings of Liang et al. (2023) regarding the high frequency of false-positive results for non-native writers, while further revealing that students are consciously aware of this systemic bias. This adds a new dimension to the "unfair burden" identified by Flowerdew (2001), as algorithmic surveillance imposes a contemporary layer of structural disadvantage on non-native speakers. Ultimately, the documented experiences of both groups confirm the warnings of Selwyn (2016) that the shift toward automated management in education serves to reinforce existing social and linguistic inequalities.

3.8 Conclusion

This chapter has worked through four data sources in sequence: the Likert questionnaire, the interview responses, the observational field notes, and the writing samples. Each source has added something the others could not provide. The Likert data established population-level patterns of awareness, adaptation, anxiety, and ideology. The open-ended responses gave those patterns specific language: a student who defines good writing as simple and containing mistakes; another who identifies the false-positive mechanism without having read the literature; a third who describes a pre-submission workflow of deliberate self-degradation. The observational data placed those individual experiences inside a larger institutional conversation that is almost entirely regulatory in its framing and almost entirely blind to the inequities it is reproducing. The writing samples converted theoretical predictions into textual evidence: three human texts flagged above 70%, three AI-aware texts scored below 20%, with the errors and imperfections of the second group functioning as the proof of humanity that the precision and correctness of the first group denied them.

Together, these findings build the empirical foundation for this study's central argument: that AI detection tools have become de facto language managers in Algerian higher education; that EFL writers bear a disproportionate and largely unacknowledged burden under the current detection regime; that the behavioral response to this regime, anti-bot language, is a real, widespread, and measurable phenomenon; and that the institutional concentration of detection on written text, to the

exclusion of all other academic domains, is not a principled policy but a technological coincidence that happens to fall hardest on the students least able to absorb its costs.

GENERAL CONCLUSION

This dissertation set out to investigate a specific and practical problem: AI detection tools in higher education are supposed to protect academic integrity, but they appear to be penalizing the wrong students. The research has confirmed that concern. More than that, it has shown exactly how the penalisation works, what it looks like in practice, and who bears the cost.

The writing sample findings are the clearest evidence. Three EFL students wrote academic texts naturally, without AI assistance, and without knowing their texts would be scanned. All three texts were classified as AI-generated by aidetector.com. The features that triggered those flags were not invented or unusual. They were the features of good EFL academic writing: balanced sentence structure, citation integration, formulaic transitions, consistent register. Three other students wrote with full awareness that detection would follow. Their texts, built around deliberate errors, awkward phrasing, grammar mistakes, and structural irregularities.

The detection tool rewards writing that contains mistakes and punishes writing that does not. A student who has mastered academic English conventions is more likely to be flagged than a student who has deliberately broken them. This is not a failure in one tool on one day. It is a structural consequence of how perplexity-and-burstiness-based detection works, and it applies to any detection system built on the same logic.

The questionnaire and interview data showed what this does to people in practice. A majority of students in this study consciously change how they write when they know a machine will read their work. They simplify their vocabulary. They engineer variation in their sentence structure to appear more human. They share avoidance strategies with their classmates . They feel anxious about writing even when they have not used AI at all . And almost three quarters of them report that writing for a detector feels like a completely different act from writing for a human reader.

Both educators confirmed these patterns from the classroom side. One described watching students write more simply than they should. The other described students using AI humanization tools to disguise their own human writing from detection software, an irony so precise it could not have been invented: the system designed to reduce AI use is generating new AI use among students who were not using AI in the first place.

The observations at three Algerian academic sites added a final layer. At every site, AI was discussed exclusively as a writing problem. No session raised the question of AI use in oral presentations, laboratory work, or research planning, because detection tools cannot reach those domains. The penalty, in every institutional conversation observed, was understood to apply only to text. And the students who produce the kind of text that detection tools find most suspicious are, as this study has shown, the EFL students who have learned to write well.

