A Comparison of AI- and Human-Generated Misuse Tagging in English and Chinese Learner Corpora

Keiko Mochizuki[1], Fukuda Sho[2], Zhang Zheng[3], Liu Qiuwen[1], Bai Fengguang[1]
[1]Tokyo University of Foreign Studies, [2]University of Toyama, [3]Bunkyo University

mkeiko@tufs.ac.jp, fukuda@las.u-toyama.ac.jp, zhangzheng.apple@icloud.com,
hitomi806672830@gmail.com, houkou6690@gmail.com

Abstract

This study examines whether GPT can help identify learner-corpus annotations that warrant re-examination. It compares existing human annotations with independent GPT judgments for 50 English and 50 Chinese verb-related examples. The data were obtained from the TUFS Learners' Error Corpora of English Searching Platform (https://corpus.icjs.jp/corpus_eng/index.php) and Chinese Searching Platform (https://corpus.icjs.jp/corpus_ch/), both accessible through the TUFS learner-corpus portal (https://corpus.icjs.jp/). GPT was provided with the learner sentence and the marked target, but not the human annotation. The model was permitted to make one minimal correction or return *no_change* when the target was acceptable. The comparison distinguishes edit operation, correction, and error category. Exact agreement was limited in both languages. Edit-operation agreement was 36.0% in English and 60.0% in Chinese; category agreement was 28.0% and 54.0%, respectively. When the sources differed, the researcher reviewed anonymized alternatives. GPT was preferred in 29 of 42 reviewed English cases and 11 of 33 reviewed Chinese cases; the human annotation was preferred in 2 and 11 cases, respectively. Many disagreements arose before label selection: one source accepted the target while the other corrected it, and several cases allowed more than one reasonable correction. The results do not demonstrate that GPT is more accurate or that either language is inherently easier. Rather, they support a narrower use of GPT as an annotation auditor that screens records and directs human attention to cases requiring re-examination.

Keywords: learner corpus; error annotation; large language model; grammatical error correction; English; Chinese; annotation audit

1. Introduction and Related Work

1.1 Learner-corpus annotation and correction variability

Error-annotated learner corpora link learner expressions to proposed corrections and explanatory categories. Annotation therefore involves three related decisions: determining whether the marked expression is erroneous in context, specifying an appropriate correction when required, and assigning the most suitable explanatory category. None of these decisions is entirely mechanical. Annotators must first determine the boundaries of the problem. They may then need to choose among several reasonable corrections, and a marked target may be locally acceptable even when the surrounding sentence remains problematic.

Previous studies demonstrate why these decisions should be reported separately. In the NUS Corpus of Learner English, agreement was only fair to moderate for error identification, classification, and correction, even when annotators were instructed to select the smallest necessary span (Dahlmeier, Ng and Wu, 2013). Chinese GEC research also recognizes that a learner sentence may admit more than one valid correction; MuCGEC therefore used multiple annotators and multiple references (Zhang et al., 2022). For this reason, the present study compares edit operation, correction, tag, and target scope separately rather than treating the stored human record as an unquestioned gold standard.

The study draws on the English and Chinese learner error resources developed through the Tokyo University of Foreign Studies project (International Center for Japanese Studies, 2016). The Chinese corpus and its cross-linguistic error taxonomy are described by Mochizuki et al. (2015). Because the two resources classify verb errors according to different principles, fine-grained tag agreement is evaluated within each language rather than by mapping English tags directly onto Chinese tags. English distinguishes Aspect, Tense, Voice, Verb Lexicon, and Agreement, whereas the Chinese analysis uses Action Verbs, Existential Verbs, Relational Verbs, Modal Auxiliary Verbs, Directional Verbs, Causative Verbs, and Stative Verbs.

1.2 LLM annotation, prompting, and minimal correction

Large language models may support corpus quality control by producing context-sensitive corrections and explanations, but their annotation performance depends on task design. AnnoLLM showed that task guidance, demonstrations, and explanations can improve LLM annotation (He et al., 2024), while large-scale evaluation has found substantial variation across models and tasks (Bavaresco et al., 2025). Together, these findings support task- and language-specific validation of prompt design; they do not justify assuming that a single prompting strategy will perform equally well across languages.

