A constraints-first approach to CAT and generative AI in translator education: a simulated client project in a master's course

MAGGIE HUI

The Hong Kong Polytechnic University, Hong Kong

This paper presents a constraints-first, CAT- and GenAI-assisted simulated client project in a 13-week master's translation course (58 students in 12 groups). Students completed a group project requiring a consideration table, sequenced use of two CAT tools, capped trials (≤10 prompts) of multiple generative models and a four-pass MQM-lite QA loop (terminology → accuracy → fluency → register/style). Analysis of one group's English-to-Chinese NVIDIA corporate-technical web-copy project documents typical GenAI risks through concrete artifacts and descriptive tallies. The study offers a replicable, standards-aligned classroom protocol supported by fully audited artifacts, while making no claims about efficiency gains or model superiority.

Keywords: Computer-assisted translation (CAT), generative AI, post-editing, translator education, simulated client project, terminology management, workflow design, strategic competence

Introduction

Professional translators today rarely work without tools (Christensen and Schjoldager, 2016). Standards such as ISO 17100 and ISO 18587, along with quality frameworks like Multidimensional Quality Metrics (MQM), now shape how translation is managed in companies. At the same time, translator education frameworks (EMT 2022 and PACTE) remind us that students need service provision, quality and strategic skills.

Yet there is still a noticeable gap. Industry expects auditable, constraint-driven workflows; many classrooms still offer only tool-familiarization exercises that lack real risk management. Recent research shows that the impressive fluency of neural systems can mask serious problems with accuracy and terminology (Koponen, 2016; Läubli, Sennrich, and Volk, 2018;

Yamada, 2019). These findings motivate staged quality checks that put terminology and accuracy first.

This article reports what happened when one master's-level student group ("Group 9") tackled a realistic English-to-Chinese (EN→CH) corporate-technical web-copy project. The source material consisted of 1,200 words from three closely related public NVIDIA blog posts announcing an "AI Blueprint for telco network configuration." Working under a demanding simulated client brief, the five students were required to strictly preserve all brand and product names (e.g., NVIDIA, NIM, AI Blueprint, BubbleRAN), consistently use a 17-item protected-term glossary they created themselves, maintain a concise corporate-technical register, avoid any unauthorized additions or summaries and provide full auditability of every AI-assisted decision.

The group employed three generative models (free ChatGPT, ChatGPT Plus and Microsoft Copilot) under a strict didactic cap of ≤10 prompts per model. They followed a carefully designed constraints-first workflow: an upfront consideration table to operationalize client requirements, sequenced use of two CAT tools (Trados Studio 2022 followed by Smartcat) for termbase and translation memory governance and a four-pass MQM-lite QA protocol (terminology → accuracy → fluency → register/style).

Drawing explicitly on Pym's (2013) emphasis on risk management as a foundational translator competence that outlasts any specific technology, the workflow prioritizes constraint identification, auditability and strategic decision-making over unchecked tool enthusiasm. In practice, this meant that typical GenAI risks, such as Copilot's tendency to insert speculative benefits not present in the source and the models' production of heavy pre-modifier structures that feel unnatural in formal Chinese, became visible and correctable at the appropriate stage.

Our aim is simply to describe what risks appeared and how the group handled them. We offer a ready-to-use classroom template, fully specified in Appendices A-C (the consideration table, the 17-item protected-term list, the three-stage prompting protocol and the MQM-lite revision log schema), that other programs can try and adapt.

The rest of this paper is organized as follows: It reviews the relevant literature, describes the research methods, presents the results from Group 9, discusses the implications, outlines the limitations and concludes.

Literature Review

Standards and quality frameworks shaping professional practice

Professional translation and post-editing are now guided by process and quality standards that spell out roles, documentation and verification steps.

ISO 17100:2015 sets out requirements for translation services, covering translator competence, project management, review processes (revision and proofreading) and the need for documented procedures (ISO, 2015). The standard separates responsibilities: the translator produces the first version, a reviser checks it against the source for accuracy and terminology, and a proofreader makes sure the target text reads correctly. This division of labor helps reduce individual bias and catch mistakes that one person might overlook.

ISO 18587:2017 focuses specifically on post-editing machine translation output. It distinguishes between light post-editing (mainly for quick understanding) and full post-editing (ready for publication) (ISO, 2017). The standard emphasizes meaning preservation, terminology consistency and proper quality control, particularly for full post-editing. It also reminds post-editors to watch out for typical machine-translation problems such as omissions, additions and misleading fluency. Together, these two standards offer a practical scaffold for designing classroom workflows: even when full commercial rigor is not possible, teachers can still adopt role separation, staged checks and audit trails.

For judging output quality, Multidimensional Quality Metrics (MQM) provides a category-based framework that makes quality issues visible and easy to classify (Lommel et al., 2014). MQM breaks quality into clear error types (for example, accuracy, terminology, fluency, style and consistency) and supports weighted scoring. In teaching situations, a simplified "MQM-lite" version (adapted from Lommel et al., 2014) that retains only the four most pedagogically relevant categories (terminology, accuracy, fluency and style/register) helps students focus on the most critical professional quality aspects while keeping the QA process manageable. The present study uses this reduced set to organize the revision logs.