The practical consequences follow directly from this. First, detection scores should not be treated as primary evidence of AI use. A score tells you what an algorithm thought of the text's statistical properties. It does not tell you whether a person wrote the text. Educators who trust the algorithm over their own professional knowledge of a student's writing history are making a category error: they are treating a probability output as a verdict.

Second, institutions that use detection tools without providing explicit guidance on those tools' known limitations are failing their students. Every student in this study who had heard that EFL writing gets flagged more often had found out through peer conversation, not through any institutional communication. That is an information failure with real academic consequences.

Third, the solution that both educators proposed, independently and without prompting, is the right one: shift assessment emphasis from the final submitted product to the process that produced it. Draft histories, annotated outlines, in-class writing exercises, oral defense components. These all provide evidence of authorship that a detection score cannot. They also happen to produce better writers. A policy that evaluates process rather than product serves academic integrity and academic development at the same time.

Three limitations shaped this research and should be clearly stated. The first is sample size. Twenty-three questionnaire participants and six writing samples is a small dataset. The patterns were internally consistent, the instrument reliability was acceptable, and the findings aligned across four different data sources. But no claim about the full Algerian EFL student population can rest on these numbers alone. The findings are suggestive, not definitive.

The second is geographic scope. The study was conducted entirely within Algerian higher education, a context defined by Arabic-French-English multilingualism, post-colonial educational structures, and a specific kind of prescriptive EFL instruction. These conditions shape the interlanguage that Algerian EFL students produce, and that interlanguage may differ from the interlanguage of EFL students in other national settings. The linguistic penalty argument is theoretically generalizable. The specific numbers may not be.

The third is timing. AI detection tools are updated continuously. The tools used in this study, and the specific linguistic features they flagged, reflect the state of detection technology at the point of data collection. As new language models are released and detection algorithms are retrained, the precise patterns of false-positive detection will shift. The broad argument should remain stable. The specific statistics will need revisiting.

A smaller practical limitation also deserves mention: the writing sample group was small by corpus linguistics standards. Three texts per group is enough to demonstrate the principle, but not enough to produce frequency data or generalizable statistics about the anti-bot register. This study identifies the register and documents its markers. It does not quantify their distribution.

Three directions follow directly from this study.The most needed is a larger replication. The same instruments applied to several hundred students across multiple Algerian universities, or across multiple national contexts, would allow the patterns observed here to be tested statistically. A comparison between Algerian EFL writing and native-speaker academic writing, scored through the same detection tools, would directly test whether the false-positive gap documented in the literature holds in this specific educational context.

The second direction is longitudinal. This study asked students how their writing had changed. It did not follow students through a semester or an academic year to observe that change as it happened. A study that collected writing samples from the same students at the beginning, middle, and end of a detection-heavy semester would show whether anti-bot adaptation accumulates over time or whether students find ways to stabilise their practice. It would also reveal whether the ideological shift documented in this study, the redefinition of good writing as safe writing, deepens as institutional pressure continues.

The third direction is institutional design. Both educators in this study proposed moving from product-based detection to process-based assessment. That shift is easy to recommend and genuinely difficult to implement. It requires changes to assessment rubrics, staff training, institutional policy, and possibly accreditation requirements. A study that worked directly with institutions to design, pilot, and evaluate process-based alternatives to AI detection would have direct policy value. It would also answer the question that this dissertation opens but does not close: not just what is wrong with current practice, but what a better practice would actually look like in the lecture hall.

Bibliography

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural language processing, and large language models in higher education and research. Research in Social and Administrative Pharmacy, 19(9), 1236–1242. https://doi.org/10.1016/j.sapharm.2023.05.016

An, Y., Yu, J. H., & James, S. (2025). Investigating higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration.International Journal of Educational Technology in Higher Education, 22(10). https://doi.org/10.1186/s41239-025-00507-3

Aydin, S., Karabacak, M., Vlachos, V., & Margetis, K. (2024). Large language models in patient education: A scoping review of applications in medicine. Frontiers in Medicine, 11. https://doi.org/10.3389/fmed.2024.1477898

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. Higher Education, 86(2), 369–385. https://doi.org/10.1007/s10734-022-00937-2

Beale, R. (2025). Adapting university policies for generative AI: Opportunities, challenges, and policy solutions in higher education [Preprint].