For GEC, overcorrection presents an additional methodological risk. Explicit minimal-edit instructions and few-shot examples can improve correction behavior, although their effectiveness remains dependent on the language and model (Karpo and Chernodub, 2026). This risk is especially relevant here because highlighting a category-specific span may lead the model to infer that a repair is required even when the span is acceptable. The prompts therefore allowed a genuine no_change response and prohibited edits outside the target.

Accordingly, the study treats GPT neither as an automatic replacement for human annotators nor as a system evaluated against unquestioned gold labels. It instead evaluates GPT in the narrower role of annotation auditor: an independent second annotator whose disagreements flag records for expert review. The design supports reproducibility through language-specific pilot testing, frozen prompts, structured output, field-by-field

mechanical comparison, and source-blinded researcher confirmation of disagreement cases.

1.3 Research questions

RQ1. Agreement between GPT judgments and existing human annotations for verb-related corrections and fine-grained tags.
RQ2. Components and causes of disagreement in operation, correction, scope, and fine tag, together with the annotation supported after contextual review.
RQ3. Descriptive similarities and differences between the English and Chinese workflows and outcomes.

2. Data and Method

2.1 Corpus project and principal contributors

The study draws on the English and Chinese components of the Learners' Error Corpora of Japanese/English/Chinese Searching Platform maintained by the International Center for Japanese Studies at Tokyo University of Foreign Studies (TUFS). The resources were developed through JSPS KAKENHI Project 25284101, conducted in the 2013-2015 fiscal years under the leadership of Keiko Mochizuki. The project combined TUFS researchers with international partners including the University of Leeds, Peking University, National Taiwan Normal University (NTNU), and Shanghai International Studies University (International Center for Japanese Studies, 2016). The project adopted a cross-referential design: learner texts in English, Chinese, and Japanese were linked to learners' first-language backgrounds, enabling recurrent errors and cross-linguistic influence to be examined across target languages.
Sano et al. (2023) provide a detailed account of corpus construction. At the time documented, the English resource contained 284 corrected and error-tagged essays by learners whose first language was Japanese or Chinese. Twenty-four native speakers of English corrected the essays, and weekly meetings established common correction and tagging procedures; Chinese-L1 data were collected with Shanghai International Studies University and NTNU. The chapter also describes the Chinese resource as comprising 369 essays by Chinese majors at TUFS across proficiency levels. Learner profiles were recorded, and 25 native speakers of Chinese produced and checked corrections and tags through weekly meetings. It also documents the distinction between Error and Modify, the Replace/Delete/Insert/Move operations, and a 74-tag fine-grained taxonomy. These figures refer to the corpus at the time of construction and should not be confused with the searchable error-record populations used for sampling in the present study.
The contributor information on the official site further shows that the resource was built collaboratively. For the Chinese component, it credits Min Quan at TUFS and Lin Yan, Wen Ting, and Zhao Yan at Beijing Language University for data collection and correction; Howard Hao-Jan Chen and Li-Ping Chang at NTNU for data construction and theoretical guidance; annotation teams led by Yamin Shen, Sho Fukuda, and Weilun Yu; and search-system development by Chia-Hou Wu, Hiroshi Sano, Go Inoue, and Yeong-il Yi. Taken together, these contributions establish the corpus as a multi-institutional resource rather than the work of a single annotator (International Center for Japanese Studies, 2016).

2.2 Corpora, analytical scope, and sampling

The unit of analysis was one previously marked target span returned by the public corpus search interface. Corpus registration determined eligibility for sampling, but it did not predetermine that the marked span was erroneous. Only Replace, Delete, and Insert/Add records were retained. English preference, style, word-order modification, and uncertain modification categories were excluded. The Chinese scope was delimited to seven semantic-functional categories in the corpus's Verb I grouping; the initially considered transitivity and reduplication categories were excluded from the final scope.
The exact search settings returned 2,132 English and 3,316 Chinese records. Sampling without replacement used the fixed seed 20260808. The English formal sample contained 10 cases from each of five categories. The Chinese formal sample contained seven cases from six categories and eight from the Modal Auxiliary Verb category, yielding a nearly balanced total of 50. Two additional records per category were reserved for prompt development. Table 1 reports the category populations and formal allocation. Development and formal cases were disjoint, and screened replacements for unsuitable English demonstrations were also excluded from the formal test.

