

# ELITE-AI

## Guidelines for Ensuring Quality in Language Acquisition with Integrated AI

---

### Quality Assurance Manual

---

Partner institutions: UNITBV (Romania) · UM (Slovenia) · UNIUD (Italy) · UNIPU (Croatia)

Erasmus+ Grant No. 2024-1-RO01-KA220-HED-000248845

[elite-ai.unitbv.ro](http://elite-ai.unitbv.ro)

## Table of Contents

|                                                                 |           |
|-----------------------------------------------------------------|-----------|
| <b>1. Purpose and Scope of This Manual .....</b>                | <b>4</b>  |
| <b>2. Theoretical Framework .....</b>                           | <b>5</b>  |
| 2.1 English for Specific Purposes .....                         | 5         |
| 2.2 Task-Based Language Teaching .....                          | 5         |
| 2.3 Learner Autonomy and AI Interaction.....                    | 6         |
| 2.4 AI-Assisted Language Learning.....                          | 6         |
| <b>3. Experimental Design and Project Structure.....</b>        | <b>8</b>  |
| 3.1 The Work Package 3 Protocol .....                           | 8         |
| 3.2 Overview of Partner Courses and AI Tool Rotations .....     | 8         |
| 3.3 Thematic Coherence Across the Consortium .....              | 9         |
| <b>4. Pedagogical Implementation Protocols .....</b>            | <b>10</b> |
| 4.1 Pre-Lesson Planning: Tool Selection and Prompt Design ..... | 10        |
| 4.2 In-Session Management: The Teacher’s Role .....             | 11        |
| 4.3 Critical Engagement with AI Output .....                    | 12        |
| 4.4 Progression in AI Tool Rotation .....                       | 12        |
| <b>5. Assessment and Evaluation Instruments .....</b>           | <b>14</b> |
| 5.1 The Observation Sheet Framework .....                       | 14        |
| 5.2 Evaluating Content Delivery Effectiveness.....              | 15        |
| 5.3 Adaptability and Differentiation .....                      | 15        |
| 5.4 Feedback Mechanisms and Self-Regulation.....                | 15        |
| 5.5 Learner Motivation and Participation Data .....             | 16        |
| <b>6. Professional Development Guidelines.....</b>              | <b>17</b> |
| 6.1 Core Competencies for AI-Integrated ESP Teaching .....      | 17        |
| 6.2 Reflective Practice and Documentation.....                  | 18        |
| 6.3 Collaborative Review within the Consortium .....            | 18        |
| <b>7. Integrated Quality Criteria.....</b>                      | <b>19</b> |
| 7.1 Pedagogical Alignment .....                                 | 19        |
| 7.2 Linguistic Accuracy and Register Appropriateness .....      | 19        |
| 7.3 Learner Agency .....                                        | 19        |
| 7.4 Documentation Consistency.....                              | 19        |
| 7.5 Ethical Implementation .....                                | 20        |
| <b>8. Summary of Recommendations .....</b>                      | <b>21</b> |
| 8.1 Pedagogical Strategy .....                                  | 21        |

|                                            |           |
|--------------------------------------------|-----------|
| <b>8.2 Technical Implementation .....</b>  | <b>21</b> |
| <b>8.3 Assessment and Evaluation .....</b> | <b>21</b> |
| <b>8.4 Professional Development.....</b>   | <b>22</b> |
| <b><i>References .....</i></b>             | <b>23</b> |

## 1. Purpose and Scope of This Manual

This manual was produced under Work Package 4, Activity A1 of the ELITE-AI project, which brings together four European universities: Transilvania University of Braşov (UNITBV, Romania), the Juraj Dobrila University of Pula (UNIPU, Croatia), the University of Maribor (UNIUM, Slovenia), and the University of Udine (UNIUD, Italy) in a shared programme of experimental AI-integrated language teaching. The key mission of the ELITE-AI project is to investigate how artificial intelligence tools can support and significantly improve English language acquisition in higher education contexts, with particular attention to English for Specific Purposes (ESP) and to the professional language demands of different specialised areas.

The present document addresses the question that is placed at the core of WP4: once AI tools have been devised into language courses and deployed in classrooms, by what criteria do instructors know whether the integration has worked? Quality assurance in AI-integrated teaching, on the one hand, and language acquisition, on the other hand, is not identical to quality assurance in traditional instruction. The variables are more numerous and interact in ways that conventional lesson observation frameworks do not fully capture – the teacher’s role shifts, the point of content generation shifts, and the learner’s relationship to language input changes substantially when part of that input is machine-generated. This manual responds to those complexities with a set of practical guidelines that are grounded in the actual teaching activities documented across the consortium’s WP3 experimental sessions.

The manual has four main functions. First, it establishes a shared theoretical foundation: a summary of the research tradition within which the project’s pedagogical choices can be situated, from ESP methodology through task-based language teaching and learner autonomy theory to the emerging literature on AI-assisted language learning. Second, it sets out protocols for the implementation of AI tools in ESP classes, covering pedagogical strategy, tool selection, and the management of the teacher-AI-learner interaction. Third, it provides instruments for assessment and evaluation, both of learner language development and of the quality of AI use itself. Fourth, it supports professional development by articulating the competencies instructors need to manage AI-integrated teaching responsibly and to document it rigorously.

All guidelines and instruments in this manual have been developed directly from the evidence gathered in WP3’s experimental teaching sequence. The lesson descriptions produced by teachers at each partner institution, covering a total of over fifty AI-integrated lessons across five courses, four-year groups, and ten distinct AI platforms, constitute the empirical base from which the recommendations here are drawn. The manual does not prescribe a single model; it distils principles from a diverse and internally varied body of practice.