Post-editing research: effort, error types and fluency-accuracy tensions

Research on post-editing (PE) has repeatedly shown that smooth surface fluency of machine translation (MT) output can hide real meaning errors. In this context, fluency refers to the linguistic quality of the target text (naturalness, grammatical correctness and readability), as opposed to accuracy, which concerns faithful preservation of the source meaning. Effort also tends to cluster around specific error types rather than simply the number of words. The three main types of post-editing effort (temporal, technical and cognitive) were first proposed by Krings (2001) and have since been widely adopted. Koponen's (2016) review synthesizes findings from both statistical and neural MT systems, showing that cognitive effort does not always align with temporal or technical effort. A fluent but inaccurate sentence may require few keystrokes yet demand substantial mental effort to identify and correct.

Yamada's (2019) classroom study of students post-editing Google Neural Machine Translation found that although quality improved, students still struggled with typical NMT problems related to accuracy and terminology. They often trusted fluent output too readily and missed subtle meaning shifts or omissions. Yamada concluded that structured QA checklists and explicit training in error spotting are essential to overcome fluency bias. On a broader scale, Läubli et al. (2018) showed that human judges sometimes prefer more fluent MT output even when its adequacy is weaker, a tendency they called "fluency bias." When evaluators compared human and machine versions side by side, they frequently rated the machine text higher in fluency but lower in accuracy, yet their overall preference still leaned toward fluency.

In this context, accuracy refers to the faithful preservation of the source text's meaning (no unauthorized additions, omissions, or semantic shifts beyond approved translation strategies), aligning with ISO 18587's core requirement for meaning preservation and MQM's accuracy category (Lommel et al., 2014). These consistent findings explain why staged QA sequences (i.e., deliberate, ordered passes that prioritize terminology and accuracy before fluency or style) are so valuable: they force students to address the most critical risks first rather than being misled by fluent but inaccurate output. In the present study, Pass 1 (terminology) and Pass 2 (accuracy) are completed first, deliberately protecting the work from fluency-masked errors.

Translator-education perspectives: competence, risk and process transparency

Competence frameworks in translator education have long stressed that technical tool skills must sit inside larger service provision, quality and strategic competences. The EMT 2022 framework consists of six areas: language and culture, translation, technology, personal and interpersonal, service provision and thematic competence (EMT Expert Group, 2022). "Technology competence" goes beyond operating tools; it includes managing translation memories, termbases and machine translation systems, plus critically evaluating their output. "Service provision competence" covers project management, quality assurance and client communication. The present study puts both into practice: technology competence through dual-CAT sequencing and capped LLM trials and a simulated client brief captured in the consideration table, an upfront tool that operationalizes client constraints and register requirements

The PACTE model similarly highlights strategic competence as the central skill that coordinates tasks, including selecting procedures, spotting errors and solving problems (PACTE Group, 2017). Kiraly's (2000) social-constructivist approach shifted translator education toward authentic, collaborative problem-solving. The present study's simulated client project,

with its consideration table, role rotation and full artifact submission, directly embodies both PACTE's emphasis on strategic competence and Kiraly's principles by positioning students as active co-creators of knowledge in a realistic workflow.

Pym (2013) makes a strong case for training that puts risk management ahead of tool enthusiasm. Responding directly to assess risks (such as terminology drift, legal liability, or client dissatisfaction) and deciding where to spend effort is a lasting skill. In the present protocol operationalizes client constraints through the upfront consideration table and imposes a didactic prompt cap that forces students to make strategic decisions about when further prompting is no longer worthwhile.

Bowker and Buitrago Ciro (2019) call for machine-translation literacy, which is the ability to understand what MT and similar technologies can and cannot do and how to use them responsibly in workflows. The present study extends this literacy to generative AI by adding prompt design, controllability checks and disciplined logging.

Taken as a whole, these three perspectives – Kiraly's emphasis on authentic collaborative tasks, Pym's focus on enduring risk management and strategic decision-making and Bowker and Buitrago Ciro's call for critical technology literacy – provide a coherent pedagogical foundation for the constraints-first, auditable workflow examined in this study.

Emerging use of generative systems in translation workflows and education

Pym and Hao (2025) suggest viewing generative AI as an augmentation of human skill rather than a replacement. They recommend structured integration linked to clear learning outcomes: use AI for drafting and paraphrasing but always require human verification; teach students to document their AI use; and focus on tasks where AI reliably saves effort while avoiding high-risk tasks (such as legally binding texts) without close oversight.

Peer-reviewed studies on large language models (LLMs) in professional translation and translator education are growing, and several patterns have already emerged. First, LLMs produce fluent drafts very quickly, yet they are prone to hallucinations (factually wrong or invented content) and unauthorized additions (material not present in the source) (e.g., Guerreiro et al., 2023). Second, their ability to follow controlled terminology differs widely depending on the model and the prompt. Third, they show some sensitivity to register but can over-adjust or under-adjust if the prompt is not specific enough.