L.	Category	Population	Dev.	Formal
EN	Aspect	220	2	10
EN	Tense	456	2	10
EN	Voice	70	2	10
EN	Verb Lexicon	1,196	2	10
EN	Agreement	190	2	10
EN	Total	2,132	10	50
ZH	Action Verbs	180	2	7
ZH	Existential Verbs	346	2	7
ZH	Relational Verbs	484	2	7
ZH	Modal Auxiliary Verbs	1,887	2	8
ZH	Directional Verbs	174	2	7
ZH	Causative Verbs	62	2	7
ZH	Stative Verbs	183	2	7
ZH	Total	3,316	14	50

Table 1: Search populations and formal sample allocation

2.3 Retrieval and reconstruction of corpus records

Corpus retrieval was automated in Google Colab with Selenium. For each language, the script applied the category and correction-type filters, traversed paginated result pages, and recorded the corpus and error identifiers embedded in preview links. The results grid displayed 40 records per page. Because Selenium's visible-text property returned empty values for some off-screen rows, extraction used the Document Object Model's *textContent* value and then validated the recovered row content. Using textContent therefore resolved a display-layer extraction problem without altering the underlying search criteria.

Each result linked to a preview containing the learner's Original text and a Revised-with-tags version. The target was reconstructed from the unique highlighted element rather than by searching for a potentially repeated identifier. The input included the target sentence and, when available, one preceding and one following sentence. *[TARGET]* and *[/TARGET]* surrounded the specified span; empty spans were preserved for insertions. Human corrections and tags were parsed from the revised representation into separate comparison fields but were not shown to GPT. The sampling category was likewise withheld to preserve independent judgment.

The audits verified the category, unique preview key, reconstructed target sentence, and extraction status of every retained record. The final English workbook contained 10 development and 50 formal records, and the Chinese workbook contained 14 development and 50 formal records. No duplicate preview key occurred within either workbook, and no development item appeared in the formal sample.

2.4 Prompt development and formal GPT annotation

Two prompt conditions were piloted in each language. Condition A supplied category definitions, output rules, a target-only instruction, and permission to return *no_change*. Condition B added one demonstration per category. Prompt selection was based on linguistic adequacy rather than on maximizing agreement with the stored annotation. For English, a semantic audit found that several initial demonstrations were not sufficiently canonical. Unused Aspect, Tense, and Voice examples were replaced, while the Verb Lexicon and Agreement examples were retained. The resulting definitions-plus-screened-examples prompt was frozen as **EN-B-v3**.

The Chinese pilot supported a different prompt choice. Across seven comparisons, Condition A was preferred twice, B once, and four were ties. Although Condition B handled one movement example correctly, it also encouraged optional insertions and repairs that appeared to be prompted by the target category itself. The definitions-only prompt **ZH-A** was therefore frozen. This language-specific outcome underscores that demonstrations should be validated for the target task rather than assumed to be beneficial (He et al., 2024). The prompts were not modified once formal testing began.

All 100 formal items were submitted as independent requests to *gpt-5.6-sol* using a structured output schema. GPT received the original marked context but not the human correction, human tag, or sampling category. For each item, the model returned an operation, a correction, a fine tag, and a short reason; Chinese also included a movement flag. Prompt identifiers and SHA-256 hashes were recorded. The English reasoning-effort setting was recorded as medium. The corresponding Chinese setting was not retained in the consolidated workbook and was therefore not reconstructed retrospectively. A one-case smoke test preceded each batch. Checkpoints, resume-by-ID logic, a three-attempt limit, and final status audits were used to prevent case loss or duplication. All 50 English and 50 Chinese requests completed successfully.

2.5 Mechanical comparison and denominator rules

Operation, exact correction, and fine tag were compared separately. In this study, **mechanical agreement** denotes identity between stored fields, not linguistic correctness. For English, the comparison retained the human database operation exactly as stored. Eleven records had noncanonical operation/correction encodings: eight were marked Replace with an empty correction and three used unusual insertion structures. These cases were retained and flagged, leaving 39 records mechanically evaluable for exact correction and full strict agreement. A surface operation reconstructed from original and revised strings was retained only as exploratory quality control.

