

ANCHOR: An English-to-Chinese Transcreation Method for Transferring Check-Worthiness Estimation Data

Yinglong Yu

yuyingling@cuc.edu.cn

Communication University of China

Fang Yuan

Communication University of China

Yu Huang

Communication University of China

Yuchang Duan

Communication University of China

Zhengyi Lyu

Communication University of China

Yu Wang

Communication University of China

Research Article

Keywords: Check-worthiness estimation, English-to-Chinese annotated data transfer, Cultural adaptation, Label preservation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11202850/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

ANCHOR: An English-to-Chinese Transcreation Method for Transferring Check-Worthiness Estimation Data

Abstract

Translating English annotated data into Chinese can supply resources for check-worthiness estimation, but entity omission, shifts in claim scope and added information during translation may make the source label inapplicable to the translated text. We propose ANCHOR (Adaptive Naturalization for Check-worthiness-Oriented Rewriting), a transcreation method for transferring English check-worthiness data into Chinese. ANCHOR first classifies the cultural dependence of each text. Highly culture-dependent texts follow a domestication path that adapts expressions while preserving factual scope, assertion strength and pragmatic function; the remaining texts follow a foreignization path built on retrieved bilingual demonstrations, candidate generation and self-supervised reranking. On three English check-worthiness datasets, NLP4IF, ClaimBuster and ViClaim, we evaluate general translation quality and check-worthiness preservation separately. Against the majority judgment of six human evaluators, cultural dependence classification reaches an accuracy of 0.8225 on 400 texts, and seven evaluators rated 100 domesticated translations at a mean of 4.386 out of 5 for meaning accuracy, cultural relevance and fluency. On 320 stratified samples, three evaluators judged whether each translation still supports checking the source claim: ANCHOR preserves the source label in 99.38% of cases, tied with DeepL for the highest among eight translation models, and preserves every negative sample. In an ablation against direct translation by GPT-5.2, ANCHOR raises this rate by 1.88 percentage points and makes fewer errors in names, quantities and entities. These results show that ANCHOR offers a practical way to reuse English annotated resources for building Chinese check-worthiness data.

Content note. Some examples presented in this paper contain profanity and offensive language. These expressions are retained where necessary to preserve the linguistic characteristics of the original data.

Keywords: Check-worthiness estimation, English-to-Chinese annotated data transfer, Cultural adaptation, Label preservation

1 Introduction

Check-worthiness estimation determines whether a text contains a claim that merits further verification, and it is the upstream step of automated fact-checking. The task jointly considers the verifiability, public relevance and potential impact of a claim, so that the limited resources of subsequent evidence retrieval and fact verification can be allocated accordingly. For political statements and social media texts, datasets such as ClaimBuster, NLP4IF and ViClaim have established annotated resources and evaluation foundations (Hassan et al. 2017; Arslan et al. 2020; Shaar et al. 2021; Giedemann et al. 2025). The Chinese fact-checking datasets CHEF and CFEVER mainly support evidence retrieval and claim veracity prediction (Hu et al. 2022; Lin et al. 2024). More recently, MultiCW has included Chinese in a multilingual check-worthiness detection benchmark and used translation to supplement some languages and classes (Hyben et al. 2026). As resource coverage expands, maintaining the correspondence between texts and their source labels during cross-lingual construction becomes a question that deserves further study.

Translating existing English annotated data into Chinese allows the task definition and label scheme of the source data to be reused, which lowers the annotation cost of building Chinese resources (Agrawal et al. 2024). However, research on cross-lingual data transfer shows that translation alters textual properties and affects downstream task performance (Artetxe et al. 2020). For check-worthiness estimation, the problem is most visible in changes of claim scope and pragmatic function. A vague emotional expression may be translated into a factual statement with an explicit target, and a concrete claim may lose its verifiable object through entity generalization or omitted details. In both cases, the correspondence between the source label and the translation changes.

Culturally loaded expressions make label preservation even harder. Political metaphors, regional references, idioms and irony usually depend on a specific context. Keeping their literal form may prevent Chinese readers from understanding their original function, whereas excessive rewriting may change entities, factual scope and assertion strength. Data transfer for check-worthiness estimation therefore has to handle semantic preservation, cultural adaptation and label consistency at the same time.

Related studies have examined these issues from different angles. MultiCW focuses on multilingual check-worthiness data construction and classification robustness (Hyben et al. 2026); Kabir et al. (2026) show that cultural prompts may aggravate label drift in mental health and sarcasm tasks; and Yuan et al. (2026) use CanMT to compare strategy constraints and multi-dimensional quality in cultural translation. These works provide the resource and problem basis for this study. We further study a generation method for transferring check-worthiness data, which combines cultural dependence classification with factual and pragmatic constraints, so that expression adaptation stays within the boundaries of the original claim.

To address the above issues, we propose ANCHOR (Adaptive Naturalization for Check-worthiness-Oriented Rewriting), a transcreation method for English-to-Chinese transfer of annotated data in the check-worthiness estimation domain. Texts with low cultural dependence are translated through a demonstration-assisted foreignization

path, and texts with high cultural dependence are adapted through a domestication path, in which requirements to preserve factual scope, assertion strength and pragmatic function bound the rewriting. The code, prompts, baseline scripts and retrieval index are available at <https://anonymous.4open.science/r/ANCHOR-AE3C/>. The main contributions of this paper are as follows:

1. We propose ANCHOR, an English-to-Chinese transcreation method whose generation objective is to keep the source label applicable to the translation. It routes each text by cultural dependence: highly culture-dependent texts are domesticated under constraints that retain facts, claim targets, assertion scope and pragmatic function, while the others are foreignized with demonstrations retrieved from a self-built Chinese–English parallel corpus of Ministry of Foreign Affairs press conferences and a self-supervised reranker.
2. We design an evaluation scheme that assesses general translation quality and check-worthiness preservation separately. General translation quality is measured with COMET-QE, COMETKiwi and an LLM-judged LLM Score; for preservation, three native Chinese evaluators judge whether each translation still supports checking the source claim on 320 samples stratified by cultural dependence and source label, which separates claim loss from the rewriting of non-factual expressions into facts.
3. Baseline and ablation experiments show that ANCHOR keeps the source label applicable while maintaining translation usability. Its label preservation rate of 99.38% ties with DeepL for the highest among eight translation models, and its LLM Score ranks second; compared with GPT-5.2 direct translation, it raises the rate by 1.88 percentage points and reduces failed units from 15 to 9.

2 Related work

2.1 Check-worthiness estimation and Chinese fact-checking resources

Check-worthiness estimation concerns whether a claim needs to enter the subsequent fact-checking process. ClaimBuster uses U.S. presidential election debates as its corpus and classifies sentences into non-factual sentences, unimportant factual sentences and check-worthy factual sentences, providing an early data foundation for the task (Hassan et al. 2017; Arslan et al. 2020); its benchmark dataset additionally provides a Groundtruth subset on which three experts reached consensus. The NLP4IF-2021 shared task targets COVID-19-related tweets in English, Arabic and Bulgarian and covers binary annotation dimensions such as whether a tweet contains a verifiable claim and whether it requires manual fact-checking (Shaar et al. 2021). ViClaim targets video transcripts; it contains 1,798 transcripts in English, German and Spanish covering six topics and annotates sentences with three claim-related labels (Giedemann et al. 2025). CLEF-2022 CheckThat! further covers Arabic, Bulgarian, Dutch, English, Spanish and Turkish (Nakov et al. 2022). These studies have gradually extended the languages and application scenarios of check-worthiness estimation.

Among Chinese fact-checking resources, CHEF builds an evidence-based fact-checking dataset from Chinese fact-checking websites (Hu et al. 2022), and CFEVER

follows the construction of FEVER and uses Chinese Wikipedia to support evidence retrieval and veracity prediction (Lin et al. 2024). TrendFact further introduces explanation generation and hotspot-aware evaluation (Zhang et al. 2025), and MedFact targets medical fact-checking (Chen et al. 2025). These resources mainly support fact verification and related tasks, which differ from the value estimation performed before a claim enters the checking process.

MultiCW directly targets check-worthiness detection, building a balanced benchmark of 123,722 samples in 16 languages including Chinese and filling gaps in some languages and classes through machine translation; its translation spot checks report problems with colloquial expressions, entities and omitted content (Hyben et al. 2026). Whereas MultiCW focuses on data integration and classification benchmarking, ANCHOR studies how to adapt culturally loaded expressions during generation while keeping the source label applicable to each transferred text.

2.2 Cross-lingual data transfer and cultural adaptation

Machine translation is a common way to build annotated data for low-resource languages, and its effectiveness depends on whether the translated text keeps the original task-relevant properties. Geng et al. (2023) examine how pseudo-data construction affects the training of translation quality estimation models; Barančíková and Bojar (2025) compare intrinsic and extrinsic evaluation and point out that improvements in translation quality metrics do not necessarily bring corresponding downstream gains. These studies indicate that cross-lingual data construction needs to examine both translation quality and task applicability.

Kabir et al. (2026) directly study semantic label drift in cross-cultural translation. They translate mental health and sarcasm data into Bengali and Greek, obtain post-translation labels through multi-LLM voting and native-speaker verification, and analyze the changes with label preservation rate, KL divergence and the Matthews correlation coefficient. Comparing a plain translation prompt with cultural prompts that introduce demographic personas, they find that explicitly invoking cultural knowledge may aggravate label drift, with effects that vary across tasks and languages. This shows that cultural adaptation by itself does not guarantee label preservation. ANCHOR takes up this problem but restricts it to English-to-Chinese check-worthiness transfer; targeting two kinds of change, namely the weakening of existing claims and the rewriting of non-factual expressions into facts, it constrains factual scope, assertion strength and pragmatic function at generation time.

Domestication and foreignization provide two orientations for cultural adaptation. The discussion can be traced back to Schleiermacher’s distinction of how a translator handles the distance between author and reader, and Venuti (2018) further elaborates the two strategies. Domestication emphasizes target-language conventions and cultural intelligibility, whereas foreignization emphasizes retaining the expressive form and cultural difference of the source text (Baker and Saldanha 2020). In transferring check-worthiness data, the choice of strategy concerns the retention of claim strength, assertion scope and verifiable objects. Cultural adaptation should focus on the expressions that affect understanding while keeping the factual information that supports the checking decision.

[Yuan et al. \(2026\)](#) propose CanMT, a culture-aware translation benchmark built from literary works that covers five languages and twelve translation directions. It evaluates translations on contextual accuracy, cultural adaptation, functional equivalence, faithfulness and naturalness, and compares semantic and communicative translation prompts. Their results show that strategy constraints change model performance across quality dimensions; the semantic translation prompt itself also requires retaining core information and avoiding the addition of unstated information. Like that work, this study values cultural expressions and their communicative function, but it further selects a processing path according to the cultural dependence of each text and examines the rewriting results against the applicability of check-worthiness labels.

Generative studies have begun to incorporate the identification of culturally relevant content into adaptation pipelines. In cross-cultural image adaptation, [Khanuja et al. \(2024\)](#) first identify culturally relevant content and then make targeted adjustments. We borrow this staged idea: cultural dependence determines the translation path, and check-worthiness preservation constrains the scope of rewriting.

2.3 LLM-based translation and demonstration retrieval

Research on LLM-based translation mainly involves prompt design, external knowledge and model fine-tuning. Prompt design guides generation through task descriptions and the organization of demonstrations. [Zhang et al. \(2023\)](#) find that the number, quality and template of demonstrations affect translation results, and [Moslem et al. \(2023\)](#) use in-context learning to introduce glossaries, example pairs and domain information into inference. [Robinson et al. \(2023\)](#) show that translation performance differs with the resource level of languages.

External knowledge can supplement lexical and contextual information in prompts. [Ghazvininejad et al. \(2023\)](#) improve translation with bilingual dictionaries and phrase-level prompts, [Lu et al. \(2024\)](#) propose chain-of-dictionary prompting, and [Iyer et al. \(2023\)](#) examine the role of LLMs in translation disambiguation. AFSP combines hybrid retrieval, few-shot prompting, candidate generation and quality ranking to select relevant demonstrations from a parallel corpus to assist translation ([Tang et al. 2025](#)). Our foreignization path follows this pipeline, builds the index with OpenAI’s general-purpose embedding model text-embedding-3-large ([OpenAI 2024](#)), and uses dense pre-filtering to reduce the number of candidates for which the full hybrid score must be computed. [Donthi et al. \(2025\)](#) establish cross-lingual idiom mappings through semantic retrieval, which provides a reference for handling culturally loaded expressions.

Model fine-tuning adapts general capabilities to translation. ALMA improves translation quality through two-stage fine-tuning on monolingual data and high-quality parallel data ([Xu et al. 2024a](#)), and contrastive preference optimization learns quality differences from preference relations between candidates ([Xu et al. 2024b](#)). Studies on domain adaptation, frozen decoders, adapter selection and memory-efficient fine-tuning discuss adaptation under different resource conditions ([Saunders and DeNeefe 2024](#); [Dhole 2025](#); [Son et al. 2025](#); [Simoulin et al. 2024](#)), and the internal semantic representations of models have also attracted attention ([Zhang et al. 2024](#)). [Stap et al. \(2024\)](#) point out that translation fine-tuning may be accompanied by

changes in general abilities and steerability. Because this study needs to impose generation constraints dynamically according to cultural dependence, ANCHOR performs adaptation through prompting and candidate ranking and keeps the parameters of the underlying generation models unchanged.

