PharmaSUG China 2026 - Paper AA141-2026

Applying AI + LLM to Reduce Manual Translation Workload in E-Data Package Submission

Megan Chen, UCB Trading (Shanghai) Co., Ltd

ABSTRACT

In 2020, NMPA issued the Guideline on the Submission of Clinical Trial Data, mandating foreign language database translation and setting minimum requirements for e-data package documents. Despite internal teams' efforts in developing in-house translation tools, much content still requires time-consuming manual translation and review. We explored AI + LLM to reduce this workload, aiming to evaluate its effectiveness, develop a quality scoring process. We selected some representative content for AI + LLM translation, with the AI + LLM workflow fine-tuned to internal and NMPA standards. Results showed 60% less translation time, 45% less review workload, 80% score overlap degree between quality scores and manual reviews, and full NMPA compliance. AI + LLM effectively reduces manual translation burden, with the quality score facilitating efficient review and optimizing e-data package submission.

KEYWORDS

AI + LLM; e-data package; Manual Translation; NMPA; AI Judge; Quality Score

INTRODUCTION

The submission of clinical trial e-data packages is subject to strict regulatory requirements from the National Medical Products Administration (NMPA). Per the 2020 Guideline on the Submission of Clinical Trial Data, all foreign language databases and documents must be translated before submission, and all translations need to meet unified regulatory standards.

Traditional manual translation is labor-intensive, low-efficiency and prone to human error. While many pharmaceutical companies and external translation agencies have started to adopt general large language models (LLMs) for translation, such off-the-shelf tools are not tailored for clinical and regulatory scenarios. Generic LLMs often produce inconsistent terminology and non-compliant expressions. Meanwhile, traditional evaluation methods cannot effectively judge the quality of regulatory-oriented translations.

This paper presents a customized AI + LLM solution dedicated to e-data package translation. We also developed an AI judging system for quality assessment, aiming to balance operational efficiency, translation quality and GxP compliance, and ultimately optimize the full translation and review workflow.

BACKGROUND

REGULATORY REQUIREMENTS FOR E-DATA PACKAGE TRANSLATION

NMPA clearly stipulates that foreign language clinical documents must be translated for official submission, with explicit rules on each part of the e-data package. High-quality and compliant translation is the prerequisite for smooth regulatory review of clinical data packages.

Table 1 summarizes the minimum translation requirements for clinical trial data to be translated from a foreign language into Chinese.

Programming related Items	Format	Requirement	Translation to Chinese
aCRF	pdf	Required	Required; at least the following content should be in Chinese: questions designed to collect data, values or codes list of efficacy variables.
SDTM dataset	XPT	Required	Required; at least the following content should be in Chinese: dataset label and

ADaM dataset	XPT	Required	variable label, adverse events terms, concomitant medications terms, medical history term, values or codes list of efficacy variables.
SDTM Define.xml	xml/pdf	Required	Required; at least the following content should be in Chinese: description/label and specification of each dataset in the database, description/label and derivation progress of variables in dataset, values or codes list of efficacy variables. ARM: titles of tables in Chinese
ADaM Define.xml (including ARM)	xml/pdf	Required	
cSDRG	pdf	Required	Required; should be submitted in Chinese
ADRG	pdf	Required	Required; should be submitted in Chinese
ADaM code	txt	Required	Not required

Table 1. The minimum translation requirements for clinical trial data from a foreign language into Chinese

PAIN POINTS OF EXISTING TRANSLATION MODES

Following the launch of our in-house translation tool, manual translation still consumes the bulk of total translation hours. Currently, both internal trial use and third-party translation vendors mainly rely on general LLMs:

1. Generic LLM translation results are average and cannot be directly used for submission, requiring comprehensive manual review and revision.
2. The classic BLEU score, a mainstream machine translation evaluation metric, only focuses on literal text matching. It ignores professional clinical terminology, fixed regulatory expressions and contextual logic, so it fails to reflect the real quality of submission documents.