For Chinese, add was normalized to insert, while movement was treated as a separate flag. Correction strings were normalized consistently, but semantically equivalent synonyms were not counted as exact string matches. Full strict agreement required all relevant fields to match simultaneously. Under these rules, 8 English and 17 Chinese records were mechanical matches. The remaining 42 English and 33 Chinese records formed the review queue because at least one relevant field differed. Mechanical matches were not subjected to separate semantic review and therefore cannot be treated as certified correct.

2.6 Source-blinded confirmation and qualitative coding

The researcher first produced provisional linguistic recommendations for the review queue. The subsequent confirmation stage used source-hidden A/B workbooks: the human and GPT alternatives were randomly assigned to A or B, and case order was shuffled. Separate key files preserved the hidden source mapping. The researcher selected A, B, BOTH, or NEITHER/undecidable after reading the original context and target. Case hashes verified that the decoded alternatives matched the reviewed content. Because the same researcher had already produced the provisional recommendations, this stage is described as **source-blinded researcher confirmation** rather than independent blind adjudication. After decoding, outcomes were GPT preferred, Human preferred, Both plausible, or Unresolved. Four Chinese cases (C006, C012, C013, and C028) remained unresolved. In the support calculation, both sources received credit for mechanical matches and Both plausible cases; after review, only the preferred source received credit, and unresolved cases were excluded from the denominator. Support is therefore not an accuracy estimate.

The final adjudication outcome and the locus of disagreement were coded separately. Diagnostic flags recorded mismatches in operation, correction, tag, movement, and no-change status. A mutually exclusive primary type was then assigned using language-specific priority rules. English source-encoding irregularities were prioritized because they can create disagreement before linguistic interpretation. Chinese context and movement/scope cases were prioritized before error-status and alternative-correction cases. These priority rules prevent double counting, but they also make the type frequencies descriptive rather than directly comparable across languages.

2.7 Error controls and data handling

No procedure can eliminate interpretive error from learner-corpus annotation. The study therefore combined preventive controls with post hoc detection (Table 2). Before model inference, fixed filters and recorded population counts defined the sampling frame, target-reconstruction audits reduced extraction artifacts, and fixed seeds with disjoint development and formal sets protected test independence. During inference and review, prompt hashes and independent structured requests prevented prompt drift and conversational carryover, checkpointing protected batch completeness, and randomized A/B presentation reduced visible source preference. Finally, reporting separated mechanical agreement, reviewed support, and unresolved cases so that none would be misrepresented as accuracy.

Stage	Primary controls
Retrieval	Fixed filters; automated pagination; population-count audit
Extraction	DOM textContent; unique preview key and highlighted-target audit
Sampling	Fixed seed; stratification; no replacement; disjoint development/test
Prompt	A/B pilot; semantic demonstration audit; no_change and target-only rules
Inference	Prompt hashes; independent requests; structured schema; preflight test
Batch	One-case smoke test; per-case checkpoint; retry and completion audit
Review	Source-hidden A/B order; separate key; case hashes; unresolved option
Reporting	Explicit denominators; support distinguished from accuracy

Table 2: Principal error-control procedures

The analysis used publicly available research-corpus records and did not recruit or contact learners. Analytical files use corpus sample identifiers rather than personal information. Model input was limited to the learner text and local context needed for the marked target. Because institutional policies differ, requirements for secondary corpus analysis and generative-AI processing should nevertheless be confirmed before submission.

3. Results

3.1 Mechanical agreement (RQ1)

The mechanical comparison showed that GPT often diverged from the stored human record (Table 3). This divergence does not, by itself, show that GPT was incorrect; it shows only that the two annotation sources recorded different decisions. Operation agreement was 36.0% in English and 60.0% in Chinese, while fine-tag agreement was 28.0% and 54.0%. Exact correction agreement was 25.6% among the 39 evaluable English records and 38.0% among all 50 Chinese records. Full strict agreement was 20.5% and 34.0%, respectively. Chinese move-flag agreement was high at 92.0%, but English had no corresponding field. Overall, the two sources often diverged at the basic stage of deciding what edit, if any, was required.

