Journal of *Intelligence*

MDPI

Article

Cognitive Predictors of AI-Related Textual Change in Upper-Intermediate L2 Russian Writing: Working Memory, Attention Control, and Critical Thinking

Tian Ran and Hong Xu *

School of Russian and Eurasian Studies, Shanghai International Studies University, 1550 Wenxiang Road, Shanghai 201620, China; 0234101603@shisu.edu.cn
* Correspondence: katiaxu@shisu.edu.cn

Abstract

Generative artificial intelligence (AI) can alter final second-language (L2) texts, yet cognitive differences may shape which changes emerge. This exploratory study examined 38 upper-intermediate learners of Russian who each produced one composition with AI assistance and one without it. Working memory, attention control, and critical thinking were measured continuously, and 11 textual outcomes were calculated from paired compositions. AI-assisted texts showed higher lemma-based lexical-diversity values, longer words, more connectives per 100 words, and substantially fewer first-person markers, while total word count, sentence measures, and formulaic-expression rate did not differ in the full sample. The higher POS-based content-word density was attenuated after first-person forms were excluded, suggesting a contribution from the change in person marking. Exploratory models suggested negative working-memory associations with lexical-diversity and connective-rate change, and negative attention-control associations with lexical-diversity change. The positive working-memory coefficient for word-count change was sensitive to HC3 estimation and leave-one-out refitting. Critical thinking showed no clear independent association in the primary change-score models, while positive coefficients for MATTR and HD-D appeared in the outcome-conditioned models. Under this supervised, student-mediated procedure, AI assistance was associated with lexical, discourse-marking, and stance differences in completed L2 texts.

Keywords: generative artificial intelligence; AI-assisted L2 writing; working memory; attention control; critical thinking; lexical diversity; individual differences; upper-intermediate Russian learners

Check for updates

Received: 22 September 2026
Revised: 28 October 2026
Accepted: 2 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Generative artificial intelligence (GenAI) has become an active component of second-language (L2) writing. A writer can now obtain candidate ideas, lexical alternatives, syntactic formulations, discourse transitions, and extended passages within seconds. These resources change the conditions under which a text is produced and raise a central question for intelligence research: why do writers differ in the textual gains they obtain from the same class of cognitive technology? Current discussion has established the educational opportunities and cognitive risks of GenAI (Barrot, 2023; Feng et al., 2025; Godwin-Jones, 2022; Kasneci et al., 2023). Research on AI-assisted L2 writing has also documented changes in writing quality, linguistic complexity, revision, error correction, and learner experience (Barrot, 2023; Feng et al., 2025; Godwin-Jones, 2022). The individual cognitive capacities that

predict these changes remain considerably less understood. This gap matters because an AI-generated suggestion becomes part of a finished composition only after the writer maintains task goals, selects among alternatives, evaluates relevance, and integrates language into an emerging text.

Writing already places simultaneous demands on planning, idea generation, linguistic encoding, transcription, monitoring, and revision (Hayes, 2012; Kellogg, 2008). These demands become especially consequential in an L2, where lexical retrieval and grammatical formulation draw on resources that can be more effortful and less automatized than in a first language (Kormos, 2012, 2023). In the present AI-assisted condition, GenAI served as a dialogue partner that could explain, illustrate, or discuss language and content. Writers still had to maintain their communicative intentions, evaluate the responses, and formulate the submitted composition themselves. AI-assisted writing can be viewed as a form of distributed performance: writers combine their own composing with information discussed through an external AI system.

Three cognitive constructs are especially relevant to this coordination. Working-memory capacity supports the temporary maintenance and manipulation of task-relevant information (Baddeley, 2003; Unsworth & Engle, 2007). In writing, it helps writers preserve goals, propositions, lexical options, and sentence plans while text is being formulated and revised. Attention control supports goal maintenance, interference management, and the selective allocation of processing resources (Lee & Engle, 2026). In an AI-assisted condition, these functions can help a writer stay aligned with the composition goal while discussing competing suggestions. Critical thinking supports the evaluation of claims, evidence, inferences, and relations among ideas (Abrami et al., 2015; Halpern & Dunn, 2021). It can influence whether information and language discussed with the AI system are evaluated, adapted, or rejected before the student incorporates them into a developing argument. The three constructs overlap at the level of complex cognition, while each contributes a distinct explanatory emphasis: active maintenance and coordination, selective regulation, and reasoned evaluation.

Recent intelligence research has placed these constructs at the center of work on complex performance and technology-mediated learning. Zhao et al. (2026) showed that cognitive inference-making contributed to L2 writing quality alongside linguistic knowledge. Al-khresheh et al. (2026) modeled working memory and attention control in ChatGPT(GPT-4)-assisted foreign-language vocabulary learning and reported a structured relation between the two cognitive resources and learning outcomes. Yang et al. (2026) linked critical-thinking disposition to AI-supported L2 learning within a social-cognitive model. Other research has examined attention control and critical thinking as individual-difference constructs (Abrami et al., 2015; Halpern & Dunn, 2021; Lee & Engle, 2026). Collectively, this work supports a construct-centered approach in which cognitive abilities are measured as continuous individual-difference variables and connected to observable performance. Research on GenAI-supported learning also shows that motivational conditions shape engagement: autonomy, relatedness, and competence can mediate positive flow experiences during informal digital learning (Wu & Wang, 2025), while perceived ease of use and attitudes toward AI predict learners' intentions to adopt it (Wu et al., 2024). These findings place cognitive differences within a broader account of how learners engage with and benefit from AI. The present study extends this account to AI-assisted L2 composition and to the linguistic properties of completed texts.

Evidence from AI-assisted writing research further indicates that GenAI can alter several dimensions of a written product. AI-generated material may be incorporated, transformed, or used as a model for subsequent writing (Woo et al., 2024). Comparisons among self-revision, machine translation, and ChatGPT assistance have identified differences in

rater judgments, linguistic complexity, and error correction (Kim & Chon, 2025). Intervention and classroom studies have reported gains in academic writing and learner engagement (Mahapatra, 2024; Song & Song, 2023), while research on feedback and collaboration has shown that outcomes depend on how learners use and integrate AI support (Banihashem et al., 2024; Nguyen et al., 2024). Corpus comparisons also show systematic lexical and collocational differences between human L2 writing and ChatGPT-generated essays (Zhang & Crosthwaite, 2025). A recent systematic review of 112 studies concluded that AI affects writing processes, written products, and learner-related outcomes across diverse L2 contexts (Feng et al., 2025). These findings establish substantial variation at the product level. They also invite an individual-differences account of who shows larger changes in production, lexical diversity, information density, sentence packaging, and discourse connectedness.

The present study develops such an account in a sample of upper-intermediate learners of Russian as an L2. Each participant produced one composition with GenAI assistance and one composition without AI assistance. The analysis treats each writer's AI–no-AI contrast as an index of AI-related textual change. Working memory, attention control, and critical thinking are modeled as continuous predictors, preserving individual variation within a relatively homogeneous upper-intermediate-proficiency group. The textual outcomes cover production, lexical diversity and concentration, sentence-level packaging, discourse connection, stance, and formulaicity. The unit of explanation is the completed, student-authored composition. Teachers monitored all AI interactions throughout the session to ensure that participants used the system only for dialogue-based assistance, without requesting complete essays or copying AI output into their handwritten compositions. This design identifies cognitive differences in the extent to which supervised AI assistance is associated with measurable changes in the final text.

The study advances research in three ways. First, it connects intelligence constructs to a consequential form of technology-mediated language performance. Second, it examines the association patterns of three cognitive resources across textual dimensions. Third, it uses a within-writer contrast, which focuses the analysis on variation in AI-related change and reduces stable between-writer differences in general writing style. The theoretical framework anticipates a comparatively broad profile for working memory across production and linguistic elaboration, a selective profile for attention control in lexical and discourse regulation, and a discourse-oriented profile for critical thinking in information packaging and explicit connection. These construct-level expectations guide the analysis through exploratory research questions.

2. Theoretical Background and Literature Review

2.1. AI-Assisted L2 Writing as Cognitively Distributed Performance

Cognitive models describe writing as a recursive coordination of planning, formulation, monitoring, and revision (Hayes, 2012; Kellogg, 2008). The writer repeatedly moves between global goals and local linguistic decisions. At one moment, the writer may be maintaining a communicative intention; at the next, selecting a lexical item, completing a clause, checking coherence, or revising a previous sentence. L2 writing intensifies the coordination problem because linguistic encoding can consume a larger share of available resources (Kormos, 2012, 2023). This creates a performance environment in which individual differences in memory and executive regulation can shape both the amount of text produced and the linguistic form of that text.

GenAI changes the distribution of resources within this system. It can provide explanations, examples, lexical alternatives, and feedback that writers evaluate during dialogue before formulating their own text (Barrot, 2023; Godwin-Jones, 2022; Li, 2025). The writer remains responsible for the task goal, selection criteria, textual continuity, and final wording.