Brown, T. B., et al. (2020). Language models are few-shot learners. arXiv. https://arxiv.org/abs/2005.14165

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20(38). https://doi.org/10.1186/s41239-023-00408-3

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20(43). https://doi.org/10.1186/s41239-023-00411-8

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20(22). https://doi.org/10.1186/s41239-023-00392-8

Fisher, E. (2010). Contemporary technology discourse and the legitimation of capitalism. European Journal of Social Theory, 13(2), 229–252. https://doi.org/10.1177/1368431010362289

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation (pp. 68–88). Routledge.

Francis, N. J., Jones, S., & Smith, D. P. (2025). Generative AI in higher education: Balancing innovation and integrity. British Journal of Biomedical Science, 81, 14048. https://doi.org/10.3389/bjbs.2024.14048

Gourlay, L. (2025). The university and the algorithmic gaze. Bloomsbury Academic.

Hu, Y., et al. (2024). Can perplexity reflect large language model's ability in long text understanding? arXiv:2405.06105.

Jelinek, F., Bahl, L. R., & Mercer, R. L. (1977). Perplexity: A measure of the difficulty of speech recognition tasks. Journal of the Acoustical Society of America, 62(S1), S63.

Jurafsky, D., & Martin, J. H. (2021). Speech and language processing (3rd ed., Chapter 3: N-gram language models).

Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Kirti Verma, Khare, P., Shukla, M., & Jain, R. (2025). A statistical and probabilistic method for natural language processing (NLP). Journal of Statistics and Mathematical Engineering, 11(3), 23–32. https://matjournals.net/engineering/index.php/JOSME/article/view/2559

Lazaridou, A., et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. NeurIPS 2021.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press.

McConvey, K., & Guha, S. (2024). “This is not a data problem”: Algorithms and power in public higher education in Canada. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). https://doi.org/10.1145/3613904.3642451

Meister, C., & Cotterell, R. (2021). Language model evaluation beyond perplexity. arXiv:2106.00085.

Metych, M. (2025, March 9). De facto. Encyclopaedia Britannica. https://www.britannica.com/topic/de-facto

Naveed, H., Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2024). A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 16, 1–72. https://doi.org/10.1145/3744746

Peláez-Sánchez, I. C., Velarde-Camaqui, D., & Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: Exploring the connection between AI and Education

4.0. Frontiers in Education, 9, 1392091. https://doi.org/10.3389/feduc.2024.1392091

Prinsloo, P. (2020). Of 'black boxes' and algorithmic decision-making in (higher) education: A commentary. Big Data & Society, 7(1). https://doi.org/10.1177/2053951720933994

Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI.

Rof, A., Bikfalvi, A., & Marques, P. (2022). Pandemic-accelerated digital transformation of a born digital higher education institution towards a customized multimode learning strategy.

Educational Technology & Society, 25(1), 124–141. https://www.jstor.org/stable/48647035

Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics in Language Teaching, 10(3), 209–231.

Wang, C., et al. (2022). Perplexity by PLM is unreliable for evaluating text quality. arXiv:2210.05892.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Sigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text.

International Journal for Educational Integrity, 19,

1–39. https://doi.org/10.1007/s40979-023-00146-z

Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., & Ozcan, A. (2024). Large language model-based chatbots in higher education. Advanced Intelligent Systems, 7. https://doi.org/10.1002/aisy.202400429

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. International Journal of Educational Technology in Higher Education, 21(21). https://doi.org/10.1186/s41239-024-00453-6

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? International Journal of Educational Technology in Higher Education, 16(39). https://doi.org/10.1186/s41239-019-0171-0

Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2023). A survey of large language models. arXiv. https://arxiv.org/abs/2303.18223

Appendices

Appendix A: Student Questionnaire

Title: Algorithmic Surveillance, Academic Writing, and the De Facto Language Manager