Measure	English	Chinese
Operation	18/50 (36.0%)	30/50 (60.0%)
Fine tag	14/50 (28.0%)	27/50 (54.0%)
Exact correction	10/39 (25.6%)	19/50 (38.0%)
Full strict	8/39 (20.5%)	17/50 (34.0%)
Move flag	-	46/50 (92.0%)

Table 3: Human-GPT mechanical agreement

The proportion of cases entering review also varied across categories in this sample. All 10 English Voice cases required review, compared with 9 Aspect, 9 Verb Lexicon, 8 Tense, and 6 Agreement cases. In Chinese, Existential Verbs produced 6 review cases among 7 observations and Modal Auxiliary Verbs 6 among 8, whereas Action Verbs produced 3 among 7. Because each category contains only 7-10 observations, these figures describe the present sample rather than stable category-level tendencies.

3.2 Confirmation outcomes (RQ2)

Table 4 reports the confirmation outcomes after contextual review of the disagreements. “GPT preferred” and “Human preferred” indicate that one proposal was judged more suitable, whereas “Both plausible” indicates that both could reasonably be accepted. Of the 42 reviewed English cases, GPT was preferred in 29 (69.0% of the review queue), Human in 2 (4.8%), and Both in 11 (26.2%). Of the 33 reviewed Chinese cases, GPT and Human were each preferred in 11 (33.3%), Both in 7 (21.2%), and 4 remained unresolved (12.1%).

Outcome	English	Chinese
Mechanical agreement; no review	8 (16.0%)	17 (34.0%)
GPT preferred	29 (58.0%)	11 (22.0%)
Human preferred	2 (4.0%)	11 (22.0%)
Both plausible	11 (22.0%)	7 (14.0%)
Unresolved	0	4 (8.0%)

Table 4: Final confirmation outcomes

Under this support definition, mechanical matches and Both outcomes counted in favor of both sources. GPT therefore received support in 48/50 English cases (96.0%), whereas the human annotation received support in 21/50 (42.0%). In Chinese, each source received support in 35/46 evaluable cases (76.1%). These figures summarize the study's internal confirmation procedure, not independently established accuracy: mechanical matches were not semantically re-reviewed, and the source-blinded decisions were made by one researcher.

3.3 Disagreement types and no-change behavior (RQ2)

The most common disagreement concerned **error status**: one source thought the marked expression needed correction, while the other accepted it as written (Table 5). English also contained 11 source-encoding irregularities. Chinese contained more cases involving alternative corrections, movement across the target boundary, or insufficient context. Purely tag-level disagreements were rare. Fine tags differed in 36/42 English and 23/33 Chinese review records, but only two English cases and one Chinese case were mainly tag-only disputes. In most cases, tag

disagreement followed an earlier disagreement over error status or repair strategy.

Primary type	English	Chinese
Error-status disagreement	22 (52.4%)	14 (42.4%)
Source-encoding irregularity	11 (26.2%)	-
Alternative correction	5 (11.9%)	7 (21.2%)
Movement or target scope	-	4 (12.1%)
Context unresolved	-	4 (12.1%)
Other tag/operation reanalysis	4 (9.5%)	4 (12.1%)

Table 5: Primary disagreement types within the review queues

GPT returned *no_change* in 28 English and 19 Chinese review cases, but the confirmation profiles differed. All 28 English cases resulted in either GPT preferred (20) or Both plausible (8). In Chinese, 9 no-change cases were Human preferred, 5 GPT preferred, 3 Both plausible, and 2 unresolved. Thus, *no_change* functioned as a productive audit signal in English, but it did not reliably indicate that a Chinese human record was questionable.

3.4 Illustrative disagreement cases

The following six cases illustrate distinct sources of disagreement. Square brackets mark the target presented to GPT. In the Chinese cases, the learner sentence is reproduced in Chinese characters and followed by an approximate English translation. The translations clarify the intended meaning without serving as corrected versions of the learner sentences.

E001: Source encoding
Learner sentence:

I [have] visited to Beijing, Shanghai, London, and Taiwan to study in my twenties and thirties.
Comparison: The human record encodes the operation as Replace but provides a blank correction, so the stored annotation is rendered as Delete. GPT directly outputs Delete. At the marked target, both analyses therefore remove have. GPT was preferred because the apparent disagreement mainly reflects the database encoding of the human record rather than a substantive difference in the proposed edit. Other problems, such as visited to, fall outside the marked target and were not corrected.