## 2. Theoretical Framework

The pedagogical choices documented in this manual sit at the intersection of several well-established research traditions. Understanding those traditions matters for quality assurance because it allows instructors and evaluators to distinguish between design decisions that are principled – even when the specific tool is new – and those that simply reproduce whatever a given platform makes convenient.

### 2.1 English for Specific Purposes

The discipline of English for Specific Purposes has, since the foundational work of Hutchinson and Waters (1987), defined itself through a methodology built on needs analysis: the systematic investigation of the communicative tasks a learner must be able to perform in a target professional or academic domain, and the design of instruction that prepares them for those tasks. Lorenzo Fiorito's (2005) formulation captures the essential distinction: the most important difference between ESP and general English instruction lies in the learners and their purposes for learning English. Because those purposes are specific and identifiable, an ESP syllabus is not merely an expanded vocabulary and grammar inventory but a mapping of the language functions, text types, vocabulary registers, and communicative conventions the learner will actually encounter.

The motivational corollary of this specificity has been documented consistently across the ESP literature. Dudley-Evans and St John (1998) note that ESP students are typically more motivated than general English students because the link between what they study and what they need to do professionally is visible and immediate – what Jordan (1997) describes as a *turbocharger effect*, in which the student's existing subject-matter knowledge accelerates the acquisition of the English labels and structures that attach to it. The ELITE-AI project's thematic frame, i.e., European student mobility and the Erasmus+ context, criminal justice and law, tourism, education, social media marketing, sport, and administration, deploys this motivational logic deliberately: every grammar exercise, every vocabulary list, and every writing task is grounded in a communicative situation students either have already faced or will face within months.

This grounding has direct quality assurance implications. An AI-integrated ESP lesson that generates vocabulary decontextualised from any target communicative task, or that produces grammar exercises whose scenarios bear no relation to the students' professional horizon, fails the primary ESP criterion regardless of the sophistication of the tools involved. Evaluators should therefore assess AI-generated content not only for linguistic accuracy but for its alignment with the documented target discourse community.

### 2.2 Task-Based Language Teaching

Task-Based Language Teaching (TBLT), as theorised by Long (2015) and practised within the tradition that runs through Ellis (2003) and Nunan (2004), holds that communicative tasks, defined as activities in which meaning is primary, there is a goal to be achieved, and success is evaluated in terms of whether that goal is achieved, are the most effective unit of language

instruction. The ELITE-AI lessons documented in this manual consistently organise language learning around tasks rather than isolated exercises: drafting an Erasmus+ motivation letter, preparing a selection interview, translating an EU directive, analysing a maxim-flouting exchange in a student dialogue. In each case, the AI tool is a resource within the task, not an end in itself.

The quality assurance implication is that the appropriate question when evaluating an AI-integrated task is not ‘did students use the AI effectively?’ in the abstract, but ‘did the AI’s involvement support the completion of a communicative task that required the target language forms?’ A session where students generate practice sentences with an AI chatbot but never integrate those sentences into any purposeful communicative act falls short of the TBLT standard, even if the generated sentences are grammatically correct.

## **2.3 Learner Autonomy and AI Interaction**

Little (1991) defines learner autonomy as the capacity to take charge of one’s own learning – to set goals, select strategies, monitor progress, and evaluate outcomes. Benson (2011) extends this to digital contexts, arguing that the key question for technology-enhanced learning is not what the tool does but whether the learner is the agent or the object of the tool’s operations. In the AI-integrated classes documented across the ELITE-AI project, the movement from teacher-directed AI use to student-initiated AI use is one of the most consistent and pedagogically significant trajectories in the data. At UNITBV, the Applied Grammar sequence records a clear progression from under-use and active by request in the first session to active by volition and high motivation by the third – a pattern that appears across multiple courses and institutions.

This trajectory has a direct bearing on quality assurance design. A quality check that only evaluates whether AI tools were used, without tracking the degree of learner initiative in their use, will miss one of the most important dimensions of the experimental design. The observation instruments described in Section 4 of this manual are built to capture learner autonomy development across a sequence, not only at a single point.

## **2.4 AI-Assisted Language Learning**

The emergence of large language models as classroom tools has generated a rapidly growing body of applied linguistics research. Godwin-Jones (2022) situates current AI-assisted language learning within a longer history of computer-mediated communication and intelligent tutoring systems, arguing that what distinguishes contemporary LLMs is the naturalness and flexibility of the language they produce – and the corresponding risk that learners treat AI output as a transparent window onto correct language use rather than as an artefact with its own biases and limitations. Huang and colleagues (2023) document the tension between AI tools’ capacity to generate extensive, personalised language input and the risk of what they term uncritical acceptance: the tendency to use AI-generated text without the evaluation that would normally accompany input from a human expert.

The ELITE-AI project's observation data reflects this tension at multiple points. In early sessions across all partner institutions, students were more likely to accept AI output without question; by later sessions, teachers recorded increasingly active verification behaviour, with students challenging AI-generated translations, questioning institutional descriptions, and flagging unnatural-sounding grammar transformations. The development of this critical stance, which Beatty (2010) characterises as a core component of digital literacy in language learning contexts, is treated in this manual as a quality outcome in its own right, not merely as a methodological precaution.