Pre-Questionnaire: AI Detection Awareness

1. Do you understand the difference between AI detection tools and AI writing tools? [Yes / No]
2. Are you aware that your university uses AI detection tools? [Yes / No]
3. Do you understand how these AI detection tools work? [Yes / No / Unsure]

Likert-Scale Items

Instructions: Please indicate your level of agreement with the following statements. (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree)

Section A: Awareness of AI Detection Tools

Q01: My institution has communicated clear AI guidelines.
Q02: Detection tools accurately distinguish L2 (EFL) writing from AI-generated text.
Q03: My institution handles detection results fairly and protects students from wrongful accusations.
Q04: I am aware of instances where students have been wrongfully flagged as AI users. Q05: AI detection tools treat all students equally, regardless of their linguistic background.
Q06: Instructors, rather than automated tools, are the most appropriate judges of whether student writing is genuinely their own.

Section B: Writing Behavior Under Algorithmic Surveillance

Q07: I consciously change my writing when I know it will be scanned by an AI detector. Q08: I intentionally use simpler vocabulary to avoid being flagged.
Q09: I vary my sentence structure specifically to appear more "human" to a detection tool.

Q10: I refuse to change my academic vocabulary for detectors, even if it might get flagged.
Q11: I deliberately introduce grammatical mistakes or informal phrasing to lower my AI detection score.

Q12: I feel my writing has become less fluent or less natural as a result of trying to avoid AI detection.
Q13: The existence of AI detection tools makes me more anxious about writing assignments, even when I have not used AI.
Q14: I have discussed strategies for avoiding AI detection flags with other students. Q15: Writing for a detector's approval feels fundamentally different from writing to communicate ideas to a human reader.

Section C: Academic Voice and Linguistic Authenticity

Q16: My academic writing genuinely reflects my own voice, ideas, and style.
Q17: Writing that satisfies a detector's criteria for "humanness" is not necessarily the same as academically excellent writing.
Q18: As an EFL writer, I feel my natural style is more likely to be flagged as AI-generated compared to a native English speaker's writing.
Q19: AI detection tools improve the overall quality of student writing.
Q20: I feel pressure to perform "algorithmically acceptable" English rather than to develop my own authentic scholarly voice.
Q21: [Reverse Coded Check] AI detection tools improve the overall quality of student writing.

Open-Ended Items

OE1: What does good academic writing mean to you now, and has your definition changed since you started university?
OE2: Describe a specific moment when you changed your writing because you knew it would be scanned by an AI detector.

Appendix B: Semi-Structured Interview Protocols

Part 1: Student Interview Guide

1. Can you describe your direct experience with using, avoiding, or adapting to AI tools in your academic writing?
2. How aware are you of AI detection tools used at this institution, and what is your emotional response to knowing your work will be scanned?
3. Have you ever been flagged by an AI detector, or have you deliberately changed your writing to avoid being flagged? Please elaborate.
4. What are your views on institutional fairness and academic integrity in the current digital landscape?
5. Do you believe that algorithmic assessment is a legitimate method for evaluating your writing? Why or why not?

Part 2: Educator Interview Guide

1. What is your understanding of the rationale behind the institution's adoption of AI detection tools?
2. Based on your experience, how accurate are these AI detection tools, particularly when assessing EFL (English as a Foreign Language) students?
3. Have you observed any changes in student writing behavior or writing quality that you relate directly to the pressure of AI detection?
4. How do you integrate or override AI detection results when making your assessment judgments?
5. What are your thoughts regarding the equity and fairness implications of using AI detectors in your courses?

Appendix C: Non-Participant Observation Protocol

Settings: Academic workshops and professional academic discussion sessions focused on AI in higher education (In-person and Virtual).