The existing standards (ISO 17100/18587) and quality taxonomies (MQM) were written before LLMs became widespread, but their process- and issue-based logic still works well. For instance, ISO 18587's requirement to "ensure that the meaning of the source text is preserved" directly tackles the unauthorized additions and omissions that are common LLM problems.

MQM's "accuracy" category can flag hallucinated content, and its "terminology" category can catch drift. Educators therefore do not need to invent entirely new frameworks; they can adapt to the ones we already have.

Gap addressed by this study

Taken together, the literature offers consistent advice: classroom workflows should mirror professional process and quality standards, fluency alone should never be accepted as proof of quality and students need to develop strategic and risk-management skills alongside tool use. What is still missing, however, are concrete, replicable classroom protocols that combine CAT assets with generative systems in an auditable and standards-congruent way. Most existing studies focus either on CAT tools alone (e.g., Kornacki, 2018), on PE without LLMs (e.g., Koponen, 2016), or on LLM use without proper CAT governance (e.g., Guerreiro et al., 2023). The present single-case study addresses this gap by describing and analyzing one such orchestration in detail. It does not claim generalizable performance advantages over human-only or MT-only baselines; instead, it offers a practical template that other programs can adapt, evaluate and refine.

Methodology

Study design and scope

This is a single-case exploratory study focusing on one student group ("Group 9") as they carried out a GenAI-supported CAT workflow translating from English into Chinese corporate-technical web copy taken from public NVIDIA blog material.

The purpose was descriptive: to document a constraints-aware classroom orchestration and to show how specific risks, such as unauthorized additions and terminology drift, became visible and were addressed. Claims about efficiency advantages over human-only or MT-only workflows, cross-group generalization and inferential statistics were deliberately kept out of scope. Group 9 was chosen because the group submitted complete artifacts as part of the course assessment and consented to in-depth analysis. It serves as a best-case illustration rather than a representative sample. Selection was based on completeness of artifacts and consent, which introduces a best-case bias in visible process quality and may under-represent typical error rates.

Participants and context

The study was conducted in a master's-level translation course at the Hong Kong Polytechnic University in the 2025 autumn semester with 58 students in

12 groups and 39 contact hours over 13 weeks. Group members rotated roles across assignments. For this project, the five members of Group 9 collaborated jointly on terminology, drafting and QA tasks.

Source text and assets

The source text consisted of approximately 1,200 words from three closely related public NVIDIA web pages (one main blog post and two related GTC announcements) concerning the "AI Blueprint for telco network configuration" within the broader GTC-related communications context. These texts were publicly available on the official NVIDIA blog and contained only text (no images or other multimodal elements).

Group 9 created their own protected-term list and a bilingual glossary of 17 items and built the termbase and translation memory during the task. No external gold-standard term list was provided.

The consideration table containing client constraints and register information and the 17-item protected-term list used by Group 9 are reproduced in Appendices A and B, respectively.

Workflow orchestration (four-pass QA loop)

The project followed an instructor-designed constraints-first, four-pass QA loop adapted from the logic of ISO 17100/18587 and MQM:

- Pass 1 for terminology setup and enforcement: Build and verify the protected-term list and termbase; enforce glossary checks; repair all terminology issues before any drafting or stylistic work.
- Pass 2 for accuracy: Perform sentence-aligned source–target comparison to ensure meaning preservation (i.e., no unauthorized additions, omissions, or semantic shifts beyond approved strategies).
- Pass 3 for fluency and cohesion: Smooth sentence-level fluency and local coherence without altering protected terms or meaning.
- Pass 4 for register and style: Align output to the enterprise web register and style sheet; conduct final QA; prepare deliverables.

The group followed the three-stage prompting protocol described in Appendix C. Full prompt and revision logs were maintained throughout for auditability.

Tools and configuration

The group used two CAT environments: Trados Studio 2022 (desktop) for building the termbase and translation memory and Smartcat (cloud) for verifying consistency and terminology across the three related texts. Although the verification step could have been performed entirely in Trados Studio,

Smartcat's cloud-based platform facilitated efficient cross-document consistency checking and real-time team collaboration. They first constructed the termbase and TM in Trados, then migrated the revised content to Smartcat to check term and TM reuse. Protected names and product lines were kept verbatim and glossary checks were used to flag any deviations. Numeric, date and tag consistency were handled through the standard CAT QA profiles.

A third-party MT plugin (Xiaoniu Translation) was used inside Trados due to version constraints; this is noted simply as a student-level detail rather than a general recommendation.

The exact settings applied by Group 9 are summarized in Table 1.