Jarvis and Achilleos (2013) draw attention to the distinction between AI as a source of input and AI as a conversational partner, arguing that the pedagogical design of AI interaction should be calibrated to which of these roles the tool is performing in any given activity. The ELITE-AI lessons deploy AI tools across both modes: as content generators (NotebookLM synthesising teacher-authored materials into podcasts and flashcards) and as interactive partners (students prompting Gemini for scenarios, querying Claude for feedback on their writing). And the quality assurance guidelines in this manual address each mode distinctly.

### 3. Experimental Design and Project Structure

Before setting out the quality assurance protocols, it is necessary to describe the experimental design within which they operate, since the design determines what counts as evidence of quality.

#### 3.1 The Work Package 3 Protocol

All teaching activities described in this manual were conducted under the WP3 Experimental Protocol (ELITE-AI Project Consortium, 2025). The protocol requires each partner institution to divide enrolled students into an experimental group and a control group at the start of each relevant course. A placement test administered before instruction begins establishes a shared baseline for both groups. The experimental group receives instruction that integrates AI tools in the ways documented in the lesson descriptions; the control group receives equivalent curricular content through traditional instructional methods, without AI tool support. A shared final assessment at the end of the semester provides the data for comparative analysis.

This design makes it possible to attribute observed differences in language outcomes to the AI-integrated treatment rather than to pre-existing differences in proficiency, since baseline equivalence is established at the outset. The quality assurance manual covers both the treatment, ensuring that AI-integrated instruction meets the standards described here, and the documentation of that treatment, since rigorous and consistent documentation is a precondition for any meaningful comparative analysis.

#### 3.2 Overview of Partner Courses and AI Tool Rotations

The table below summarises the courses, year groups, and AI tools documented across the consortium's WP3 experimental activities. The tool rotation column lists the platforms used in each course sequence; it should be read not as a catalogue of options but as evidence of deliberate differentiation, since each tool was selected because its specific affordances matched a stage of the instructional sequence.

| Institution | Course                              | Year Group  | AI Tools Used                                                                     | Lesson Count |
|-------------|-------------------------------------|-------------|-----------------------------------------------------------------------------------|--------------|
| UNITBV      | Applied Grammar                     | BA 1st Year | NotebookLM, Gemini, Google AI Studio, Diffit                                      | 4            |
| UNITBV      | ESP                                 | BA 2nd Year | NotebookLM, Claude, Google AI Studio, Diffit                                      | 4            |
| UNITBV      | Pragmatics                          | BA 3rd Year | NotebookLM, Gemini, Claude, Google AI Studio                                      | 6            |
| UNITBV      | Practical Grammar & Vocabulary      | BA 1st Year | ChatGPT, Google AI Studio (ESP Monster), Claude                                   | 15           |
| UNIPU       | English for Early and Preschool Ed. | BA 1st Year | ChatGPT, Zeno Chat, Gemini, Perplexity, Coral AI, Grammarly, Luvvoice, Mentimeter | 22           |

| Institution | Course                                  | Year Group  | AI Tools Used                              | Lesson Count |
|-------------|-----------------------------------------|-------------|--------------------------------------------|--------------|
| UNIUM       | Terminology of Criminal Justice         | BA 2nd Year | ChatGPT, Copilot                           | 20           |
| UNIUM       | English for Criminal Justice & Security | BA 2nd Year | Perplexity AI, ChatGPT, ChatGPT 5.0        | 29           |
| UNIUD       | English for Tourism I                   | BA 1st Year | Copilot, ChatGPT, Grammarly, Classtime.com | 8            |

**Table 1. AI-integrated courses documented across the ELITE-AI consortium, fall semester 2025/2026**

### 3.3 Thematic Coherence Across the Consortium

While the courses vary considerably in subject matter, from Applied Grammar to Tourism communication to Criminal Justice terminology, they share a structural commitment to contextualised language learning.

At UNITBV, the unifying frame is European student mobility and the Erasmus+ discourse community. This means that grammatical structures (conditional sentences, passive voice, topicalisation) and vocabulary sets (learning agreement, traineeship, host institution, credit recognition) appear not in isolation but in texts and tasks that students would plausibly encounter as mobility candidates or participants.

At UNIPU, the thematic frame is professional pedagogy (education theory, learning styles, the discourse of early childhood teaching), which gives the language content immediate professional relevance for students who will work as preschool teachers.

At the University of Maribor's two courses, the frame is criminal justice and security communication, with the AI tools deployed within document simulation tasks, institutional description tasks, and formal writing tasks directly relevant to policing, security administration, and legal communication.

At UNIUD, the frame is tourism communication, with tasks organised around the genres students will use as tourism professionals: emails, promotional leaflets, destination analyses, cover letters, and business correspondence.

This diversity across institutions makes the current quality assurance template very complex, so what this manual is able to provide is a set of principles and instruments that can be applied flexibly within any of these contextual frames.

## 4. Pedagogical Implementation Protocols

The protocols in this section translate the theoretical principles of Section 2 into practical guidance for the design and delivery of AI-integrated language lessons. They are organised around four decision points that instructors face before, during, and after each session.

### 4.1 Pre-Lesson Planning: Tool Selection and Prompt Design

The selection of an AI tool for a given lesson should be driven by a single question: what does this tool can contribute that the lesson genuinely needs? The evidence from the WP3 lessons demonstrates that this is a question with a clear answer in each case and is directly coordinated with the Repository created in WP2, as well as with interactive lesson plan designed still in WP2, on skills and lesson stages, each having been attributed the best AI tool by choice of stage aim.