AI AND LLM: ADVANTAGES, LIMITATIONS AND APPLICATION VALUE

As the core technology of artificial intelligence, LLMs have powerful natural language processing capabilities and can handle various cross-lingual translation tasks efficiently, greatly reducing repetitive manual work. However, general LLMs have obvious limitations in professional fields: lacking training on pharmaceutical terminology, they cannot fully adapt to the high requirements of clinical submission documents.

Given the mandatory translation requirements for e-data packages, applying LLMs is an inevitable trend to improve efficiency. To make up for the defects of general models and traditional evaluation methods, we developed a customized LLM translation tool and a supporting AI judging system.

OBJECTIVES

1. Apply customized AI + LLM to reduce the manual translation and review workload of e-data package textual content.
2. Establish an AI judging system based on internal regulatory submission experience and customizable scoring criteria to replace the inapplicable BLEU score, and realize automatic scoring and revision comment output.
3. Verify the translation quality, compliance and practical value of the integrated solution.
4. Provide a feasible solution for internal promotion and large-scale application within the company.

METHODOLOGY

TEST DATA SELECTION

We selected method and comment sections from define documents as test samples. This content primarily consists of programming language constructs that sit at the intersection of natural language and machine language—incorporating both structured code syntax and descriptive human-readable annotations. These sections are particularly challenging for generic translation tools because they require simultaneous preservation of technical syntax accuracy and contextual semantic clarity. Complex texts were included to verify the performance of the model in difficult scenarios.

CUSTOMIZED LLM TRANSLATION WORKFLOW

Our translation workflow leverages the company's official designated translation platform. This platform is an industry-leading enterprise AI translation solution designed for strict compliance. It features robust data security, adaptive NMT supporting 150+ languages and flexible deployment, forming a dedicated translation pipeline for e-data package documents.

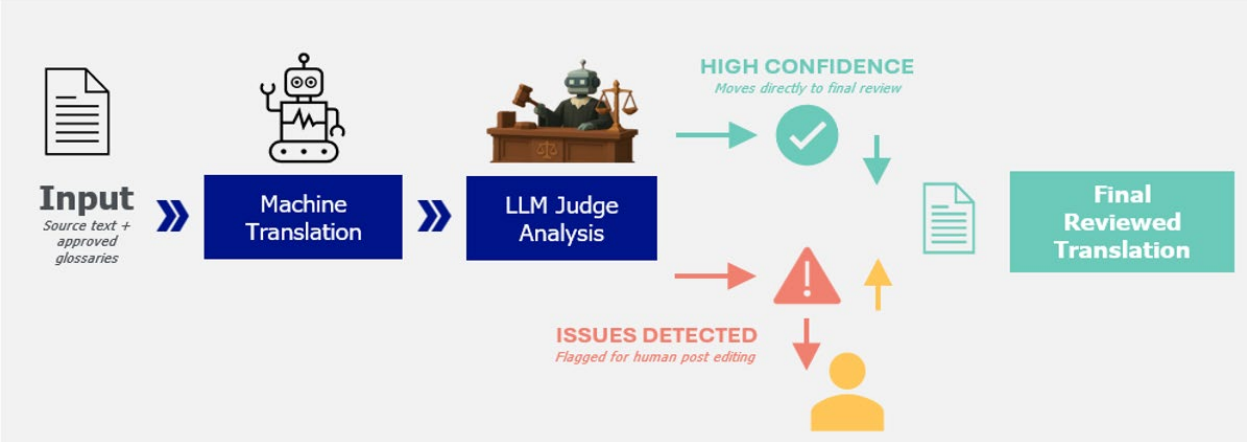
AI JUDGE SYSTEM DEVELOPMENT

A custom AI judging module was built for quality assessment, abandoning the traditional BLEU score. The evaluation framework was designed based on years of accumulated internal experience in regulatory submission translation, capturing domain-specific quality dimensions that generic metrics overlook. The evaluation dimensions cover terminology accuracy, semantic integrity, regulatory compliance and readability, all formulated according to NMPA guidelines and internal standards.

Crucially, the scoring criteria are fully customizable. Translation teams can adjust dimension weights, add new compliance checkpoints, or refine terminology rules as NMPA guidelines evolve or as organizational standards are updated. This flexibility ensures the system evolves alongside internal expertise rather than relying on fixed, external reference-based scoring.