Each response expands the set of possibilities for discussion. That expansion can support performance when the writer coordinates it effectively, while the teacher's monitoring keeps the interaction within the prescribed consultation procedure.

Prior studies in which model-generated material was available illustrate several possible modes of conversion, including insertion, paraphrase, use as an organizational template, or use as a springboard for new text (Nguyen et al., 2024; Woo et al., 2024). In the present supervised procedure, students were prohibited from requesting a complete composition or copying AI output into the submitted text. They could consult the system about vocabulary, translation, ideas, and grammatical questions, but they had to formulate the submitted composition themselves under teacher monitoring. ChatGPT support has been associated with changes in complexity, error correction, writing scores, and motivation (Kim & Chon, 2025; Mahapatra, 2024; Song & Song, 2023). AI-generated feedback can also direct revision toward particular linguistic and rhetorical features (Banihashem et al., 2024). These effects vary across learners and tasks, which suggests that access to AI provides an opportunity structure whose realized value depends on the writer's capacity to coordinate, control, and evaluate the available material.

This account follows a broader individual-differences tradition that explains complex performance through measurable cognitive mechanisms. Lee and Engle (2026) emphasized attention control as a mechanism underlying performance across cognitively demanding domains. Applied to AI-assisted L2 writing, the same logic predicts stronger cognitive relations for textual outcomes that directly express the operations supported by each construct. Working memory should have wide relevance because active maintenance and coordination operate throughout composition. Attention control should become especially visible where the writer must select, inhibit, and sustain coherence. Critical thinking should become visible where the final text reflects evaluation, informational concentration, and explicit logical connection.

2.2. Product-Level Dimensions of AI-Related Textual Change

The present framework organizes the completed-text indicators into five functional domains, drawing on multidimensional accounts of complexity, accuracy, and fluency in L2 performance (Housen & Kuiken, 2009; Norris & Ortega, 2009). The first domain is production, represented primarily by word count. Sentence-level packaging is treated separately through sentences per 100 words and mean sentence length. Word count captures how much language reaches the final text, although it leaves the quality and organization of that language open. Production is cognitively relevant because sustained goal maintenance, retrieval, and formulation support the continuation of text generation.

The second domain is lexical differentiation and concentration. Lexical diversity was calculated on Russian lemmas, reducing inflation from inflectional variants. Guiraud's R relates the number of lemma types to the square root of token count (Jarvis, 2002), moving-average type–token ratio (MATTR) averages local diversity across 50-word windows, and HD-D estimates the expected contribution of lemma types in a 42-word sample (McCarthy & Jarvis, 2010). Mean word length provides a form-based index, while POS-based content-word density estimates the concentration of nouns, verbs, adjectives, adverbs, participles, and gerunds. Together, these measures capture complementary aspects of lexical range, local repetition, form, and informational concentration.

The third domain is sentence-level packaging, represented by sentences per 100 words and mean sentence length. Sentence rate captures the frequency of sentence boundaries relative to text length, whereas mean sentence length reflects how much material is organized within a sentence. These measures can change through clausal elaboration, phrasal expansion, coordination, or simple accumulation, so they remain broad indicators of pack-

aging across several possible syntactic structures (Ortega, 2003). Their relevance to the current cognitive account is clear: maintaining and integrating a larger set of elements within a sentence requires coordination across lexical, grammatical, and propositional levels. AI suggestions may supply longer or more elaborated structures, while the writer's cognitive resources influence whether those structures are integrated coherently into the final composition.

The fourth domain is discourse connectedness, represented by connective rate. Connectives make semantic relations more explicit and can signal addition, contrast, cause, sequence, or conclusion. Their frequency per 100 words offers a product-level indicator of overt relational marking. Cohesion research shows that connective use captures one component of textual connectedness and must be interpreted alongside the content and distribution of the relations expressed (Crossley et al., 2016).

The fifth domain includes stance and formulaicity. First-person marking combined pronouns and possessives with finite first-person verbs and was expressed per 100 words. Formulaic expressions were matched as contiguous lemma sequences from a fixed list and were also rate-normalized. These indices respond to genre conventions, stylistic preference, and available phraseology, so their cognitive associations may be narrower than those of production and lexical diversity (Bestgen, 2017).

This multidimensional organization follows the broader complexity–accuracy–fluency tradition while retaining the distinct meaning of each indicator (Housen & Kuiken, 2009; Norris & Ortega, 2009). It also reflects findings that writing development and writing quality can show different linguistic profiles (Crossley & McNamara, 2014). The current outcome set is intentionally broad enough to reveal whether the three cognitive predictors have common or differentiated associations with AI-related textual change. A cognitive profile spanning several domains would indicate a general coordination function. A profile concentrated in lexical or discourse measures would indicate a more selective regulatory or evaluative function.

2.3. Working Memory and AI-Related Change in L2 Writing

Working memory is a limited-capacity system that supports temporary storage, active maintenance, and manipulation during complex cognition (Baddeley, 2003). Individual-differences research further links working-memory capacity to the ability to keep goal-relevant representations accessible under interference (Unsworth & Engle, 2007). Writing draws on these functions continuously. Writers must maintain the intended message while retrieving words, construct a sentence while preserving its relation to prior discourse, and monitor the accumulating text while planning what comes next (Hayes, 2012; Kellogg, 2008). L2 formulation makes this coordination especially demanding because lexical and grammatical choices may require more deliberate processing (Kormos, 2012, 2023).

The empirical L2 writing literature generally supports a relationship between working memory and written performance, with the size and location of the relationship varying across measures and task conditions. Li's systematic review of 16 studies identified associations between working memory and several aspects of L2 writing, while also showing that measurement choices, proficiency, and task demands shape the observed pattern (Li, 2023). Michel et al. found that working memory contributed to young L2 learners' written performance (Michel et al., 2019). Manchón et al. demonstrated independent and interactive effects of working memory, L2 proficiency, and task complexity (Manchón et al., 2023). Révész et al. linked working memory to the temporal distribution of pausing and revision (Révész et al., 2023). Mavrou's study of L2 Spanish writing connected working memory and executive functions to complexity, accuracy, and fluency measures; updating emerged as an

important predictor of subordination density and accuracy, and the study also identified trade-offs between fluency and lexical complexity (Mavrou, 2020).

These findings support a broad working-memory profile in AI-assisted writing. At the production level, working memory can sustain message generation and preserve the next writing goal. At the lexical level, it can keep alternatives active for comparison and integration. The lemma-based diversity indices separate complementary aspects of lexical variety: Guiraud's R scales types to overall length, MATTR tracks local repetition across fixed windows, and HD-D estimates diversity in a fixed-size sample. At the discourse level, it can maintain prior propositions while the writer selects a connective that marks their relation.

AI assistance adds a source of explanations, examples, and language discussed in dialogue to these established writing demands. The cognitive benefit of this pool depends on the writer's capacity to keep the composition goal and existing context active while processing suggested material. Al-khresheh et al. (2026) reported direct and indirect relations among working memory, attention control, and vocabulary retention in ChatGPT-assisted foreign-language learning. Their study concerned vocabulary learning and used a structural model, yet its central mechanism transfers meaningfully to AI-assisted composition: externally supplied language still requires maintenance, selection, and integration. GenAI may ease lexical retrieval and sentence formulation, while working memory supports the integration of these contributions into coherent text.

Working memory is therefore expected to show the widest construct-level association across AI-related product changes. The strongest candidates include production, lexical differentiation, lexical concentration, sentence packaging, and connective use. This expectation concerns the breadth of the profile. The direction and magnitude of each indicator-level association are estimated empirically because AI can support expansion, compression, or reorganization depending on how the writer uses the available language.

2.4. Attention Control and Selective Regulation of AI-Supported Text

Attention control refers to the capacity to maintain goals, resist interference, inhibit dominant responses, and allocate attention to task-relevant information (Lee & Engle, 2026). It is closely connected to executive-function accounts that distinguish inhibition, shifting, and updating while recognizing their shared variance (Miyake et al., 2000). In complex skill performance, attention control helps individuals sustain the current objective and move efficiently between competing operations. Segalowitz and Frenkiel-Fishman linked attention shifting and control to L2 proficiency, providing early evidence that attentional regulation contributes to complex language performance (Segalowitz & Frenkiel-Fishman, 2005).

Recent intelligence research has strengthened the case for treating attention control as a central individual-difference mechanism. Lee and Engle (2026) reviewed evidence across learning, problem solving, retrieval, multitasking, and real-world performance and argued that attention control captures goal maintenance and interference management underlying many working-memory relations. Kane et al. (2001) linked controlled attention to individual differences in working-memory capacity. Shipstead et al. (2015) further showed that working-memory capacity reflects both the scope and the control of attention. This work suggests that attention control becomes especially important when a task presents several plausible representations and requires the individual to preserve a higher-order relation among them.