Thus, NotebookLM's capacity to synthesise multiple source documents into podcast, video, flashcard, and mind-map formats from a single uploaded workspace made it the appropriate tool for theoretical introduction sessions where the priority was building a rich, multi-modal knowledge base from teacher-authored materials. Claude's responsive scenario generation made it well suited to practice sessions where students needed a context-sensitive engine for producing the communicative situations they would analyse. Gemini's document generation capacity, evidenced in both the practical grammar sequences and in the ESP sequence's application letter work, made it the right tool for sessions where the output was a polished, print-ready artefact. Diffit's differentiated text generation, as well as worksheet design, made it appropriate for reading sessions where students were working at different points within the same broad proficiency band, as well as for modern digital drilling.

The implication for pre-lesson planning is that tool selection should precede prompt design, and both should follow needs analysis. Before writing a single prompt, the instructor should be able to answer: What communicative task will students perform in this lesson? What language forms, vocabulary register, or genre conventions does that task require? What does the AI tool need to produce for students to have the input or support they need? Only after those questions have answers does prompt design become possible, and the quality of the prompt determines the quality of everything that follows.

**Pre-Lesson Planning Checklist**

1. Identify the target communicative task and the language forms it requires.
2. Select the AI tool whose output type best matches the lesson's input requirements.
3. Draft the instructional prompt, specifying: student level, thematic context, target structures, and required output format.
4. Review AI-generated outputs before the session for accuracy, register appropriateness, and alignment with the ESP context.
5. Prepare critical evaluation prompts to use during the session if students do not spontaneously question AI outputs.

**4.2 In-Session Management: The Teacher's Role**

One of the most important findings from the WP3 observation data is that the teacher's role in AI-integrated sessions is more demanding, not less, than in traditional instruction. This finding is consistent with the broader AI-in-education literature: Selwyn (2019) argues that 'the teacher as AI mediator' requires a dual competence, i.e., sufficient familiarity with the tool to anticipate and contextualise its outputs, and sufficient pedagogical authority to ensure that the tool serves the lesson's aims rather than replacing them. Luckin and colleagues (2016) describe this role as 'intelligent teaching', in which the teacher's primary function shifts from information delivery to the orchestration of student-AI-content interaction.

Across the ELITE-AI sessions, three distinct teacher roles appear in the observation data, and each is associated with a different phase of the experimental sequence. In early sessions, instructors operated as curators: they assembled source materials for AI tools, reviewed the generated outputs in advance, and presented those outputs to students as structured instructional inputs – podcasts, videos, flashcard sets – with explanatory commentary. This curation mode corresponds to what the observation sheets record as under-use of AI, in the sense that students were not yet operating the tools themselves; but it is not a lesser form of AI integration. It is the appropriate mode for a first session where students are simultaneously acquiring new conceptual content and orienting to an unfamiliar interaction model.

In middle-phase sessions, instructors operated as scaffolders: they provided the contextual prompt that students then took to the AI tool, modelling what a productive interaction with the tool might look like before students attempted it independently. The UNIPU sessions' use of teacher-prepared prompts before students approached the chatbot, documented in the Montessori and Vygotsky lesson, produced noticeably more focused information retrieval than

earlier open-ended chatbot use, a pattern that maps directly onto Vygotsky's (1978) 'zone of proximal development': the scaffold makes possible what the learner cannot yet do alone.

In later sessions, instructors operated as evaluators and surveyors, stepping back from the generation of AI input and focusing their attention on facilitating critical engagement with the outputs students themselves had prompted. This is the mode associated with active by volition participation in the observation data, and it is the mode that most directly develops the learner autonomy and AI literacy competencies that the ELITE-AI project identifies as primary learning outcomes.

### 4.3 Critical Engagement with AI Output

The requirement that students engage critically with AI-generated content is stated throughout the ELITE-AI protocol and is treated in this manual as a non-negotiable component of quality implementation. The theoretical basis for this requirement is twofold: first, there is the accuracy risk, which refers to the fact that AI language models are known to produce confident-sounding text that contains factual errors, outdated information, or linguistically plausible but contextually inappropriate formulations. In specialist language domains, such as legal terminology, institutional descriptions, translation studies vocabulary, this risk is sensitive, as the UNIUM sessions' experience with AI-generated summaries of criminal justice institutions made clear. Second, there is the pedagogical risk, which involves the situation when students accept AI output without evaluation, thus depriving themselves of any metalinguistic engagement, consisting of the analysis of why a particular formulation is correct, effective, or appropriate, i.e., the mechanism through which language is actually acquired rather than merely reproduced.

Subsequently, critical engagement cannot be assumed, it must be designed into the task. The most reliable design mechanism, evidenced across multiple courses in the WP3 data, is the verification task: asking students to compare AI-generated content against an authoritative source (official institutional descriptions, authentic published texts, the teacher's own annotated model), to identify discrepancies, and to articulate why the AI's formulation falls short. At UNIUM, this comparison with official Slovenian legislation and institutional sources was built into every lesson from the second session onwards. At UNITBV, it was built into the ESP Class 2 application letter task, where students adapted Claude's template to their own situation and were therefore obliged to evaluate every sentence the tool had produced.

### 4.4 Progression in AI Tool Rotation

One of the most consistent findings across the ELITE-AI consortium is that a deliberate rotation of AI tools, rather than a sustained use of a single platform, produces better pedagogical outcomes than single-tool approaches. The rationale is threefold: exposure to different tools develops a broader and more transferable AI literacy; different tools' diverse output styles and interaction modes highlight various dimensions of language use; and the comparative analysis that arises naturally when students have used more than one tool – 'What did Gemini give us that ChatGPT didn't?' – is itself a high-order critical thinking exercise.