The system can not only output objective quality scores, but also automatically generate targeted revision comments for problematic content.

Display 1 illustrates the workflow of using one LLM for translation and another LLM as a judge to assess translation quality.



Display 1. The AI + LLM translation workflow

PERFORMANCE EVALUATION

We compared the efficiency and quality between pure manual translation and our customized AI + LLM solution. We calculated the time cost, review workload, consistency between AI scores and expert manual review.

RESULTS

1. Efficiency Improvement: Compared with traditional manual translation, our solution reduced overall

translation time by 60% and subsequent manual review workload by 45%.

2. Quality Evaluation Reliability: The overlap degree between scores from our AI judge system and professional manual review reached more than 80%. The automatic revision comments accurately located problems and guided revision work.
3. Regulatory Compliance: All translations generated by our customized LLM fully met NMPA's minimum requirements, with no critical compliance errors.

Figure 1 shows the distribution of quality scores across BLEU, CHRF, BERTSCORE, DEEPSEEK_JUDGE, and human judgments. BLEU and CHRF cluster near zero, fundamentally misaligning with human scores (peak 80-90). DEEPSEEK_JUDGE exhibits extreme variance, indicating instability of generic LLM-based evaluation. Only BERTSCORE partially overlaps with human judgment yet remains systematically lower. This divergence underscores the necessity of a domain-customized AI Judge system.

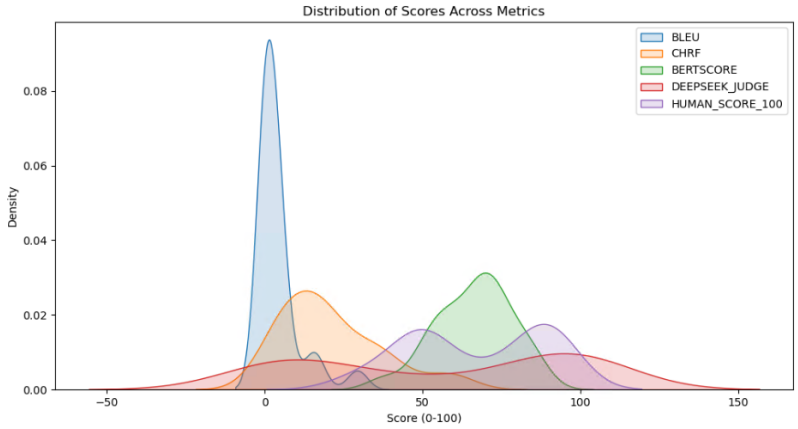


Figure 1. Distribution of Scores Across Metrics

Figure 2 shows Score Distribution Comparison for Chinese to English translation across KIMI and 2 human judgments.

