Meaning-Making Process and Error Dynamics in ChatGPT-Mediated Translation

Iulia Mihalache
Université du Québec en Outaouais

iulia.mihalache@uqo.ca

María-José Varela Salinas
Universidad de Málaga

mjvs@uma.es

Abstract

This study examines errors in a ChatGPT-mediated translation of a German economic text on inflation into Spanish, post-edited by 20 translation students. The German–Spanish language pair represents an under-studied combination in post-editing research, where English-pivot pairs predominate. The analysis classifies 132 annotated instances by error origin (ChatGPT-generated versus student-introduced during post-editing) and by linguistic category. Results show that terminology is the highest-risk domain across the entire workflow (34.1%), followed by tense/aspect (15.2%) and style (13.6%). ChatGPT-related errors account for 50.8% of all instances, while student-introduced errors through over-editing represent 21.2%. A further 28.0% are preferential changes: acceptable reformulations of segments already adequate to task norms. Students tend to trust fluent machine output even when it contains subtle semantic distortions, yet they also over-edit segments that are already acceptable. The findings highlight three didactic priorities: developing LLM-based MT literacy, strengthening decision-making strategies in post-editing, and fostering genre- and domain-sensitive editing competence.

Keywords: post-editing, machine translation, ChatGPT, translator training, error analysis

1 Introduction

This research presents a comprehensive analysis of errors observed in the ChatGPT-mediated translation of an economic text on inflation from German to Spanish, followed by post-editing by translation students. It focuses on two key aspects: first, the types of errors according to their origin (ChatGPT-4.5 versus student post-editing decisions), and second, the linguistic categories most affected across all error types. The analysis relies on a translation task carried out by 20 students enrolled in a Spanish-German/German-Spanish translation course on specialized texts.

Empirical research on translator trainees in MT-mediated workflows has reported two recurrent challenges: students tend to under-edit fluent machine output, missing adequacy errors that do not disrupt surface grammaticality, and they tend to over-edit segments that are already acceptable, introducing preferential changes that may be detrimental to quality (Mossop, 2020; Setkowicz-Ryszka, 2025). Most empirical work in this area has so far focused on neural MT systems and on language pairs that include English. The German–Spanish language pair remains markedly underrepresented in research on machine translation post-editing (MTPE), and this underrepresentation is both quantitative and structural. Successive editions of the WMT Automatic Post-Editing shared task (2016–2023) have evaluated EN–DE, DE–EN, EN–RU, EN–ZH and EN–MR but never DE–ES (Chatterjee et al., 2016; Chatterjee et al., 2019; Bhattacharyya et al., 2022), and recent large-scale post-editing corpora such as LangMark (Velazquez et al., 2025) and DivEMT (Guerberof Arenas et al., 2022) are built exclusively with English as source language, structurally excluding non-English pairs from standard benchmarks. Bibliometric reviews of the field (Lu, 2022; Chen & Zhou, 2025)

likewise identify English-pivoted pairs as the dominant locus of empirical post-editing studies, and the English-centric character of multilingual training data has been documented as a systemic feature of the field rather than a contingent gap (Wang, 2023). A growing body of DE–ES scholarship has nonetheless addressed machine translation and post-editing from the complementary perspectives of contrastive linguistics and translation pedagogy, with particular attention to specialised domains, terminology and didactic integration (Roiss, 2021; Castillo Bernal, 2022; Varela Salinas & Burbat, 2023; Holgado-Sáez & Martínez Martínez, 2025). The present study contributes to this strand by combining error-category annotation with a workflow-sensitive distinction between MT-generated and post-editor-introduced errors in a DE–ES economic text, thereby connecting the contrastive-pedagogical tradition with the methodological apparatus prevalent in international MTPE research.

The contribution of the present study is twofold. First, it documents how trainees behave when they post-edit the highly fluent output of an LLM in a specialised economic register, where surface naturalness can hide problems of domain and adequacy. Second, it analyses not only which errors persist in the post-edited text, but also how new errors are introduced during revision, looking at both error propagation and over-editing.

2 Theoretical framework

The essence of intelligence, according to Wang (2008), "is the principle of adapting to the environment while working with insufficient knowledge and resources". Accordingly, an intelligent system does not equate with optimal or perfect reasoning as the "intelligent" behavior is shaped by real-world limitations such as finite processing capacity and time pressure. Moreover, intelligent systems learn from experience, handling uncertainty, which means they make errors. From this point of view, are they different from humans?

In human-only translation, too, uncertainty is intrinsic to the act of translating (de León and Cardona Guerra, 2022), a complex meaning-making process involving factors such as the translator's subjectivity and depending, among others, on the translator's skills. Even slight changes at the outset of the process can lead to significantly different target texts, argue Hassani and Malekshahi (2024) who lean on chaos theory to explain the complexity of the translation process. As Pym (2010: 93) notes in translation studies, this mirrors the observer effect in physics, where the act of observation changes the phenomenon being observed. The more translators aesthetically engage with their texts, experiencing pleasure and perceiving beauty in navigating textual ambiguity, being guided not only by conscious thought but also by their "gut", the more the process is unpredictable. Robinson (1991) calls this somatic translation—a mode of embodied responsiveness which enables translators to endure and even embrace uncertainty. Much like psychotherapists (Sarasso et al., 2022: 1), translators work within a dynamic of uncertainty where aesthetic perception transforms instability of meanings into creative insight.

When translators take on the role of post-editors, however, the pressure for speed becomes unavoidable, which may diminish their capacity to closely observe and reflect on the source and target texts, while at the same time increasing the probability they make more errors, unless they develop what Krüger (2018: 122) calls "text reception competence,"[1] "adequate text selection competence,"[2] and "text adaptation/optimization competence"[3] as new competences needed in technology-mediated translation. Post-editing guidelines in the translation industry often explicitly instruct translators to avoid creativity or "optimization" as long as the machine correctly rendered the "meaning." This approach could

[1] This refers to a translator's ability, for example when receiving target-language translation data, to assess the quality of these data with regard to their potential integration into the target text to be produced. (Krüger 2018: 122; our translation)

[2] […] choose, from predominantly target-language digital translation data that are closely related to the target text, those data that—taking contextual and efficiency-related aspects into account—can be integrated into the target text to be produced with the least possible effort of revision or optimization. (Krüger 2018: 122; our translation)

[3] Text adaptation/optimization competence includes a recombination and recontextualization competence. […] I deliberately distinguish between adaptation and optimization […]. Certain translation data—such as human translations from uncontaminated translation memories—generally only need to be embedded coherently and contextually into the target text, without requiring quality improvement. By contrast, machine translation output is usually defective to some degree and therefore needs to be optimized in terms of quality as part of its integration into the target text. (Krüger 2018: 122-123; our translation)

either have negative effects, if it constrains translators' agency, or positive ones, if it successfully minimizes the introduction of new errors in the target text.