2.4 Evaluating translation quality and check-worthiness preservation

Machine translation quality evaluation usually focuses on semantic accuracy, completeness and fluency. COMET learns quality prediction from multilingual representations and human scores (Rei et al. 2020), and COMETKiwi targets reference-free quality estimation (Rei et al. 2022). LLMs can also act as translation judges through prompting, and related studies discuss their system-ranking ability and judging biases (Kocmi and Federmann 2023; Qian et al. 2024). Bennett et al. (2025) further examine the applicability of different evaluation metrics.

Human evaluation complements automatic metrics in judging specific errors and contextual information. Based on the MQM framework, Freitag et al. (2021) compare expert and crowd annotations and find that evaluator background and evaluation conditions affect system rankings. CanMT evaluates multi-dimensional quality scores of cultural translation, whereas Kabir et al. (2026) compare source labels with labels re-annotated after translation (Yuan et al. 2026). Beyond general translation quality, this study uses human evaluation to judge whether a translation still supports the original check-worthiness label.

3 Methodology

ANCHOR targets English-to-Chinese transfer of annotated data in the check-worthiness estimation domain. Given an English text x^{en} , the method outputs a Chinese text $\hat{x}^{zh}$, with the goal that the original check-worthiness label still applies to the transferred text. To this end, generation must preserve the claim target, factual scope, assertion strength and pragmatic function: an existing verifiable claim must not lose its checking target through entity omission or generalized reference, and opinions, emotions and rhetorical expressions must not become new factual claims through added information.

The overall pipeline is shown in Fig. 1. Cultural dependence classification first identifies how strongly an expression depends on a specific cultural context and selects the foreignization or domestication path. The foreignization path organizes translation through retrieved demonstrations, candidate generation and reranking; the domestication path constrains the scope of cultural adaptation in the prompt. Both paths output one Chinese text, which is then evaluated for general translation quality and for the applicability of the source label.

In contrast to MultiCW, which focuses on corpus integration, and CanMT, which compares strategy prompts (Hyben et al. 2026; Yuan et al. 2026), this study applies the processing strategy to each individual text to be transferred. Cultural dependence determines how expressions are adapted, and factual and pragmatic requirements

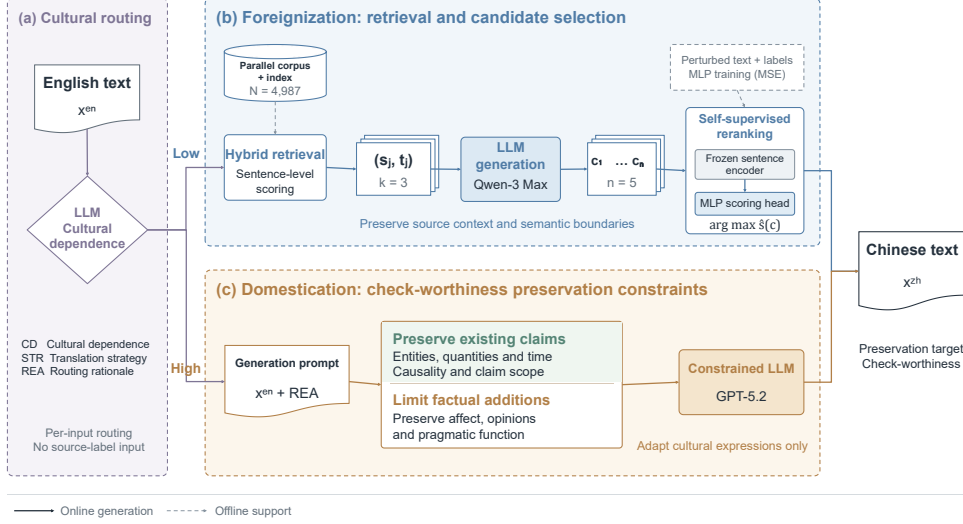


Fig. 1 Overall architecture of ANCHOR. (a) Cultural dependence classification and path selection; (b) the foreignization path based on demonstration retrieval, candidate generation and self-supervised reranking; (c) the domestication path, which adapts expressions under check-worthiness preservation constraints. Dashed arrows indicate offline support. The foreignization experiments use hybrid retrieval over the full corpus.

define the boundary of rewriting. The following subsections describe the objective, operation and output of each component.

3.1 Cultural dependence classification and path selection

Cultural dependence classification identifies contextual information that affects whether Chinese readers can understand the original pragmatic function. Following the idea of first identifying culturally relevant content and then adapting it (Khanuja et al. 2024), the classifier uses GPT-5.2 with temperature 0. It takes the English text and the strategy prompt P_{sel} as input and outputs the cultural dependence level, the translation strategy and the rationale:

$$(CD, STR, REA) = \text{LLM}(x^{en}, P_{sel}). \quad (1)$$

Here $CD \in \{\text{High}, \text{Low}\}$, corresponding to domestication and foreignization, respectively. Expressions that depend on a specific social background, regional reference, idiom or cultural metaphor are classified as High; general emotional expressions, everyday complaints and metaphors that are understandable across cultures are classified as Low. The rationale records the expression features that need handling and provides contextual grounds for the domestication prompt. The classification prompt is given in Appendix A.1.

3.2 Foreignization path: demonstration-assisted candidate generation and selection

The foreignization path handles texts with low cultural dependence, aiming to keep the context, register and semantic boundaries that can be transferred directly. Following AFSP (Adaptive Few-shot Prompting) (Tang et al. 2025), we organize a few-shot translation pipeline that assists generation with relevant bilingual demonstrations and then selects the output from multiple candidates.

The following paragraphs describe the parallel corpus, vector index, hybrid retrieval, candidate generation and reranking in turn. The generation model is a replaceable component of this pipeline and is determined by the model selection experiment in Section 4.2.

3.2.1 Building a domain parallel corpus

The parallel corpus was collected from the Chinese and English transcripts of the regular press conferences published on the official website of the Ministry of Foreign Affairs of China in 2024–2025¹. These transcripts have corresponding bilingual versions and cover topics such as public affairs and international relations together with related terminology, which makes them suitable for building domain translation demonstrations. After collection and alignment, 4,987 bilingual entries were obtained.

The construction process extracts the URL, date, title and body paragraphs of the bilingual pages, pairs pages by publication date, and then aligns Chinese and English paragraphs by position. The aligned entries are stored in JSONL format and denoted as

$$\mathcal{D} = \{(s_i, t_i, d_i)\}_{i=1}^N, \quad s_i \in \mathcal{L}_{zh}, t_i \in \mathcal{L}_{en}, \quad (2)$$

where s_i and t_i are a Chinese paragraph and its English counterpart, d_i is the date identifier and N is the number of aligned entries. For English-to-Chinese retrieval, the English side is used as the matching target of the query, and the Chinese side serves as the demonstration translation.

3.2.2 Building an offline vector index

We use text-embedding-3-large to build dense representations of the corpus (OpenAI 2024), which decouples embedding computation from translation generation. The index is built offline, and online retrieval only needs to encode the current query.

For each English entry t_i in the corpus, a normalized embedding is computed:

$$\mathbf{e}_i = \text{Normalize}(\text{Embedding}(t_i)) \in \mathbb{R}^H, \quad (3)$$

where $\text{Embedding}(\cdot)$ denotes the text-embedding-3-large model and $\text{Normalize}(\cdot)$ performs L2 normalization:

$$\text{Normalize}(\mathbf{v}) = \frac{\mathbf{v}}{\|\mathbf{v}\|_2 + \epsilon}, \quad \epsilon = 10^{-12}. \quad (4)$$

¹https://www.fmprc.gov.cn/fyrbt_673021/jzhsl_673025/

The corpus vectors form a matrix $\mathbf{E} \in \mathbb{R}^{N \times H}$ with $H = 3072$. The index also stores the bilingual texts and date metadata so that demonstration pairs can be recovered from retrieval results.

Separate indexes are built for the Chinese and English sides to support both translation directions. Queries and the corpus index use the same embedding model; if the model is changed, the corpus must be re-encoded to keep the representation space consistent.

3.2.3 Two-stage hybrid demonstration retrieval

Hybrid retrieval score

Beyond sentence-level semantic matching, hybrid retrieval introduces two additional scores. Let the input query be q with normalized sentence vector $\mathbf{q} \in \mathbb{R}^H$, and let its length L_q be the number of units obtained by whitespace splitting, which for English text equals the number of words; corpus entry i corresponds to $\mathbf{e}_i$ and L_i . In the English-to-Chinese direction the English side serves as both query and matching target, so L_q and L_i are computed on English text. The relevance score between the query and entry i is

$$\text{score}(q, i) = \alpha_1 \cdot S_{\text{dense}}(q, i) + \alpha_2 \cdot S_{\text{sparse}}(q, i) + \alpha_3 \cdot S_{\text{multi}}(q, i), \quad (5)$$

where $\alpha_1, \alpha_2, \alpha_3$ are mixing weights satisfying $\alpha_1 + \alpha_2 + \alpha_3 = 1$; we set $(\alpha_1, \alpha_2, \alpha_3) = (0.4, 0.4, 0.2)$. The dense score captures global semantic similarity through cosine similarity:

$$S_{\text{dense}}(q, i) = \mathbf{q}^\top \mathbf{e}_i, \quad (6)$$

where $\mathbf{e}_i$ is the L2-normalized sentence vector of entry i . The sparse score multiplies the activations obtained by projecting onto a weight vector and scales them by the number of split units:

$$S_{\text{sparse}}(q, i) = \min(L_q, L_i) \cdot \text{ReLU}(\mathbf{q}^\top \mathbf{w}_s) \cdot \text{ReLU}(\mathbf{e}_i^\top \mathbf{w}_s), \quad (7)$$

where $\mathbf{w}_s \in \mathbb{R}^H$ is a sparse weight vector. The multi-vector score computes similarity in a projected space:

$$S_{\text{multi}}(q, i) = (\mathbf{q} \mathbf{W}_m)^\top (\mathbf{e}_i \mathbf{W}_m), \quad (8)$$

where $\mathbf{W}_m \in \mathbb{R}^{H \times H}$ is a projection matrix. Both $\mathbf{w}_s$ and $\mathbf{W}_m$ are randomly initialized from a Gaussian distribution with standard deviation 0.02 and kept fixed during retrieval without training; the random seed is fixed to 42 so that retrieval results are reproducible.

Both additional scores are derived from sentence vectors. Beyond dense similarity, they introduce similarity signals under different projections and a length-dependent scaling, and they act as complements to sentence-level information. Note that defining length by whitespace splitting relies on languages that separate words with spaces. In our English-to-Chinese direction, the length is computed on the English side and ranges from 1 to 315 (median 55). If a language without spaces between words, such as Chinese, were used on the query side, this length would degenerate to a constant, and

Eq. (7) would reduce to a length-independent product; a tokenizer for that language would then be needed to count length.

Two-stage retrieval with dense pre-filtering

The dense score can be computed over the whole corpus with a single matrix multiplication, whereas Eqs. (7) and (8) are computed entry by entry in the implementation; the $H \times H$ projection by $\mathbf{W}_m$ is the main cost. An optional acceleration scheme uses two-stage cascaded retrieval, which restricts the full hybrid score to a candidate subset obtained by dense retrieval.

The first stage computes the dense scores between the query and all corpus entries by matrix multiplication:

$$\mathbf{s}_{\text{dense}} = \mathbf{E} \cdot \mathbf{q} \in \mathbb{R}^N, \quad (9)$$

where $\mathbf{E} \in \mathbb{R}^{N \times H}$ contains the vector representations of all N corpus entries. The top M entries are then selected as candidates:

$$\mathcal{C}_M = \text{argTopK}(\mathbf{s}_{\text{dense}}, M), \quad M \ll N. \quad (10)$$

The acceleration experiment sets $M = 512$, while the translation experiments in the main text use $M \geq N$. Candidate selection uses a partition-based selection algorithm and does not need to sort all N scores.

The second stage computes the full hybrid score for the entries in $\mathcal{C}_M$ and reranks them accordingly:

$$\text{score}(q, i) = \alpha_1 \cdot S_{\text{dense}}(q, i) + \alpha_2 \cdot S_{\text{sparse}}(q, i) + \alpha_3 \cdot S_{\text{multi}}(q, i), \quad \forall i \in \mathcal{C}_M. \quad (11)$$

The dense scores obtained in the first stage are reused directly in the second stage. Pre-filtering reduces the number of entries for which the two additional scores must be computed from N to M , while dense scoring over the full corpus is still required. Under the acceleration configuration, the full hybrid score covers 512 candidates of our 4,987 corpus entries. When $M \geq N$, two-stage retrieval degenerates into full enumeration and yields the same result as computing the full hybrid score directly. Actual running time also depends on embedding computation and hardware. The effect of the candidate number M on retrieval consistency and time is reported in Appendix C. Following the demonstration configuration of AFSP, we take the top $k = 3$ reranked entries as few-shot demonstrations.

3.2.4 LLM-based translation candidate generation

The k retrieved bilingual demonstrations and the input query q are combined into a generation prompt:

$$P_{\text{for}} = \text{Template}(\{(s_j^{\text{demo}}, t_j^{\text{demo}})\}_{j=1}^k, q), \quad (12)$$

where s_j^{demo} and t_j^{demo} are the Chinese and English sides of the retrieved demonstrations; in English-to-Chinese translation, t_j^{demo} serves as the demonstration source and s_j^{demo} as the demonstration translation. The template is given in Appendix A.2.

From the same prompt, n candidate translations are generated:

$$\{c_1, c_2, \dots, c_n\} = \text{LLM}(P_{for}, n). \quad (13)$$

We set $n = 5$, and the reranker then selects the output from the candidate set.

3.2.5 Self-supervised reranker

The reranker is trained on the parallel corpus and its perturbed texts and learns a text scoring function from pseudo-labels, which is used to compare the candidate translations generated by the LLM.