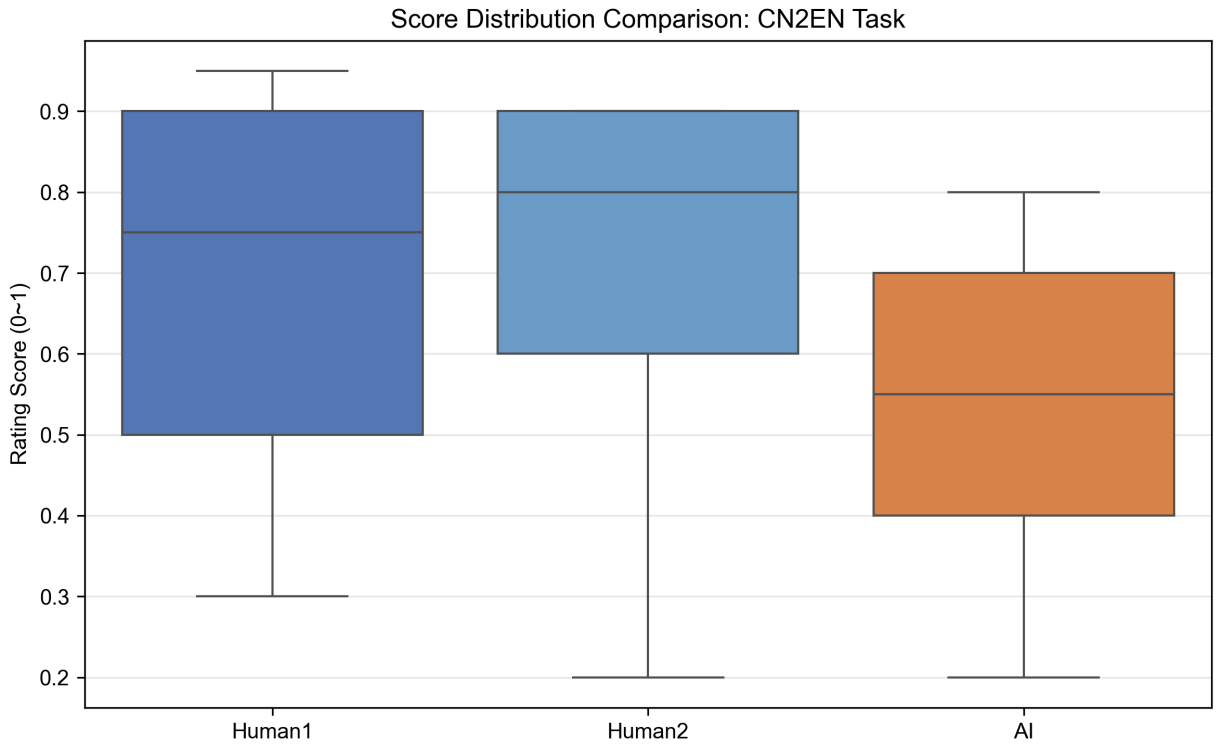


Figure 2. Score Distribution Comparison for Chinese to English translation

Figure 3 shows Score Distribution Comparison for English to Chinese translation across KIMI and 2 human judgments.