Translation is nevertheless a decision-making activity (e.g., Levý, 2000; Kußmaul, 1986; Wilss, 2007; Angelone, 2010), one that constantly negotiates between alternative solutions and multiple interpretations, "a constant state of process and becoming" (Bohm, 1980/2002) which demands time. Concurrently, AI systems aim at replicating human cognitive functions, such as problem-solving, and decision-making skills (Veselovsky et al., 2021), by making processes less time-consuming. The view of translation as a decision-driven process illustrates that meaning is not fixed or extracted from the sociocultural, institutional, or political "turmoil" (Kristeva, 1984: 13). Rather, meaning emerges as an "unlimited and unbounded generating process," (Kristeva, 1984: 13) revealing the semiotic body's unconscious drives—what Kristeva terms signifiance, which also immerses the translator in a discovery process inherently prone to errors. And we could argue that being learners, translation students are likely the ones most fully engaged in a process of discovery.

In AI-mediated translation, when translators postedit machine-generated content, they must negotiate with the system, often plunged into a "cognitive state of indecision" (Angelone, 2010: 18), navigating uncertainty and making real-time decisions. One could argue these human decisions will be better because the artificial intelligence decisions, which come before the intervention of the post-editor translator, are algorithmically based on statistical patterns rather than on conscious deliberation and experiential understanding, as well as on a limited number of knowledge resources. But AI systems are constantly being refined, with GenAI often described as displaying human-like capacities such as creativity, social awareness, and emotional responsiveness, enabling it to perform complex, non-routine tasks (Brynjolfsson et al., 2025). In some cases, GenAI even outperforms humans: for example, Vinchon et al. (2024) found that GenAI's divergent thinking—defined as the ability to generate multiple possible solutions—shows that machines may be more creative than humans. In 2025, Daria Sinitsyna, Lead Computational Linguist at Intento, published a report after having tested 12 of the latest Large Language Models (LLMs) for translation, including GPT-4.5 preview of OpenAI. The tests involved translations from English to Spanish and English to German, with two types of prompts (one regular and one extended), and two types of tests: general domain translations (for both EN>ES and EN>DE), and specialized translations in in the legal and healthcare fields (only for EN>DE). The report highlights, among other conclusions, how neural machine translation (NMT) engines remain the leaders in translation performance, even if the authors recognize the tests have limitations because of the minimal number of languages and fields chosen, with GPT-4.5 (preview) being the best-performing new model. Claude 3.5 Sonnet stands out across domains, especially in EN>ES general translation where it ranks highest. Gemini 2.0 variants excel in healthcare, and DeepSeek-V3 demonstrates strong overall performance, according to this report.

Although Sinitsyna's study is just an example which shows the advantages of using technology in specific translation contexts, research shows that optimal outcomes are achieved when humans and AI do what they do best. For translators, this means they can reduce the likelihood of mistakes when they are working in areas where their competence is highest or where they have already developed strong skills; indeed, according to Vaccaro et al. (2024), when humans are less accurate at a task compared to the AI system, their ability to decide when to trust the algorithms and when to trust their own judgment is impaired.

This echoes Pym, who pointed out that "the revolutionary potential of the technologies can only be realized if they connect with significantly developed human skill sets" (Pym 2011: online). Various authors have already identified a range of skills specific to post-editing[4] —drawing also attention to the fact that human translators acquire

[4] These skills include "basic MT technology concepts, MT evaluation techniques, statistical MT (SMT) training, pre-editing and controlled language, monolingual PE, understanding various levels of PE (light and full), creating guidelines, MT evaluation, output error identification, when to discard unusable segments and continuous PE practice" (Guerberof Arenas 2020: 348). Nitzke and Hansen-Schirra (2021) propose a post-editing competence model grounded in translation competence and extended with skills such as error handling, MT engineering, and specialized consulting. Similarly, Ginovart and Oliver (2020) identify three core skills—decision-making, error identification, and adherence to PE guidelines—and outline a broader set of 11 related competences which are: the capacity to decide when to edit or

post-editing skills gradually, the level of comfort being attained at the end of 100,000 post-edited words (Vasconcellos 1987: 145). However, it is equally important to emphasize the need for a broader development of human expertise combined with a critical assessment of technology to avoid falling into what Winner (2014) calls "technological somnambulism". As Baumgarten (2025) argues, technology is never neutral but reflects dominant social values, making any critique of translation technology inseparable from a critique of the hegemonic systems that shape its design and use: the capitalist value system, the legal and institutional contexts, the political constraints, the individuals involved and the people who benefit most.

Human expertise can be seen as a corollary to post-editing skills, and it remains essential not only for post-editing but for translation practice more generally. Translation industry analysists themselves recommend investing heavily in talent development and upskilling (Beninatto and Varga 2024: online). Seen as a learnable skill (Klein 20024), human expertise is grounded in powerful intuition that enables individuals to make effective decisions in complex situations. It also encompasses analytical subtlety and tacit knowledge which allows humans to recognize patterns, observe anomalies, sense familiarity or build mental models of how situations work. Naikar et al. (2025: 11) argue that AI systems may gradually weaken these abilities by limiting opportunities for insight, sensemaking, discovery, and critical engagement. In post-editing, this cognitive weakening can be observed in inappropriate post-editing practices such as producing "deceptively fluent but incorrect translations", using "inconsistent or non-compliant terminology", or generating "neural babble" (RWS n.d.: online), but also maybe in letting oneself overly influenced by technology, rather than making sure technology acts in one's own interest.

3 Method

3.1 Research design and objectives

This study is an observational, descriptive analysis of AI-mediated translation in translator training. It examines errors in a ChatGPT-generated German into Spanish translation—distinguishing between errors made by the system and those made by the students' post-editing decisions—and classifies them. The study therefore complements prior classroom MT research by focusing not only on the presence of errors in the final text, but also on how errors emerge, persist, or are introduced during post-editing by translation learners. The dataset is consciously small-scale and exploratory: working with 20 trainees on a single short text allows close, qualitatively informed observation of post-editing behaviour, at the cost of statistical generalisability. Claims throughout the paper are framed accordingly. The text dataset comprises 20 student identifiers (TIC 1–TIC 20). All participants completed the same post-editing task using the translation generated by ChatGPT-4.5 as starting point. For anonymization, students were assigned identifiers TIC 1–TIC 20.