Table 1: CAT Tool Settings and QA Configuration Used by Group 9

Parameter	Setting
Desktop CAT	Trados Studio 2022
Cloud CAT	Smartcat *(exact version not recorded)*
TB/Glossary enforcement mode	Strict enforcement (manual override allowed)
Matching	Exact matching, case-sensitive for branded terms
Variants handled via	Manual addition to glossary; no regex used
Numbers and Measurements	Flag mismatches; unit consistency enforced for numerals
Tags/inline codes	Must match count and order (default QA profile)
Punctuation/spacing	Full-width for Chinese commas/periods; half-width for embedded English; non-breaking space before %
Terminology	Flag deviations from protected list; forbid paraphrasing of brand/product family names
Capitalization	Sentence-initial and proper-name rules; brand capitalization locked
Dates/numerics	Source pattern (e.g., "2024-03-25") → target "2024年3月25日"
Repeated segments	Enforce consistency on repetitions; alert on divergent translations
Custom regex	None used in this project

Notes: All matches were reviewed manually; auto-acceptance was disabled. Match leverage (as reported by Smartcat): 101 % context matches: 12; 100 % exact matches: 8; repetitions: 23; fuzzy bands (95-99 %): 6; fuzzy bands (85-94 %): 11.

Generative model usage, data collection and analysis

Three models were compared: the free version of ChatGPT (powered by GPT-4o mini), ChatGPT Plus (powered by GPT-4o) and Microsoft Copilot (powered by a GPT-4-based model as of autumn 2025). The group applied the same three core prompts to each model: (1) baseline translation, (2) glossary-constrained revision and (3) consideration-table-constrained refinement. Copilot required three additional targeted prompts in this case.

The full three-stage prompting protocol and the additional prompts required by Copilot are provided in Appendix C.

A didactic prompt cap of ≤10 prompts per model was applied as a teaching constraint. Every prompt, model output and human edit was logged with timestamps.

Collected artifacts included the consideration table, protected-term list, TB/TMX exports, CAT QA notes, prompt log excerpts, segment-level before/after examples and the final bilingual deliverable.

Only descriptive tallies were produced (no composite scores). Categories were protected-term deviations, unauthorized additions, numeric/date mismatches, terminology inconsistencies (non-protected items) and fluency issues. Operational definitions were applied consistently in the revision log. The group logged observations; the instructor reviewed them for face validity. No inter-rater statistics were calculated because of the single-group, pedagogical nature of the study.

Ethics, replicability and deviations from ideal practice

The study was conducted under the ethical guidelines of The Hong Kong Polytechnic University for classroom-based pedagogical research. Students provided informed consent for the use of their anonymized artifacts in this analysis. Students were permitted to use GenAI for drafting and paraphrasing under strict constraints, but they remained fully responsible for meaning preservation and terminology compliance. Final submissions could include GenAI-originated text only if it passed Pass 1 (terminology) and Pass 2 (accuracy). Prompt and revision logs were submitted to the instructor as a mandatory requirement. External sharing of artifacts was limited to anonymized excerpts in accordance with institutional policy.

Templates and schemas for replication are provided in Appendices A-C. These include the consideration table (Appendix A), the protected-term list example (Appendix B) and the three-stage prompting protocol with the MQM-lite revision log schema (Appendix C). The main text also provides detailed descriptions of the CAT tool settings and annotated examples.

Deviations from ideal research practice were: (a) no human-only or MT-only baseline on the same text; (b) single-group design with no cross-group replication or inferential statistics; (c) no inter-rater reliability check for coding (all tallies are descriptive and pedagogical); and (d) time-on-task was estimated by the group (roughly 8-10 person-hours, excluding CAT training) but not measured systematically.

Results

Task description

Group 9 successfully completed the English-to-Chinese corporate-technical web-copy project. The simulated client brief was demanding and realistic: it required strict preservation of all brand and product names (such as NVIDIA, NIM, AI Blueprint and BubbleRAN), consistent use of the 17 approved Chinese terms, a concise corporate-technical register suitable for an official website, no unauthorized additions or summaries and complete auditability of every AI-assisted decision. The five group members collaborated jointly on terminology management, drafting and quality assurance. They estimated that the entire workflow took roughly 8-10 person-hours (excluding initial CAT tool training), although time-on-task was not measured systematically. As this was the students' first extended collaborative exercise combining CAT tools and generative AI, the project prioritized learning, tool familiarization, reflection and process discipline over translation speed; the study makes no claims about efficiency gains compared with human-only or MT-only translation.

Outcomes from the CAT tools

Even before any generative AI was introduced, the group's own termbase and glossary proved highly effective at catching basic inconsistencies. In Trados, the initial QA pass (conducted on human-translated segments during termbase and TM creation) identified several issues that would otherwise have carried through to later stages: mixed use of full-width and half-width semicolons in lists, inconsistent numeric formatting (for example, "2 million" versus "200万") and two accidental lower-case renderings of the branded product name "BubbleRAN." All of these were corrected at an early stage, demonstrating the value of building solid assets upfront.

The first text was used to construct the termbase and TM. After migrating the revised content to Smartcat, the group translated the remaining two texts from the same set of three. The platform reported 12 context matches (101 %), 8 exact matches (100 %), 23 repetitions and several fuzzy matches in the 85-99 % range. Segments containing protected terms were reused correctly and consistently across all three texts. No residual inconsistencies were reported. Although similar consistency could have been achieved in Trados alone, the dual-CAT approach (Trados for asset construction followed by Smartcat for cloud-based verification across texts) facilitated efficient cross-document checking and real-time team collaboration.