The UNIPU sequence, with eight distinct tools across twelve sessions, is the most ambitious implementation of this principle in the project, and the observation data supports its effectiveness: the progression from partially effective critical evaluation in early sessions to more consistent verification behaviour by the collocation and formal writing lessons was attributed by the teacher to the increasing sophistication students had developed in distinguishing between tools' different response styles and limitations.

For instructors designing a tool rotation, the practical guidance can refer to varying the tool at the point where the lesson's instructional demands change significantly. When moving from a theory-building session to an applied practice session, a tool change signals the shift in mode and brings fresh learning opportunities with it. When the same kind of task recurs across sessions, as formal writing tasks do throughout the UNIUM criminal justice sequence, consistency of tool allows students to develop depth of facility with a single interface before encountering another.

## 5. Assessment and Evaluation Instruments

This section describes the instruments used to assess the quality of AI-integrated instruction across the ELITE-AI project and to evaluate learner development within experimental sessions. The instruments fall into two categories: those used by instructors to document session quality (observation sheets and lesson records), and those used to assess learner language and AI literacy development (participation indicators, critical evaluation ratings, and the shared final assessment).

### 5.1 The Observation Sheet Framework

The observation sheet is the primary documentation instrument in the WP3 protocol. It records, for each session, the lesson stage (Leand-in, Teaching, Practice, Feedback or pairs), the degree of AI use (under-use, appropriate, or over-use), student participation indicators, target language use, and quality-of-AI-use ratings across four dimensions. The four quality dimensions are: (a) communication with the AI – the clarity and precision with which students or teachers formulate requests to the tool; (b) prompt effectiveness – the degree to which the prompts used elicit the kinds of output the lesson requires; (c) interpretation of AI feedback – the accuracy with which students or teachers read and apply the tool’s outputs; and (d) critical evaluation – the degree to which outputs are assessed rather than accepted.

The participation indicators are designed to capture the qualitative shift from passive reception to active initiation that the ELITE-AI project identifies as the developmental arc of successful AI integration. Three degrees of participation are distinguished across the dataset: active by request (students engage when directed, do not initiate); active by request with some volition (students occasionally prompt the tool independently); and active by volition (students initiate AI interactions without teacher direction). This progression maps onto Holec’s (1981) conceptualisation of learner autonomy development and provides a basis for longitudinal comparison across a teaching sequence.

**Table 2. Observation Sheet Quality Dimensions and Rating Scale**

| Dimension                        | Rating Scale                                                   |
|----------------------------------|----------------------------------------------------------------|
| Communication with AI            | <b>Effective / Partially Effective / Developing</b>            |
| Prompt Effectiveness             | <b>Effective / Partially Effective / Developing</b>            |
| Interpretation of AI Output      | <b>Effective / Partially Effective / Developing</b>            |
| Critical Evaluation of AI Output | <b>Active &amp; Discriminating / Aware (general) / Limited</b> |
| Student Participation Degree     | <b>Active by Volition / Active by Request / Passive</b>        |
| Target Language Use              | <b>Increased / Maintained / Diminished</b>                     |
| Degree of AI Use                 | <b>Under-use / Appropriate / Over-use</b>                      |

## 5.2 Evaluating Content Delivery Effectiveness

Content delivery effectiveness in AI-integrated sessions is evaluated against three criteria drawn from the ESP framework: alignment with target needs (does the AI-generated content address the communicative demands of the students' target discourse community?), linguistic accuracy (is the language the AI produces correct at the levels of grammar, register, and vocabulary?), and pedagogical integration (is the AI output embedded within a task structure that requires active language processing rather than passive reception?).

In practice, evaluators should approach content delivery quality by tracing the instructional sequence backwards: what was the communicative task the lesson aimed to develop? What AI-generated input or support was provided? Did that input, in terms of both content and language, serve the task adequately? The UNITBV Applied Grammar sessions provide a clear model of positive alignment: the Google AI Studio-generated sentence transformation exercises in Class 3 specified both the grammatical structures required (conditional clauses, wish, passive voice, topicalisation) and the communicative context (Erasmus+ mobility vocabulary), so that grammar practice and vocabulary consolidation occurred simultaneously within a task that was directly relevant to students' professional horizon.

Negative alignment, referring to content that is linguistically accurate but contextually mismatched, is less visible but equally problematic. AI tools tend to default to generic examples when contextual parameters are not specified in the prompt. An evaluator reviewing AI-generated materials should check not only that the language is correct but that the examples and scenarios are drawn from the target domain rather than from a generic English teaching register.

## 5.3 Adaptability and Differentiation

One of the genuine advantages AI tools offer in language teaching is the potential for differentiated materials, meaning outputs calibrated to different proficiency levels within the same group without proportionate increases in teacher preparation time. Diffit's explicit differentiation function, deployed in both the UNITBV ESP and Applied Grammar sequences, is the clearest example in the WP3 data: the platform processes a single input text or exercise set and produces multiple entry points organised by complexity level, making it possible to serve students at different points within a broad B2 band with the same underlying content.

Quality assessment of differentiation should verify that the calibration is accurate, that, for example, materials labelled appropriate for upper-B2 students do not, on inspection, contain C1 vocabulary or syntactic complexity without any contextual support, and that lower-B2 materials do not oversimplify to the point of removing the productive challenge that makes language learning possible. Vygotsky's (1978) 'zone of proximal development' provides the operative criterion: appropriately differentiated materials should be pitched at the level where students can succeed with effort and support, not at the level where success requires no effort at all.