3.2 Participants and instructional context

The participants were 20 upper-level undergraduate students enrolled in a specialised translation course covering the German–Spanish/Spanish–German pair within a Translation and Interpreting BA programme. All students were native speakers of Spanish. The BA programme establishes B2 as the target level of German, although in practice the trainees' proficiency was closer to B1. They had received introductory exposure to machine translation and post-editing in earlier modules of the curriculum but had no extensive prior PE experience. Participation was part of regular classroom activity; the task was formative, did not contribute to course grades, and students consented to the anonymised use of their post-edits for research purposes. The eight students who introduced no annotated changes (TIC 13–TIC 20) were not asked retrospectively about their decision-making, since the present study was designed as a product-focused observation; this point is taken up in Section 5 as a limitation.

3.3 Source text, MT system and prompt

The source material is a semi-specialized German economic news text on inflation. The full source text (*Inflation schwächer als erwartet*, dpa; 127

discard (translating from scratch) an MT result; the capacity to post-edit according to PE guidelines; the capacity to post-edit up to human quality (full PE); the capacity to post-edit to a good enough quality (light PE), the capacity to pre-edit a source language according to a controlled language model; the capacity to train and tune an MT engine; the capacity to identify MT output errors; the capacity to apply the right correction strategy; the capacity to advise when MTPE is appropriate for a text or project; the capacity to provide feedback for the MT solution engineers, and the capacity to learn about new technologies.

words) is reproduced in Appendix A, and the raw ChatGPT-4.5 output used as PE baseline is reproduced in Appendix B. The task domain was selected because economic reporting combines general-language fluency with domain-specific terminology, numeracy, and genre-sensitive conventions that typically generate post-editing challenges even when students translate into their L1.

Each student generated the baseline Spanish output individually by submitting the German source text to ChatGPT-4.5 with the same minimal prompt provided by the instructor: "Translate the following text into Spanish while preserving the journalistic register." Although the prompt was identical for all participants, the resulting outputs differed slightly across students owing to the non-deterministic behaviour of the model; the analysis therefore considers each student's own ChatGPT output as their individual starting point. The prompt was deliberately kept short and non-engineered to approximate a realistic use of an LLM by trainees with only basic knowledge of prompt design.

3.4 Task design and resources

Students were told to post-edit the full ChatGPT output rather than retranslate from scratch. In line with the mentioned DeepL study design, post-editing criteria emphasized accuracy/adequacy, fluency, and genre-appropriate style. The task was completed individually in a 45-minute classroom session. Students were allowed to consult online dictionaries, terminology databases and parallel texts in the Spanish economic press but were explicitly instructed not to use any other MT engine or LLM during the task. Instructions were communicated orally at the start of the session and made available on the course platform; their content can be summarised as: (i) generate a Spanish translation of the source text by submitting it to ChatGPT-4.5 with the prompt provided; (ii) compare the source text with the resulting Spanish output; (iii) post-edit that output so that it is accurate, fluent, and consistent with the journalistic register; (iv) consult dictionaries, terminology resources and parallel texts whenever necessary; and (v) do not use any further MT engine or LLM during post-editing.

3.5 Annotation and error attribution

Error annotation was carried out by the authors, who are trained translation scholars and instructors; one of them taught the course in which the data were collected. In this dataset, *ChatGPT errors* refers exclusively to problems present in the raw ChatGPT output, as established by the authors through independent analysis against the source text and target-genre conventions. The label thus reflects the authors' assessment of the machine output, not the students' own error detection.

Each segment was reviewed by comparing source text, raw ChatGPT output and each student's post-edited version, allowing the five behavioural categories used in Section 4 to be coded.

3.6 Cohort composition

Notably, eight students (TIC 13–TIC 20) show no annotated errors. Consequently, all coded instances in this dataset (N = 132) refer to the remaining 12 students (TIC 1–TIC 12). Where relevant, we report descriptive statistics for both the full cohort (n = 20) and the subset with recorded instances (n = 12).

These eight students completed the task and submitted post-edited outputs in which no segment fell into any of the five behavioural categories coded in this study: their post-edits left no uncorrected ChatGPT errors, introduced no new errors, and did not contain preferential reformulations. In other words, their interventions on the ChatGPT output were judged correct and necessary in every case where they intervened.

The present work adopts a bounded classroom task and an error-taxonomy approach grounded in observable output data rather than process measures, in continuity with the authors' earlier research on post-editing of economic texts in the DE–ES pair. The analysis is framed as descriptive and limited to product evidence — that is, without claims about cognitive effort or revision sequence — and focuses on the weaknesses of the workflow that can inform pedagogical intervention.

4 Analysis: ChatGPT vs Student Behavior by Category

Five behavioural categories were coded: (i) *ChatGPT error* detected and corrected by the student; (ii) *ChatGPT error carried into the post-edited text*; (iii) *change correct ChatGPT proposal to an incorrect alternative* (over-editing leading to error introduction); (iv) *both proposals incorrect* (the student replaces a ChatGPT error with a different error); and (v) *change correct ChatGPT proposal to another correct alternative*

(preferential change in the sense of Mossop, 2020). Throughout the analysis we use *over-editing* as an umbrella term for unjustified student interventions on segments that did not require revision, and reserve *preferential change* for category (v), where the student modification is not technically erroneous but adds no value relative to task norms.

4.1 Detected ChatGPT errors (corrected during post-editing)

We first describe the distribution of ChatGPT-originated errors that students detected and corrected during post-editing (instances coded as ChatGPT error). Considering only the rows labelled "ChatGPT error", the system mainly fails in terminology (10 instances), orthotypography (7), style (5), syntax (4) and tense (3). It produces very few errors in lexical choice (1), prepositions (2) and meaning transfer (2), and no pure spelling errors.

For this economic text, ChatGPT generates output that is formally accurate at the level of spelling but not fully aligned with the terminological and orthotypographical norms of the domain. The system tends to propose fluent, grammatically acceptable translations that nevertheless contain suboptimal terms, formatting inconsistencies or stylistic mismatches with German financial news discourse.

4.2 Student PE behaviour

When students fail to correct ChatGPT errors, that is, in "ChatGPT error carried into post-edited text", they mostly retain problems in tense (9 cases), terminology (7) and meaning transfer (3). These are precisely the categories that require more global interpretative decisions rather than local surface corrections.