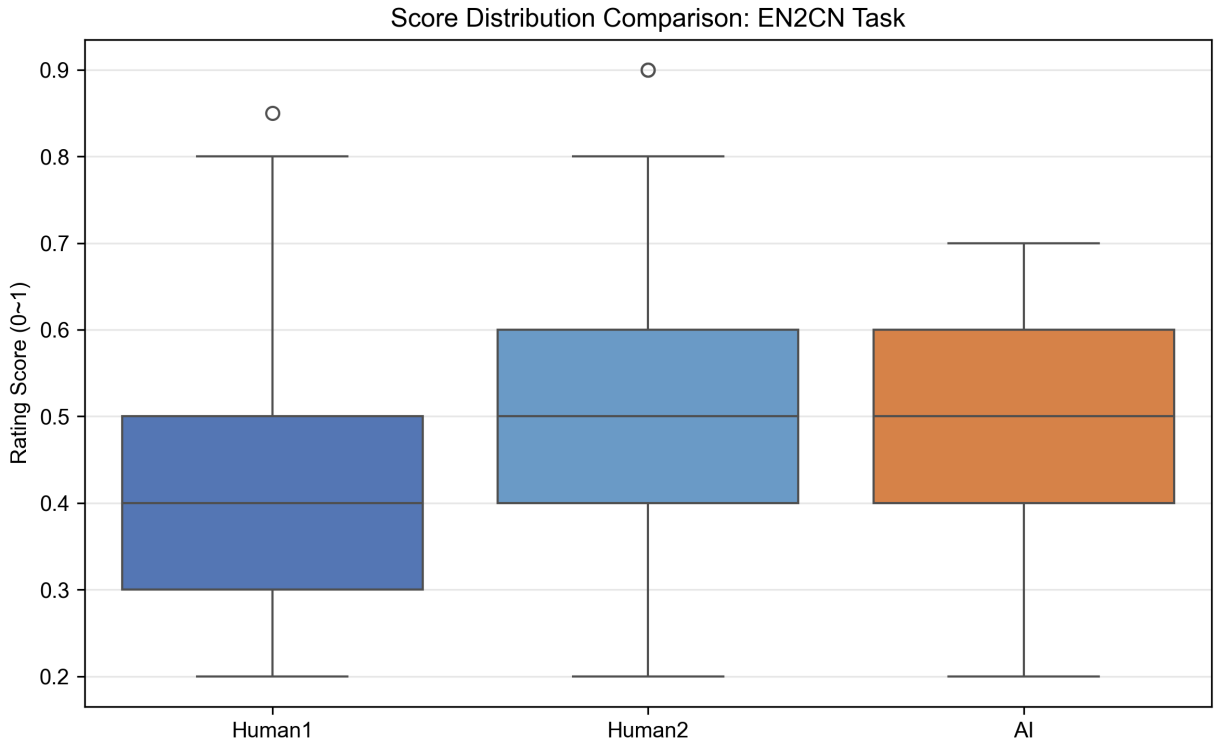


Figure 3. Score Distribution Comparison for English to Chinese translation

Both human annotators and the AI model consistently assign higher quality scores to Chinese-to-English translations, while English-to-Chinese translations are mostly rated at medium-low levels. This indicates the overall translation quality of Chinese-to-English is superior to English-to-Chinese under the unified 0–1 scoring standard.

Across both translation directions, around 80% of the AI score range overlaps with human manual score intervals shown in boxplots. The consistent overlapping interquartile ranges visually verify that the AI model reproduces the majority of human scoring thresholds, with only extreme low-score human samples falling outside the AI prediction scope.

DISCUSSION

CORE ADVANTAGES OF THE PROPOSED SOLUTION

1. Optimized Quality Evaluation: The self-developed AI judge replaces the impractical BLEU score, realizing standardized, automated and traceable quality control.
2. One-stop Closed-loop Workflow: Integrates translation, scoring and intelligent comment generation in a single system, avoiding scattered files and fragmented processes.
3. Data Security & Controllability: In-house deployment eliminates risks of confidential clinical data leakage.

EXISTING CHALLENGES

To comply with GxP requirements, all AI + LLM translation outputs currently require 100% human review before final export. While our solution significantly reduces translation time and review workload, the mandatory full manual review step remains a prerequisite for regulatory submission. Future iterations will explore how to further streamline this compliance-driven workflow without compromising quality assurance standards.

FUTURE DIRECTIONS

Looking ahead, we plan to extend the AI + LLM translation and LLM scoring pipeline to additional submission document types, including review guides and annotated Case Report Forms (aCRFs). These documents present distinct linguistic and structural challenges compared to DEFINE content, and their integration will require further refinement of our domain-specific rules and evaluation dimensions.

Additionally, we will continue to explore the feasibility of reducing or optimizing the current 100% human review requirement. As the AI Judge system's consistency with expert reviews improves and regulatory guidance on AI-assisted translation evolves, we aim to establish a risk-based review framework—potentially enabling targeted human review for high-risk segments while allowing qualified low-risk content to proceed with reduced manual oversight. This evolution will be pursued in close alignment with GxP compliance standards and regulatory expectations.

CONCLUSION

LLMs are effective tools to solve the heavy translation work of e-data packages, while general AI translation and traditional evaluation metrics have obvious shortcomings for regulatory submission scenarios. Our customized AI + LLM solution, equipped with an GxP-compliant AI judge system, effectively improves translation efficiency and guarantees document compliance.

This in-house tool has prominent advantages in professionalism, workflow integration and data security. It perfectly fits the daily work of clinical data package submission. We recommend promoting this solution across relevant departments to comprehensively upgrade the translation and quality control workflow for regulatory submissions.

REFERENCES

NMPA. (2020). Guideline on the Submission of Clinical Trial Data. Available at https://www.cde.org.cn/zdyz/domesticinfopage?zdyzIdCODE=776d02bd9234511f00da866a30760de1.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “Bleu: a Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. Available at https://aclanthology.org/P02-1040.pdf

ACKNOWLEDGMENTS

Sincere gratitude goes to Annie Xu and Naoki Isogawa for their consistent support across the full lifecycle of this project. We also greatly appreciate Neil Pearson, Karina Nazarova and Helena Andres-Terre for their contributions to the tool’s development, testing and ongoing support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Megan Chen
UCB
+86 15652386321
Megan.chen2@ucb.com