Reranker architecture

The reranker consists of a frozen sentence encoder and a trainable scoring head. The encoder is implemented with Sentence-Transformers and denoted $f_\theta(\cdot)$. During training, θ is fixed and only the parameters of the scoring head are updated, which reuses pretrained sentence representations and reduces back-propagation cost.

A two-layer multilayer perceptron (MLP) $g_\phi(\cdot)$ maps sentence representations to scalar scores:

$$g_\phi(\mathbf{h}) = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2, \quad (14)$$

where $\mathbf{W}_1 \in \mathbb{R}^{128 \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times 128}$, $\mathbf{b}_1 \in \mathbb{R}^{128}$ and $\mathbf{b}_2 \in \mathbb{R}$. The ranking score of a candidate translation c is

$$\hat{s}(c) = g_\phi(f_\theta(c)) \in \mathbb{R}. \quad (15)$$

Self-supervised training with perturbation pseudo-labels

Following studies on pseudo-data construction and quality evaluation (Geng et al. 2023; Bennett et al. 2025), we construct original and perturbed samples from target-side texts. Let z_i denote a target-language text used for reranker training and $|z_i|$ its number of tokens; in English-to-Chinese translation, $z_i = s_i$, i.e., the Chinese side. Each text yields several “text-pseudo-label” pairs.

The pseudo-label of an original text is 1.0:

$$(z_i, y = 1.0). \quad (16)$$

The first perturbation swaps the first and third tokens to simulate local word-order errors, with pseudo-label 0.8:

$$(z_i^{\text{swap}}, y = 0.8), \quad z_i^{\text{swap}} = \text{Swap}(z_i, 1, 3). \quad (17)$$

The second perturbation deletes the last token to simulate truncation, also with pseudo-label 0.8. Using 1-based indices,

$$(z_i^{\text{drop}}, y = 0.8), \quad z_i^{\text{drop}} = (z_{i,1}, \dots, z_{i,|z_i|-1}). \quad (18)$$

Both perturbations are applied only to texts with more than three tokens. The pseudo-labels represent predefined perturbation levels.

The scoring head is trained with the mean squared error (MSE) between predicted scores and pseudo-labels, where $\mathcal{B}$ is a training batch:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(c,y) \in \mathcal{B}} (\hat{s}(c) - y)^2. \quad (19)$$

Training uses the Adam optimizer with a learning rate of 10^{-4} and a batch size of 16. Up to 3,000 texts are sampled from the corpus with a fixed random seed, and training runs for two epochs. At inference, each of the n candidates is scored, and the highest-scoring translation is output:

$$\hat{y} = \arg \max_{c \in \{c_1, \dots, c_n\}} \hat{s}(c). \quad (20)$$

3.3 Domestication path: expression adaptation under factual and pragmatic constraints

The domestication path handles texts with high cultural dependence and conveys idioms, metaphors and cultural references with expressions of similar function in the Chinese context. The generation model receives the English text, the strategy rationale and the domestication template; its selection is described in Section 4.3:

$$\hat{x}^{zh} = \text{LLM}(x^{en}, \text{REA}, P_{gen}), \quad (21)$$

where P_{gen} turns the task requirements into generation instructions. We limit the scope of rewriting in four respects:

1. retain the verifiable facts already present in the source and introduce no new factual claims;
2. retain person, organization and place names that serve as claim targets, as well as numbers, times, causal relations and assertion scope;
3. maintain the emotional tone and pragmatic function, and avoid expanding opinions, emotions or small talk into verifiable details;
4. render cultural expressions that affect understanding with Chinese expressions of similar function while retaining the factual elements above.

These rules constrain, respectively, the retention of existing claims and the adaptation of non-factual expressions. Whereas Kabir et al. (2026) use cultural persona prompts to examine how cultural knowledge affects label drift, this study sets generation constraints around the information on which the checking decision depends. The full prompt template is given in Appendix A.4.

4 Experiments

The experiments address three questions: whether ANCHOR preserves the source check-worthiness label better than the baseline translation models and GPT-5.2 direct translation; which textual errors account for the preservation difference between

ANCHOR and direct translation; and how ANCHOR performs in general translation quality on each path, and whether quality rankings agree with label preservation. Before answering them, we validate cultural dependence classification and determine the underlying model of each path through model selection experiments. We then introduce the overall evaluation setup, report check-worthiness preservation in the baseline and ablation experiments together with the failure types, and finally report general translation quality.

4.1 Validation of cultural dependence classification

Cultural dependence classification determines the translation path of a text. This experiment uses human verification labels as the reference to evaluate the classification performance of this step and the agreement among evaluators.

4.1.1 Data and verification setup

The experiment includes 200 English samples each from the Groundtruth subset of ClaimBuster (Arslan et al. 2020) and from ViClaim (Giedemann et al. 2025). From each source, 100 High and 100 Low samples were drawn according to the model’s cultural dependence output, 400 in total. This stratification is based on model predictions, so the class proportions after human verification are determined by the actual judgments.

Six master’s students with a humanities and social sciences background who had passed the College English Test Band 6 (CET-6) took part in the verification; they are denoted C_1 to C_6 . Before the formal evaluation, the authors explained the definition of cultural dependence and the judgment criteria. The evaluators independently labeled each text as High or Low, corresponding to the domestication or foreignization strategy.

The class receiving at least four votes for a sample is its reference label; when the two classes each receive three votes, the authors adjudicate. Let $C_{i,j}$ be the judgment of evaluator C_j on sample i and y_i^{author} the adjudication result. The final reference label y_i^* is defined as

$$y_i^* = \begin{cases} \text{Majority}(C_{i,1}, \dots, C_{i,6}), & \text{if the six judgments are not tied,} \\ y_i^{\text{author}}, & \text{if the six judgments are tied 3:3.} \end{cases} \quad (22)$$

With high cultural dependence as the positive class, we report accuracy, balanced accuracy, Macro-F1, High-class precision, recall and F1, and Cohen’s κ . In addition, Fleiss’ κ , pairwise Cohen’s κ and observed agreement are used to evaluate the reliability among C_1 to C_6 .

4.1.2 Results and analysis

The six evaluators reached a majority on 380 samples, of which 309 agreed with the model and 71 did not; the remaining 20 samples were 3:3 ties, all of which agreed with the model after adjudication. Overall, the model classified 329 of the 400 samples correctly and 71 incorrectly, as shown in Table 1.

Table 1 Validation results of ANCHOR’s cultural dependence classification

Accuracy	Balanced accuracy	Macro-F1	Cohen’s κ
0.8225	0.8690	0.8167	0.6450

The model obtains an accuracy of 0.8225, a balanced accuracy of 0.8690, a Macro-F1 of 0.8167 and a Cohen’s κ of 0.6450. With High as the positive class, there are 129 true positives, 71 false positives and 0 false negatives, corresponding to a precision of 0.6450, a recall of 1.0000 and an F1 of 0.7842. Errors concentrate on texts whose human reference label is Low but which the model classified as High.

The observed agreement of the six evaluators is 0.8835 and Fleiss’ κ is 0.7169; the mean Cohen’s κ over the 15 evaluator pairs is 0.7174 (Table 2).

Table 2 Agreement of the human cultural dependence verification

Scope	Agreement	κ
All six evaluators	0.8835	0.7169 (Fleiss’ κ)
Mean over 15 pairs	0.8835	0.7174 (Cohen’s κ)

Under this sample and adjudication rule, the High-class recall is 1.0000 and the precision is 0.6450. The 71 false positives indicate that some texts with low cultural dependence were routed to the domestication path, which may add unnecessary rewriting.

4.2 Model selection for the foreignization path

Model selection for the foreignization path is carried out under a unified retrieval, prompting and reranking configuration. The translation quality of four candidate models is compared to determine the underlying model used for subsequent generation.

4.2.1 Data

The evaluation uses 504 texts from the English training set of NLP4IF-2021 (Shaar et al. 2021). The corpus comes from social media and contains annotation dimensions such as verifiable claim, need for manual fact-checking, public interest and potential harm, which correspond to the data transfer task of this study.

4.2.2 Evaluation metrics

We use COMET-QE (Rei et al. 2020) and COMETKiwi (Rei et al. 2022) for reference-free quality estimation. Both metrics predict translation quality from the source and the translation.

LLM-assisted evaluation gives a composite score of 0–100 covering meaning fidelity, terminology accuracy, fluency, register match and pragmatic effect, denoted LLM Score. Following studies on LLM-based translation judging (Kocmi and Federmann 2023; Qian et al. 2024), GPT-5.2 is used as the judge. The scoring template is given in Appendix A.3. LLM Score and the two QE metrics are reported separately.

4.2.3 Experimental details

The candidate models are GPT-5.2, Seed-1.6, DeepSeek-3.2 and Qwen-3 Max. Each model translates the same 504 English texts, with $k = 3$ retrieved demonstrations and $n = 5$ candidate translations.

All foreignization results reported in this paper are obtained with $M \geq N$, i.e., the hybrid score covers all 4,987 corpus entries. As described in Section 3, this configuration is equivalent in results to two-stage cascaded retrieval; dense pre-filtering ($M = 512$) is an optional acceleration that controls the second-stage cost when the corpus grows. Its effect on retrieval consistency and time is reported in Appendix C.

4.2.4 Results

Table 3 Foreignization quality of the candidate models

Model	COMET-QE WMT20	COMETKiwi QE WMT22	LLM Score (0–100)
GPT-5.2	0.0899	0.7563	91.1869
Seed-1.6	0.0953	0.7362	88.4048
DeepSeek-3.2	0.1275	0.7484	89.8451
Qwen-3 Max	0.1306	0.7574	91.0883

The results of the four models are shown in Table 3 and Fig. 2. Qwen-3 Max obtains the highest COMET-QE score of 0.1306, followed by DeepSeek-3.2 with 0.1275, a difference of 0.0031; Seed-1.6 and GPT-5.2 obtain 0.0953 and 0.0899, respectively. On COMETKiwi, Qwen-3 Max and GPT-5.2 score 0.7574 and 0.7563, a mean difference of 0.0011, while DeepSeek-3.2 and Seed-1.6 score 0.7484 and 0.7362. GPT-5.2 has the highest LLM Score of 91.1869 and Qwen-3 Max follows with 91.0883, a difference of 0.0986; DeepSeek-3.2 and Seed-1.6 score 89.8451 and 88.4048. The QE metrics and the LLM Score thus rank the top model differently.

Sample analysis shows that some translations insufficiently preserve colloquialisms, profanity, irony, puns and emotional intensity. For example, when the source “this is f**king bullsh*t” is rendered as “胡说八道” or “一派胡言” (both roughly “non-sense”), the emotional intensity and register change. Other errors involve proper names and formatting, including the handling of “Dollar General” and the omission of URL placeholders, hashtags and ellipses. These changes affect the complete presentation of source information. When translating “No offense but the corona virus disappearing before April actually sucks”, Seed-1.6 output a refusal and irrelevant content, and this

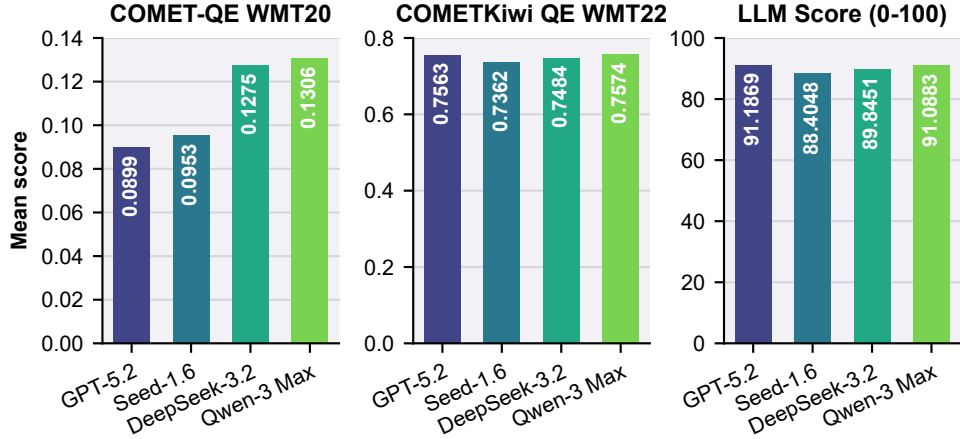


Fig. 2 Visualization of the foreignization quality of the candidate models

sample received an LLM Score of 5. The example shows that unfulfilled translation requests affect corpus coverage and average quality.

Taken together, Qwen-3 Max has the highest means on both QE metrics, and its LLM Score is less than 0.1 below that of GPT-5.2. We therefore select Qwen-3 Max as the underlying model of the foreignization path in the subsequent overall evaluation.

4.3 Model selection for the domestication path

The domestication path conveys the pragmatic function of the source through cultural adaptation of expressions, and its generation requirements differ from those of the foreignization path. This subsection therefore compares candidate models under a unified domestication prompt and determines the underlying model of this path from the results.

4.3.1 Data

We randomly sampled 500 texts each from the English corpora of ClaimBuster and ViClaim, covering presidential debates and colloquial online video speech, respectively, for domestication model selection.

4.3.2 Evaluation metrics and setup

The candidate models are the same as in foreignization selection: GPT-5.2, Seed-1.6, DeepSeek-3.2 and Qwen-3 Max. The four models process the same inputs under the same domestication prompt configuration, and translation quality is compared with COMET-QE, COMETKiwi and LLM Score. The LLM Score is judged by GPT-5.2, and the QE metrics and LLM Score are reported separately.

After the model is determined, we examine information changes in cultural adaptation with raw output cases and use a human questionnaire to assess meaning accuracy,

cultural relevance and fluency. The cases and the questionnaire setup and results are given in Sections 4.3.4 and 4.3.5.