Generative AI performance under the capped prompt sequence

The group evaluated the three models (ChatGPT free version, ChatGPT Plus and Microsoft Copilot) using the identical three core prompts. This controlled prompting allowed direct comparison of model behavior under the same conditions.

ChatGPT Plus reached a quality level acceptable for post-editing (i.e., requiring only minor terminology and accuracy fixes before passing the four-pass QA loop) after only the third prompt and required no additional prompting. The model explicitly confirmed that it had received the glossary and even asked clarifying questions, such as whether "AI Blueprint" should always be rendered as "人工智能蓝图" even in headings. ChatGPT (free) also achieved acceptable quality after the third prompt, although its responses to multiple simultaneous constraints were slightly less precise than those of the Plus version. In contrast, Copilot still contained unauthorized additions and term drift after the third prompt, so the group issued three additional targeted prompts to bring it into line. These additional prompts specifically targeted the remaining unauthorized additions, terminology drift and style deviations that persisted after the core three-stage sequence (see Appendix C.2 and the concrete examples in the "Risk Profiles by Category" Section). The process stopped at six prompts for Copilot, well within the didactic ≤10-prompt cap. Every prompt, model output and subsequent human edit was logged with timestamps and snapshots for full auditability.

Risk profiles by category

Terminology obedience

Comparison of the final model outputs against the 17-item glossary revealed clear differences in controllability. ChatGPT Plus showed the strongest adherence, replacing only 3 terms (17.6 % substitution rate) and producing zero unauthorized substitutions. ChatGPT (free) replaced 8 terms (47.1 %) but still recorded zero unauthorized substitutions. Copilot replaced 7 terms (41.2 %) and introduced 3 unauthorized substitutions, such as rendering "agentic AI" as "代理式人工智能" instead of the approved "智能体式人工智能", "autonomous networks" as "自主网络" instead of "自治网络" and "incident ticket" as "事件工单" instead of "故障工单". Some of the 17 glossary terms appeared more than once across the texts; thus the substitution rate refers to unique glossary items. All terminology deviations were corrected manually during Pass 1.

Example (Terminology – Copilot):

- Source: "…leveraging agentic AI capabilities…"

- Copilot output (after the core three prompts): "…利用**代理式**人工智能的能力…" (lit. "utilise **agent-style** artificial intelligence's capability")
- Approved glossary term: "**智能体式**人工智能" (lit. "**Agentic** AI"; the protected official NVIDIA term)
- Commentary: Copilot substituted a more common but unauthorized rendering. This deviation was caught and corrected in Pass 1 through strict glossary enforcement. It illustrates model-specific controllability challenges even under identical prompting conditions.

Unauthorized additions

The dedicated accuracy pass (Pass 2) identified six unauthorized additions across the three model outputs. Five originated from Copilot, including speculative benefits not present in the source and an invented subtitle "方案价值" (Solution Value). One minor addition came from ChatGPT (free). ChatGPT Plus produced none. All six additions were removed during Pass 2.

Example (Unauthorized Addition – primarily Copilot):

- Source: "The AI Blueprint provides a recipe for building autonomous networks."
- Copilot output: "The AI Blueprint provides a recipe for building autonomous networks. This solution will accelerate the implementation of generative AI across telecom operations and reduce operational costs significantly."
- Commentary: The added speculative benefits and cost-reduction claims were hallucinations not present in the source text. They were excised in Pass 2 through careful sentence-aligned accuracy checking. This example clearly demonstrates why a dedicated accuracy pass must precede any fluency or style editing.

Mistranslations and modifier scope errors

Both Copilot and ChatGPT (free) occasionally mishandled modifier scope. In one sentence describing announcements at NVIDIA GTC Paris, the models incorrectly attached the relative clause "showcasing…" to the conference location rather than to the announcements themselves. The group corrected this manually in Pass 2. A parallel terminology error occurred with "incident ticket", where both Copilot and ChatGPT Plus initially used the wrong glossary form.

Fluency and register issues

Even after the dedicated fluency and register passes, all three models required substantial human restructuring to achieve natural, enterprise-level Chinese.

The dominant problem was heavy pre-modification. This refers to long strings of attributes placed before the subject, which feels awkward in formal Chinese web copy.

Example (Heavy Pre-modification for all models, EN→CH specific):

- Source: "Norway-based Telenor Group, which serves over 200 million customers globally, is the first telco to integrate the AI Blueprint…"
- Model outputs (all three): "总部位于挪威、服务全球超过2亿客户的Telenor集团是首个将人工智能蓝图集成到电信网络配置中的电信运营商……"
- Human revision (Passes 3–4): "Telenor 集团总部位于挪威，全球服务客户超过两亿。该集团是首个集成人工智能蓝图的电信运营商……"
- Commentary: The models faithfully reproduced the English-style complex subject with stacked pre-modifiers, resulting in a sentence that feels unnatural and overloaded in Chinese. Breaking it into shorter, clearer clauses significantly improved readability and register appropriateness. This structural challenge is particularly notable in EN→CH post-editing and further justifies placing terminology and accuracy checks before fluency and register passes.