AI-assisted writing has precisely this structure. A fluent AI response can capture attention even when its content fits the text only partially. Several lexical or syntactic alternatives can compete for selection. The writer may also shift repeatedly between reading the AI response, evaluating it, and continuing the composition. Stronger attention control

can preserve the writing goal during these shifts, inhibit locally attractive material that disrupts coherence, and support sustained monitoring of the emerging text. Its effect should therefore be selective and strongest for outcomes that depend on regulation across competing alternatives.

Lexical diversity is one such outcome. Maintaining diversity across a text requires sensitivity to repetition and continued access to alternative expressions. Guiraud's R, MATTR, and HD-D capture this process at different scales without treating Russian inflectional forms as separate lexical items. POS-based content-word density provides a complementary measure of informational concentration, and connective rate indexes overt relational marking.

This profile distinguishes attention control from the broader maintenance function of working memory. The two constructs can correlate and operate jointly (Al-khresheh et al., 2026; Kane et al., 2001; Lee & Engle, 2026; Miyake et al., 2000). Their product-level signatures can still differ. Working memory can support the availability of multiple representations across many writing operations. Attention control can determine which representation receives priority and whether the current discourse goal survives interference from competing AI-generated language. The present study therefore examines whether attention control is concentrated in lexical stability, informational concentration, sentence packaging, and discourse connection.

2.5. Critical Thinking and the Evaluation of AI-Generated Language

Critical thinking involves the purposeful evaluation of claims, evidence, assumptions, inferences, and conclusions (Abrami et al., 2015; Halpern & Dunn, 2021). Halpern and Dunn placed critical thinking within an intelligence framework oriented toward solving consequential real-world problems (Halpern & Dunn, 2021). Abrami et al.'s meta-analysis linked critical-thinking development to explicit strategy instruction and guided practice (Abrami et al., 2015). Dwyer et al. proposed an integrated framework in which critical thinking involves metacognitive regulation, motivation, and the evaluation of evidence (Dwyer et al., 2014). This construct is highly relevant to GenAI because AI-generated language often arrives in a fluent form that can invite rapid acceptance.

In L2 writing, critical thinking has been associated with argumentative writing and writing performance. Pei et al. (2017) reported a relationship between critical thinking and argumentative writing among Chinese EFL learners. Indah (2017) found that critical thinking and topic familiarity were associated with Indonesian EFL learners' writing performance. In a technology-supported writing context, Liu et al. (2023) showed that an automatic evaluation and peer-assessment approach could engage critical-thinking tendency and deeper reflection. These studies connect critical thinking to the organization and evaluation of written content, while related work links it to cognitive ability, resistance to reasoning bias, and real-world decision making (Butler et al., 2017; Stanovich & West, 2008; Yang et al., 2026).

Critical thinking can shape an AI-assisted composition at three decision points. First, the writer evaluates the relevance and credibility of information and language discussed with the AI system. Second, the writer decides whether a proposed formulation expresses the intended relation accurately. Third, the writer integrates accepted material into a coherent line of thought. These decisions can leave product-level traces even when process data are unavailable. Lexical density can reflect the concentration of substantive content. Mean sentence length can reflect the packaging of related propositions into more elaborated units. Connective rate can reflect the explicit marking of relations among propositions. Each measure is a textual correlate of organization and informational structure; together they provide a focused window on how evaluative reasoning may be expressed in the completed composition.

Connectives have particular theoretical value for critical thinking. Reasoning depends on relations such as cause, contrast, qualification, sequence, and conclusion. A writer who evaluates the logical status of adjacent propositions can select markers that make those relations explicit. Critical thinking may therefore predict AI-related change in connective use when AI expands the set of candidate ideas and formulations. Lexical density and sentence packaging provide complementary information about whether those ideas are compressed into content-rich language and integrated into larger propositional units.

Research on GenAI-supported learning also highlights differences in how learners engage with AI. Yang et al. (2026) identified critical-thinking disposition as a pathway connecting AI-related learner characteristics and creativity in an L2 context. Nguyen et al. (2024) observed distinct interaction patterns in AI-assisted academic writing, with iterative collaboration associated with stronger writing performance. These findings motivate closer examination of the evaluative resources that writers bring to AI-supported tasks. The present study brings that construct into the linguistic analysis of L2 writing products. It examines whether critical-thinking ability predicts a concentrated pattern in information density, sentence-level elaboration, and discourse connection.

2.6. An Integrated Cognitive Profile of AI-Related Textual Change

The preceding sections identify complementary functions that converge during composition. Working memory maintains and coordinates representations across the text, attention control regulates competition among alternatives, and critical thinking evaluates the relevance and logical fit of candidate content. Together, these functions provide an integrated account of AI-assisted writing. A writer can draw on all three resources within the same sentence: working memory keeps the intended proposition and prior context accessible, attention control filters competing expressions, and critical thinking assesses whether the selected formulation advances a coherent relation.

The framework also predicts differentiation. Working memory should have the broadest reach because maintenance and coordination support production, lexical selection, sentence construction, and discourse continuity. Attention control should show a more selective profile centered on lexical regulation, information density, sentence packaging, and connective choice. Critical thinking should show its clearest profile in outcomes that express informational and relational organization, especially lexical density, mean sentence length, and connectives. Production volume, mean word length, first-person use, and formulaic phrase frequency remain useful comparison outcomes because they test the boundaries of each cognitive profile.

Two complementary mechanisms can generate these differentiated profiles. Cognitive amplification occurs when a stronger ability enables the writer to exploit a larger share of the linguistic and informational resources made available by AI. Cognitive compensation occurs when external support reduces a demand that constrains writers with a lower level of a relevant ability, allowing those writers to show a larger AI-related gain. A positive association between ability and beneficial textual change would be compatible with amplification, whereas larger gains among writers with lower ability would be compatible with compensation. The mechanism that dominates can vary across textual outcomes because production, lexical diversification, sentence packaging, and discourse marking impose different coordination, selection, and evaluation demands.

Restricting the sample to upper-intermediate L2 learners strengthens this design. A relatively narrow proficiency range reduces the extent to which broad linguistic-knowledge differences dominate the associations. Continuous cognitive scores preserve meaningful individual variation within that upper-intermediate group. The within-writer AI–no-AI contrast further isolates how each participant's textual profile changes when GenAI becomes

available. Together, these design choices support an exploratory examination of cognitive heterogeneity in AI-assisted L2 writing.

The study addresses the following research questions:

1. To what extent does working-memory capacity predict individual variation in AI-related changes across the production, lexical, sentence-level, discourse, stance, and formulaic characteristics of upper-intermediate L2 writing (AI minus no-AI)?
2. To what extent does attention control predict individual variation in these AI-related textual changes (AI minus no-AI)?
3. To what extent does critical-thinking ability predict individual variation in these AI-related textual changes (AI minus no-AI)?
4. What descriptive patterns of association are observed for working memory, attention control, and critical thinking across dimensions of AI-related textual change?

3. Method

3.1. Research Design

The study used a within-participant repeated-measures design combined with continuous individual-difference predictors. Each participant completed one L2 Russian composition with generative-AI assistance and one composition under a no-AI condition. The two texts were paired by participant identifier, yielding one AI–no-AI contrast for each textual measure and each writer. Working memory, attention control, and critical thinking were measured independently and entered as continuous predictors of these within-writer textual changes. This design retained the full range of cognitive scores and focused the analysis on heterogeneity in AI-related change among learners within the same upper-intermediate proficiency band.

3.2. Participants and Upper-Intermediate Proficiency Screening

Participants were recruited from undergraduate students enrolled in the Russian program at a university in Shanghai. The initial pool comprised 55 students. Ten were excluded because they did not meet the study's Russian-proficiency criterion, two because their prior Russian-study background did not match the target entry profile, three because teacher-monitored records showed that they had asked DeepSeek to generate a complete composition in the AI-assisted condition, and two because they used AI assistance during the no-AI condition. The resulting sample comprised 38 upper-intermediate learners of Russian as an L2, identified as H01–H38. Participants were 21–23 years old ($M = 22.00$, $SD = 0.62$), with 19 women and 19 men; this distribution arose naturally from recruitment rather than quota assignment. All participants had studied Russian for four years as part of their program. Their Russian proficiency was established through the prespecified background and performance criteria collected before the experimental writing sessions.

Course instructors administered and scored an onsite Test of Russian as a Foreign Language (TORFL)-based assessment for internal proficiency screening. Thirty-six participants were classified at B2 and two at C1. On a five-point self-rating scale covering listening, speaking, reading, writing, and translation, participants reported a mean Russian proficiency of 4.14 ($SD = 0.14$, range = 3.90–4.40). Before the experimental sessions, participants completed a separate Russian composition as a homework assignment. Two course instructors rated the compositions holistically on a 20-point scale using professional judgments informed by their teaching and marking experience. The final baseline-writing score was the mean of the two ratings. The mean baseline writing score was 16.184 ($SD = 2.179$, range = 10–19). This screening composition used a task separate from the two experimental compositions and therefore provided an independent descriptive measure of L2 writing performance; it was not entered into either inferential model family. Results from the national

Russian-major Band 4 and Band 8 examinations (ТРЯ-4 and ТРЯ-8) were collected as additional language-background information; all 38 participants were recorded as having passed both examinations. Every retained participant satisfied the study's upper-intermediate proficiency threshold.