When students degrade correct output ("change correct ChatGPT proposal to an incorrect" and "both proposals incorrect"), they introduce 35 errors, concentrated again in terminology, tense, orthotypography and syntax. The per-category breakdown of these 35 over-editing instances is reported in Appendix C, Table C1, where it is contrasted with the distribution of ChatGPT-originated errors and of preferential changes; this allows the reader to compare, for each linguistic category, the proportion of student interventions that improved, preserved or degraded the baseline output. In other words, students sometimes replace adequate machine solutions with less appropriate alternatives, particularly when negotiating specialist vocabulary or complex clause structure. Illustrative examples for the most affected categories—terminology, tense/aspect and meaning transfer—are provided in Appendix C, Table C2.

By contrast, when they replace a correct proposal with another correct one ("change correct ChatGPT proposal to another correct"), their interventions consist of preferential changes that focus on terminology (16 cases) and style (12), and only marginally on other categories. These changes are not classified as errors, but they are conceptually closer to over-editing than to value-adding revision, since they modify segments already adequate to the task.

Behavioural category	**N**	**%**
ChatGPT error (corrected)	37	28.0%
Correct → another correct (preferential)	37	28.0%
Correct → incorrect (over-editing)	28	21.2%
ChatGPT error carried over	23	17.4%
Both proposals incorrect	7	5.3%

Table 1. *Distribution of annotated instances by behavioural category (N = 132).*

Linguistic category	**N**	**%**
Terminology	45	34.1%
Gram (tense/aspect)	20	15.2%
Style	18	13.6%
Gram (syntax)	12	9.1%
Orthotypography — over-editing (correct rendering replaced by an incorrect one)	10	7.6%
Meaning transfer	8	6.1%
Lexicon	8	6.1%
Typographical rules	5	3.8%
Gram (prepositions)	3	2.3%
Spelling	3	2.3%

Table 2. Distribution of annotated instances by linguistic category (N = 132).

Table 1 reports the distribution of annotated instances by behavioural category (N = 132). Importantly, the category *change correct ChatGPT proposal to another correct* refers to acceptable reformulations rather than errors. When focusing on system-originated problems, ChatGPT-related issues account for 67 instances (50.8%): 37 were detected and corrected during post-editing, 23 were carried over into the post-edited product, and 7 remained incorrect after an unsuccessful student alternative. Student-introduced errors through over-editing on a correct

ChatGPT proposal account for 28 instances (21.2%). The remaining 37 instances (28.0%) reflect preferential changes (correct → correct).

Students display an ambivalent relationship with AI output: they tend to trust fluent segments even when they contain subtle semantic or aspectual problems, yet they also over-edit segments that are already acceptable. Their tendency to intervene in terminological and stylistic choices indicates a desire to assert authorial control or translator's agency, but it also reveals gaps in their ability to distinguish between necessary corrections and optional improvements, especially in the context of domain-specific German.

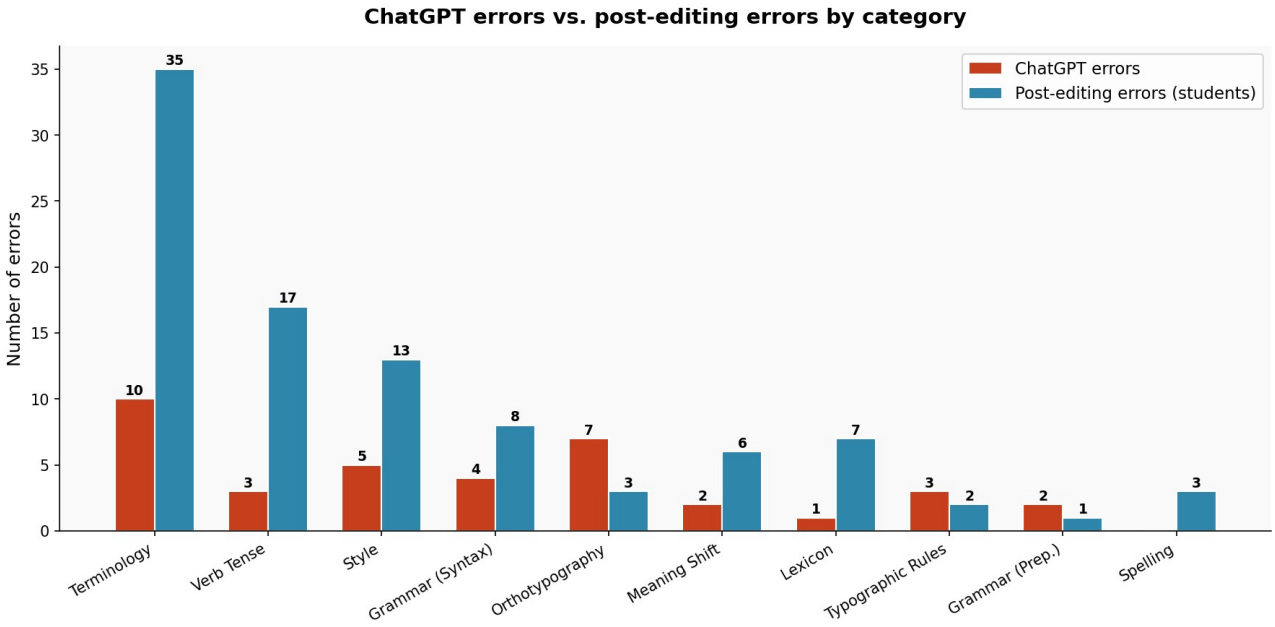

Figure 2. Category-level comparison: ChatGPT errors vs. PE errors.

However inter-student dispersion is substantial. In the full cohort (n = 20), total recorded instances range from 0 to 18 (M = 6.6, median = 5.5), reflecting the presence of an apparent error-free subgroup (TIC 13–TIC 20). Restricting the analysis to students with recorded instances (n = 12), totals range from 4 to 18 (M = 11.0, median = 10.5).