4.3.3 Results and selection

Table 4 Domestication quality of the candidate models

Model	COMET-QE WMT20	COMETKiwi QE WMT22	LLM Score (0–100)
GPT-5.2	0.2488	0.7772	88.3948
Seed-1.6	0.1000	0.7759	86.8520
DeepSeek-3.2	−0.2036	0.7570	87.8050
Qwen-3 Max	−0.2364	0.7723	86.5110

The aggregated results of the four candidate models are shown in Table 4.

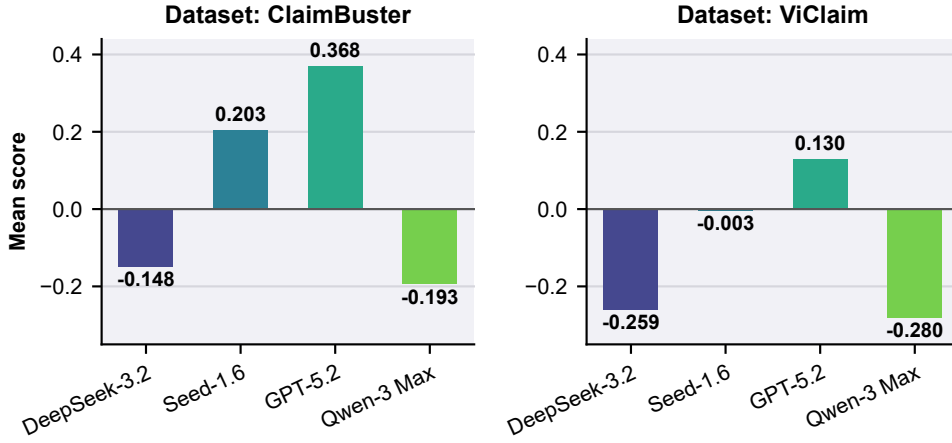


Fig. 3 COMET-QE WMT20 results of the candidate models for domestication

Figure 3 gives the COMET-QE results by dataset. GPT-5.2 scores 0.3681 on ClaimBuster and 0.1296 on ViClaim, the highest on both; Seed-1.6 scores 0.2030 and −0.0031. DeepSeek-3.2 scores −0.1484 and −0.2588, and Qwen-3 Max scores −0.1929 and −0.2800 on the two datasets. All four models score lower on ViClaim than on ClaimBuster, showing a data-source difference under this metric.

The COMETKiwi results are shown in Fig. 4. On ClaimBuster, GPT-5.2, Seed-1.6, Qwen-3 Max and DeepSeek-3.2 score 0.8108, 0.8045, 0.7983 and 0.7863, respectively, with a difference of 0.0245 between the highest and lowest means. On ViClaim, Seed-1.6, Qwen-3 Max, GPT-5.2 and DeepSeek-3.2 score 0.7472, 0.7464, 0.7436 and 0.7277; the means of the first three differ by less than 0.004, and the ranking differs from that on ClaimBuster.

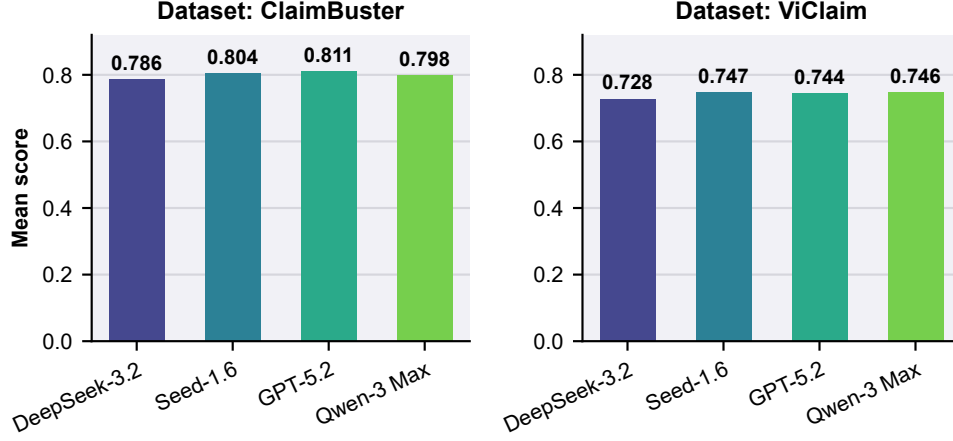


Fig. 4 COMETKiwi QE WMT22 results of the candidate models for domestication

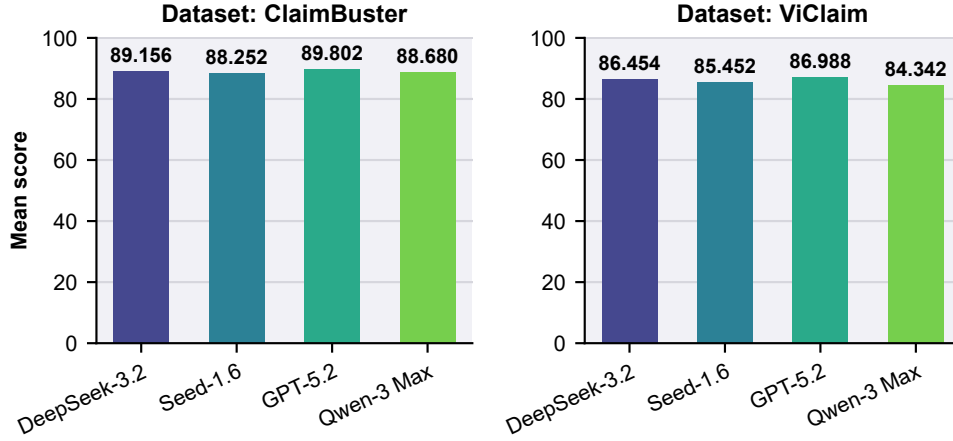


Fig. 5 LLM Score results of the candidate models for domestication

The LLM Scores are shown in Fig. 5. On ClaimBuster, GPT-5.2 and DeepSeek-3.2 score 89.8016 and 89.1560, and Qwen-3 Max and Seed-1.6 score 88.6800 and 88.2520. On ViClaim, GPT-5.2, DeepSeek-3.2, Seed-1.6 and Qwen-3 Max score 86.9880, 86.4540, 85.4520 and 84.3420, all lower than on ClaimBuster.

Over the two datasets, GPT-5.2 has the highest mean on all three metrics, namely 0.2488, 0.7772 and 88.3948. We therefore select GPT-5.2 as the underlying model of the domestication path. Since the LLM judge and this candidate are the same model, the selection also relies on the QE metrics and human analysis.

4.3.4 Raw outputs of the domestication path

To further analyze information changes in cultural adaptation on the domestication path, Tables 5 and 6 show raw domestication outputs. The Chinese outputs are reproduced verbatim, each followed by an English gloss in italics.

Table 5 Representative domestication outputs (ID = 3197)

Item	Text
Source	We’re currently staying in one of the most obese states in America, sweet home Alabama.
GPT-5.2	我们现在待在一个出了名“胖友”多的地方，哎呀这熟悉的家乡味儿。 <i>We are now staying in a place famous for its many “chubby friends”; oh, this familiar taste of home.</i>
Qwen-3 Max	我们现在待在一个特别能吃的地方，简直就是美食天堂。 <i>We are now staying in a place where people eat a lot; it is simply a food paradise.</i>
DeepSeek-3.2	我们这会儿待的地方，在全国体重榜上可是名列前茅的，真是好山好水好福气啊。 <i>The place we are staying now ranks near the top of the national weight chart; truly fine hills, fine waters and good fortune.</i>
Seed-1.6	我们现在待的这地方呀，就是大家口中那种容易长膘的地儿，哈哈，毕竟咱这儿的老话都说“金窝银窝不如自家草窝”嘛。 <i>The place we are staying now is the kind of place people say makes you put on weight, haha; after all, as the old saying goes here, “a golden or silver nest is no match for one’s own straw nest”.</i>
Analysis	GPT-5.2 keeps the teasing tone but omits America and Alabama; Qwen-3 Max changes the obesity description; DeepSeek-3.2 and Seed-1.6 add extra rhetoric. All four outputs contain information changes.

Sample 3197 contains a description of obesity, a regional reference to America and Alabama, and the cultural expression “sweet home Alabama”. GPT-5.2 keeps the teasing tone but omits the country and state names; Qwen-3 Max changes the obesity description into an eating preference; the other two outputs add extra rhetoric. Natural expression does not imply that the factual objects and descriptive scope are fully retained.

Sample 230 involves the governance of Florida and the rhetoric of “the dirty little secret”. GPT-5.2 generalizes the place name to “this place” and adds “to win”, which the source does not state; Qwen-3 Max keeps the place name and the governance topic; DeepSeek-3.2 does not complete the sentence, and Seed-1.6 also generalizes the regional reference. Adapting cultural expressions therefore requires checking the claim target, factual scope and information addition at the same time.

Table 6 Representative domestication outputs (ID = 230)

Item	Text
Source	The dirty little secret about Florida is to really, well govern the state.
GPT-5.2	说到底，这地方的“潜台词”就是：想赢，关键还是得把治理这事儿真做好。 <i>In the end, the “subtext” of this place is: if you want to win, the key is still to really get governance right.</i>
Qwen-3 Max	佛罗里达州的那些小秘密，其实就是如何真正治理好这个州。 <i>Florida’s little secrets are really about how to truly govern the state well.</i>
DeepSeek-3.2	关于佛罗里达，大家心照不宣的一点就是，真要把州管好，其实..... <i>About Florida, the tacit understanding is that, to really run the state well, actually...</i>
Seed-1.6	说起某个地方的小秘密啊，其实就是好好管事儿呗。 <i>Speaking of a certain place’s little secret, it is just about managing things well.</i>
Analysis	GPT-5.2 generalizes the place name and adds “to win”; Qwen-3 Max keeps the place name and the governance topic; DeepSeek-3.2 leaves the output unfinished; Seed-1.6 generalizes the regional reference.

4.3.5 Three-dimensional human evaluation of domestication outputs

After the domestication model was selected, the questionnaire evaluated domesticated translations generated by GPT-5.2. For sampling, texts classified as highly culture-dependent in the ClaimBuster Groundtruth subset and in ViClaim were first merged into a candidate pool of 549 texts, 424 from ViClaim and 125 from ClaimBuster; 100 texts were then drawn from the pool with a fixed random seed (42), regardless of source. The sample composition therefore follows the proportions of the two sources in the pool, with 74 texts from ViClaim and 26 from ClaimBuster. Seven graduate students (denoted V_1 to V_7) rated meaning accuracy, cultural relevance and fluency on a 1–5 scale and recorded problems. Three of them have experience in translation, linguistics or cross-lingual text processing, and four regularly read news, online comments and texts on public issues. The questionnaire displayed the source, the translation, the data source and the strategy category, while hiding the model and automatic scores; the focus of each dimension is given in Table 7.

The 100 questionnaire samples from the domestication selection stage were rated by seven evaluators, yielding 700 evaluation records (Table 8). The means of meaning accuracy, cultural relevance and fluency are 4.146, 4.304 and 4.707, respectively, with an overall mean of 4.386.

The scores of each evaluator and the overall means are visualized in Fig. 6. The individual means of the seven evaluators range from 4.077 to 4.740, and the ordering of cultural relevance and meaning accuracy for V_4 , V_5 and V_7 differs from that of

Table 7 Dimensions of the human questionnaire

Dimension	Scale	Focus
Meaning accuracy	1–5	Whether the core meaning and pragmatic orientation of the source are kept
Cultural relevance	1–5	Whether reasonable adaptation to the target-language culture is made
Fluency	1–5	Whether the Chinese expression is natural, smooth and readable

Table 8 Mean questionnaire scores by evaluator and overall

Evaluator	Meaning accuracy	Cultural relevance	Fluency	Evaluator mean
V_1	4.440	4.840	4.940	4.740
V_2	4.110	4.390	4.920	4.473
V_3	3.810	4.020	4.680	4.170
V_4	3.770	3.740	4.720	4.077
V_5	4.740	4.690	4.760	4.730
V_6	3.920	4.480	4.760	4.387
V_7	4.230	3.970	4.170	4.123
Overall	4.146	4.304	4.707	4.386

the overall means. Among the three dimensions, fluency has the highest mean and meaning accuracy the lowest, with a difference of 0.561.

4.4 Overall evaluation setup

4.4.1 Selected configuration and compared models

Based on the model selection results in Sections 4.2 and 4.3, the overall evaluation fixes Qwen-3 Max for generating foreignization candidates and GPT-5.2 for domestication. The foreignization path keeps $k = 3$ retrieved demonstrations, $n = 5$ candidates and full-corpus hybrid scoring; the domestication path keeps the prompt constraints described above. With the models fixed, the following experiments compare the complete ANCHOR pipeline with other translation models.

The baseline experiment compares ANCHOR with seven translation models: Baidu Translate, Google Translate, DeepL, OPUS-MT en-zh, NLLB-200 distilled 600M, Hy-MT2-1.8B and TranslateGemma-4B-IT. All models process the same English inputs.

The ablation experiment additionally uses GPT-5.2 direct translation, which translates English into Simplified Chinese with a minimal instruction and outputs only the translation. This control uses the same model and the same temperature (0.3) as ANCHOR. The ablation experiment compares the complete method with direct translation, and its results are reported separately.