ТРЯ-4 and ТРЯ-8 are curriculum-based national proficiency examinations for undergraduate Russian majors in China, normally taken in the second and fourth years, respectively. They are organized nationally through the Russian testing group of the foreign-language teaching assessment office and administered at participating universities. They assess attainment within the Chinese Russian-major curriculum.

3.3. Cognitive Measures

3.3.1. Working Memory

Working-memory capacity was assessed with an operation-span task and a symmetry-span task. Both are complex-span measures that combine temporary storage with an interleaved processing requirement (Redick et al., 2012; Unsworth et al., 2005). In the operation-span task, participants solved processing items while maintaining verbal items for subsequent serial recall. The raw score represented correctly recalled items, with a possible range of 0–75. In the symmetry-span task, participants made symmetry judgments while maintaining spatial locations for ordered recall. The raw score had a possible range of 0–42.

The two raw scores were standardized within the sample. The working-memory composite was calculated as the mean of the operation-span and symmetry-span z scores and then standardized again for regression analysis. Higher values represented stronger working-memory performance. The operation-span and symmetry-span scores were moderately correlated in the present sample ($r = 0.55$), supporting their combination as indicators of a shared complex-span construct.

3.3.2. Attention Control

Attention control was assessed with a Stroop task and a task-switching paradigm. The Stroop task measured the ability to manage interference from a competing response (Stroop, 1935). Two indicators were retained: interference cost in milliseconds and response accuracy. Smaller interference costs and higher accuracy indicated stronger performance. The task-switching paradigm measured the ability to shift between task rules while maintaining the currently relevant response set (Monsell, 2003). Switch cost in milliseconds and switching accuracy were retained, with smaller costs and higher accuracy indicating stronger control.

The four indicators were placed on a common directional scale before aggregation. Stroop interference and task-switch cost were reverse-coded, after which all four indicators were standardized. The attention-control composite was the mean of reverse-coded Stroop interference, Stroop accuracy, reverse-coded task-switch cost, and task-switch accuracy. The resulting composite was standardized for regression analysis, so higher values consistently represented more efficient attention control. In this sample, the standardized four-indicator composite had internal consistency of $\alpha = 0.804$ (standardized Cronbach's alpha).

3.3.3. Critical Thinking

Critical-thinking performance was assessed with Form D of the 40-item Watson–Glaser II Critical Thinking Appraisal (WG-II; Watson & Glaser, 2010). The test samples the evaluation of inferences, assumptions, deductions, interpretations, and arguments. It was administered in Chinese using a translation prepared by the research team. The total score was treated as an individual-difference measure of current evaluative performance: performance reflects accumulated learning and the efficiency with which participants

apply reasoning skills to unfamiliar items. Correct responses were summed to produce a raw score from 0 to 40, which was standardized within the sample and entered as a continuous predictor.

Table 1 presents the descriptive statistics for the observed cognitive-task scores.

Table 1. Descriptive statistics for the cognitive measures.

Construct and Observed Measure	Possible Range	M	SD	Observed Range
Working memory: operation span	0–75	50.95	8.06	30–64
Working memory: symmetry span	0–42	28.00	4.79	18–38
Attention control: Stroop interference (ms)	—	96.08	18.98	46–131
Attention control: Stroop accuracy	0–1	0.925	0.022	0.876–0.976
Attention control: task-switch cost (ms)	—	135.92	27.96	81–179
Attention control: task-switch accuracy	0–1	0.912	0.027	0.857–0.960
Critical thinking: Watson–Glaser total	0–40	25.95	3.56	17–34

3.4. Writing Tasks and Procedure

Participants completed the language-background questionnaire, the onsite proficiency assessments, and the three cognitive assessment components before the two experimental writing sessions. The experimental sessions followed the same general instruction and submission framework. Each participant produced one Russian composition in an AI-assisted condition and one in a no-AI condition. All formally included compositions focused on the same topic—the writer's hobby—which was held constant across the AI-assisted and no-AI conditions rather than manipulated experimentally. The writing prompt and instructions for both conditions are provided in the Supplementary Materials. Condition order was balanced with a two-sequence design (AI → no-AI versus no-AI → AI): participants were randomly allocated in blocks to the two sequences, with half completing AI-assisted writing first and half completing no-AI writing first. The two sessions were separated by one month.

In the no-AI condition, participants completed the writing task onsite without access to generative-AI assistance. In the AI-assisted condition, participants accessed the DeepSeek web interface using the DeepSeek-R1 model. The teacher remained present and monitored the interaction in real time. Students could ask questions about vocabulary, request translations of words or sentences, seek ideas for their composition, and ask for explanations or checks of grammatical and other language problems. They were explicitly prohibited from requesting a complete composition, and they were not permitted to copy AI output into the submission. Teacher monitoring included live observation during the session and random checks of the saved dialogue logs after the experiment. The final composition was handwritten and submitted by the student. Both tasks were completed within 40 min under standardized instructions. The dialogue logs were retained for procedural oversight but were outside the present textual and regression analyses.

One author transcribed the handwritten compositions into Word documents, and another author checked the transcriptions against the handwritten originals. Both transcription and checking were conducted without knowledge of the writing condition. Spelling, grammatical forms, and punctuation were retained as written. The main analyses used these transcribed texts, while the conservative spelling-sensitivity analysis used a separate corrected token sequence. Documents were linked using the participant identifiers embed-

ded in their filenames. Each participant contributed one document per condition, yielding 76 texts.

3.5. Text Preparation and Product Measures

Text extraction and metric calculation were performed locally in Python 3.12. Russian word tokens retained internal hyphens, were converted to lowercase, and normalized ë to e. Pymorphy3 supplied the highest-ranked lemma and part-of-speech analysis for each token, and Razdel segmented sentences. Main lexical-diversity values were calculated from the original token sequence after lemmatization. A separate conservative spelling-sensitivity pass corrected only context-checked, unambiguous misspellings and accidental word joins; proper names and ambiguous forms were retained. Ellipsis sequences, unparsed-token rates, and surface-form variation within lemmas were recorded as pipeline diagnostics.

Eleven measures were calculated from every completed composition. They were selected to represent production, lexical differentiation and concentration, sentence-level packaging, discourse connectedness, stance, and formulaicity. Table 2 provides their operational definitions.

Table 2. Textual outcome measures.

Domain	Measure	Operational Definition
Production	Word count	Number of Russian word tokens in the essay body
Lexical diversity	Lemma-based Guiraud's R	Lemma types divided by the square root of word tokens (Jarvis, 2002)
Lexical diversity	Lemma-based MATTR (50)	Mean lemma TTR across moving 50-word windows
Lexical diversity	Lemma-based HD-D (42)	Expected lemma-type contribution in a 42-word sample (McCarthy & Jarvis, 2010)
Lexical form	Mean word length	Mean number of token characters per word token, using the tokenization rules described above
Sentence packaging	Sentences per 100 words	Razdel sentence units divided by word count and multiplied by 100
Sentence packaging	Mean sentence length	Word tokens divided by sentence units
Lexical concentration	POS-based content-word density	Nouns, verbs, adjectives, adverbs, comparatives, participles, and gerunds divided by word tokens
Stance	First-person markers per 100 words	First-person pronouns/possessives and finite first-person verbs per 100 words
Discourse connection	Connectives per 100 words	Fixed-list connective occurrences per 100 words
Formulaicity	Formulaic expressions per 100 words	Fixed-list contiguous lemma-sequence matches per 100 words

Guiraud's R, MATTR, and HD-D were retained as complementary lemma-based lexical-diversity indices. MATTR used a 50-word moving window, and HD-D used a 42-word sample. Bidirectional lemma-based MTLD was also computed with the standard 0.72 threshold as a diagnostic, but it was not a primary outcome because 15 compositions produced no complete factor and nine were shorter than 100 words. POS-based lexical density classified nouns, finite and infinitive verbs, adjectives, adverbs, comparatives, participles, and gerunds as content words. Connectives, first-person forms, and formulaic expressions were reported per 100 words; finite first-person verbs were included in the stance measure,

and contiguous lemma-sequence matching captured inflectional variants of formulaic expressions. Density and mean word length were additionally recomputed after first-person markers were removed. The fixed inventories and statistical analysis code are supplied in Supplementary S6.

3.6. Statistical Analysis

The analysis proceeded at two levels. First, condition-level descriptive statistics were calculated for all 11 textual measures. Paired AI–no-AI differences were summarized with the mean difference, its 95% confidence interval, the paired-samples t statistic, and Cohen's dz. Positive differences indicated a higher value in the AI-assisted composition. Exact two-sided p values were reported, with .05 used as a nominal threshold in exploratory comparisons.