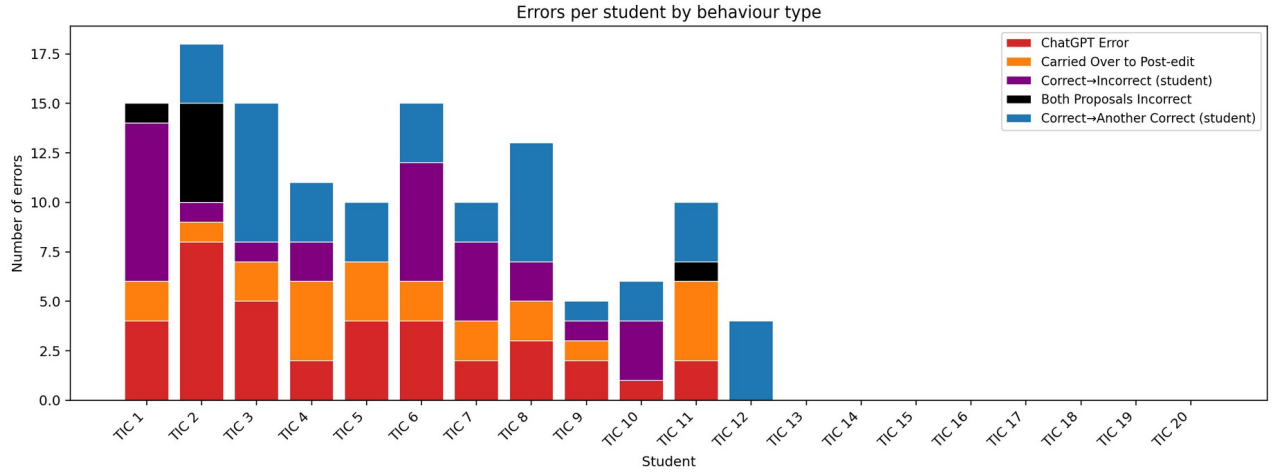

Figure 3. Errors per student by behaviour type. Students TIC 13–TIC 20 show no annotated errors.

5 Discussion

5.1 Interpreting the error distribution

The findings suggest that PE competence is at least as critical as baseline LLM-mediated translation quality in determining the final error profile of student translations. Students do not merely "repair" the output of ChatGPT; they also introduce new problems, particularly when they attempt to modify segments that are already acceptable according to task norms. The relatively high proportion of uncorrected errors further indicates that students do not always recognize problematic segments, especially when the output is fluent and superficially natural.

Two distinct mechanisms can be identified in the way errors enter the post-edited product. First, errors persist when students fail to detect existing ChatGPT errors and carry them into the final text, particularly in categories that require more global interpretative decisions, such as tense/aspect and meaning transfer (23 instances; category ii in § 4). Second, new errors are introduced when students replace a correct ChatGPT proposal with an incorrect alternative, a pattern especially visible in terminology and tense/aspect (28 instances; category iii). Beyond these two main mechanisms, 7 instances (category iv, both proposals incorrect) correspond to mixed trajectories in which a ChatGPT error was replaced by a different student-introduced error: the original error does not survive, but a new one takes its place. The two main mechanisms are roughly comparable in magnitude, suggesting that the final error profile is shaped as much by students' PE behaviour as by the quality of the LLM proposal itself. The 37 ChatGPT errors that students successfully detected and corrected (category i) do not enter the final text, but they remain analytically informative as evidence of the model's weaknesses and of students' detection capacity. Finally, 37 preferential changes (category v) do not affect the error profile and are discussed in § 5.2. The distribution of errors across linguistic categories is consistent with advanced L2 learners working with a specialized news text on economy (inflation): students are able to control basic grammar and spelling, but they struggle with domain-specific terminology, tense-aspect choices and stylistic alignment with the genre of financial journalism. The predominance of terminological errors confirms that the integration of LLM-based MT does not automatically solve specialist vocabulary issues and that students still need explicit support in identifying and validating domain-appropriate equivalents.

Terminology is the highest-risk domain across the entire workflow. While students correct a part of ChatGPT's terminological errors, the effect remains limited because a considerable number of terminological issues are left unrepaired in the final post-edited product and because students additionally introduce new terminology errors through over-editing. This suggests that terminological decision-making is not only a tool-output problem but also a competence issue: fluent target-language output may reduce perceived need for documentation, and domain plausibility can further make non-standard or context-inappropriate term choices more difficult to discover.

5.2 Implications for translation pedagogy

Student errors are produced by two mechanisms of similar magnitude: propagation and over-editing. Approximately half of final errors reflect unrepaired model output, whereas the remainder are errors introduced by students when replacing correct suggestions with incorrect alternatives. This dual-risk pattern indicates that pedagogical interventions should not focus exclusively on "detecting MT/LLM errors," but also on controlling revision behavior—i.e., when not to edit, and how to justify edits using explicit quality criteria.

The high frequency of replacing a correct version by another correct version suggests a form of preferential change without added value, conceptually close to over-editing. This may reflect low trust in machine output, a usual preference for reformulation, or not applying the conventional revision practices to PE. In PE contexts, such changes can be counterproductive because they increase cognitive load and create opportunities for new adequacy errors. A process-oriented classroom intervention could therefore require brief justifications for edits (accuracy, terminology, grammar, register, and cohesion) and evaluate unjustified rewrites as negative evidence of PE competence.

In this dataset, most tense/aspect errors produced by ChatGPT are not corrected, and additional tense/aspect errors are introduced during editing. Many of these errors involve mismatches in the rendering of German *Perfekt* forms within the Spanish indicative past system,

where the choice between *pretérito indefinido* and *pretérito perfecto compuesto* must be aligned with the explicit temporal anchoring of the news item (a specific month or quarter) and with the genre conventions of Spanish financial journalism. In some cases, the LLM selects an aspectually inappropriate form that goes undetected; in others, students replace an aspectually adequate form with one that breaks the temporal cohesion of the article. Selected examples are provided in Appendix C. This points to a specific didactic need: students should perform a dedicated "tense/aspect pass" during adequacy checking, focusing on temporal anchoring and aspectual framing in Spanish texts. It is a kind of error that seems strange but is becoming increasingly frequent amongst the students, as other colleagues commented too.

The substantial inter-student variability suggests heterogeneous PE strategies. Importantly, higher activity (more annotations and changes) is not necessarily equivalent to lower quality; some students may simply engage more strongly in reformulation, whereas others show higher rates of propagation or introduced errors. This variability suggests further research with triangulation with qualitative data (screen recordings, revision logs, retrospective interviews) to identify strategy profiles.

Finally, the presence of a error-free subgroup (TIC 13–TIC 20) is pedagogically informative. If interpreted as error-free outcomes, it suggests that a non-trivial proportion of trainees can avoid both error propagation and quality-degrading over-editing, making these cases valuable as "positive exemplars" for modelling effective PE behavior. This pattern also increases distributional skew, reinforcing the need to treat PE competence as heterogeneous and to design instruction that supports both low-performing and high-performing trajectories (e.g., peer modelling, strategy comparison, and structured decision-justification tasks).

A specific limitation of the present design is that the eight students who introduced no annotated changes (TIC 13–TIC 20) were not interviewed retrospectively, so we cannot distinguish between cases of confident non-intervention, under-engagement with the task, or strategic acceptance of fluent output. Disambiguating these profiles is left to future research.