## 5.4 Feedback Mechanisms and Self-Regulation

Nicol and Macfarlane-Dick (2006) identify seven principles of good feedback practice in higher education: it clarifies what good performance looks like, it delivers high-quality information about learning, it encourages teacher and peer dialogue, it encourages positive motivational beliefs, it provides opportunities to close the gap between current and desired performance, it provides information to teachers about the quality of instruction, and, critically, it helps students develop self-assessment. AI tools, when used well, can contribute to several of these functions.

The UNITBV Practical Grammar sequence provides a model of AI-mediated feedback used to develop self-regulation. In Lesson 6 (Speaking and Functional Language), students typed up representative exchanges from their role-plays and submitted them to ChatGPT for commentary on register, vocabulary, and functional coherence. The AI's feedback did not replace the teacher's but extended its reach: sixteen students could receive individualised written commentary on their spoken output simultaneously, something a single instructor could not provide. The quality of this feedback loop depends on two conditions: the prompt must specify the evaluative criteria the AI should apply (register appropriateness, vocabulary precision, structural coherence), and students must be supported in reading the AI's commentary critically rather than treating it as authoritative.

Peer feedback, used in several UNIUD sessions and in the UNITBV Grammar Class 4 peer review checklist, provides a complementary mechanism. The combination of AI-generated commentary and peer assessment creates two distinct perspectives on the same output, and the comparison between them – why did the AI flag this, while my peer thought it was fine? – is itself a metalinguistic exercise of considerable value.

## 5.5 Learner Motivation and Participation Data

Motivation is recorded in the ELITE-AI observation sheets as 'developing', 'high', or 'fluctuating', and the patterns in the WP3 data are instructive. Across nearly all partner courses, the session with the highest motivation rating is the one with the most direct personal relevance: at UNITBV, the ESP Class 2 on Erasmus+ selection process preparation received 'high motivation' because several students in the experimental group were themselves preparing applications; at UNIPU, the learning styles and pedagogical theory sessions received higher engagement than the accuracy-focused grammar drills; at UNIUD, all eight English for Tourism sessions maintained 'high motivation', a finding the teacher attributed to the combination of scaffolded prompt design instruction and iterative output refinement that produced a genuine sense of productive agency.

These patterns are consistent with Deci and Ryan's (2000) 'Self-Determination Theory' (SDT), as applied to L2 motivation by Dörnyei (2001), which identifies three psychological needs that must be met for sustained motivational engagement: competence (the sense of being able to do the task), autonomy (the sense of choosing one's approach), and relatedness (the sense that the task connects to one's goals and community). AI tools, when well-integrated, can contribute to all three: they provide immediate competence support; they expand the range of choices available to the learner in how to approach a task; and they make personal relevance visible by

generating content in the specific vocabulary and scenarios of the learner's target discourse community.

## 6. Professional Development Guidelines

The integration of AI tools in ESP teaching places new demands on instructors that are not adequately addressed by conventional English language teacher education. This section identifies the competency areas that the ELITE-AI project's experimental sequence has shown to be essential and describes the professional development activities most likely to build them.

### 6.1 Core Competencies for AI-Integrated ESP Teaching

Based on the evidence gathered across the WP3 sessions, five competency areas are identified as essential for effective AI-integrated ESP instruction.

The first one refers to the needs analysis and tool selection, more precisely to the ability to identify target communicative needs with sufficient precision to select the AI tool and design the prompt that will generate appropriate support. This competency is grounded in the existing ESP tradition but extended by the requirement to match communicative needs to the specific competences of different AI platforms, a matching task that requires familiarity with a range of tools. For this, a database like one created in WP2 becomes extremely useful and practical.

The second one considers the prompt design, i.e., the ability to construct prompts that elicit the kinds of output the lesson requires. The ELITE-AI data consistently shows that vague or underdeveloped prompts produce generic outputs that fail the ESP criterion of contextual relevance. Effective prompt design refers back to the communicative frame of teaching and learning, when very detailed task environments were created for a specific task, incorporating student level, thematic context, target language forms, communicative register, and output format, all in a single, coherent request. This is a skill that develops with practice and benefits from documented examples of effective and ineffective prompts.

The third one reflects upon the critical evaluation of AI output and refers to the ability to assess AI-generated content for linguistic accuracy, contextual appropriateness, and potential misinformation before presenting it to the students, and to model the same evaluative habit for learners. This competency is particularly critical in specialist language domains where the consequences of terminological error can be significant; thus, legal, medical, or institutional terminology requires a higher verification standard than general vocabulary work.

The fourth one includes session documentation, meaning the ability to complete observation sheets and lesson records with sufficient consistency and detail to support both cross-institutional comparison, in case of extended research across multiple partners, and the instructor's own reflective practice. Systematic documentation is not a bureaucratic requirement, but the mechanism through which the project's experimental design generates knowledge.

The sixth one integrates mediation between AI tools and learners, developing the pedagogical and interpersonal skill to position AI-generated content productively within the instructional

sequence, neither overstating its authority nor dismissing its utility. This mediation role is most demanding in early sessions, where students are simultaneously learning new content, learning a new interaction model, and developing the critical viewpoint that will protect them from uncritical acceptance, all at once.