Second, the cognitive analysis examined whether participants' working memory, attention control, and critical thinking predicted the extent of their AI-related textual change. For participant i and textual measure j, the outcome was defined as

$$\Delta Y_{ij} = Y_{ij,\mathrm{AI}} - Y_{ij,\mathrm{noAI}}.$$

A separate multiple linear regression was estimated for each of the 11 textual outcomes:

$$\Delta Y_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{WM}_i) + \beta_{2j}(\mathrm{AC}_i) + \beta_{3j}(\mathrm{CT}_i) + \varepsilon_{ij}.$$

WM, AC, and CT were standardized before model estimation and retained as continuous variables throughout. The three-predictor model served as the primary analysis because the sample had already been restricted to a narrow upper-intermediate proficiency band. Each primary model used all 38 participants.

As a secondary analysis, each AI-condition textual outcome was regressed on the corresponding no-AI outcome and the same three standardized cognitive predictors: Y_i,AI = γ0 + γ1Y_i,noAI + γ2WM_i + γ3AC_i + γ4CT_i + ε_i. These outcome-specific models estimate associations conditional on the observed no-AI value. The independent screening-writing score was reported descriptively and was not included in either model family.

For every cognitive coefficient, the analysis reported the standardized coefficient, standard error, t statistic, exact two-sided p value, 95% confidence interval, and model R^2. All analyses were exploratory, and p values were reported without adjustment for multiple comparisons. Values below .05 were treated as nominal evidence for further investigation. Effect directions, confidence intervals, model fit, predictor collinearity, and sensitivity results were considered jointly when interpreting the coefficient profiles.

HC3 heteroskedasticity-consistent standard errors and leave-one-out refits were used to assess the sensitivity of the primary cognitive coefficients. Each leave-one-out refit used 37 participants, omitting a different participant in turn across 38 refits. Coefficient directions, magnitudes, confidence intervals, and p values were examined across these checks; all 38 participants remained in the primary analyses. RQ4 was addressed descriptively by summarizing the estimated association patterns across cognitive predictors and textual outcomes.

Pearson correlations between the AI and no-AI values were calculated for each outcome. A text-length sensitivity analysis retained writers whose compositions contained at least 100 Russian word tokens in both conditions (30 paired observations). Paired condition comparisons and the primary cognitive models were estimated in this subset, with continuous predictors and outcome differences standardized within the analytic sample. A separate sentence-level sensitivity analysis used 36 writer pairs, excluding the two pairs whose no-AI compositions contained only one or two detected sentence units. All 38 writers remained in the primary analyses.

3.7. Ethics and Data Handling

The study protocol was approved by the Ethics Committee of the host university. Participants provided informed consent before data collection, and participation was voluntary and uncompensated. Participant names were replaced with the identifiers H01–H38. The analytic dataset linked language-background information, cognitive scores, and essay-derived measures through these identifiers. Final reports contain aggregate statistics and no personally identifying information.

4. Results

4.1. Data Completeness and Predictor Diagnostics

All 38 participants contributed complete language-background, cognitive, and writing data, yielding 76 experimental compositions. The correlations among the three cognitive composites were modest: working memory correlated with attention control at r = 0.290 and with critical thinking at r = 0.313, and attention control correlated with critical thinking at r = 0.233. Variance-inflation factors in the primary three-predictor models ranged from 1.12 to 1.17, indicating that the working-memory, attention-control, and critical-thinking coefficients could be estimated with little inflation from predictor overlap.

4.2. Textual Differences Between the AI-Assisted and No-AI Conditions

Table 3 reports paired comparisons based on Russian lemma and POS analyses. In these exploratory, unadjusted comparisons, AI-assisted compositions showed higher Guiraud's R, $p = .008$, MATTR, $p = .048$, and HD-D, $p = .046$; longer words, $p < .001$; higher POS-based content-word density, $p = .023$; and more connectives per 100 words, $p < .001$. First-person markers per 100 words decreased sharply, $p < .001$. Word count, sentence rate, mean sentence length, and formulaic-expression rate did not meet the nominal .05 threshold.

The largest standardized differences were the reduction in first-person marking, dz = −1.43, the increase in mean word length, dz = 1.19, and the increase in connective rate, dz = 0.84. The lemma-diversity differences were positive across all three indices: dz = 0.45 for Guiraud's R, 0.33 for MATTR, and 0.34 for HD-D. Total word count differed by 5.95 words, 95% CI [−4.37, 16.26], $p = .250$.

Supplementary diagnostics support the operationalizations. No ellipsis sequences occurred in either condition. A conservative spelling pass applied 14 corrections in AI essays and 29 in no-AI essays; the Guiraud difference remained similar after correction (Mdiff = 0.373, $p = .007$). AI essays contained 2.53 more surface-form types in excess of lemma types, $p = .024$, whereas the proportion of lemmas realized in multiple forms did not differ, $p = .496$; primary diversity measures therefore used lemmas rather than surface forms. Lemma matching recovered 11 additional formulaic-expression occurrences in AI essays and 10 in no-AI essays, without producing a condition difference. After first-person markers were excluded, the mean-word-length difference remained 0.499 letters, 95% CI [0.343, 0.656], $p < .001$, whereas the POS–density difference was 0.008, 95% CI [−0.008, 0.024], $p = .300$. Bidirectional lemma MTLD was higher in the AI condition, but 8 AI and 7 no-AI essays produced no complete factor, so MTLD was retained as a diagnostic rather than a primary outcome.

The sentence counts in the two no-AI compositions with the fewest detected boundaries were confirmed on inspection: one contained 87 words in a single sentence unit and the other contained 93 words in two sentence units. In the 36-pair sentence-level sensitivity analysis, AI compositions had longer mean sentence units (mean difference = 1.759 words, 95% CI [0.346, 3.172], $p = .016$) and fewer sentences per 100 words (mean difference = −1.476, 95% CI [−2.464, −0.489], $p = .005$). The corresponding full-sample estimates were

−1.301 words, 95% CI [−6.042, 3.441], and −.963 sentences per 100 words, 95% CI [−2.148, 0.221], respectively (Supplementary S4i).

Table 3. Paired comparisons of textual outcomes in the AI-assisted and no-AI conditions.

Outcome	AI, M (SD)	No AI, M (SD)	AI − No AI, M [95% CI]	t(37)	*p*	dz
Word count	140.50 (37.15)	134.55 (28.77)	5.95 [−4.37, 16.26]	1.17	.250	0.19
Lemma-based Guiraud's R	7.973 (1.091)	7.610 (1.091)	0.363 [0.100, 0.626]	2.79	.008	0.45
Lemma-based MATTR (50 words)	0.820 (0.054)	0.799 (0.055)	0.021 [0.000, 0.041]	2.04	.048	0.33
Lemma-based HD-D (42 words)	0.834 (0.047)	0.815 (0.055)	0.018 [0.000, 0.037]	2.07	.046	0.34
Mean word length (letters)	5.521 (0.452)	4.927 (0.348)	0.594 [0.430, 0.759]	7.31	<.001	1.19
Sentences per 100 words	8.78 (2.31)	9.75 (3.41)	−0.96 [−2.15, 0.22]	−1.65	.108	−0.27
Mean sentence length (words)	12.27 (3.68)	13.57 (13.90)	−1.30 [−6.04, 3.44]	−0.56	.582	−0.09
POS-based content-word density	0.689 (0.036)	0.668 (0.040)	0.022 [0.003, 0.040]	2.37	.023	0.38
First-person markers per 100 words	5.70 (2.68)	13.92 (6.05)	−8.22 [−10.11, −6.33]	−8.81	<.001	−1.43
Connectives per 100 words	9.07 (2.56)	6.85 (1.74)	2.22 [1.35, 3.10]	5.16	<.001	0.84
Formulaic expressions per 100 words	0.89 (0.82)	0.93 (0.99)	−0.04 [−0.43, 0.34]	−0.23	.819	−0.04

Note. N = 38 paired observations. Positive differences indicate higher values in the AI-assisted condition. MATTR = moving-average type–token ratio; HD-D = hypergeometric distribution diversity; dz = paired-samples standardized mean difference. All *p* values are two-sided and unadjusted for multiple comparisons.

4.3. Primary Cognitive Models

The primary regressions used working memory, attention control, and critical thinking as simultaneous standardized predictors of each AI–no-AI difference score. Model R^2 values ranged from 0.013 for formulaic-expression rate to 0.303 for HD-D. The formulaic-expression and sentence-level models explained little variation (R^2 = 0.013–0.056), whereas the lexical-diversity models explained 24–30% of the observed change-score variance. Table 4 presents the complete coefficient set.