Comparing with the forthcoming study on PE of DeepL-driven MT, the analysis foregrounded the persistence of domain-sensitive problem areas—most notably terminology and meaning-related adequacy checks—within a journalistic economic register. In the present ChatGPT-based dataset, the high-risk domains are the same: terminology is the most frequent category, and errors in tense/aspect and meaning transfer are also frequent and less repaired. The persistence of meaning-transfer and tense/aspect problems in the post-edited product is consistent with the observation that, as MT output becomes more fluent, adequacy issues become less salient. As Setkowicz-Ryszka notes, "adequacy errors are more difficult to detect in the seemingly correct output," (Setkowicz-Ryszka, 2025: 8) which helps explain why students may fail to identify problems that do not disrupt surface grammaticality. In parallel, the error trajectories observed in this study support the claim that PE is not only about correcting: students do not only miss errors but can also create them. This pattern closely matches the phenomenon described in literature as preferential changes, whereby revisions that are not required by the task or guidelines may be detrimental. Indeed, "preferential changes may actually introduce errors into the text" (Mossop, 2020: 179, 201 in Setkowicz-Ryszka). In our data, we can see this risk when students replace correct machine proposals with inferior alternatives, especially in terminology and tense/aspect, indicating that revision discipline and decision informed by evidence thresholds are central components of post-editing competence.

Taken together, the two studies suggest that, in German-Spanish economic news translation, the biggest risks are not primarily surface-level spelling or basic grammar, but rather decision-intensive domains where correct solutions depend on documentation, genre conventions, and refined semantic alignment.

Beyond category distributions, both studies point to a shared behavioral challenge: PE does not automatically "fix" the translated text but can also worsen output quality when students intervene without stable decision criteria. The DeepL study (forthcoming) documented instances where MT errors were carried over into the final text and cases where students replaced correct MT output with inferior alternatives. This new study shows the same dynamism: a substantial proportion of errors in the post-edited text results from propagation of system-originated issues, while a similar proportion stems from student-introduced errors (over-editing). This cross-tool

alignment indicates that the pedagogical target is not tool-specific troubleshooting alone, but a broader competence: knowing what to change, what to keep, and how to justify interventions under task constraints.

A difference between the studies concerns the perceived "baseline" affordances of the systems. DeepL output often resembles conventional NMT behavior: locally plausible phrasing with systematic domain vulnerabilities that students may or may not detect. ChatGPT, by contrast, tends to produce highly fluent target-language prose, which can reduce overt surface errors but also increase the risk of "plausible fluency" masking adequacy problems. This difference matters pedagogically: whereas DeepL output may prompt attention to more visible MT artefacts, ChatGPT output may tempt students into either acceptance without verification (leading to error propagation) or reformulation without improvement (increasing over-editing). Yet, despite this contrast in surface presentation, the competence demands converge. Both datasets support the need for an adequacy-first PE protocol: terminology verification, tense/aspect checking, and meaning transfer validation come first, before stylistic polishing.

6 Conclusions

The combined analysis of error types and linguistic categories highlights three didactic priorities for translator training in MT-supported environments that integrate LLMs.

First, literacy in LLM-generated MT and error awareness should be developed explicitly. Although ChatGPT provides a generally fluent raw output with comparatively few surface-level problems (e.g., spelling and basic grammatical issues), this fluency can make it more difficult to detect higher-risk weaknesses in texts on specialized topics. Training should therefore focus on terminological adequacy, tense/aspect choices, and subtle meaning-transfer distortions—through guided analysis of raw LLM- output, with special attention to segments that read well but are semantically or pragmatically misleading.

Second, decision-making strategies in PE are crucial. Students sometimes under-edit by allowing ChatGPT errors to persist in the final text, but sometimes over-edit by replacing correct machine proposals with lower-quality alternatives. Classroom tasks should therefore target decision thresholds by requiring students to justify interventions as necessary, optional, or counterproductive in relation to task specifications, communicative purpose, and possible client expectations. A structured two-pass QA protocol—adequacy-first (terminology, tense/aspect, meaning transfer) and form-second (orthotypography, punctuation, stylistic smoothing)—can help reduce both error propagation and unnecessary rewriting.

Third, genre- and domain-sensitive editing strategies are highly important. The most frequent and consequential problems are related to terminology, tense/aspect, and genre-aligned style. One practical approach is to elaborate small bilingual glossaries and style sheets, then use them as normative reference points when evaluating both ChatGPT output and human revisions.

The findings suggest that PE training should not be reduced to correcting "obvious mistakes." Rather, it should cultivate a strategic, norm-oriented competence that enables students to leverage the advantages of LLMs while systematically compensating for their limitations in specialized communication. This includes both error detection and revision discipline.

The present results should be read in light of clear limitations: a single short text (127 words), a single LLM (ChatGPT-4.5), 20 trainees of which only 12 produced annotated instances, and a product-based design that does not capture process data. The marked inter-student variability observed in the dataset should be further investigated combining quantitative error annotation with process data (e.g., screen recordings, keystroke logs, or retrospective protocols) to explain why certain PE strategies lead to error propagation or error introduction.

Finally, the fact that eight students present no annotated errors highlights marked heterogeneity in PE outcomes and supports differentiated, strategy-focused pedagogical interventions. Future research should also test whether comparable patterns emerge across text types and LLM systems (including replication on additional German–Spanish source texts and on contrasting LLM/NMT baselines) and evaluate targeted pedagogical interventions designed to strengthen critical PE skills in translator trainees.

References

Angelone, E. 2010. Uncertainty, uncertainty management and metacognitive problem solving in the translation task. In G. M. Shreve & E. Angelone

(Eds.), Translation and Cognition. John Benjamins, 17–40.

Baumgarten, Stefan (2025). Welcome to the Machine? Prolegomena zu einer kritischen Translationswissenschaft in Zeiten ungebremster Maschinisierung. Yearbook of Translational Hermeneutics 5.1, 59-85.Bohm, David. 1980/2002. Wholeness and the Implicate Order. Ark.

Bhattacharyya, P., Chatterjee, R., Freitag, M., Kanojia, D., Negri, M., & Turchi, M. (2022). Findings of the WMT 2022 Shared Task on Automatic Post-Editing. In Proceedings of the Seventh Conference on Machine Translation (WMT) (pp. 109–117). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.5

Beninatto, R., and L. K. Varga (2024). The Language Industry Curve. Nimdzi Insights. https://www.nimdzi.com/the-language-industry-curve/.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. The Quarterly Journal of Economics, 140(2):889–942.