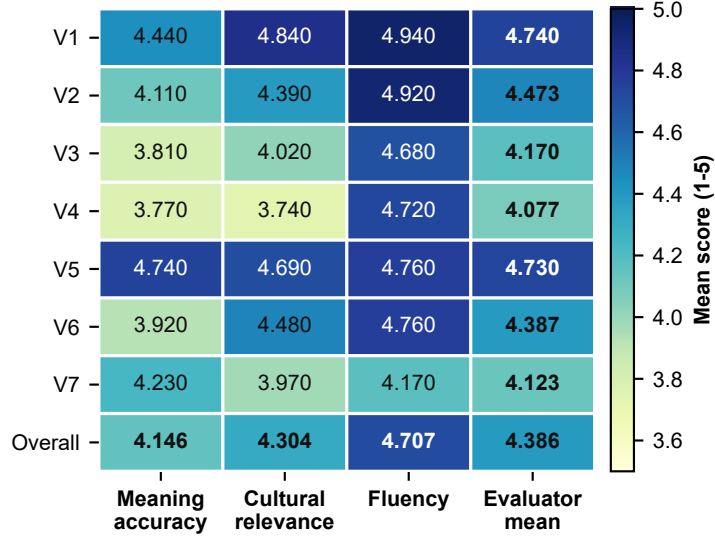


Fig. 6 Mean questionnaire scores by evaluator and overall

4.4.2 Sample use and evaluation scope

The general translation quality evaluation includes 1,504 inputs from NLP4IF, ClaimBuster and ViClaim. Baidu Translate returned a content policy refusal code for 5 of them; after removing these 5 texts from the common evaluation set of all models, 1,499 texts remain, distributed as shown in Table 9. By the path actually taken in ANCHOR, 258 are domesticated and 1,241 are foreignized; all 501 NLP4IF texts follow the foreignization path. This set has the same inputs as the model selection experiments in Sections 4.2 and 4.3, so that the complete ANCHOR pipeline with the selected underlying models and each baseline model are compared on the same inputs. Model selection only determines the underlying model of each path among the four candidate LLMs; baseline outputs were not involved in any selection or tuning, and the reranker selects among candidates using its self-supervised score only, without access to any evaluation metric. The human evaluation samples for check-worthiness preservation are drawn separately, as described below.

Table 9 Data for the general translation quality evaluation

Dataset	Original size	Size used
NLP4IF	504	501
ClaimBuster	500	500
ViClaim	500	498
Total	1,504	1,499

This set contains 1,196 negative samples, and its ClaimBuster and ViClaim parts are all negative. To cover both classes, the check-worthiness preservation evaluation draws another 1,250 texts from the three English corpora, with equal numbers of positive and negative samples within each source: 198/198 for NLP4IF, 210/210 for ClaimBuster and 217/217 for ViClaim. From these, 320 texts are drawn by the strata “cultural dependence $\times$ source label” for human evaluation in the baseline and ablation experiments: 160 each for High and Low, each stratum containing 70 positive and 90 negative samples, for a total of 140 positive and 180 negative samples.

Stratified sampling keeps the workload of item-by-item bilingual judgment manageable while covering both classes and both cultural dependence groups. In the baseline experiment, the 320 source texts correspond to 2,560 model output units, each rated by three evaluators, requiring 7,680 unit judgments; in the ablation experiment, the 320 source texts correspond to 640 output units, requiring 1,920 initial judgments plus subsequent review. Under a limited annotation budget, the sampling design affects the quality of evaluation estimates (Saldías Fuentes et al. 2022). We therefore use a stratified sample for detailed human evaluation; general translation quality and check-worthiness preservation use different sample sets and are each reported according to their evaluation objectives.

4.4.3 General translation quality metrics

General translation quality is measured with COMET-QE (Rei et al. 2020), COMETKiwi (Rei et al. 2022) and LLM Score. The first two are reference-free quality estimates based on the source and the translation. LLM Score is a composite 0–100 score given by GPT-5.2 over meaning fidelity, entities and terminology, fluency, register and pragmatic effect; the prompt is given in Appendix A.3, and the evaluation follows related work on translation judging (Kocmi and Federmann 2023; Qian et al. 2024).

4.4.4 Check-worthiness preservation rules in the baseline and ablation experiments

Evaluators view the English source, the source check-worthiness label and an anonymous Chinese translation, and judge whether the translation retains the same check-worthiness property. Expanding an opinion or emotion into a verifiable fact, or generalizing an explicit claim into a statement without a checking target, is judged as not preserved. Candidate identities are hidden and the order is randomized per sample, following human translation evaluation settings that provide judgment criteria and anonymization (Freitag et al. 2021).

Baseline experiment: evaluation of eight models

Three native Chinese master’s students with fact-checking experience evaluated independently after unified training and calibration. Let model m have n_v samples in group v , and let $k_{m,i,r} \in \{0, 1\}$ be the preservation judgment of evaluator R_r ($r = 1, 2, 3$).

The label preservation rate (LPR) is

$$\text{LPR}_{m,v} = \frac{100}{n_v} \sum_{i \in v} \left(\frac{1}{3} \sum_{r=1}^3 k_{m,i,r} \right). \quad (23)$$

The three judgments are first averaged within a unit and then aggregated with equal weight per sample. Evaluators also rated factual scope, pragmatic function, cultural appropriateness and fluency on a 0–4 scale; the first three and the preservation judgment are converted with equal weight into a composite task score. The preservation judgment has a Fleiss’ κ of 0.58, a three-way full agreement of 95.6% and a mean pairwise observed agreement of 97.1%; the Krippendorff’s α of the four ratings ranges from 0.7984 to 0.8463. Full statistics are given in Appendix B.

Ablation experiment: direct translation control

The ablation experiment uses the same 320 stratified samples to compare ANCHOR with GPT-5.2 direct translation. Model identities were hidden and the order shuffled. Another three graduate students with a humanities and social sciences background and fact-checking experience each judged whether the translation could still support checking the same claim, and the initial judgments were reviewed to form final judgments. The two experiments differ in evaluators and aggregation: the baseline experiment averages three judgments, whereas the ablation experiment uses the final reviewed judgments. ANCHOR’s preservation rates in the two experiments are therefore not directly comparable; however, the model comparison within each experiment is made under the same evaluators and the same rules.

4.5 Overall and group results of check-worthiness preservation

In the baseline experiment, ANCHOR and DeepL both obtain an overall label preservation rate of 99.38%, tied for the highest (Table 10). Google and Hy-MT2 obtain 98.75% and 98.85%, close to the top two, while OPUS-MT and NLLB-600M obtain 91.88% and 87.29%.

Table 10 Check-worthiness label preservation rates (%) under anonymous blind evaluation in the baseline experiment

Model	Overall $n=320$	Check-worthy $n=140$	Not check-worthy $n=180$	High cultural dependence $n=160$	High× Not check-worthy $n=90$
Baidu Translate	97.71	95.71	99.26	97.50	99.26
Google Translate	98.75	98.33	99.07	98.33	98.52
DeepL	99.38	99.29	99.44	99.17	98.89
OPUS-MT	91.88	83.57	98.33	89.79	98.15
NLLB-600M	87.29	74.29	97.41	84.79	98.15
Hy-MT2	98.85	97.62	99.81	98.33	99.63
TranslateGemma	97.81	95.95	99.26	96.67	98.89
ANCHOR	99.38	98.57	100.00	98.75	100.00

The group results reveal class-level differences between ANCHOR and DeepL. DeepL reaches 99.29% on the positive class, higher than ANCHOR’s 98.57%, and 99.17% on the High cultural dependence group, also higher than ANCHOR’s 98.75%. ANCHOR reaches 100.00% on both the negative class and the High×negative group, compared with DeepL’s 99.44% and 98.89%. Thus, two models with the same overall preservation rate differ in how well they preserve the check-worthiness property of retained claims and of non-factual expressions.

The positive and negative classes need to be examined separately. The negative-class preservation rates of OPUS-MT and NLLB-600M are 98.33% and 97.41%, whereas their positive-class rates are 83.57% and 74.29%; reporting only the high negative-class rates would hide the loss of verifiable claims. ANCHOR achieves high preservation rates on both classes; its 100.00% on the negative class means that all three evaluators judged every one of the 180 negative samples as preserved.

4.6 Ablation experiment

4.6.1 Label preservation results

This subsection compares check-worthiness preservation between the ANCHOR method and direct translation with an LLM, answering what the complete transcreation pipeline gains over GPT-5.2 direct translation. The human evaluation in the ablation experiment uses the same 320 stratified source texts, and the results are shown in Table 11.

Table 11 Label preservation rates (%) in the human evaluation of the ablation experiment; differences are in percentage points

Model	Overall <i>n</i> =320	Check-worthy <i>n</i> =140	Not check-worthy <i>n</i> =180	High cultural dependence <i>n</i> =160	High× Not check-worthy <i>n</i> =90
ANCHOR	97.19	97.14	97.22	98.12	97.78
GPT-5.2	95.31	96.43	94.44	95.00	94.44
Difference	+1.88	+0.71	+2.78	+3.12	+3.33

ANCHOR’s overall preservation rate is 97.19% (311/320), compared with 95.31% (305/320) for GPT-5.2 direct translation, a difference of 1.88 percentage points. On the High cultural dependence group the rates are 98.12% and 95.00%, a difference of 3.12 percentage points; on the High×negative group they are 97.78% and 94.44%, a difference of 3.33 percentage points. Differences are computed from unrounded proportions and then rounded to two decimals, so they may differ slightly from subtracting the rounded percentages in the table.

The difference in preservation rate corresponds to 9 and 15 failed units, respectively, six fewer for ANCHOR. The error distribution further shows 2 versus 6 proper-name reference errors and 0 versus 2 quantity or entity rewrites, whereas ANCHOR has 2 added conclusions against 1 for direct translation. The overall gain mainly accompanies

fewer errors in handling proper names, quantities and entities, while rewriting still introduces extra information.

4.6.2 Discussion of ablation results: failure types and information changes

Table 12 classifies the final failure records of the ablation experiment. Failed units are deduplicated by “source text $\times$ model”; error types are counted separately for each problem occurring in a unit, and one unit may involve several types, so the sum of type counts does not equal the number of failed units.

Table 12 Error type counts and deduplicated failed units in the ablation experiment

Error type	ANCHOR	GPT-5.2
Incorrect proper-name reference	2	6
Quantity or entity rewritten	0	2
Assertion direction reversed	2	2
Conclusion absent from source added	2	1
Verifiable information added	1	1
Keyword mistranslated	2	2
Qualifier lost	0	1
Vocative structure broken	0	1
Hashtag meaning mistranslated	1	0
Failed units (deduplicated)	9	15

The clearest difference between the two models lies in handling proper names, quantities and entities. ANCHOR makes 2 proper-name reference errors against 6 for direct translation, and 0 against 2 quantity or entity rewrites. Instances from direct translation include expanding the source number “22” into “22万死亡” (220,000 deaths) and interpreting the equipment model “Challenger 2” as “挑战者” (“Challenger”). These changes alter the quantity or object to be checked and tie the difference in preservation rate to concrete textual problems.

ANCHOR still exhibits information addition and changes in propositional direction. It has 2 added conclusions, more than the 1 of direct translation; for example, after an unfinished source sentence, it adds “使各城市陷入困境” (“plunging the cities into difficulty”). Both models have 2 reversed assertion directions: ANCHOR renders “under control of” as “掌控” (“to control”) and “blocking sanctions” as “实施封锁制裁” (“to impose blockade sanctions”), while the direct translation case involves a misreading of the direction of irony. Such errors change causal conclusions, active–passive relations or attitudinal direction, and each may affect the original checking decision.

The remaining types include 2 keyword mistranslations and 1 addition of verifiable information for each model, and 1 lost qualifier and 1 broken vocative structure for direct translation. ANCHOR additionally has 1 mistranslated hashtag meaning.

4.7 General translation quality and path-level performance

4.7.1 General translation quality of the eight models

On the 1,499 common samples, ANCHOR obtains an LLM Score of 90.63, ranking second among the eight models, below DeepL’s 91.78 and slightly above Google’s 90.49 (Table 13). Its COMET-QE of 0.2095 is lower than those of Google, DeepL and the two open-weight translation LLMs; its COMETKiwi of 0.7708 is also below Baidu Translate and ranks sixth among the eight models. All three metrics are higher than those of OPUS-MT and NLLB-600M. The metrics rank the models differently.

Table 13 General translation quality on the 1,499 common samples

Model	Category	COMET-QE	COMETKiwi	LLM Score (0–100)
Baidu Translate	Commercial	0.1234	0.7839	84.49
Google Translate	Commercial	0.2739	0.8066	90.49
DeepL	Commercial	0.2722	0.8094	91.78
OPUS-MT	NMT	−0.0261	0.7220	73.80
NLLB-600M	NMT	−0.0182	0.6954	72.98
Hy-MT2	Open-weight translation LLM	0.2405	0.8080	88.95
TranslateGemma	Open-weight translation LLM	0.2284	0.7834	87.12
ANCHOR	Proposed	0.2095	0.7708	90.63

Figure 7 shows the score distribution of the models under the three metrics, which order the models differently.

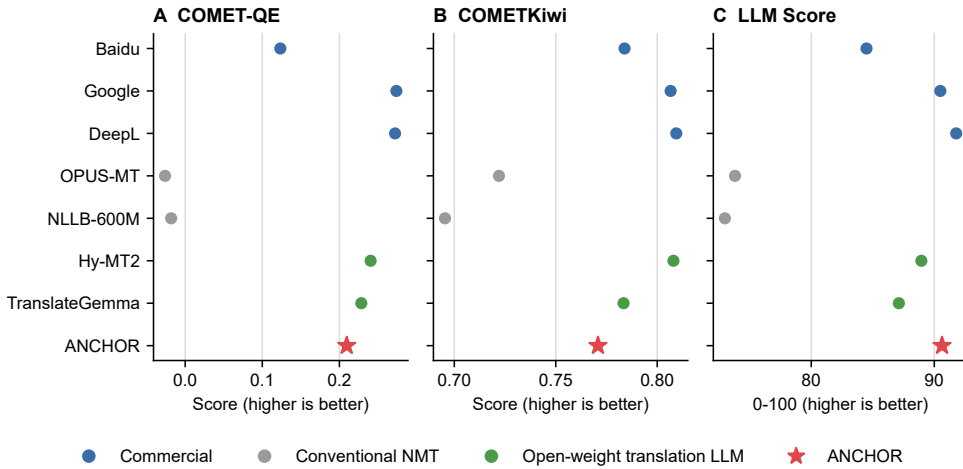


Fig. 7 Comparison of the eight models on the three general translation quality metrics

4.7.2 Quality by data source

The results by data source are shown in Table 14.