E032: Alternative acceptable repairs
Learner sentence:

When you study, are you alone or [study] with friends?
Comparison: The human annotation deletes study, producing Are you alone or with friends? GPT replaces study with studying, producing Are you alone or studying with friends? The human correction contrasts two study arrangements, whereas the GPT correction contrasts being alone with the activity of studying with friends. Both sentences are grammatically acceptable in the available context. This case therefore illustrates why exact correction-string agreement may classify functionally acceptable alternatives as disagreement.

E012: Same correction but different fine-grained tags

Learner sentence:

I was afraid that I could not [made] myself understood due to my poor English.

Comparison: Both the human annotation and GPT replace *made* with *make*, producing *I could not make myself understood*. They therefore agree on the correction and operation but assign different fine-grained tags. The human annotation assigns the Tense tag, whereas GPT assigns Verb Lexicon. GPT's explanation correctly states that a modal verb such as *could* must be followed by the base form of the verb. However, the lexical verb itself is not being replaced by a different lexical item: *made* and *make* are inflected forms of the same verb. Within the available corpus taxonomy, Tense was therefore preferred over Verb Lexicon. This case demonstrates that a model may identify the appropriate correction and explain the grammatical rule correctly while still selecting a questionable error category.

C026: Near-synonymous insertions
Learner sentence:

我们[ ]卖给六十岁以上的人这所瑞典的房子，因为他们结束工作以后希望住在安静的地方。
We [ ] sell this house in Sweden to people aged sixty or older, because they hope to live in a quiet place after finishing work.
Comparison: The human annotation inserts 打算 ('plan/intend to'), whereas GPT inserts 想 ('want/intend to'). Both express intended future action and receive the same Modal Auxiliary Verb tag.

C044: Debatable tag boundary
Learner sentence:

第四个，善于调动积极性也是为了组织别人需要的因素，因为领导一定[要]别人的积极的协调。
Fourth, being able to motivate others is also a factor needed for organizing people, because a leader certainly [wants] others' active cooperation.
Comparison: Both annotations replace 要 ('want/need') with 需要 ('need'). The human annotation assigns the Stative Verb tag, whereas GPT assigns the Modal Auxiliary Verb tag. Both tags were judged plausible, indicating a category-boundary disagreement rather than a correction disagreement.

C047: Target-scope mismatch
Learner sentence:

不用说全部的企业采用这样的方法，但是不管提高不提高价格，企业的刻苦将来一定[增加]。
Needless to say, not all companies adopt this method, but whether or not prices are raised, companies' efforts will certainly [increase] in the future.
Comparison: The human annotation replaces 增加 ('increase') with 多 ('greater/more'), whereas GPT returns no_change because 增加 can function as the predicate and the main collocational problem lies in 企业的刻苦 ('companies' hard work/efforts'), outside the marked target. This is a scope issue.

Taken together, these cases show that a mismatch cannot automatically be interpreted as a model error. The differences may reflect source encoding (E001), functionally equivalent repairs (E032 and C026), or the

limits of the marked target (C047). E012 and C044 further demonstrate that correction agreement does not guarantee tag agreement. In E012, the human tag was preferred because the error concerned verb form rather than lexical selection; in C044, by contrast, both tags remained plausible because the category boundary itself was debatable. The four unresolved Chinese cases similarly required broader rewriting or more context than the local target allowed.

3.5 Descriptive cross-language patterns (RQ3)

English produced a larger review queue than Chinese (42/50, 84.0%, versus 33/50, 66.0%) and fewer no-review mechanical matches. Within the English review queue, outcomes strongly favored GPT, whereas the Chinese reviewed cases were evenly split between GPT and Human preferred. GPT received support in every English Agreement, Aspect, and Voice case and in 9/10 cases in each of the Tense and Verb Lexicon categories. In Chinese, the sources received equal support in 5/6 evaluable Action Verbs cases, while different profiles appeared for Relational Verbs (GPT 6/7; Human 3/7) and Directional Verbs (GPT 4/7; Human 7/7).