## 6.2 Reflective Practice and Documentation

Schön's (1983) distinction between 'reflection-in-action' (the adjustments a practitioner makes in the moment), which resembles the 'Design-Based Research' (DBR) approach (2003), the 'adaptive design' method (Fisher, 1935), the 'emergent design' frame (Lincoln & Guba, 1985) or the 'action research' technique (Lewin, 1946), and 'reflection-on-action' (the retrospective analysis of what happened and why) provides a useful framework for understanding what the ELITE-AI observation sheet system is designed to support. Completing an observation sheet after each session is an act of reflection-on-action: the instructor is encouraged to notice patterns that might not be visible in the flow of a ninety-minute lesson: the fact that motivation dropped in the grammar drill session but recovered in the scenario generation session; the fact that peer interaction frequency increased when students were given free choice of AI tool; the fact that critical evaluation was more active in sessions where the AI's limitations were most visible.

The longitudinal analysis of observation sheet data across a teaching sequence, comparing participation, motivation, and quality of AI use from the first session to the last, is the most powerful reflective instrument available within the project. Teachers who compare their own Class 1 and Class 5 observation sheets, for example, and trace the changes in student autonomy and AI literacy are engaging in exactly the kind of evidence-based professional development that the project's comparative methodology is designed to generate.

## 6.3 Collaborative Review within the Consortium

The multi-institutional character of the ELITE-AI project creates an asset that single-institution AI integration initiatives do not have: a community of practice, in Wenger's (1998) sense, whose members are engaged in the same fundamental activity, adapting AI tools to language teaching, across different disciplinary contexts, student populations, and national educational traditions. The variation across the consortium is not noise in the data, but an analytically valuable instrument. In this way, the fact that UNIPU's group's eight-tool rotation system produced broadly consistent quality development with a B2 first-year group in early childhood education, while the UNIUM's group's more focused two-tool approach produced substantively similar critical evaluation trajectories with a B2 second-year group in criminal justice is a finding about the robustness of the underlying pedagogical principles rather than the particular effectiveness of any single tool.

Consortium-level professional development should therefore include structured opportunities for cross-institutional lesson observation (in the form of shared video recordings or live observation visits), discussion of divergent findings (cases where observation data from one institution contradicts the general pattern, and analysis of why), and collaborative prompt

development workshops, in which instructors from different disciplines pool their experience of what kinds of prompts produce what kinds of output for their specific contextual frames.

## 7. Integrated Quality Criteria

The protocols and instruments described in previous sections can be organised into a set of integrated quality criteria that apply across all partner institutions, course types, and AI tools. These criteria provide the evaluative framework for the quality assurance activities described in the WP4 manual as a whole.

### 7.1 Pedagogical Alignment

Every AI-empowered activity should be derivable from a documented communicative need. The chain of alignment runs from target discourse community (what communicative tasks will students need to perform?) through lesson task (what communicative task does this session develop?) through AI tool function (what does the tool produce that serves this task?) to assessment criterion (how do instructors know whether students can now perform the target task?). An activity that breaks this chain at any point, where the AI generates content that is linguistically interesting but not connected to any communicative task, or where the task exists but no assessment criterion has been specified, fails the pedagogical alignment criterion.

### 7.2 Linguistic Accuracy and Register Appropriateness

AI-generated language content should meet the same standards of accuracy and register appropriateness as any other instructional material. In practice this means that instructors must review AI outputs before sessions, that students must be given the tools and the habit of verifying AI language against authoritative sources, and that the observation sheet's record of critical evaluation behaviour is treated as a quality indicator rather than merely as an observation of student behaviour. A session in which AI-generated content contained identifiable errors and students did not notice them represents a quality failure even if every other dimension of the lesson was well executed.

### 7.3 Learner Agency

Quality AI integration is characterised by increasing learner agency in the interaction with AI tools across a teaching sequence. A course in which students are using AI tools in exactly the same way in the final session as in the first, receiving teacher-curated outputs rather than generating their own prompts, has not achieved the autonomy development that the ELITE-AI project identifies as a primary learning outcome. This does not mean that teacher-directed AI use is a failure; it is, as the WP3 data shows, often the appropriate mode for early sessions. The quality criterion is progression, as the observation record should show movement along the agency spectrum, from passive reception through progressive prompting to independent initiation, across the teaching sequence as a whole.

### 7.4 Documentation Consistency

The validity of the comparative analysis between experimental and control groups, and between different partner institutions, depends on the consistency and completeness of the observation documentation. This manual treats incomplete or inconsistent documentation as a quality issue, not merely an administrative one. An observation sheet that records motivation as ‘high’ without noting the session’s task type, student group size, or lesson stage provides insufficient information for cross-institutional comparison. The minimum documentation standard for each session is: observation sheet completed with all relevant indicators; lesson materials (AI prompts, generated outputs, worksheets) filed or linked; any deviations from the WP3 protocol noted and explained.

## 7.5 Ethical Implementation

AI tools in educational settings raise ethical questions that instructors need to be able to address with students directly, not merely manage administratively. The most pressing of these in the ELITE-AI context are: transparency about what AI tools do and do not know (the temporal limitation that explains why AI-generated institutional descriptions may be outdated, the probabilistic nature of language generation that explains why AI output is plausible rather than necessarily accurate); data privacy (what information students share with AI platforms and what those platforms do with it); and the relationship between AI assistance and academic integrity (the distinction between using AI as a support for language learning and using it as a substitute for it, which is the distinction the observation sheets track when they record over-use of AI as a quality concern). An AI-integrated lesson that does not address these questions, at least implicitly through its task design, is incomplete as a quality-assured implementation.