Table 14 General translation quality by evaluation data

Model	NLP4IF			ClaimBuster/ViClaim		
	COMET-QE	COMETKiwi	LLM Score	COMET-QE	COMETKiwi	LLM Score
Baidu Translate	-0.0786	0.7300	79.88	0.2245	0.8109	86.79
Google Translate	0.1350	0.7663	89.31	0.3435	0.8268	91.08
DeepL	0.1270	0.7770	91.16	0.3449	0.8256	92.09
OPUS-MT	-0.2440	0.6313	63.35	0.0830	0.7674	79.03
NLLB-600M	-0.1910	0.6235	66.54	0.0683	0.7313	76.20
Hy-MT2	0.1044	0.7860	87.36	0.3086	0.8190	89.74
TranslateGemma	0.0783	0.7406	85.46	0.3036	0.8049	87.95
ANCHOR	0.1310	0.7583	91.08	0.2487	0.7770	90.40

On the NLP4IF samples, which all follow the foreignization path, ANCHOR and DeepL obtain LLM Scores of 91.08 and 91.16, COMET-QE scores of 0.1310 and 0.1270, and COMETKiwi scores of 0.7583 and 0.7770; the first two metrics are close, while DeepL has the higher COMETKiwi. On the ClaimBuster and ViClaim samples, which mix both paths, ANCHOR’s LLM Score of 90.40 is below DeepL’s 92.09, and their COMET-QE scores are 0.2487 and 0.3449. Splitting the ablation results by the path actually taken further reveals the quality gap of the domestication group.

4.7.3 General translation quality in the ablation experiment

On the same 1,499 source texts, the differences between ANCHOR and GPT-5.2 direct translation are -0.0009 on COMET-QE, -0.0311 on COMETKiwi and $+2.28$ on LLM Score (Table 15).

In the foreignization group, COMET-QE and LLM Score are higher, with differences of $+0.0258$ and $+3.23$, while COMETKiwi is lower, with a difference of -0.0109 . In the domestication group, all three metrics are below those of direct translation: the differences are -0.1296 on COMET-QE, -0.1283 on COMETKiwi and -2.25 on LLM Score.

Overall, the general translation quality results of the ablation experiment point in the same direction as the by-source results of the baseline experiment: ANCHOR performs better on the foreignization path and worse on the domestication path. In the baseline experiment, ANCHOR is close to DeepL on the NLP4IF samples, which all follow the foreignization path, and falls behind DeepL on the ClaimBuster and ViClaim samples, which mix both paths. After splitting by the path actually taken, the foreignization group of the ablation experiment scores higher than GPT-5.2 direct translation on COMET-QE and LLM Score, whereas the domestication group scores lower on all three metrics. Because the foreignization group accounts for 1,241 of the 1,499 samples, the overall LLM Score remains 2.28 higher, and the overall COMETKiwi gap comes mainly from the domestication group.

Table 15 General translation quality of ANCHOR and GPT-5.2 direct translation

Model	COMET-QE	COMETKiwi	LLM Score (0–100)
<i>Overall (n=1,499)</i>			
ANCHOR	0.2095	0.7708	90.63
GPT-5.2 direct translation	0.2104	0.8019	88.35
Difference	−0.0009	−0.0311	+2.28
<i>Domestication (n=258)</i>			
ANCHOR	0.0311	0.6606	83.71
GPT-5.2 direct translation	0.1606	0.7889	85.96
Difference	−0.1296	−0.1283	−2.25
<i>Foreignization (n=1,241)</i>			
ANCHOR	0.2466	0.7937	92.07
GPT-5.2 direct translation	0.2208	0.8046	88.84
Difference	+0.0258	−0.0109	+3.23

4.7.4 Relationship between general translation quality and check-worthiness preservation

The two evaluations examine different objects: general translation quality measures the linguistic and semantic performance of translations, whereas label preservation measures whether a translation can still support the checking decision of the source. They rank the same models differently. Google’s COMET-QE of 0.2739 is higher than ANCHOR’s 0.2095, but its overall label preservation rate in the baseline experiment, 98.75%, is lower than ANCHOR’s 99.38%. OPUS-MT and NLLB-600M reach negative-class preservation rates of 98.33% and 97.41%, yet they are the two lowest of the eight models on all three general translation quality metrics, with negative COMET-QE scores. The ranking by quality metrics is therefore not consistent with the degree of check-worthiness preservation, and the two evaluations need to be reported separately.

For check-worthiness data transfer, label preservation is the primary evaluation objective, and general translation quality indicates how usable a translation is linguistically and semantically. ANCHOR’s COMET-QE and COMETKiwi are below those of Google and DeepL, and its LLM Score of 90.63 ranks second among the eight models; its overall label preservation rate is tied with DeepL for the highest, and it reaches 100.00% on the negative class and the High×negative group. That ANCHOR does not rank first in general translation quality should be interpreted with the generation path in mind: the domestication path rewrites cultural expressions for pragmatic function, which lowers the sentence-level correspondence between source and translation and may lead COMET-QE and COMETKiwi to underestimate quality.

4.8 Discussion

Gains in check-worthiness preservation

In the baseline experiment, ANCHOR and DeepL have the same overall preservation rate; the class breakdown shows that DeepL is higher on the positive class and the High

group, whereas ANCHOR is higher on the negative class and the High $\times$ negative group. In the ablation experiment, ANCHOR’s overall preservation rate is 1.88 percentage points higher than that of direct translation, with six fewer failed units and fewer errors in handling proper names, quantities and entities. These results describe the performance of the transcreation pipeline at two levels: the applicability of the source label and concrete textual changes.

Information boundaries of cultural adaptation

The domestication quality metrics and cases jointly indicate that keeping expressions natural also requires keeping the information needed for checking. The domestication examples contain entity omission and generalized reference, and the ablation experiment also shows added conclusions and changes in assertion direction. ANCHOR has more added-conclusion errors than direct translation, which shows that the preservation requirements in the generation prompt cannot yet guarantee that every output meets the objective. Future improvements should focus on checking whether adjustments of cultural expressions change the claim target, factual scope and assertion direction.

Role of the evaluation design

Reporting positive and negative classes, cultural dependence groups and error types together helps distinguish the loss of existing claims from information added to non-factual expressions. General translation quality provides a complementary check of linguistic and semantic performance, and the domestication questionnaire provides observations on meaning accuracy, cultural relevance and fluency. Together, these evaluations serve the judgment of annotated data transfer.

5 Conclusion

This paper addresses the problem that source labels may no longer apply when English check-worthiness estimation data are transferred into Chinese, and proposes the method ANCHOR. Entity omission, changes in claim scope and information addition during translation alter whether a translation can support the original checking decision, and culturally loaded expressions make it hard to balance literal retention against excessive rewriting. ANCHOR first classifies the cultural dependence of a text: texts with high cultural dependence follow the domestication path, which adapts Chinese expressions under requirements to preserve factual scope, assertion strength and pragmatic function; texts with low cultural dependence follow the foreignization path, which retrieves demonstrations from a Chinese–English parallel corpus of Ministry of Foreign Affairs press conferences and obtains the translation through candidate generation and self-supervised reranking. With the applicability of the source label to the Chinese text as its objective, the method provides a generation and evaluation pipeline for reusing English annotated resources.

For evaluation, this study examines general translation quality and check-worthiness preservation separately. General translation quality is measured with

COMET-QE, COMETKiwi and LLM Score on 1,499 common samples. For check-worthiness preservation, 320 source texts are sampled by the strata “cultural dependence $\times$ source label”, three evaluators judge item by item whether each translation can support checking the source claim, and label preservation rates are reported by positive class, negative class and cultural dependence group. This design makes it possible to present the two risks of claim loss and the rewriting of non-factual expressions into facts separately. Two components of the generation pipeline were also examined by humans: with the majority vote of six evaluators as the reference label, cultural dependence classification achieves an accuracy of 0.8225 on 400 samples, with a High-class recall of 1.0000 and precision of 0.6450; seven evaluators rated 100 domesticated translations for meaning accuracy, cultural relevance and fluency on a 1–5 scale, giving an overall mean of 4.386.

In the baseline experiment, ANCHOR and seven baseline models were evaluated on 2,560 output units corresponding to 320 source texts. ANCHOR’s overall label preservation rate is 99.38%, tied with DeepL; DeepL is higher on the positive class and the High cultural dependence group, whereas ANCHOR is higher on the negative class and the High $\times$ negative group. In the ablation experiment, the preservation rates of ANCHOR and GPT-5.2 direct translation are 97.19% and 95.31%, with 9 and 15 failed units, respectively. The reduction of errors in proper names, quantities and entities explains the textual changes behind the overall difference, while added conclusions and changes in assertion direction reveal the information boundaries that rewriting still needs to control.

The general translation quality results complement this assessment. ANCHOR’s LLM Score of 90.63 ranks second among the eight models. When split by the path actually taken, the foreignization group has higher COMET-QE and LLM Score and lower COMETKiwi than direct translation, and the domestication group is lower on all three metrics; this agrees in direction with the by-source results of the baseline experiment, i.e., ANCHOR performs better on the foreignization path and worse on the domestication path. Entity omission, generalized reference and content addition in the raw outputs of the domestication path show that natural expression, cultural adaptation and the retention of factual information need to be checked separately. When evaluating English-to-Chinese transfer of annotated data, combining general translation quality, label preservation rates and concrete errors presents a method’s performance better than reporting overall scores alone.

6 Limitations

This study still has limitations. While preserving check-worthiness, there is room to improve general translation quality. In addition, the experiments cover only the English-to-Chinese direction. The cultural dependence criteria, the retrieval demonstration corpus and the domestication prompt are all designed around English–Chinese data; when transferred to other language pairs, linguistic morphology, cultural distance and available retrieval resources all change. Check-worthiness preservation currently relies on human judgment: the 320 stratified samples and eight models form 2,560 evaluation units, each judged by three evaluators, which is costly. A more objective way of

evaluating label preservation, for example automatically judging whether label applicability changes with cross-lingual representation models or Chinese check-worthiness classifiers, would help scale up evaluation and reduce the influence of inter-rater differences, and its calibration against human judgment deserves exploration. Furthermore, combining this construction approach with augmentation methods such as back-translation, template rewriting and synthetic data could extend corpora in domain, topic and temporal coverage, build Chinese check-worthiness estimation datasets and further support research on Chinese fact-checking.

Appendix A Prompts

The prompts are experimental inputs and are reproduced verbatim. Chinese text inside the prompts is part of the original input; an English translation of the Chinese LLM Score prompt is provided after it for readability.

A.1 Cultural dependence classification prompt

```

You are an expert in cross-lingual text transcreation for fact-checking
→ value preservation. Your goal is to convert "Fact-Checking" English
→ texts into Chinese.

Your task now is ONLY to determine the Cultural Dependence and the
Translation Strategy for the given text. Do NOT translate yet.

### Core Concept
The goal is NOT literal translation, but translating the judgment of
→ "whether it is worth fact-checking".
- Domestication (归化): Used when the text relies on specific cultural
→ backgrounds (e.g., American humor, political metaphors, slang).
→ Replace with Chinese-context equivalents (functional equivalence) so
→ it feels natural to a Chinese speaker.
- Foreignization (异化): Used when the text's non-native feel does not
→ affect the fact-checking judgment. Keep the original context and just
→ translate the language.

### Definitions
1. High Cultural Dependence (Strategy: Domestication)
* Relies on specific cultural context, humor, local allusions, or
→ political/social metaphors.
* Requires adjustment to fit the target culture (Chinese) so the reader
→ "gets" the meaning and emotion immediately.
* Example 1 (American Pandemic Satire -> Chinese Pandemic Context):
→ "cough in a rich person's face" -> "see if anyone around is taken for
→ testing". (Preserves satire/emotion without the specific class
→ conflict which might be less resonant or distracting).
* Example 2 (School Metaphor): "recess" -> "連坐" (Collective
→ punishment concept).