These differences should be interpreted descriptively rather than as evidence of a language effect. English and Chinese used different frozen prompts, category systems, and output fields. English also contained 11 noncanonical encodings, whereas Chinese included movement and four unresolved cases. Small category samples further limit inference. The comparison instead indicates that AI-assisted annotation depends jointly on linguistic judgment, prompt design, target anchoring, source representation, and taxonomy granularity.

4. Discussion

4.1 Interpreting mechanical agreement

RQ1 examined the degree to which GPT reproduced the existing records. The results indicate limited correspondence with the stored records, particularly for English. However, the agreement rates combine several distinct situations. The exact-correction metric counts two different but acceptable repairs as a mismatch. The English denominator excludes records whose correction field could not be evaluated normally, and the Chinese movement flag records an additional dimension of restructuring. Mechanical agreement therefore measures similarity to the database record rather than linguistic correctness.

The lower fine-tag agreement rates suggest that applying a corpus taxonomy is a distinct task from producing a natural correction. The qualitative analysis, however, adds an important qualification: most tag mismatches did not occur after the sources had agreed on the edit. They arose because the sources disagreed on whether an error existed or how the construction should be repaired. Consequently, a low tag-agreement rate should not be described simply as GPT selecting the wrong label. Error-status judgment and span selection precede classification in the decision chain.

4.2 No-change, alternative corrections, and target scope

The predominance of error-status disagreement confirms the importance of allowing *no_change*. Without that option, a category-specific marked span can prime the model to produce an unnecessary repair. The Chinese pilot showed that demonstrations sometimes intensified such priming; the pilot therefore compared definitions-only and definitions-plus-examples conditions rather than assuming few-shot superiority.

Nevertheless, no-change outcomes require human review. The strong English and mixed Chinese confirmation patterns show that the same model response can have different evidential value across datasets. Possible explanations include different proportions of locally acceptable spans, the 11 English encoding irregularities, Chinese semantic and discourse requirements, and the different prompt conditions. The present design cannot isolate these factors.

The Both plausible cases further support the view that existing corrections should not be treated as unique gold answers. The availability of multiple lexical or structural repairs is an inherent feature of learner correction, not merely an evaluation nuisance. An audit workflow should distinguish exact-string identity from functional equivalence. It should also provide explicit labels for "target acceptable but problem elsewhere," movement across the marked boundary, and insufficient context. These fields would preserve the benefits of minimal editing without forcing a globally ill-formed sentence into a local binary judgment.

4.3 Cross-language comparison and the auditor role

The English-Chinese contrast provides only a descriptive answer to RQ3. A causal language comparison would require larger samples, repeated model runs, comparable taxonomies or a shared higher-level schema, and prompt conditions that can be held constant without distorting the language-specific analysis. The observed differences are inseparable from variation in source encoding, movement representation, prompt examples, and unresolved-case rates.

The results nevertheless support a practical human-centered workflow. In this workflow, the corpus annotation remains hidden during model inference, and the model returns a structured minimal edit with a no-change option. Operation, correction, scope, and tag are then compared separately, and disagreement cases are routed to linguistic review. Source-encoding irregularities, no-change responses, movement, and alternative corrections require different review questions. Mechanical matches may receive lower priority, but should not be certified automatically because they were not independently validated here.

Although this role is narrower than replacement annotation, it remains practically valuable. The model brought database inconsistencies, alternative repairs, and scope problems to attention, while the review also identified model errors. The pilot also shows that prompt examples must be semantically canonical, excluded from the formal sample, and tested for category priming before the prompt is frozen. Under these constraints, GPT can serve as an annotation auditor: a preliminary filter that directs human attention to potentially unstable records.

5. Limitations and Future Research

The study used 50 cases per language and a single output from a single model for each case. Category-level patterns are therefore based on only 7-10 formal observations, or fewer after unresolved cases. Run-to-run stability, model-family differences, and temperature or reasoning settings were not experimentally compared. The Chinese reasoning-effort setting was not preserved in the consolidated workbook. Future studies should record the model snapshot where available, complete request parameters, SDK and API versions, the execution date, and repeated-run agreement.