Quantitative summary of post-editing effort

Analysis of Group 9's revision logs (as reviewed by the instructor) recorded a total of 78 edits across the three texts. The distribution by MQM-lite category was as follows:

Table 2: MQM-lite Edit Distribution

MQM-lite Category	Number of Edits	Percentage of Total
Terminology (protected-term deviations)	24	31%
Accuracy (unauthorized additions/omissions)	11	14%
Fluency (ungrammatical or non-idiomatic)	28	36%
Style/Register (pre-modifications, etc.)	15	19%
Total	**78**	**100%**

Notes: Fluency edits were the most numerous, followed by terminology edits. Accuracy edits totaled 11 (6 unauthorized additions and 5 omission or modifier-scope errors). These accuracy issues originated only from Copilot, while ChatGPT Plus produced none.

Discussion

Visibility of typical risks under a staged workflow

The Group 9 case shows that a constraints-first, four-pass workflow (terminology → accuracy → fluency → register/style) makes typical GenAI risks visible and manageable at an early stage. First, protected-term deviations were caught in Pass 1 because the glossary was enforced before any stylistic polishing began. Second, unauthorized additions, a common problem with generative models especially Copilot, were identified in Pass 2 through careful sentence-aligned accuracy checks. Third, numeric/date mismatches and modifier scope errors (for example, incorrectly attaching the relative clause "showcasing…" to "GTC Paris" instead of to the announcements) were flagged and corrected in the same pass. Without this deliberate staging, a substantial number of these errors would have survived into the final deliverable, hidden behind otherwise fluent output. This finding supports Yamada's (2019) observation that NMT's surface fluency can mask adequacy problems and suggests it extends the same insight to LLM-assisted drafting.

Controllability and prompt sequencing

Across the three models, the sequence of prompts proved far more important than the precise wording of any individual prompt. When the group first tried refining the style before locking down terminology, the models, especially Copilot, frequently reintroduced term drift (for example, changing "人工智能蓝图" back to "AI Blueprint"). Once the sequence was reordered to place terminology and accuracy first, the amount of drift decreased markedly. This finding highlights a clear pedagogical principle: terminology and meaning should be treated as non-negotiable invariants, while fluency and register can be adjusted as later, more flexible layers. For instructors, this principle is far more practical and transferable than relying on model-specific prompting techniques.

Prompt cap as a teaching constraint, not a performance metric

The ≤10-prompt cap was introduced purely as a teaching tool. It forced students to budget their interactions carefully and to decide which constraints were worth enforcing. In practice, ChatGPT Plus reached an "acceptable for post-editing" state within three prompts, while Copilot needed six. The cap prevented endless trial-and-error tweaking and encouraged students to move on to manual editing when a model proved uncooperative. This mirrors real-world professional practice, in which translators routinely decide when further machine assistance is no longer cost-effective and must shift to post-editing (Pym 2013). Importantly, the cap is not presented as a claim about model

quality; it is simply a classroom device to build strategic competence (PACTE Group, 2017) and risk-management skills (Pym, 2013).

Alignment with professional process logic

The workflow closely echoes the role separation described in ISO 17100 (translator → reviser → editor) and ISO 18587's strong emphasis on meaning preservation. However, the classroom setting differs from commercial practice in several important ways: grading incentives, fixed deadlines and the lack of formal client-vendor contracts. For this reason, we do not claim that the student workflow is equivalent to vendor-grade practice. Nevertheless, the underlying logic of staging quality checks and maintaining full audit trails through prompt logs and revision logs is highly transferable. Students learned to justify every significant change by explicitly linking it to an MQM-lite category (e.g., "this was corrected under Accuracy because it introduced an unauthorized addition"), a skill that is directly relevant to professional quality management and audit requirements (Lommel et al., 2014).

Pedagogical hypotheses for future testing

The present single-case study suggests several pedagogical hypotheses that could be examined in future replications with larger samples:

- H1 (Sequencing effect): In EN→CH corporate-technical translation, applying terminology and accuracy gates before style-related prompts reduces the reintroduction of term and meaning errors compared with a style-first sequence.
- H2 (Asset effect): Groups that construct a minimal protected-term list and termbase upfront produce fewer protected-term deviations than groups that do not.
- H3 (Audit effect): Requiring prompt and revision logs improves students' ability to correctly classify their own edits into MQM-lite categories (accuracy, terminology, fluency, style) by the end of the semester.

Recommendations for instructors

This case yields four practical recommendations for instructors:

(1) Front-load constraints. Require students to complete a consideration table before any drafting or prompting and treat it as the single source of truth for all QA decisions. This simple step forces students to articulate client needs and register expectations explicitly, making subsequent tool and AI use far more purposeful.

(2) Sequence, do not just prompt. Enforce the order: terminology → accuracy → fluency → register/style. Allow style-related prompts

only after protected terms have been locked. In practice, this dramatically reduces the reintroduction of term drift and unauthorized additions.