Working memory was positively associated with word-count change, β = 0.487, 95% CI [0.151, 0.824], *p* = .006. Its coefficients were negative for MATTR change, β = −0.411, 95% CI [−0.739, −0.082], *p* = .016, HD-D change, β = −0.385, 95% CI [−0.700, −0.069], *p* = .018, and connective-rate change, β = −0.365, 95% CI [−0.716, −0.013], *p* = .043. These estimates describe a positive word-count coefficient alongside negative coefficients for rate- and sample-controlled lexical diversity and explicit connective use.

Attention control was negatively associated with Guiraud's R change, β = −0.504, 95% CI [−0.822, −0.185], *p* = .003, and HD-D change, β = −0.335, 95% CI [−0.643, −0.027], *p* = .034. Critical thinking showed no clear independent association in the primary change-score models.

Table 4. Primary regressions predicting AI-related textual change from working memory, attention control, and critical thinking.

Outcome	Predictor	Standardized β	95% CI	*p*	Model R^2
Word count	Working memory	0.487	[0.151, 0.824]	.006	0.208
	Attention control	−0.188	[−0.516, 0.140]	.252	
	Critical thinking	−0.092	[−0.422, 0.239]	.577	
Lemma-based Guiraud's R	Working memory	−0.051	[−0.378, 0.275]	.750	0.254
	Attention control	−0.504	[−0.822, −0.185]	.003	
	Critical thinking	0.118	[−0.203, 0.439]	.461	
Lemma-based MATTR (50 words)	Working memory	−0.411	[−0.739, −0.082]	.016	0.243
	Attention control	−0.229	[−0.550, 0.092]	.156	
	Critical thinking	0.156	[−0.168, 0.479]	.335	
Lemma-based HD-D (42 words)	Working memory	−0.385	[−0.700, −0.069]	.018	0.303
	Attention control	−0.335	[−0.643, −0.027]	.034	
	Critical thinking	0.111	[−0.199, 0.421]	.471	
Mean word length (letters)	Working memory	−0.030	[−0.385, 0.324]	.862	0.121
	Attention control	−0.314	[−0.660, 0.032]	.074	
	Critical thinking	−0.067	[−0.415, 0.281]	.699	
Sentences per 100 words	Working memory	0.008	[−0.359, 0.374]	.967	0.056
	Attention control	0.168	[−0.191, 0.526]	.348	
	Critical thinking	0.128	[−0.233, 0.489]	.477	
Mean sentence length (words)	Working memory	−0.170	[−0.541, 0.200]	.356	0.039
	Attention control	−0.057	[−0.418, 0.305]	.752	
	Critical thinking	0.139	[−0.225, 0.504]	.442	
POS-based content-word density	Working memory	−0.138	[−0.487, 0.211]	.428	0.145
	Attention control	−0.240	[−0.581, 0.101]	.162	
	Critical thinking	−0.144	[−0.487, 0.200]	.401	
First-person markers per 100 words	Working memory	−0.010	[−0.348, 0.328]	.952	0.198
	Attention control	0.267	[−0.064, 0.597]	.110	
	Critical thinking	0.304	[−0.029, 0.637]	.072	
Connectives per 100 words	Working memory	−0.365	[−0.716, −0.013]	.043	0.132
	Attention control	0.207	[−0.136, 0.551]	.229	
	Critical thinking	0.131	[−0.215, 0.477]	.448	
Formulaic expressions per 100 words	Working memory	−0.016	[−0.391, 0.359]	.930	0.013
	Attention control	−0.062	[−0.428, 0.304]	.733	
	Critical thinking	−0.074	[−0.443, 0.296]	.688	

Note. Each row is a coefficient from an outcome-specific primary model containing all three cognitive predictors (N = 38). Outcomes are AI minus no AI. Predictors and outcome differences were standardized for β. MATTR = moving-average type–token ratio; HD-D = hypergeometric distribution diversity. Confidence intervals and *p* values in this table use conventional OLS standard errors; *p* values are two-sided and unadjusted. HC3 and leave-one-out results are reported in Supplementary S4e.

Together, the primary models yielded six cognitive coefficients with unadjusted p values below .05: four for working memory and two for attention control. These nominal associations were examined further through sensitivity analyses.

Sensitivity analyses indicated differing levels of precision across these associations. The working-memory coefficient for word-count change had an HC3 95% CI of [−0.221, 1.196], $p = .171$; the HC3 p value for connective-rate change was .082. Leave-one-out p values reached .104 for working memory–MATTR, 0.204 for working memory–connective rate, and .077 for attention control–HD-D. Working memory–HD-D and attention control–Guiraud's R retained p values below .05 under both checks. All six coefficient directions were preserved across leave-one-out refits (Supplementary S4e).

4.4. Outcome-Specific Models

Table 5 summarizes the cognitive coefficients from the outcome-specific models; full model estimates are provided in Supplementary S4c. In these secondary exploratory models, working memory was positively associated with AI-condition word count after conditioning on no-AI word count, attention control, and critical thinking, β = 0.383, 95% CI [0.099, 0.666], $p = .010$. Its coefficient for POS-based content-word density was negative, β = −0.434, 95% CI [−0.773, −0.094], $p = .014$. Attention control was negatively associated with AI-condition Guiraud's R, β = −0.289, 95% CI [−0.549, −0.028], $p = .031$. Critical thinking was positively associated with MATTR, β = 0.340, 95% CI [0.00003, 0.67954], $p = .04998$, and HD-D, β = 0.373, 95% CI [0.036, 0.710], $p = .031$. No other cognitive coefficient in these models had an unadjusted p value below .05. These estimates describe associations conditional on the observed value of the same textual measure in the no-AI condition.

Table 5. Cognitive coefficients in outcome-specific models predicting AI-condition textual outcomes.

Textual Outcome	Working Memory β (p)	Attention Control β (p)	Critical Thinking β (p)	Model R^2
Word count	0.383 (.010)	−0.155 (.257)	0.014 (.927)	0.466
Guiraud's R	−0.012 (.918)	−0.289 (.031)	0.169 (.202)	0.620
MATTR	−0.263 (.142)	0.143 (.475)	0.340 (.04998)	0.251
HD-D	−0.249 (.139)	−0.025 (.891)	0.373 (.031)	0.324
Mean word length (letters)	−0.140 (.452)	−0.113 (.570)	0.137 (.487)	0.097
Sentences per 100 words	−0.232 (.194)	−0.141 (.431)	−0.125 (.480)	0.167
Mean sentence length (words)	0.187 (.307)	0.165 (.347)	0.141 (.426)	0.124
POS-based content-word density	−0.434 (.014)	−0.068 (.685)	−0.058 (.725)	0.241
First-person markers per 100 words	0.099 (.579)	−0.126 (.502)	−0.005 (.977)	0.126
Connectives per 100 words	−0.284 (.113)	0.255 (.135)	0.129 (.445)	0.193
Formulaic expressions per 100 words	−0.049 (.789)	−0.212 (.241)	−0.021 (.905)	0.080

Note. Supplementary S4c reports all coefficients, 95% confidence intervals, and model statistics. Cells show standardized β coefficients with two-sided, unadjusted p values in parentheses. Each model includes the corresponding no-AI outcome, working memory, attention control, and critical thinking (N = 38).

4.5. Text-Length Sensitivity and Cross-Condition Correlations

The text-length sensitivity analysis included 30 writer pairs. All six cognitive coefficients with unadjusted p values below .05 in the full sample retained their direction and

remained below that threshold: working memory with word count, β = 0.524, 95% CI [0.131, 0.917], $p = .011$; MATTR, β = −0.407, 95% CI [−0.770, −0.044], $p = .029$; HD-D, β = −0.358, 95% CI [−0.702, −0.014], $p = .042$; and connective rate, β = −0.543, 95% CI [−0.916, −0.170], $p = .006$; attention control with Guiraud's R, β = −0.539, 95% CI [−0.894, −0.185], $p = .004$; and HD-D, β = −0.419, 95% CI [−0.745, −0.092], $p = .014$. The complete coefficient comparison is reported in Supplementary S4h.

In this 30-pair subset, AI-condition Guiraud's R, MATTR, and HD-D remained higher (mean differences = 0.489, 0.02384, and 0.02305; $p = .0014$, .035, and .027, respectively). Mean word length, POS-based density, connective rate, and first-person marking retained the same condition directions, while word count and formulaic-expression rate remained above the .05 threshold. Sentence units were longer in the AI condition (mean difference = 1.870 words, 95% CI [0.210, 3.530], $p = .029$), and sentence rate was lower (mean difference = −1.574 per 100 words, 95% CI [−2.675, −0.473], $p = .007$). Full paired comparisons are provided in Supplementary S4g.

Across all 38 writers, cross-condition correlations were r = 0.572 for word count, 0.731 for Guiraud's R, 0.334 for MATTR, 0.428 for HD-D, 0.237 for mean word length, 0.252 for sentence rate, −0.014 for mean sentence length, −0.101 for POS-based content-word density, 0.331 for first-person marking, 0.284 for connective rate, and 0.155 for formulaic-expression rate. Supplementary S4f reports confidence intervals and the corresponding correlations in the 30-pair subset.