Observational Focus Points:

1- Institutional Talk: Document how institutional actors discuss AI tools and detection practices in academic writing.
2- Regulatory Framing: Note the emphasis on procedural administration of detection results (e.g., misconduct thresholds, documentation).
3- Missing Discourses: Identify absences in the conversation regarding the limitations of detection accuracy, EFL vulnerabilities, and pedagogical costs.
4- Compliance vs. Resistance: Record mentions of student behavioral adaptations, including the emergence of anti-bot language or deliberate error injection.

Appendix D: Writing Sample Collection Prompts

Group 1: Naturalistic Writing Condition

Task:

Write a complete academic text (a short argumentative or expository essay) on a topic related to your field of study or any topic considered academic.
Parameters:
Prohibited from using AI writing tools (e.g., ChatGPT, Gemini).
Permitted to use the internet for research and AI-assisted grammar correction tools for minor edits.
Non-disclosure condition: Participants must NOT be informed that their texts will be analyzed using AI detection tools until after the collection (deferred consent).

Group 2: AI-Aware Writing Condition

Task:

Write an academic text using any resources chosen.
Parameters:
Permitted to use AI writing tools, LLMs, paraphrasing tools, or any combination.
Transparency condition: Participants MUST be explicitly told before writing that their texts will be rigorously examined using an AI detection tool to measure behavioral response.

Collected Texts Group 01:

Text 01 :

Ethics is often used in connection with the activities of organisations and with professional codes of conduct: for instance, medical and business ethics that often set a number of rules and guidelines stating how employees are expected to behave in their workplaces. Morality, however, is more often used in connection with the ways in which individuals conduct their personal, private lives, often in relation to personal financial probity, lawful conduct and acceptable standards of interpersonal behaviour. Morality, therefore, is “the character of being in accord with the principles or standards of right conduct” (Jackson, 1984, p. 319). Simply, morals define personal character while ethics stress a social system in which those morals are applied.

Text 02 :

Quantitative analysis was used to analyze the frequency of each pattern of code-switching, while qualitative analysis was used to determine the functions of code-switching. The results show several patterns of code-switching into English: code-switching from Arabic to English without translation; Arabic lexical items followed by an English translation; English lexical items followed by an Arabic translation; translation from Arabic into English in two turns; and metadiscursive code-switching. English lexical items are introduced through code-switching in each episode. English words without translation account for the highest percentage of
code-switching. In the code-switching to English, some English units are permanent, while some are context units that depend on the episode topic: these include basic formulaic and non-formulaic expressions. Lexical items for greeting, appreciation, and evaluation are the most frequent pragmatic functions of code-switching. Further research is recommended on code-switching in other TV animated series in other languages to determine the patterns of codeswitching and the part of speech that is the focus of switching.

Text 03

A dialect is no longer considered a language variety that is solely characterized by a regional limitedness that can be made geographically visible through a linguistically informed tracing of isoglosses on a map.Rather, for Labov (1966), the lines defining a dialect may exist not always between geographical regions necessarily but between social classes, levels of education, ethnicities, genders, ages, or other sociolinguistic factors.That is to say: if the statuses of standard varieties may be dictated by governments, those of nonstandard varieties are dictated by societies and while this latter form of dictation may be less observable, it is just as structured.Dendane (1994), however, suggests that the application of these findings to an Arabic context necessitates an added consideration of Bilingualism and Diglossia.This may be because both of these two phenomena are concerned with the convergence of language varieties at both a micro and macro level which is bound to impact the systems dictating them, however institutional.

Collected Texts Group 02:

Text 1 :

The concept of creativity has long been vaguely defined by scholars due to its multifaceted nature; however, approximative definitions proposed through years of research came to draw a clearer image of it. For instance, from previous perspectives on the concept creativity was equated with divergent thinking (Guilford, 1950); yet, such definitions were limited, opening space to vagueness rather than clarity which demanded further expansion of the concept.

According to Rogers (1954, p.250) creativity is “the emergence in the action of a novel relational product, growing out of the uniqueness of the individual on the one hand, and the materials, events, people, or circumstances of his life on the other”. This definition highlights that creativity is an action and process that comprises significant variables, such as novelty, uniqueness of the individual, availability of resources, time for growth and environment, which upon interacting with one another, produces creativity.