Castillo Bernal, P. (2022). La traducción literaria asistida por ordenador aplicada a la novela histórica (alemán-español): entrenamiento y comparación de sistemas de traducción automática. Quaderns de Filologia: Estudis Lingüístics, 27, 41–58. https://doi.org/10.7203/qf.0.24624

Chatterjee, R., de Souza, J. G. C., Negri, M., & Turchi, M. (2016). The FBK Participation in the WMT 2016 Automatic Post-editing Shared Task. In Proceedings of the First Conference on Machine Translation (Vol. 2, pp. 745–750). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2377

Chatterjee, R., Federmann, C., Negri, M., & Turchi, M. (2019). Findings of the WMT 2019 Shared Task on Automatic Post-Editing. In Proceedings of the Fourth Conference on Machine Translation (Vol. 3, pp. 11–28). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5402

Chen, S., & Zhou, T. (2025). The scale-integration paradox: A contrastive bibliometric analysis of global and Chinese machine translation post-editing research. Cogent Arts & Humanities, 12(1), Article 2584420. https://doi.org/10.1080/23311983.2025.2584420

Ginovart Cid, C., & Oliver, A. (2020). The post-editor's skill set according to industry, trainers and linguists. Portsiel J, editor. Maschinelle Übersetzung für Übersetzungsprofis. Berlin: BDÜ Weiterbildungs-und Fachverlagsgesellschaft mbh; 2020. Online at: https://repositori.upf.edu/bitstream/handle/10230/46285/ginovart_bdu_postedi-tor.pdf?sequence=1&isAllowed=y.

Guerberof Arenas, A. (2020). Pre-editing and post-editing. In E. Angelone, M. Ehrensberger-Dow, & G. Massey (Eds.), The Bloomsbury Companion to Language Industry Studies, London/New York/Oxford/New Delhi/Sydney, Bloomsbury Academic, p. 333–360.

Guerberof Arenas, A., Toral, A., Bisazza, A., & Sarti, G. (2022). DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7795–7816). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.532

Hassani, Ghodrat, and Mohammad Malekshahi. 2024. Navigating the continuum: Exploring the value of pluralism in translation uncertainty research. Journal of Language Horizons, 8(1):93–114.

Holgado-Sáez, C., & Martínez Martínez, S. (2025). Análisis de errores terminológicos de unidades léxicas aisladas en el lenguaje jurídico-administrativo nacionalsocialista: traducción automática neuronal vs. traducción humana (alemán-español). Estudios de Lingüística de la Universidad de Alicante (ELUA), 44. https://doi.org/10.14198/elua.27618

Kristeva, Julia. 1984. Revolution in Poetic Language. Translated by Leon S. Roudiez. Columbia University Press, New York.

Krüger, Ralph. 2018. Technologieinduzierte Verschiebungen in der Tektonik der Translationskompetenz. trans-kom, 11(1):101–137.

Kußmaul, Paul. 1986. Übersetzen als Entscheidungsprozess. Die Rolle der Fehleranalyse in der Übersetzungsdidaktik. In M. Snell-Hornby (Ed.), Übersetzungswissenschaft. Eine Neuorientierung. Francke, 206–229.

Levý, Jiří. 2000 [1967]. Translation as a decision process. In L. Venuti (Ed.), The Translation Studies Reader. Routledge, 148–159.

Lu, Y. (2022). A bibliometric analysis of machine translation post-editing from 2012 to 2021. In 2022 International Conference on Electronics, Computers and Natural Language Processing in Information Retrieval (ECNLPIR) (pp. 65–69). IEEE. https://doi.org/10.1109/ECNL-PIR57021.2022.00020

Martín de León, Celia, and José M. Cardona Guerra. 2022. Spoiled for choice?: Uncertainty facing options in translation. Translation & Interpreting, 14(2):50–67.

Mossop, Brian. 2020. Revising and Editing for Translators. Routledge.

Naikar, Neelam, Robert Hoffman, Emilie M. Roth, Gary Klein, Laura G. Militello, and Cindy Dominguez (2025). "Should we Make AI More

Tool-like or Teammate-Like?" Journal of Cognitive Engineering and Decision Making, vol. 0(0) 1–21.

Nitzke, J., & Hansen-Schirra, S. (2021). A short guide to post-editing (Translation and Multilingual Natural Language Processing 16). Berlin: Language Science Press.

Pym, Anthony. 2010. Exploring Translation Theories. Routledge, London.

Pym, A. (2011). Democratizing translation technologies – The role of humanistic research. Paper presented at the Luspio Translation Automation Conference, Rome, April 5, 2011.

Robinson, Douglas. 1991. The Translator's Turn. The Johns Hopkins University Press, Baltimore and London.

Roiss, S. (2021). Y las máquinas rompieron a traducir... Consideraciones didácticas en relación con la traducción automática de referencias culturales en el ámbito jurídico. TRANS: Revista de Traductología, 25, 491–505. https://doi.org/10.24310/trans.2021.v1i25.11978

RWS. (n.d.). Post-editing machine translation (MTPE): Resources. https://www.rws.com/localization/products/resources/post-editing-machine-translation/.

Sarasso, Pietro, Gianni Francesetti, Jan Roubal, Michela Gecele, Irene Ronga, Marco Neppi-Modona, and Katiuscia Sacco. 2022. Beauty and uncertainty as transformative factors. Frontiers in Human Neuroscience, 16:906188.

Setkowicz-Ryszka, Anna. 2025. Post-editing of Machine Translation Output in the Training of Legal Translators. Doctoral dissertation, University of Łódź.

Sinitsyna, Daria. 2025. Generative AI for translation in 2025. Intento blog, 30 March 2025. https://inten.to/blog/generative-ai-for-translation-in-2025/.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8:2293–2303.

Varela Salinas, M. J. , & Burbat, R. (2023). Google Translate and DeepL: Breaking taboos in translator training. Observational study and analysis. Ibérica, (45), 243-266.

Vasconcellos, M. (1987). "Postediting on-screen: machine translation from Spanish into English". In A profession on the move: proceedings of translating and the computer 8, ed. C. Picken, 133-146. London: Aslib.

Velazquez, D., Grace, M., Karageorgos, K., Carin, L., Schliem, A., Zaikis, D., & Wechsler, R. (2025). LangMark: A multilingual dataset for automatic post-editing. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1569

Veselovsky, V., et al. 2021. [Full bibliographic reference to be completed — cited in Section 2 of the original paper but missing from its reference list.]