(3) Cap and compare. Use a modest prompt budget (e.g., ≤10 per model) so students can discover differences in model controllability and learn to value systems that respect constraints rather than endlessly tweaking prompts.

(4) Assess process, not just product. Base grading on artifacts such as prompt logs, revision logs tagged with MQM-lite categories and TB/TMX files. This rewards disciplined workflows and strategic decision-making rather than final polish alone and mirrors professional audit requirements.

We make no claims that any particular model is universally superior, that the observed error counts would replicate in other texts or language pairs, or that this workflow is more efficient than human-only translation. These boundaries are essential for interpreting the case correctly.

Limitations

The single-case study, while providing a detailed illustration of the protocol in action, has several limitations. The genre (corporate-technical web copy) and language direction (EN→CH) may limit generalizability; error profiles could differ substantially in legal, medical, literary, or other language pairs. The classroom context, with its time pressures and pedagogical focus, differs noticeably from commercial practice. The outcomes should therefore be viewed as pedagogical observations rather than vendor-grade evidence.

Descriptive tallies rely on the group's self-reported revision logs, reviewed by the instructor for face validity only, with no inter-rater reliability assessment. Model behavior is time-sensitive due to provider updates and stochastic sampling. Full artifacts cannot be publicly released due to institutional policies, limiting replication to the provided templates, settings and anonymized excerpts.

These constraints mean the study documents visible risks and workflow manageability in one pedagogical instance but does not support broader quantitative inferences or cross-context claims. The methodological choices underlying these constraints, such as the single-group design, absence of baselines and lack of inter-rater reliability, were already explained and justified in the "Ethics, Replicability and Deviations from Ideal Practice" Section.

Conclusions

This single case demonstrates that a constraints-first workflow combining CAT assets with a fixed prompt sequence makes typical risks visible at an early stage and correctable before stylistic work begins. These risks include protected-term deviations, unauthorized additions and numeric mismatches. The glossary-first setup and dedicated accuracy pass played a central role in surfacing and repairing these issues, while achieving register-appropriate Chinese still required substantial human restructuring. The main practical contribution is a replicable classroom protocol supported by concrete artifacts (consideration table, TB/TMX files, prompt and revision logs and CAT-QA settings) that other programs can readily adapt and test.

Further research should strengthen the evidence base through baseline comparisons, rigorous inter-rater reliability assessment on stratified samples and replication across different groups, genres and language directions using identical CAT-QA settings.

Declaration of AI use

In preparing this manuscript, the author used Grok for language editing, literature search and checking referencing format. All AI-assisted content was reviewed, edited and verified by the author, who takes full responsibility for the accuracy and integrity of the final text.

References

Bowker, L. and J. Buitrago Ciro. 2019. Machine translation and global research: Towards improved machine translation literacy in the scholarly community. Emerald Publishing.

Christensen, T. P. and Schjoldager, A. 2016. “Computer-aided translation tools – the uptake and use by Danish translation service providers”. The Journal of Specialised Translation, 25, 89-105.

EMT Expert Group. 2022. European Master’s in Translation: EMT competence framework 2022. European Commission.

Guerreiro, N. M., Alves, D., Voita, E., and Martins, A. F. T. 2023. “Hallucinations in large multilingual translation models”. Transactions of the Association for Computational Linguistics, 11, 1500-1517.

ISO. 2015. ISO 17100:2015 – Translation services: Requirements for translation services. International Organization for Standardization.

ISO. 2017. ISO 18587:2017 – Translation services: Post-editing of machine translation output – Requirements. International Organization for Standardization.

Kiraly, D. 2000. A social constructivist approach to translator education: Empowerment from theory to practice. St. Jerome.

Koponen, M. 2016. “Is machine translation post-editing worth the effort? A survey of research into post-editing and effort”. The Journal of Specialised Translation 25: 131-148.

Kornacki, M. 2018. Computer-Assisted Translation (CAT) Tools in the Translator Training Process. Peter Lang.

Krings, H. P. 2001. Repairing texts: Empirical investigations of machine translation post-editing processes. Kent State University Press.

Läubli, S., R. Sennrich, and M. Volk. 2018. “Has machine translation achieved human parity? A case for adequate evaluation”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 4791-4796. Association for Computational Linguistics.

Lommel, A., Uszkoreit, H., and Burchardt, A. 2014. “Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics”. Revista Tradumàtica: Tecnologies de la Traducció, 12, 455–463.

PACTE Group. 2017. “PACTE translation competence model: A holistic, dynamic model of translation competence”. In A. Hurtado Albir (ed.) Researching translation competence by PACTE Group, 35-42. John Benjamins.

Pym, A. 2013. “Translation skill-sets in a machine-translation age”. Meta, 58(3), 487–503.

Pym, A. and Y. Hao. 2025. How to augment language skills: Generative AI and machine translation in language learning and translator training. Routledge.

Yamada, M. 2019. “The impact of Google Neural Machine Translation on post-editing by student translators”. The Journal of Specialised Translation 31: 87-106.