5. Discussion

5.1. Overview of the Findings

This study examined how three cognitive measures related to AI-associated textual change among upper-intermediate L2 writers. The condition comparisons showed differences in lemma-based lexical diversity, word length, POS-based content-word density, connective rate, and first-person marking, while total length, sentence measures, and formulaic-expression rate did not show clear condition differences. The exploratory cognitive models yielded a positive working-memory coefficient for word-count change and negative coefficients for several rate-controlled measures, alongside negative attention-control coefficients for lexical-diversity change. Several associations were sensitive to the standard-error method or individual omissions. Critical thinking showed no clear independent association in the primary change-score models, while positive MATTR and HD-D coefficients appeared in the secondary outcome-specific models.

The baseline score provided a descriptive measure of participants' writing performance and was not entered into either inferential model family.

5.2. AI-Related Lexical, Discourse-Marking, and Stance Differences

The exploratory condition comparisons suggest that supervised AI assistance was associated with differences in the internal composition of student-authored texts. Three lemma-based indices—Guiraud's R, MATTR, and HD-D—showed higher diversity in the AI condition, and the result remained for Guiraud's R after the conservative spelling-sensitivity correction. Mean word length also increased substantially. These findings align with evidence that AI support can alter the lexical characteristics of L2 written products (Feng et al., 2025; Kim & Chon, 2025; Zhang & Crosthwaite, 2025). Because the students wrote the submissions themselves and copying AI output was prohibited under teacher monitoring, the contrasts characterize writing produced with supervised AI consultation rather than direct insertion of model-generated prose.

In the full sample, sentences per 100 words and mean sentence length did not differ clearly between conditions. The sentence-level and text-length sensitivity analyses yielded

longer sentence units and fewer sentences per 100 words in the AI condition. This pattern indicates that the sentence-level comparison depends on the inclusion of compositions with very few detected sentence boundaries.

The marked reduction in first-person marking indicates a shift in visible authorial stance. Both explicit first-person pronouns or possessives and finite first-person verbs were less frequent in the AI condition. Connective rate increased, whereas neither exact-form nor lemma-matched formulaic-expression rate changed. A diagnostic analysis that removed first-person items from the denominator and numerator reduced the content-word-density contrast to a nonsignificant difference, indicating that the density result partly reflects the change in first-person marking. Mean word length remained higher after the same exclusion, Mdiff = 0.499 letters, 95% CI [0.343, 0.656], dz = 1.05, $p < .001$, so the word-length result cannot be attributed to the loss of short first-person forms alone.

5.3. Working Memory and Productive Expansion

Working memory yielded nominal associations with several textual outcomes. The positive word-count coefficient was consistent with a possible production-amplification account involving active maintenance when writers coordinate ideas and language discussed during AI consultation (Baddeley, 2003; Hayes, 2012; Kellogg, 2008; Kormos, 2012, 2023; Unsworth & Engle, 2007). Its coefficients for MATTR, HD-D, and connective-rate change were negative. The word-count and connective-rate associations were sensitive to HC3 estimation, and several associations varied in precision across leave-one-out refits. Read together, these estimates suggest a possible combination of output expansion and smaller changes in fixed-window and fixed-sample lexical diversity and explicit relational marking.

This configuration suggests a possible production–diversity trade-off for further investigation. Writers with greater working-memory capacity may use AI consultation to sustain more content, while writers with lower scores may show larger relative changes in fixed-window and fixed-sample lexical variety. The connective-rate coefficient points in the same relative direction. These measures do not imply lower absolute performance among higher-working-memory writers; they describe how each writer's AI condition differed from that writer's no-AI condition.

Production amplification remains a possible interpretation of the positive word-count coefficient, although its sensitivity to HC3 estimation and leave-one-out refitting makes this interpretation tentative. The estimated pattern also varies with whether change is defined as additional output or as diversity and discourse marking relative to output.

5.4. Attention Control and Lexical-Diversity Change

Attention control showed negative associations with change in Guiraud's R and HD-D. Lower-scoring participants displayed larger lexical-diversity increases when AI was available. The Guiraud's R association was more consistent across the sensitivity checks, whereas the HD-D association was sensitive to individual omissions. This exploratory pattern is compatible with compensation or scaffolding accounts in which dialogue-based support reduces the burden of internally generating and regulating lexical alternatives (Barrot, 2023; Godwin-Jones, 2022; Kim & Chon, 2025; Mahapatra, 2024).

The outcome-specific model also yielded a negative Guiraud's R coefficient. After AI-condition Guiraud's R was predicted from the corresponding no-AI value and all three cognitive measures, attention control remained negatively associated with the AI outcome, $\beta = -0.289$, $p = .031$. The Guiraud relation remained after conditioning on the observed no-AI value, but this adjustment does not rule out regression-to-the-mean effects arising from measurement error or carryover from the repeated topic. The HD-D association

did not persist in that model, suggesting that the conditional association was localized to Guiraud's R.

This interpretation also clarifies how attention control can matter when students consult AI without copying its output. Goal maintenance and interference control remain important as writers move between a dialogue response and their own developing text (Kane et al., 2001; Lee & Engle, 2026; Miyake et al., 2000; Segalowitz & Frenkiel-Fishman, 2005; Shipstead et al., 2015). In the present data, the observable relation was concentrated in lexical differentiation.

5.5. Critical Thinking and Product-Level Textual Change

Critical thinking showed no clear independent association with AI-minus-no-AI change in the primary models. In the outcome-specific models, however, higher Watson–Glaser scores were associated with higher AI-condition MATTR and HD-D after accounting for the corresponding no-AI values and the other cognitive predictors. These specifications address different aspects of performance: the primary models concern the observed within-writer difference, whereas the secondary models concern AI performance conditional on no-AI performance. The positive conditional coefficients therefore locate the critical-thinking pattern in adjusted lexical diversity rather than a consistent association across both model families. Critical thinking may contribute to the evaluation of lexical alternatives during AI consultation; this interpretation provides a hypothesis for future process-based research.

5.6. Implications for Cognitive Accounts of AI-Assisted Writing

The exploratory findings provide a descriptive account of cognitive associations across textual outcomes and model specifications. The primary estimates suggested working-memory associations with production and lexical-diversity change and attention-control associations with lexical-diversity change. The outcome-conditioned models also yielded positive critical-thinking coefficients for MATTR and HD-D and a negative attention-control coefficient for Guiraud's R. The consistency of these patterns varied across sensitivity checks. They provide hypotheses for further tests of amplification and compensation in AI-assisted writing.

Continuous measurement retained individual variation in cognitive abilities. The observed slopes describe graded relations within a deliberately restricted upper-intermediate proficiency group. Dividing participants into high and low groups would discard variation, reduce power, and obscure variation across textual outcomes. Future studies can strengthen this approach with larger samples, latent-variable measurement based on multiple tasks per construct, and preregistered outcome families. Such designs would permit direct comparisons of amplification and compensation across lexical, syntactic, and discourse outcomes.

For L2 writing pedagogy and assessment, comparable text length can conceal changes in lexical diversity, discourse marking, and stance. Instruction can begin with a brief goal statement before AI use, require learners to compare two or three alternative formulations, and ask them to annotate which claims or propositions each suggestion supports. A final revision pass can mark connective relations and identify language retained, adapted, or rejected. These activities make goal maintenance, lexical comparison, and propositional fit visible in classroom practice.

5.7. Scope, Future Research, and Conclusion

Several design features define the scope of our conclusions. The sample contained 38 upper-intermediate learners of Russian, and each participant contributed one composition per condition on the same hobby-introduction prompt. The two-sequence order was intended to be balanced by random block allocation, but participant-level sequence effects

were not modeled. The repeated-topic design and one-month interval reduce immediate repetition concerns but cannot rule out carryover. The narrow proficiency band places lower-proficiency populations outside the present generalization. Replication with larger samples, repeated texts per condition, process records, and outcome-specific baselines would provide more precise writer-level estimates.

The sample of 38 writers limits the precision of the cognitive coefficients and the ability to detect smaller associations. The analyses were exploratory and retained unadjusted p values to identify candidate patterns for further study alongside effect sizes, confidence intervals, and sensitivity results. With 11 condition comparisons and 33 primary cognitive coefficients, this approach leaves an increased risk of false-positive findings and does not control family-wise error or the false discovery rate. Associations near the nominal .05 threshold, and those sensitive to HC3 estimation or individual omissions, therefore require particular caution. Larger, preregistered studies with clearly specified outcome families are needed to evaluate their reproducibility.