Source-blinded confirmation was conducted by one researcher who had previously seen provisional recommendations. Randomized A/B labels reduced visible source preference, but the procedure was not independent blind adjudication and does not provide an estimate of inter-annotator reliability. Multiple trained reviewers should independently evaluate all disagreement cases and a sample of mechanical matches. This would test both the confirmation decisions and the assumption that mechanical matches support both sources.

Cross-language comparability remains limited by different taxonomies, prompts, schemas, and database representations. A larger study could retain each corpus's fine tags while adding a common higher-level layer for error status, operation, local scope, and functional correction equivalence. The current context was limited to at most three sentences. Longer context may resolve discourse-dependent cases but may also increase off-target rewriting; future work should therefore manipulate context length explicitly and pair it with an *insufficient_context* response.

Finally, the exact-correction metric is deliberately strict and does not capture synonymous or structurally different repairs. Future work should test a rubric distinguishing orthographic identity, lemma-level equivalence, grammatical-function equivalence, and broader paraphrase. The literature review is focused rather than systematic. Before final publication, it should be extended to include additional English- and Chinese-specific annotation research.

6. Conclusion

This study compared existing human annotations with independent GPT judgments for 100 verb-related examples. The limited exact agreement did not point to a single source of error. Instead, it reflected human-record encoding problems, multiple reasonable corrections, problems extending beyond the highlighted target, and GPT errors. The most common dispute concerned whether the target required correction rather than label selection. GPT received stronger support in the English review queue, whereas Chinese outcomes were more evenly distributed. Because the languages differed in prompts, category systems, output fields, and database conventions, this contrast does not establish superior GPT performance for English.

The principal practical implication is not that GPT should replace human annotators, but that it can serve as an **annotation auditor** or screening tool. In this role, GPT independently evaluates a marked passage, provides a rationale, and flags disagreements for review. A responsible workflow should permit *no_change*, restrict edits to the marked target, freeze prompts using separate development examples, and record structured outputs. It should also conceal the annotation source during confirmation, retain unresolved cases, and report denominators clearly. Used in this limited role, GPT can direct scarce human review time to records most likely to require re-examination.

7. Data Availability Statement

The analysis is based on finalized English and Chinese adjudication workbooks, the cross-language comparison workbook, and the qualitative disagreement-analysis workbook dated 2026-08-09. These files preserve case-level model outputs, existing human records, mechanical comparisons, confirmation outcomes, qualitative types, and methodological constraints. Public corpus records remain subject to the source portal's terms of use.

8. Bibliographical References

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. (2025). LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 238-255). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-short.20

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22-31). Association for Computational Linguistics. https://aclanthology.org/W13-1703/

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. (2024). AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of NAACL 2024: Industry Track* (pp. 165-190). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-industry.15

Karpo, K., and Chernodub, A. (2026). How far can prompting go for minimal-edit Ukrainian grammatical error correction? In *Proceedings of the Fifth Ukrainian Natural Language Processing Conference* (pp. 136-154). Association for Computational Linguistics. https://aclanthology.org/2026.unlp-1.13/

Mochizuki, K., Sano, H., Shen, Y.-M., and Wu, C.-H. (2015). Cross-linguistic error types of misused Chinese based on learners' corpora. *International Journal of Computational Linguistics and Chinese Language Processing, 20*(1). https://aclanthology.org/O15-2007/

Sano, H., Yi, Y.-I., Wu, C.-H., Inoue, G., Shen, Y.-M., Oyanagi, N., and Mochizuki, K. (2023). Cross-referentiality of multilingual error learner corpora of Chinese, English and Japanese for second language

acquisition of Chinese grammar. In H. H.-J. Chen, K. Mochizuki, and H. Tao (Eds.), *Learner corpora: Construction and explorations in Chinese and related languages* (pp. 69-106). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5731-4_5

Zhang, Y., Li, Z., Bao, Z., Li, J., Zhang, B., Li, C., Huang, F., and Zhang, M. (2022). MuCGEC: A multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of NAACL 2022* (pp. 3118-3130). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.227

9. Language Resource References

International Center for Japanese Studies, Tokyo University of Foreign Studies. (2016). *English, Japanese, and Chinese web-based learner error corpora*. https://corpus.icjs.jp/