```

```

* *Example 3 (Slang/Idioms)*: English catchphrases -> Chinese internet
  ↳ slang (e.g., "ghosting" -> "放鸽子").

2. **Low Cultural Dependence (Strategy: Foreignization)**
* Content is universal and easy to understand across cultures.
* Includes pure emotional venting, universal complaints, or abstract
  ↳ metaphors.
* *Example 1 (Venting)*: "this is fucking bullshit" -> "这他妈全是狗屁"
  ↳ (Universal emotion).
* *Example 2 (Universal Complaint)*: "every app update makes things
  ↳ worse" (Universal experience).
* *Example 3 (Abstract Metaphor)*: "brain running on 1% battery"
  ↳ (Universal metaphor).

### Task
Analyze the provided English text and determine:
1. **Cultural Dependence**: High or Low.
2. **Strategy**: Domestication (if High) or Foreignization (if Low).
3. **Reasoning**: Brief explanation of why.

### Output Format
Output STRICTLY in JSON format:
{
  "cultural_dependence": "High" or "Low",
  "strategy": "Domestication" or "Foreignization",
  "reasoning": "..."
}

### Text to Analyze:
"{text}"

```

A.2 LLM-based translation candidate generation prompt

```

"You are a professional translator. I will give you one or more examples
  ↳ of text fragments, "
  ↳ "where the first one is in {src_lang} and the second one is the
  ↳ translation in {tgt_lang}."
  ↳ "\nThese sentences will be displayed below.\n"
  ↳ "{examples}\n"
  ↳ "After the example pairs, I will provide a {src_lang} sentence and I would
  ↳ like you to translate it into {tgt_lang}."
  ↳ "Please provide only the translation result without any additional
  ↳ comments.\n\n"
  ↳ "{src_lang} text: {query}\n"
  ↳ "Translation:"

```

A.3 LLM Score evaluation prompt

The original prompt, written in Chinese, is:

你是严谨的翻译质量评审员。你的任务：只根据英文原文(src)与中文译文(mt)，在无参考译文情况下进行质量估计(QE)打分。

请从以下维度综合判断：

- 1) 含义保真/信息完整度（是否漏译、误译、增译）
- 2) 术语/实体/数字准确性
- 3) 中文流畅性与可读性
- 4) 风格与语域是否匹配（如口语、粗口、讽刺等是否被合理保留）
- 5) 语用/修辞效果（如幽默、反讽是否被保留）

输出必须是 JSON，字段：

```
{
  "score": 0-100 的整数,
  "errors": ["要点式列出主要问题(若无则空数组)"],
  "brief": "一句话总体评价"
}
```

评分解释（供你内部使用）：

90-100：几乎可直接发布；75-89：小瑕疵不影响理解；60-74：明显问题但大体可用；40-59：错误较多；0-39：严重错误/不通顺/不相关。

English translation (for readability only; the Chinese version above was used in all experiments):

You are a rigorous translation quality reviewer. Your task: based only on
→ the English source (src) and the Chinese translation (mt), give a
→ quality estimation (QE) score without a reference translation.

Judge comprehensively along the following dimensions:

- 1) Meaning fidelity / information completeness (omission, mistranslation,
→ addition)
- 2) Accuracy of terminology / entities / numbers
- 3) Fluency and readability of the Chinese
- 4) Whether style and register match (e.g., whether colloquialisms, profanity
→ and sarcasm are reasonably preserved)
- 5) Pragmatic / rhetorical effect (e.g., whether humour and irony are
→ preserved)

The output must be JSON with the fields:

```
{
  "score": an integer from 0 to 100,
  "errors": ["list the main problems as bullet points (empty array if none)"],
  "brief": "a one-sentence overall assessment"
}
```

Score interpretation (for your internal use):

90-100: almost publishable as is; 75-89: minor flaws that do not affect
→ understanding; 60-74: obvious problems but broadly usable; 40-59: many
→ errors; 0-39: serious errors / not fluent / irrelevant.

A.4 Domestication prompt

You are an expert in cross-lingual transcreation for preserving
→ fact-checking value. Convert English texts into Chinese while keeping
→ the SAME check-worthiness judgment across languages and cultures.

Task: Generate ONLY the final Chinese text according to the given strategy,
→ without changing how check-worthy the content appears.

Inputs provided:

- Strategy: {strategy} (Domestication / Foreignization)
- Reason: {reason} (Why this strategy fits; use it as guidance)
- Text: "{text}"

Core objective:

- Preserve the original judgment in BOTH directions. If the source contains
→ a verifiable, publicly relevant claim, a Chinese reader must still
→ recognise it as something worth checking. If the source is pure opinion,
→ venting, humour or small talk, a Chinese reader must still see it as not
→ worth checking.
- Translate the judgment, not the surface form. Never invent checkable facts
→ that were absent, and never dissolve checkable facts that were present.

Factual anchors — preserve verbatim whenever they exist in the source:

- Named people, organisations, institutions, brands, and place names that
→ are the subject of a claim
- Numbers, quantities, percentages, money amounts, dates and time references
- The asserted proposition itself, including its scope, its subject, and its
→ assertion strength
- Causal and quantified relations ("X caused Y", "X increased by N")

These are NOT culture-specific decoration. Localise the framing around them,
→ never the anchors themselves. Replacing a named entity with a generic
→ paraphrase, or a number with a vague quantifier, destroys
→ check-worthiness and is a failure even if the Chinese reads naturally.

Strategy rules:

- Domestication (归化): Replace culture-specific elements (humor, metaphors,
→ slang, local institutions, holidays, sports references) with
→ Chinese-context equivalents that convey the same function and emotion.
→ Make it sound native and natural in Chinese. Apply this to the
→ culturally loaded framing only; carry every factual anchor through
→ unchanged.
- Foreignization (异化): Keep the original context and flavor if it does not
→ affect the judgment; render the language naturally in Chinese.

Strict constraints:

- Do NOT introduce new verifiable facts that the source lacks
- Do NOT remove, generalise or blur verifiable facts that the source
→ contains
- Do NOT raise public relevance or perceived risk for non-informational text

- Do NOT lower public relevance or verifiability for text that carries a
→ real claim
- Match the information density of the source: keep it low when the source
→ is non-informational, keep it high when the source is factual
- Maintain the original emotional and attitudinal function
- Keep links/hashtags if they help context, but do not localize them into
→ Chinese-specific political references

Examples and cautions — non-informational sources (keep them
→ non-checkable):

- English: "The government keeps changing the rules and expecting us to just
→ keep up."
 - Bad literal (risk ↑): "政府不断修改政策, 却指望我们立刻适应。" (容易被误
→ 读为可核查政策声明)
 - Good domestication (risk ↔): "各种规定一变再变, 说实话真的有点跟不上节
→ 奏。"
- English: "You know, I saw a movie — 'Crocodile Dundee.'"
 - Bad literal (risk ↑): "你知道吗, 我看了一部叫《鳄鱼邓迪》的电影。" (片名和
→ 文化背景在中文语境里陌生、信息量高)
 - Good domestication (risk ↔): "你知道吗, 前两天刚看了一部特别火的冒险
→ 片。" (保留\随口举个大众电影例子"的功能, 不增加可核查细节)
- English: "That answer was about as clear as Boston harbor."
 - Bad literal (risk ↑): "那个回答就像波士顿港一样清澈。" (比喻依赖美国地理与
→ 历史联想, 中文读者难以自然理解)
 - Good domestication (risk ↔): "那个回答糊得跟一锅浆糊一样。" (用中文常见比
→ 喻表达\很不清楚", 保持调侃、不可核性质)
- English: "Go out there and buy yourself a gun."
 - Bad literal (risk ↑): "出去给自己买把枪。" (直指\买枪", 在中文语境中容易被
→ 误读为严肃可核查主张)
 - Good domestication (risk ↔): "你就自己想办法给自己多加点底气吧。" (弱化具
→ 体行为, 保留\鼓动/打气"的语用功能)
- Pure venting like "this is fucking bullshit" → 强烈否定情绪表达
→ (Foreignization)
- Universal metaphors like "1% battery" → 保留隐喻 (Foreignization)

Examples and cautions — factual sources (keep them checkable):

- English: "Senator Cruz voted against the infrastructure bill three times
→ last year."
 - Bad domestication (verifiability ↓): "有些议员去年好几次都没支持那个基建
→ 方案。" (人名、次数、议案全部被泛化, 中文读者已无从核查)
 - Good domestication (verifiability ↔): "克鲁兹参议员去年三次投票反对那项
→ 基础设施法案。" (人名、次数、议案原样保留, 仅调整表述习惯)
- English: "Unemployment in Ohio dropped to 3.2 percent, the lowest since
→ 1999."
 - Bad domestication (verifiability ↓): "那个州的失业率降到了近年来的低
→ 点。" (地名、数字、年份被抹去, 可核查性消失)
 - Good domestication (verifiability ↔): "俄亥俄州失业率降至 3.2%, 为 1999
→ 年以来最低水平。" (事实锚点全部保留)

```

- English: "They shut down the local hospital right after the Fourth of July
↪ parade."
- Good domestication (verifiability ↔): "他们在国庆游行刚结束后就关停了当
↪ 地那家医院。" (把 Fourth of July 归化为中文读者可理解的\国庆游行", 但\关
↪ 停当地医院"这一可核查主张原样保留)

Output format:
{{
  "text_cn": "...
}}
```

Appendix B Inter-rater agreement of the check-worthiness preservation evaluation

This appendix reports agreement details of the human check-worthiness preservation evaluation in the main text. The evaluation covers 320 source samples, each paired with anonymous candidate translations from eight models, giving 2,560 “sample × model” evaluation units. Three volunteers (denoted R_1 to R_3) each independently completed all 2,560 units, yielding 7,680 rating records with no missing values. The three evaluators used identical guidelines and scoring criteria; only the presentation order of candidates differed among them.

B.1 Agreement of the preservation judgment

The preservation judgment is binary (preserved / not preserved). The marginal preservation rates of the three evaluators are 95.90%, 97.07% and 96.17%, and preserved judgments account for 96.38% of all records. The three evaluators fully agree on 2,448 units (95.62%); the remaining 112 units (4.38%) are 2:1 splits, of which 72 are “two preserved, one not preserved” and 40 are “one preserved, two not preserved”. The pairwise observed agreements are 97.50% (R_1 – R_2), 96.99% (R_1 – R_3) and 96.76% (R_2 – R_3), and the mean pairwise agreement (i.e., $\bar{P}$ in the computation of Fleiss’ κ) is 97.08%.

Fleiss’ κ is 0.5820, with observed agreement $\bar{P} = 0.9708$ and expected agreement $P_e = 0.9302$. Preserved judgments account for 96.38%, which corresponds to a high expected agreement. We therefore report κ , three-way full agreement and pairwise observed agreement together to present agreement under different measures.

The disagreement rate is 8.12% for positive-class units ($n = 1,120$) and 1.46% for negative-class units ($n = 1,440$); it is 5.23% for High cultural dependence units and 3.52% for Low. In this sample, disagreements occur more often on positive-class and High cultural dependence texts.

B.2 Agreement of the four-dimension ratings

Krippendorff’s α of the four 0–4 dimensions is computed on an interval scale, as shown in Table B2. The four α values range from 0.7984 to 0.8463, and the composite task score reaches 0.8488. Units with identical scores from all three evaluators account for

Table B1 Inter-rater agreement of the check-worthiness preservation judgment

Measure	Value
Evaluation units	2,560
Three-way full agreement	95.62%
Observed agreement (mean pairwise)	97.08%
Expected agreement P_e	0.9302
Fleiss' κ	0.5820
Share of preserved judgments	96.38%

65.59%–68.95%, units with a score range of at most 1 for 93.40%–95.78%, and units with a range of 2 or more for 4.22%–6.60%. The pairwise Pearson correlations of the composite task score are 0.857 (R_1 – R_2), 0.847 (R_1 – R_3) and 0.847 (R_2 – R_3).

Table B2 Inter-rater agreement of the four-dimension ratings (interval-scale Krippendorff's α)

Dimension	α	Identical scores	Range ≤ 1	Mean
Factual scope	0.8425	65.90%	93.48%	3.383
Pragmatic function	0.7984	67.54%	94.92%	3.516
Cultural appropriateness	0.8235	65.59%	93.40%	3.399
Chinese fluency	0.8463	68.95%	95.78%	3.441
Composite task score (0–100)	0.8488	—	—	88.46

B.3 Marginal distributions of the evaluators

Table B3 gives the marginal statistics of the evaluators. Their preservation rates differ by 1.17 percentage points, the means of the four dimensions differ by no more than 0.046 points, and the composite task scores differ by 0.70 points. For aggregation, the three judgments are first averaged within each “sample $\times$ model” unit and then summarized per source sample.

Table B3 Marginal statistics of the three evaluators

Evaluator	Preservation rate	Factual scope	Pragmatic function	Cultural appropriateness	Chinese fluency	Task score (0–100)
R_1	95.90%	3.380	3.525	3.398	3.445	88.37
R_2	97.07%	3.396	3.514	3.423	3.442	88.85
R_3	96.17%	3.373	3.507	3.377	3.437	88.15
Mean	96.38%	3.383	3.516	3.399	3.441	88.46

Appendix C Ablation of the dense pre-filtering candidate number

This appendix reports the effect of the candidate number M in two-stage retrieval, clarifying the trade-off of the dense pre-filtering design in Section 3.

C.1 Setup

The experiment is run on the index of the 4,987-entry parallel corpus used by the foreignization path, with embedding dimension $H = 3072$ and L2-normalized vectors. With a fixed random seed, 200 entries are sampled from the English side of the corpus as queries; to prevent a query from matching its own entry, the dense score of that entry is set to $-\infty$. The mixing weights are $(\alpha_1, \alpha_2, \alpha_3) = (0.4, 0.4, 0.2)$, and the projection parameters $\mathbf{w}_s$ and $\mathbf{W}_m$ are randomly initialized and fixed as in Section 3.2 (seed 42), identical for all values of M so that the settings are comparable.

Taking the retrieval results with $M \geq N$ (i.e., the full hybrid score over the whole corpus) as the reference, we compare retrieval consistency under each M , measured by the number of queries whose top-3 demonstration sequence is identical, the number of queries whose top-1 demonstration is identical, and the mean overlap ratio of the top-3 demonstration sets (overlap@3).

Consistency statistics compute the scores of all queries and candidates with vectorized versions of Eqs. (7) and (8); running time is measured entry by entry with the actual retrieval implementation, and the two differ numerically only at the level of floating-point error (10^{-16}). The test hardware is an Intel Core i9-10900K CPU running NumPy 1.26.4 on a single thread.