Appendix A. Group report: consideration table

Consideration	Remarks
Genre	Technical informative text: website
Orientation	Information-centered
Intended Text Function (main)	Informative*** Vocative** (hidden)
User-friendliness	***
Intended Readership (geographical)	People from China
Intended Readership (others)	Tech enthusiasts and potential customers
Register: Field	Technical texts
Register: Mode	Written and published on the NVIDIA official website
Register: Tenor	Tech giant and its users
Overall Register	Neutral-to-Formal
Translation Purpose	Promotion of generative AI in NVIDIA Support NVIDIA's branding and communication goals in the local market
Translation Approach	Communicative Translation
Translation Strategy	Paraphrase, explicitation, generalization, limited omission/addition (elaboration) and adjustment (e.g., sentence order) **Except for literalism**

Appendix B. Group report: protected-term list example

This is the exact protected-term list used by Group 9, which was created in Excel, imported into Trados and Smartcat and enforced strictly in all AI prompts.

No	English Term	Chinese Term	Notes / Protection Rule
1	AI Blueprint	人工智能蓝图	Protected brand term: never leave in English
2	AI	人工智能	Replace all standalone occurrences
3	NIM microservices	NIM 微服务	Keep “NIM” in English
4	Telco network configuration	电信网络配置	Core technical phrase
5	Agentic AI	智能体式人工智能	Official NVIDIA term
6	GTC Paris	GTC 巴黎	Event name: keep English acronym
7	5G O-RAN	5G O-RAN	Keep English acronym
8	Autonomous network	自治网络	Preferred translation
9	NVIDIA AI Enterprise	NVIDIA AI Enterprise	Brand name: no translation
10	AI Refinery	AI Refinery	Platform name: keep English
11	BubbleRAN	BubbleRAN	Partner platform: keep English
12	Telenor Group	Telenor 集团	Company name: keep English
13	Accenture	埃森哲	First occurrence only
14	Incident tickets	故障工单	Standard industry term
15	Deployment tools	部署工具	Consistent technical term
16	Accelerated computing	加速计算	Core NVIDIA technology term
17	Generative AI	生成式人工智能	Official Chinese equivalent

Appendix C. Group report: prompt design and methodology

C.1 Core three-stage prompting protocol

The group applied the same controlled three-stage prompting sequence to all three models using the versions accessible to end-users at the time of data collection (September-December 2025): ChatGPT (free tier) powered by GPT-4o mini, ChatGPT Plus powered by GPT-4o and Microsoft Copilot powered by the GPT-4-based model then current.

The sequence was designed to mirror real-world professional constraints:

(1) Baseline translation: raw source text with minimal instruction.
(2) Terminology-constrained revision: revision using the bilingual glossary created in the CAT-tool stage.
(3) Consideration-table refinement: full alignment with the client Consideration Table (terminology accuracy, accuracy, corporate-technical register, logical clarity and conciseness).

The exact prompts used in each stage are shown in Table C1.

Table C1: Core three-stage prompting protocol

Stage	Prompt	Purpose
1. Baseline	Translate the following text into Chinese.	Obtain each model's default, unguided output.
2. Terminology Control	Please revise the above translation according to the provided bilingual glossary.	Test adherence to domain-specific terminology developed in Trados/Smartcat.
3. Consideration-Table Refinement	Please refine and improve the translation according to the Consideration Table. Ensure terminology accuracy, corporate-technical register, logical clarity and conciseness.	Align output with all client constraints and professional standards.

Notes: These three core prompts were applied identically to all models. Copilot required three additional targeted prompts to resolve persistent deviations (see C.2 below).

C.2 Additional targeted prompts (Copilot only)

Because Copilot continued to deviate after the core prompts, the group issued three supplementary prompts:

(4) Please adjust the style further to match the language style of NVIDIA's official Chinese website.
(5) When making revisions, do not deviate from the glossary and Consideration Table I provided earlier. Please verify and correct accordingly.
(6) Please change "AI Blueprint" to "人工智能蓝图" and change "AI" to "人工智能".

C.3 Revision log schema (MQM-lite categories)

All human edits were recorded in a standardized revision log using the following MQM-lite categories. Each edit was tagged with the model that produced the original error and the pass in which it was corrected.

Category	**Definition**	**Example**
Terminology	Deviation from the protected-term list or glossary (including unauthorized substitution or drift)	"AI Blueprint" rendered as "AI 蓝图" instead of the required "人工智能蓝图"
Accuracy	Unauthorized addition, omission, or scope error that changes meaning	Addition of "加速生成式人工智能的落地应用" not present in the source text
Fluency	Ungrammatical, awkward, or non-idiomatic Chinese	Nested parentheses or pre-modifier stacking that violates natural Chinese word order
Style/Register	Failure to match the required corporate-technical register of NVIDIA's official Chinese website	Overly colloquial phrasing or inconsistent formality

Usage instructions for replicability

- Log every prompt, model output and human edit with timestamps.
- Tag each edit with the MQM-lite category above.
- Record which model introduced the error, and in which pass it was fixed.
- Export the log as an Excel or CSV file together with the final TMX/TB for full auditability.