Text 02:

One of the most observable instance of two different language varieties coming into contact is bilingualism, multilingualism or plurilingualism which Myers-Scotton (2006) reports are all different terms researchers prefer for the same linguistic phenomenon: “speaking at least two languages” (p. 2) or more precisely being able to speak them sufficiently enough to casually converse even in a limited manner. The act, however, that a bilingual does in speaking is known as codeswitching which refers to using morphemes from both varieties in the same conversation, turn or clause. This act is considered unpredictable and foreign unlike borrowing which is also the use of morphemes from different varieties, but it can be done by a monolingual. Additionally, it must be noted that borrowing and codeswitching is not limited to standard varieties…

Text 03:

In recent years, anxiety and depression have emerged as central concerns in both academic and public discourse.Their increasing visibility prompts a crucial question: Does the modern age cultivate greater psychological vulnerability, or has the expansion of media and mental health awareness simply brought to light conditions that have always existed but remained largely unspoken?What was once submerged beneath the surface of social consciousness is now openly examined and debated.As Michel Foucault argued, societies do not merely discover phenomena, they also construct new ways of speaking about them, which suggest that the language through which psychological suffering are understood may be just as important as the conditions themselves.From this perspective, the prominence of anxiety and depression may signify not only a change in mental health trends, but also a transformation in how psychological suffering is perceived, categorized and discussed within different social and cultural contexts that often varies according to historical circumstances.In other words, the growing attention given to these issues may reflect both an actual increase in psychological distress and a greater willingness to acknowledge and talk about experiences that were often hidden or ignored in the past, and which for many individuals were not always recognized as legitimate forms of suffering.

ملخص:

تبحث هذه الدراسة الأكاديمية في أثر أدوات كشف الذكاء الاصطناعي والمراقبة الخوارزمية على سلوكيات الكتابة، والصوت الأكاديمي، والأصالة اللغوية لدى طلاب اللغة الإنجليزية كلغة ثانية أو أجنبية. وتعتمد الدراسة على منهجية مختلطة تكاملية تجمع بين الاستبيانات الإحصائية الكمية، والمقابلات شبه المنظمة مع الطلاب والمعلمين، والملاحظات الميدانية. وتكشف النتائج عما يُعرف بـ "مفارقة الصوت"، حيث يتأثر سلوك الطلاب بين التكيف والمقاومة الخوارزمية؛ كما أظهر التحليل التطبيقي للعينات الكتابية عبر أداة الكشف نمطين رئيسيين: كتابات بشرية تُصنّف خطأً كذكاء اصطناعي، وكتابات يتنّبى فيها الطلاب عيوباً لغوية استراتيجية ومقصودة للظهور بمظهر بشري وتجاوز الفلاتر الرقمية، مما يسلط الضوء على التداعيات والقيود العملية لهذه الأدوات في البيئة الأكاديمية.

Résumé

Cette étude académique examine l'impact des outils de détection d'intelligence artificielle et de la surveillance algorithmique sur les comportements d'écriture, la voix académique et l'authenticité linguistique des étudiants d'anglais langue seconde ou étrangère. La recherche s'appuie sur une méthodologie mixte intégrative combinant des questionnaires statistiques quantitatifs, des entretiens semi-directifs avec des étudiants et des enseignants, ainsi que des observations sur le terrain. Les résultats révèlent ce que l'on appelle le « paradoxe de la voix », où le comportement des étudiants oscille entre l'adaptation et la résistance algorithmique. En outre, l'analyse appliquée des échantillons d'écriture à l'aide de l'outil de détection a mis en évidence deux tendances principales : des textes rédigés par des humains mais classés à tort comme étant générés par l'IA, et des textes dans lesquels les étudiants introduisent intentionnellement des imperfections linguistiques stratégiques pour paraître plus humains et déjouer les filtres numériques. Ces éléments soulignent les implications pratiques et les .limites inhérentes à l'utilisation de ces outils dans le milieu académique