C.2 Results and analysis

The results are shown in Table C4. For all values of M from 2,048 down to 128, the top-1 demonstration is identical to the full-corpus reference (200/200) and overlap@3 is 1.0000; the ordered top-3 sequence differs only for isolated queries in which the second and third positions are swapped, leaving the demonstration set unchanged. For the corpus size and query distribution of this study, dense pre-filtering does not change the demonstrations that actually enter the few-shot prompt.

Table C4 Effect of the candidate number M on retrieval consistency and per-query time

M	Top-3 identical	Top-1 identical	overlap@3	Time (s)
$N = 4,987$	Reference	Reference	1.0000	129.7
2,048	199/200	200/200	1.0000	—
1,024	199/200	200/200	1.0000	—
512	199/200	200/200	1.0000	12.7
256	200/200	200/200	1.0000	—
128	199/200	200/200	1.0000	—

In terms of time, single-query retrieval takes 12.7 seconds with $M = 512$ and 129.7 seconds with full-corpus computation, a difference of about 10.2 times. The difference arises because the number of entries for which the two additional scores are computed in the second stage drops from N to M , consistent with the analysis in the main text. The times are measurements under the single-threaded CPU setting above and will change on other hardware or with parallel implementations, but the relative relation between two-stage and full-corpus computation is determined by N/M .

Consistency is maintained even at $M = 128$, which agrees with the definition of the two additional scores: S_{sparse} and S_{multi} are both derived from sentence vectors and are highly correlated with the dense score, so dense pre-filtering rarely excludes entries that would win under the full score. If an encoder that outputs token-level hidden states were used instead, the two scores would carry discriminative information across positions, and the effect of M would need to be re-evaluated.

References

- Agrawal A, Fazili B, Jyothi P (2024) Translation errors significantly impact low-resource languages in cross-lingual learning. In: Graham Y, Purver M (eds) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, St. Julian’s, Malta, pp 319–329, <https://doi.org/10.18653/v1/2024.eacl-short.28>, URL <https://aclanthology.org/2024.eacl-short.28/>
- Arslan F, Hassan N, Li C, et al (2020) A benchmark dataset of check-worthy factual claims. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 14. AAAI Press, pp 821–829, <https://doi.org/10.1609/icwsm.v14i1.7346>
- Artetxe M, Labaka G, Agirre E (2020) Translation artifacts in cross-lingual transfer learning. In: Webber B, Cohn T, He Y, et al (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 7674–7684, <https://doi.org/10.18653/v1/2020.emnlp-main.618>, URL <https://aclanthology.org/2020.emnlp-main.618/>
- Baker M, Saldanha G (eds) (2020) Routledge Encyclopedia of Translation Studies, third edition edn. Routledge, London New York, <https://doi.org/10.4324/9781315678627>
- Barančíková P, Bojar O (2025) Intrinsic vs. Extrinsic Evaluation of Czech Sentence Embeddings: Semantic Relevance Doesn’t Help with MT Evaluation. In: Bouillon P, Gerlach J, Girletti S, et al (eds) Proceedings of Machine Translation Summit XX: Volume 1. European Association for Machine Translation, Geneva, Switzerland, pp 265–275, URL <https://aclanthology.org/2025.mtsummit-1.20/>
- Bennett ER, Han H, Yang X, et al (2025) Evaluating Evaluation Metrics for Ancient Chinese to English Machine Translation. In: Anderson A, Gordin S, Li B, et al (eds) Proceedings of the Second Workshop on Ancient Language Processing. Association

- for Computational Linguistics, The Albuquerque Convention Center, Laguna, pp 71–76, <https://doi.org/10.18653/v1/2025.alp-1.9>, URL <https://aclanthology.org/2025.alp-1.9/>
- Chen T, Wang Z, Miao Y, et al (2025) MedFact: A large-scale chinese dataset for evidence-based medical fact-checking of LLM responses. In: Christodoulopoulos C, Chakraborty T, Rose C, et al (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 32340–32353, <https://doi.org/10.18653/v1/2025.emnlp-main.1646>, URL <https://aclanthology.org/2025.emnlp-main.1646/>
- Dhole KD (2025) A Multi-Encoder Frozen-Decoder Approach for Fine-Tuning Large Language Models. <https://doi.org/10.48550/arXiv.2501.07818>, URL <http://arxiv.org/abs/2501.07818>, arXiv:2501.07818
- Donthi S, Spencer M, Patel OB, et al (2025) Improving LLM Abilities in Idiomatic Translation. In: Hettiarachchi H, Ranasinghe T, Rayson P, et al (eds) Proceedings of the First Workshop on Language Models for Low-Resource Languages. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 175–181, URL <https://aclanthology.org/2025.loreslm-1.13/>
- Freitag M, Foster G, Grangier D, et al (2021) Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics 9:1460–1474. <https://doi.org/10.1162/tacl.a.00437>, URL <https://aclanthology.org/2021.tacl-1.87/>
- Geng X, Zhang Y, Lai Z, et al (2023) Improved Pseudo Data for Machine Translation Quality Estimation with Constrained Beam Search. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 12434–12447, <https://doi.org/10.18653/v1/2023.emnlp-main.764>, URL <https://aclanthology.org/2023.emnlp-main.764/>
- Ghazvininejad M, Gonen H, Zettlemoyer L (2023) Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. <https://doi.org/10.48550/arXiv.2302.07856>, URL <http://arxiv.org/abs/2302.07856>, arXiv:2302.07856
- Giedemann P, von Däniken P, Deriu JM, et al (2025) ViClaim: A multilingual multilabel dataset for automatic claim detection in videos. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 397–413, <https://doi.org/10.18653/v1/2025.emnlp-main.21>, URL <https://aclanthology.org/2025.emnlp-main.21/>
- Hassan N, Arslan F, Li C, et al (2017) Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Halifax NS Canada, pp 1803–1812, <https://doi.org/10.1145/3097983>.

- 3098131, URL <https://dl.acm.org/doi/10.1145/3097983.3098131>
- Hu X, Guo Z, Wu G, et al (2022) CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. URL <http://arxiv.org/abs/2206.11863>, arXiv:2206.11863
- Hyben M, Kula S, Cegin J, et al (2026) MultiCW: A large-scale balanced benchmark dataset for training robust check-worthiness detection models. In: Findings of the Association for Computational Linguistics: EACL 2026. Association for Computational Linguistics, Rabat, Morocco, pp 3737–3754, <https://doi.org/10.18653/v1/2026.findings-eacl.194>, URL <https://aclanthology.org/2026.findings-eacl.194/>
- Iyer V, Chen P, Birch A (2023) Towards Effective Disambiguation for Machine Translation with Large Language Models. In: Koehn P, Haddow B, Kocmi T, et al (eds) Proceedings of the Eighth Conference on Machine Translation. Association for Computational Linguistics, Singapore, pp 482–495, <https://doi.org/10.18653/v1/2023.wmt-1.44>, URL <https://aclanthology.org/2023.wmt-1.44/>
- Kabir M, Ahmed T, Rahman MM, et al (2026) Semantic label drift in cross-cultural translation. In: Proceedings of the Fifteenth Language Resources and Evaluation Conference. ELRA Language Resource Association, Palma de Mallorca, Spain, pp 3714–3724, <https://doi.org/10.63317/5ae9txdv2s3g>, URL <https://aclanthology.org/2026.lrec-1.297/>
- Khanuja S, Ramamoorthy S, Song Y, et al (2024) An image speaks a thousand words, but can everyone listen? On image transcreation for cultural relevance. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 10258–10279, URL <https://aclanthology.org/2024.emnlp-main.573>
- Kocmi T, Federmann C (2023) Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In: Nurminen M, Brenner J, Koponen M, et al (eds) Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, Tampere, Finland, pp 193–203, URL <https://aclanthology.org/2023.eamt-1.19/>
- Lin YJ, Lin CY, Yeh CJ, et al (2024) CFEVER: A Chinese Fact Extraction and VERification Dataset. Proceedings of the AAAI Conference on Artificial Intelligence 38(17):18626–18634. <https://doi.org/10.1609/aaai.v38i17.29825>, URL <http://arxiv.org/abs/2402.13025>
- Lu H, Yang H, Huang H, et al (2024) Chain-of-Dictionary Prompting Elicits Translation in Large Language Models. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 958–976, <https://doi.org/10.18653/v1/2024.emnlp-main.55>, URL <https://aclanthology.org/2024.emnlp-main.55/>

- Moslem Y, Haque R, Kelleher JD, et al (2023) Adaptive Machine Translation with Large Language Models. In: Nurminen M, Brenner J, Koponen M, et al (eds) Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, Tampere, Finland, pp 227–237, URL <https://aclanthology.org/2023.eamt-1.22/>
- Nakov P, Barrón-Cedeño A, Da San Martino G, et al (2022) Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol 3180. CEUR-WS.org, Bologna, Italy, pp 368–392, URL <https://ceur-ws.org/Vol-3180/paper-28.pdf>
- OpenAI (2024) New embedding models and api updates. URL <https://openai.com/index/new-embedding-models-and-api-updates/>
- Qian S, Orăsan C, Kanojia D, et al (2024) Are large language models state-of-the-art quality estimators for machine translation of user-generated content? URL <https://arxiv.org/abs/2410.06338>, arXiv:2410.06338
- Rei R, Stewart C, Farinha AC, et al (2020) Comet: A neural framework for mt evaluation. In: Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp), pp 2685–2702
- Rei R, Treviso M, Guerreiro NM, et al (2022) Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp 634–645
- Robinson N, Ogayo P, Mortensen DR, et al (2023) ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In: Koehn P, Haddow B, Kočmi T, et al (eds) Proceedings of the Eighth Conference on Machine Translation. Association for Computational Linguistics, Singapore, pp 392–418, <https://doi.org/10.18653/v1/2023.wmt-1.40>, URL <https://aclanthology.org/2023.wmt-1.40/>
- Saldías Fuentes B, Foster G, Freitag M, et al (2022) Toward more effective human evaluation for machine translation. In: Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval). Association for Computational Linguistics, Dublin, Ireland, pp 76–89, <https://doi.org/10.18653/v1/2022.humeval-1.7>, URL <https://aclanthology.org/2022.humeval-1.7/>
- Saunders D, DeNeefe S (2024) Domain adapted machine translation: What does catastrophic forgetting forget and why? In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 12660–12671, <https://doi.org/10.18653/v1/2024.emnlp-main.704>, URL <https://aclanthology.org/2024.emnlp-main.704/>

- Shaar S, Alam F, Da San Martino G, et al (2021) Findings of the NLP4IF-2021 Shared Tasks on Fighting the COVID-19 Infodemic and Censorship Detection. In: Feldman A, Da San Martino G, Leberknight C, et al (eds) Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. Association for Computational Linguistics, Online, pp 82–92, <https://doi.org/10.18653/v1/2021.nlp4if-1.12>, URL <https://aclanthology.org/2021.nlp4if-1.12/>
- Simoulin A, Park N, Liu X, et al (2024) Memory-Efficient Fine-Tuning of Transformers via Token Selection. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 21565–21580, <https://doi.org/10.18653/v1/2024.emnlp-main.1202>, URL <https://aclanthology.org/2024.emnlp-main.1202/>
- Son H, Son Y, Kim C, et al (2025) Not All Adapters Matter: Selective Adapter Freezing for Memory-Efficient Fine-Tuning of Language Models. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 9479–9496, <https://doi.org/10.18653/v1/2025.naacl-long.480>, URL <https://aclanthology.org/2025.naacl-long.480/>
- Stap D, Hasler E, Byrne B, et al (2024) The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. In: Ku LW, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 6189–6206, <https://doi.org/10.18653/v1/2024.acl-long.336>, URL <https://aclanthology.org/2024.acl-long.336/>
- Tang L, Qin J, Ye W, et al (2025) Adaptive Few-shot Prompting for Machine Translation with Pre-trained Language Models. Proceedings of the AAAI Conference on Artificial Intelligence 39(24):25255–25263. <https://doi.org/10.1609/aaai.v39i24.34712>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/34712>
- Venuti L (2018) The Translator’s Invisibility: A History of Translation, third edition edn. Routledge Translation Classics, Routledge, New York, <https://doi.org/10.4324/9781315098746>
- Xu H, Kim YJ, Mohamed Nabil Aly Aly Sharaf A, et al (2024a) A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. International Conference on Learning Representations 2024:2747–2767. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b9536e186a77feff516893a5f393f7a-Abstract-Conference.html
- Xu H, Sharaf A, Chen Y, et al (2024b) Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. <https://doi.org/10.>

48550/arXiv.2401.08417, URL <http://arxiv.org/abs/2401.08417>, arXiv:2401.08417

- Yuan Z, Ye Y, Feng X, et al (2026) Culture-aware machine translation in large language models: Benchmarking and investigation. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, San Diego, California, United States, pp 29636–29661, <https://doi.org/10.18653/v1/2026.acl-long.1367>, URL <https://aclanthology.org/2026.acl-long.1367/>
- Zhang B, Haddow B, Birch A (2023) Prompting large language model for machine translation: A case study. In: Krause A, Brunskill E, Cho K, et al (eds) Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 202. PMLR, pp 41092–41110, URL <https://proceedings.mlr.press/v202/zhang23m.html>
- Zhang L, Burian D, John V, et al (2024) Unveiling Semantic Information in Sentence Embeddings. In: Bonial C, Bonn J, Hwang JD (eds) Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024. ELRA and ICCL, Torino, Italia, pp 39–47, URL <https://aclanthology.org/2024.dmr-1.5/>
- Zhang X, Wang X, Lu Y, et al (2025) TrendFact: A Benchmark for Explainable Hotspot Perception in Fact-Checking with Natural Language Explanation. <https://doi.org/10.48550/arXiv.2410.15135>, URL <http://arxiv.org/abs/2410.15135>, arXiv:2410.15135