The linguistic measures also vary in resolution. The short compositions and use of a single text per condition may limit the reliability of the textual measures and their difference scores. The 100-word sensitivity analysis assesses dependence on text length but does not establish difference-score reliability. Lemmatization reduces inflation from Russian inflection, and rate normalization separates discourse and stance counts from text length. MTLD was retained only as a diagnostic because many short compositions formed no complete factor. No ellipsis sequences occurred in the analyzed essays. Conservative spelling correction changed the magnitude of Guiraud's R only minimally, and lemma-based formulaic matching recovered contiguous inflectional variants without changing the null condition result. The formulaic-expression measure does not capture discontinuous templates and therefore remains a bounded inventory measure. The fixed inventories, diagnostic outputs, and statistical analysis code accompany the manuscript.

Within these boundaries, the study documents a product-level pattern of AI-associated change in upper-intermediate L2 writing under supervised, student-mediated assistance. AI consultation was associated with higher lemma-based lexical diversity, longer words, more connective marking, and less first-person marking without a reliable increase in total word count. The POS–density contrast was attenuated after first-person forms were removed, suggesting that person-marking composition contributes to this difference. The cognitive analyses identified exploratory association patterns that varied across textual outcomes and model specifications, with several primary coefficients sensitive to HC3 estimation or individual omissions. These findings motivate further investigation of how cognitive resources relate to AI-assisted writing.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jintelligence14110286/s1, The supporting information comprises six files: S1, Language Background (Anonymized); S2, Cognitive Abilities (Anonymized); S3, Recomputed Essay Metrics; S4, Cognitive–Essay Statistics (worksheets S4a, condition effects; S4b, primary models; S4c, outcome-specific models; S4d, pipeline diagnostics; S4e, influence sensitivity; S4f, cross-condition correlations; S4g, short-text paired comparisons; S4h, short-text cognitive models; and S4i, sentence-level sensitivity); S5, Russian Writing Task Instructions; and S6, Analysis Code and Methods.

Author Contributions: T.R.: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Software, Writing—original draft, Writing—review and editing. H.X.: Conceptualization, Methodology, Supervision, Project administration, Resources, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the open project Cognitive Neural Mechanism of Russian Vocabulary Learning (2021KFKT001) of Shanghai Key Laboratory of Brain-Machine Intelligence for Information Behavior.

Institutional Review Board Statement: The study protocol was approved by the Ethics Committee of Shanghai International Studies University. The study involved adult participants and no invasive or physiological procedures.

Informed Consent Statement: Written informed consent was obtained from all participants before data collection. Participation was voluntary and uncompensated.

Data Availability Statement: De-identified language-background and cognitive-ability data, essay-derived measures, statistical results, and analysis code are provided in Supplementary S1–S5. Essay-level source files are not publicly shared because of participant privacy.

Acknowledgments: The authors would like to thank all the participants in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AI	artificial intelligence
GenAI	generative artificial intelligence
L2	second language
MATTR	moving-average type–token ratio
HD-D	hypergeometric distribution diversity
MTLD	measure of textual lexical diversity (diagnostic only)

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*(2), 275–314. [CrossRef]

Al-khresheh, M. H., Almayez, M., & Alruwaili, S. F. (2026). Working memory, attention control, and vocabulary retention in AI (ChatGPT)-assisted foreign language learning: A structural cognitive modelling approach. *Journal of Intelligence*, *14*(4), 62. [CrossRef] [PubMed]

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. [CrossRef] [PubMed]

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1), 23. [CrossRef]

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. [CrossRef]

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65–78. [CrossRef]

Butler, H. A., Pentoney, C., & Bong, M. (2017). Predicting real-world outcomes: Critical thinking ability is a better predictor of life decisions than intelligence. *Thinking Skills and Creativity*, *25*, 38–46. [CrossRef]

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. [CrossRef] [PubMed]

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. [CrossRef]

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, *12*, 43–52. [CrossRef]

Feng, H., Li, K., & Zhang, L. J. (2025). What does AI bring to second language writing? A systematic review (2014–2024). *Language Learning & Technology*, *29*(1), 1–27. [CrossRef]

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, *26*(2), 5–24. [CrossRef]

Halpern, D. F., & Dunn, D. S. (2021). Critical thinking: A model of intelligence for solving real-world problems. *Journal of Intelligence*, *9*(2), 22. [CrossRef] [PubMed]

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388. [CrossRef]

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. [CrossRef]

Indah, R. N. (2017). Critical thinking, writing performance and topic familiarity of Indonesian EFL learners. *Journal of Language Teaching and Research*, *8*(2), 229–236. [CrossRef]

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. [CrossRef]

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*(2), 169–183. [CrossRef] [PubMed]

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. [CrossRef]

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26. [CrossRef]

Kim, M., & Chon, Y. V. (2025). The impact of self-revision, machine translation, and ChatGPT on L2 writing: Raters' assessments, linguistic complexity, and error correction. *Assessing Writing*, *65*, 100950. [CrossRef]

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing*, *21*(4), 390–403. [CrossRef]

Kormos, J. (2023). The role of cognitive factors in second language writing and writing to learn a second language. *Studies in Second Language Acquisition*, *45*(3), 622–646. [CrossRef]

Lee, Y., & Engle, R. (2026). Beyond working memory capacity: Attention control as the underlying mechanism of cognitive abilities. *Journal of Intelligence*, *14*(2), 22. [CrossRef] [PubMed]

Li, S. (2023). Working memory and second language writing: A systematic review. *Studies in Second Language Acquisition*, *45*(3), 647–679. [CrossRef]

Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature*, *2*(1), 122–152. [CrossRef]

Liu, C.-C., Liu, S.-J., Hwang, G.-J., Tu, Y.-F., Wang, Y., & Wang, N. (2023). Engaging EFL students' critical thinking tendency and in-depth reflection in technology-based writing contexts: A peer assessment-incorporated automatic evaluation approach. *Education and Information Technologies*, *28*(10), 13027–13052. [CrossRef]

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), 9. [CrossRef]

Manchón, R. M., McBride, S., Mellado Martínez, M. D., & Vasylets, O. (2023). Working memory, L2 proficiency, and task complexity: Independent and interactive effects on L2 written performance. *Studies in Second Language Acquisition*, *45*(3), 737–764. [CrossRef]

Mavrou, I. (2020). Working memory, executive functions, and emotional intelligence in second language writing. *Journal of Second Language Writing*, *50*, 100758. [CrossRef]

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. [CrossRef] [PubMed]

Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learners' written performances. *Journal of Second Language Writing*, *45*, 31–45. [CrossRef]

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. [CrossRef] [PubMed]

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. [CrossRef] [PubMed]

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human–AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864. [CrossRef]

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. [CrossRef]

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. [CrossRef]

Pei, Z., Zheng, C., Zhang, M., & Liu, F. (2017). Critical thinking and argumentative writing: Inspecting the association among EFL learners in China. *English Language Teaching*, *10*(10), 31–42. [CrossRef]

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171. [CrossRef]

Révész, A., Michel, M., & Lee, M. (2023). Exploring the relationship of working memory to the temporal distribution of pausing and revision behaviors during L2 writing. *Studies in Second Language Acquisition*, *45*(3), 680–709. [CrossRef]

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition*, *33*(4), 644–653. [CrossRef] [PubMed]

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2015). Working memory capacity and the scope and control of attention. *Attention, Perception, & Psychophysics*, *77*(6), 1863–1880. [CrossRef] [PubMed]

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843. [CrossRef] [PubMed]

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695. [CrossRef] [PubMed]

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. [CrossRef]

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. [CrossRef] [PubMed]

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. [CrossRef] [PubMed]

Watson, G., & Glaser, E. M. (2010). *Watson–Glaser II critical thinking appraisal: Technical manual and user's guide*. NCS Pearson.

Woo, D. J., Susanto, H., Yeung, C. H., Guo, K., & Fung, A. K. Y. (2024). Exploring AI-generated text in student writing: How does AI help? *Language Learning & Technology*, *28*(2), 183–209. [CrossRef]

Wu, H., & Wang, Y. (2025). Disclosing Chinese college students' flow experience in GenAI-assisted informal digital learning of English: A self-determination theory perspective. *Learning and Motivation*, *90*, 102134. [CrossRef]

Wu, H., Wang, Y., & Wang, Y. (2024). "To use or not to use?" A mixed-methods study on the determinants of EFL college learners' behavioral intention to use AI in the distributed learning context. *International Review of Research in Open and Distributed Learning*, *25*(3), 158–178. [CrossRef]

Yang, Y., Zhang, S., & Li, Y. (2026). The influence of AI on critical thinking and creativity in L2 learning contexts: A social cognitive perspective. *Journal of Intelligence*, *14*(5), 78. [CrossRef] [PubMed]

Zhang, M., & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching*, *64*(2), 733–760. [CrossRef]

Zhao, A., Chen, F., & Li, X. (2026). The concurrent and longitudinal contributions of linguistic and cognitive skills to L2 writing quality. *Journal of Intelligence*, *14*(1), 11. [CrossRef] [PubMed]

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