Vinchon, Florent, V. Gironnay, and Todd Lubart. 2024. GenAI creativity in narrative tasks. Journal of Intelligence, 12(12):125.

Wang, Pei. 2008. What do you mean by "AI"? In P. Wang, B. Goertzel, and S. Franklin (Eds.), Artificial General Intelligence 2008. IOS Press, 362–373.

Wang, S. (2023). Towards English-centric zero-shot neural machine translation: The analysis and solution. In 2023 International Conference on Sensors, Electronics and Computer Engineering (ICSECE) (pp. 1577–1582). IEEE. https://doi.org/10.1109/ICSECE58870.2023.10263323

Wilss, Wolfram. 2007. The impact of uncertainty on translation. In R. Jääskeläinen, T. Puurtinen, and H. Stotesbury (Eds.), Text, Processes, and Corpora. Joensuu University, 163–170.

Winner, L. (2014). Technologies as Forms of Life. In: Sandler, R.L. (eds), Ethics and Emerging Technologies. Palgrave Macmillan, London.

Appendix A. Source text (German)

Inflation schwächer als erwartet

dpa. 28.09.2017. In Spanien sind die Verbraucherpreise im September schwächer als erwartet gestiegen. Das nach europäischer Methode errechnete Preisniveau (HVPI) lag 1,9 Prozent höher als im Vorjahresmonat, wie das Statistikamt INE am Donnerstag in Madrid nach einer ersten Schätzung mitteilte.

Volkswirte hatten im Schnitt mit einer Inflation von 2,0 Prozent gerechnet. Im August hatte die Inflationsrate bei 2,0 Prozent gelegen und im Juli bei 1,7 Prozent.

Im Monatsvergleich stiegen die spanischen Verbraucherpreise im September um 0,6 Prozent. Auch in dieser Abgrenzung hatten Experten einen stärkeren Preisanstieg erwartet.

Am Freitag stehen Inflationsdaten aus der Eurozone auf dem Programm. Im gemeinsamen Währungsraum lag die Teuerung im August bei 1,5 Prozent und damit weiter deutlich unter der von der Europäischen Zentralbank (EZB) angepeilten Zielmarke von knapp zwei Prozent.

Appendix B. Raw ChatGPT-4.5 baseline output (Spanish)

Inflación más débil de lo esperado

dpa. En España, los precios de los consumidores aumentaron en septiembre menos de lo esperado. El nivel de precios calculado según el método europeo (Índice Armonizado de Precios al Consumo, HVPI) fue un 1,9 por ciento más alto que en el mismo mes del año anterior, según informó el jueves en Madrid el Instituto Nacional de Estadística (INE) tras una primera estimación. Los economistas habían esperado una inflación del 2,0 por ciento. En agosto, la tasa de inflación fue del 2,0 por ciento y en julio del 1,7 por ciento.

En comparación con el mes anterior, los precios de los consumidores en España subieron un 0,6 por ciento en septiembre. También en este caso, los expertos esperaban un aumento de precios más fuerte.

El viernes se publicarán los datos de inflación de la zona euro. En la zona del euro, la inflación en agosto fue del 1,5 por ciento, continuando significativamente por debajo del objetivo del 2 por ciento que tiene el Banco Central Europeo (BCE).

Appendix C. Detailed error breakdown and illustrative examples

Table C1. Distribution of annotated instances by linguistic category and origin (N = 132). ChatGPT errors aggregate cases where the source of the error is the MT system, irrespective of whether the student corrected it (37 corrected + 23 carried over + 7 cases where both proposals were incorrect). Over-editing refers to instances where students replaced a correct ChatGPT proposal with an incorrect alternative (correct→incorrect). Preferential changes refers to instances where students replaced a correct proposal with another correct one (correct→correct). Categories are ordered by total frequency.

Linguistic category	ChatGPT errors	Over-editing	Preferential changes	Total
Terminology	20	9	16	45
Gram (tense/aspect)	14	4	2	20
Style	5	1	12	18
Gram (syntax)	6	3	3	12
Orthotypography	7	3	0	10
Meaning transfer	5	3	0	8
Lexicon	3	2	3	8
Typographical rules	3	2	0	5
Spelling	2	1	0	3
Gram (prepositions)	2	0	1	3
Total	**67**	**28**	**37**	**132**

Table C2. Illustrative examples by linguistic category and behavioural origin. Each row shows the German source segment, the ChatGPT raw output, and the student-post-edited version, classified according to the behavioural typology in Table C1.

Category	DE source	ChatGPT MT	Student PE
Tense/aspect — ChatGPT error carried over	*Auch in dieser Abgrenzung hatten Experten einen stärkeren Preisanstieg erwartet.*	También en esta comparación, los expertos esperaban un aumento mayor.	También en esta medida los expertos esperaban un incremento de precios mayor.
Tense/aspect — ChatGPT error corrected by the student	*In Spanien sind die Verbraucherpreise im September schwächer als erwartet gestiegen.*	En España, los precios al consumidor subieron en septiembre menos de lo esperado.	En España, los precios al consumidor han subido este mes de septiembre menos de lo esperado.

Syntax / calque — ChatGPT error carried over	*Auch in dieser Abgrenzung hatten Experten einen stärkeren Preisanstieg erwartet.*	También en esta comparación, los expertos esperaban un aumento mayor.	También en esta comparación, los expertos esperaban un aumento mayor de los precios.
Terminology — over-editing (correct rendering replaced by an incorrect one)	*nach einer ersten Schätzung*	en una primera estimación	basándose en un cálculo aproximado
Terminology — over-editing (correct rendering replaced by an incorrect one)	*die Verbraucherpreise*	los precios al consumidor	los precios al por menor
Orthotypography — over-editing (correct rendering replaced by an incorrect one)	*1,9 Prozent*	1,9 por ciento	1.9 %
Orthotypography — over-editing (correct rendering replaced by an incorrect [incoherent] one)	*1,5 Prozent … knapp zwei Prozent*	1,5 % … casi el 2 %	1,5 % … casi el dos por ciento
Style — preferential change (correct rendering replaced by another correct one)	*das Statistikamt INE*	la oficina de estadística INE	el INE
Style — preferential change (correct rendering replaced by another correct one)	*das Statistikamt INE*	la oficina de estadística INE	el Instituto Nacional de Estadística (INE)
Tense/aspect — ChatGPT error corrected by the student	*Auch in dieser Abgrenzung hatten Experten einen stärkeren Preisanstieg erwartet.*	También en esta comparación, los expertos esperaban un aumento mayor.	También en esta medición los expertos habían anticipado un incremento más acusado.