## 8. Summary of Recommendations

The following recommendations distil the core guidance of this manual into a form that can be referenced quickly by instructors and evaluators. They are organised by the four activity areas specified in the WP4 mandate: pedagogical strategy, technical implementation, assessment and evaluation, and professional development.

### 8.1 Pedagogical Strategy

1. Ground every AI-integrated activity in a documented communicative need derived from target discourse analysis.
2. Select AI tools for their specific learning opportunities relative to the lesson stage, not for their general reputation or familiarity.
3. Rotate AI tools across a teaching sequence to develop breadth of AI literacy alongside depth of facility with individual platforms.
4. Design AI interaction to progress from teacher-curated input in early sessions to student-initiated prompting in later sessions.
5. Build verification tasks into every session where AI-generated content is used, requiring students to compare AI outputs against authoritative sources.

### 8.2 Technical Implementation

1. Specify student level, thematic context, target language forms, and required output format in every instructional prompt.
2. Review all AI-generated materials before sessions for linguistic accuracy, register appropriateness, and contextual alignment.
3. Document prompt text and AI outputs as part of the lesson record to enable comparison and refinement across the teaching sequence.
4. Ensure students have reliable access to the relevant platforms, with alternative activities prepared for connectivity failures.

### 8.3 Assessment and Evaluation

1. Complete the WP3 observation sheet (<https://elite-ai.unitbv.ro/wp-content/uploads/2026/04/Teacher-Observation-Sheet.pdf>) after every session, recording all required indicators.
2. Track the quality-of-AI-use dimensions (communication, prompt effectiveness, interpretation, critical evaluation) across the teaching sequence, not only at single time points.
3. Use the shared baseline (placement test) and final assessment to evaluate language development outcomes, controlling for initial proficiency differences.

4. Treat learner agency progression – from active by request to active by volition – as a quality outcome to be tracked and reported.
5. Document and report instances of over-use as well as effective use, since over-reliance on AI for content generation rather than language support is a quality concern equivalent in significance to under-use.

## 8.4 Professional Development

1. Allocate dedicated time before the start of each semester for prompt design preparation and AI tool familiarisation.
2. Use the comparative analysis of own observation sheet data across a teaching sequence as the primary instrument of reflective practice.
3. Participate in consortium-level professional development activities, including cross-institutional lesson observation and collaborative prompt development workshops.
4. Address the ethical dimensions of AI use – transparency, data privacy, academic integrity – explicitly in at least one session per course, and implicitly through task design throughout.

## References

- Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.
- Beatty, K. (2010). *Teaching and Researching Computer-Assisted Language Learning* (2nd ed.). Pearson Longman.
- Benson, P. (2011). *Teaching and Researching Autonomy* (2nd ed.). Pearson Longman.
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 219–251.
- Design-Based Research Collective. (2003). *Design-based research: An emerging paradigm for educational inquiry*. *Educational Researcher*, 32(1), 5–8. <https://doi.org/10.3102/0013189X032001005>
- Dörnyei, Z. (2001). *Motivational Strategies in the Language Classroom*. Cambridge University Press.
- Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for Specific Purposes: A Multi-Disciplinary Approach*. Cambridge University Press.
- ELITE-AI Project Consortium. (2025). WP3 Experimental Protocol. Internal project document.
- Ellis, R. (2003). *Task-Based Language Learning and Teaching*. Oxford University Press.
- Fiorito, L. (2005). *Teaching English for Specific Purposes (ESP)*. UsingEnglish.com. <https://www.usingenglish.com/articles/teaching-english-for-specific-purposes-esp.html>
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd
- Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5-24. <https://www.lltjournal.org/item/10197>
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics*, Vol. 3: *Speech Acts* (pp. 41-58). Academic Press.
- Holec, H. (1981). *Autonomy and Foreign Language Learning*. Pergamon. [First published 1979, Council of Europe.]
- Honey, P., & Mumford, A. (1982). *The Manual of Learning Styles*. Peter Honey.
- Huang, J., Saleh, S., & Liu, Y. (2023). A review on artificial intelligence in education: Chatbots and their applications in language learning. *Sustainability*, 15(12), 9525. <https://doi.org/10.3390/su15129525>
- Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes: A Learning-Centred Approach*. Cambridge University Press.
- Jarvis, H., & Achilleos, M. (2013). From computer assisted language learning (CALL) to mobile assisted language use. *TESL-EJ*, 16(4). <http://tesl-ej.org/wordpress/issues/volume16/ej64/ej64a2/>

- Jordan, R. R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers*. Cambridge University Press.
- Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, 2(4), 34–46. <https://doi.org/10.1111/j.1540-4560.1946.tb02295.x>
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications.
- Little, D. (1991). *Learner Autonomy 1: Definitions, Issues and Problems*. Authentik.
- Long, M. H. (2015). *Second Language Acquisition and Task-Based Language Teaching*. Wiley-Blackwell.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Pearson. <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/news/news-announcements/2016/Intelligence-Unleashed-Publication.pdf>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Nunan, D. (2004). *Task-Based Language Teaching*. Cambridge University Press.
- Schön, D. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Selwyn, N. (2019). *Should Robots Replace Teachers? AI and the Future of Education*. Polity Press.
- Thomas, J. (1995). *Meaning in Interaction: An Introduction to Pragmatics*. Longman.
- Verschueren, J. (1999). *Understanding Pragmatics*. Arnold.
- Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds. & Trans.). Harvard University Press.
- Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.